



Politecnico
di Torino



SAPIENZA
UNIVERSITÀ DI ROMA

UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II
ISTITUTO DI INQUINAMENTO ATMOSFERICO – CONSIGLIO NAZIONALE DELLE
RICERCHE



NATIONAL PH.D. PROGRAM IN
ARTIFICIAL INTELLIGENCE

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Application of Machine Learning Techniques and Geospatial Data for High-Resolution Intra-Urban Air Quality Monitoring and Assessment

Supervisor

Dott. Paolo FAZZINI

Co-Supervisor

Dott. Francesco PETRACCHINI

Candidate

Marica DE LUCIA

Abstract

Air pollution represents one of the greatest environmental and health threats of our time. Accurate monitoring and prediction of particulate matter concentrations and other pollutants are fundamental for developing effective mitigation strategies. Artificial intelligence offers an innovative approach to improve existing monitoring data and develop advanced predictive models for atmospheric particulate matter (PM), including its chemical speciation, potentially contributing to filling these important knowledge gaps. Methods: This study developed the HReBLOCK (High Resolution Ensemble Block Level Of City Knowledge) framework to address three different tasks. The first two tasks enabled the prediction of PM₁₀ and PM_{2.5} concentrations in the metropolitan area of Bari at an intra-urban spatial resolution of 300m, using machine learning techniques to optimize satellite AOD data, meteorological and geographical parameters. The third task developed the foundations for PM_{2.5} speciation (EC, OC, SO₄, NO₃, DUST) in the context of the PTA-Rome of the NASA MAIA mission. The models were validated with data from ground monitoring stations and enriched with explainability techniques. The AOD improvement model increased Sentinel-3 Synergy's correlation with AERONET ground data by 47%. For particulate prediction, the framework achieved robust performance in spatial cross-validation, with determination coefficients (R^2) of 0.81 ± 0.01 for PM₁₀ and 0.80 ± 0.04 for PM_{2.5}, and RMSE values of approximately $4.90 \pm 0.31 \mu\text{g}/\text{m}^3$ and $2.90 \pm 0.04 \mu\text{g}/\text{m}^3$ respectively. Regarding PM_{2.5} speciation, a comparison was conducted with the GRM algorithm of the NASA MAIA mission, demonstrating that the addition of spatio-temporal variables allowed for achieving comparable results for each chemical component, with R^2 values ranging from 0.53 to 0.70. The approach developed through artificial intelligence techniques significantly expands the capabilities of atmospheric pollution monitoring and prediction, extending them also to areas without ground monitoring stations. The high spatial resolution maps generated represent an essential tool for environmental and health authorities, enabling targeted interventions to protect the most vulnerable populations. The ability to estimate not only the overall concentrations of particulate matter but also its chemical composition constitutes a fundamental advancement towards more precise epidemiological studies, capable of assessing the specific hazards of different pollutant components and deepening the understanding of atmospheric pollution effects on public health.

Contents

Lista degli Acronimi	xi
1 Introduction	1
1.1 Atmospheric Pollutants	3
1.2 Monitoring Networks and Satellite Data	6
1.3 Predicting PM: a review	8
1.3.1 Prediction methods	9
1.4 Thesis Structure	11
2 Application Area	13
2.1 MAIA Mission	13
2.1.1 Instrument	14
2.1.2 Target Area	17
2.1.3 Monitoring Network	19
2.2 Structure and Management of MAIA Data Products	21
2.2.1 Level 1 Products	22
2.2.2 Level 2 Products	22
2.2.3 Level 4 Products	22
2.2.4 Ancillary Products	23
2.3 APEMAIA project	25
3 Methodology	27
3.1 HReBLOCK framework	28
3.2 Improved AOD methodology	31
3.2.1 Materials	31
3.3 PM prediction methodology	35
3.3.1 Methodological Framework	35
3.3.2 Study Area	35
3.3.3 Materials	36
3.4 Machine Learning technique	45
3.5 Ensemble Methods	45
3.5.1 Random Forest	45
3.5.2 Gradient Boosting	48

3.5.3	XGBoost	50
3.6	Feature Selection	50
3.6.1	Recursive Feature Elimination	51
3.6.2	Boruta	51
3.6.3	Shap values	52
4	Results	55
4.1	Task 1: AOD Improved	55
4.1.1	Data preparation	57
4.1.2	Data exploration	65
4.1.3	Machine Learning Approaches and Performance	69
4.2	Task 2: PM prediction	74
4.2.1	Data preparation	74
4.2.2	Machine Learning Approaches and Performance	77
4.3	Particulate Matter 2.5 maps	101
5	PM species estimations	107
5.1	Study Area	107
5.2	Dataset	108
5.2.1	PM Species data	111
5.3	Random Forest and GRM model comparison	120
6	Discussion and Conclusions	125
6.0.1	Applicability and Future Perspectives	127

List of Figures

1.1	Leading risk factors for disease burden, 2021, IHME	2
1.2	Anthropic pollution.	3
1.3	Saharian dust over Italy.	4
1.4	PM species classification	4
2.1	Maia Digital Camera parts	16
2.2	Maia instrument main components	18
2.3	Target area defined for MAIA Mission	19
2.4	MAIA Project Surface Monitoring Network	20
2.5	MAIA Simulated Data Products.	24
2.6	APEMAIA project logo.	25
3.1	HReBLOCK framework.	29
3.2	Task 1 and task 2 framework.	30
3.3	Bari study area for Task 1.	37
3.4	PM ₁₀ measuring stations.	38
3.5	PM _{2.5} measuring stations.	39
3.6	Land Cover classification over Bari area.	42
3.7	Building use classification over Bari area.	44
4.1	Frequency Distributions of Metereological Variables	59
4.2	Correlation between MODIS MAIAC AOD (550 nm) and AERONET ground-based measurements.	61
4.3	Correlation analysis between satellite Sentinel-3 retrievals and AERONET ground-based measurements for Synergy A and B	62
4.4	Correlation between Sentinel-3 A Synergy and Sentinel-3 B Synergy.	64
4.5	Correlation between Sentinel-3 Synergy fusion AOD and AERONET ground-based measurements.	65
4.6	Average monthly AOD data products availability.	66
4.7	Time series of AOD products at the "Roma la Sapienza" station	67
4.8	Correlation analysis in AOD improvement task	68
4.9	Random Forest Variable importance.	69
4.10	Boruta variables importance.	70
4.11	Predicted vs AERONET AOD with Ensemble - LOYO and LOSO	72

4.12 AOD Predicted vs AERONET AOD with Ensemble - 10 fold CV	74
4.13 PM ₁₀ and PM _{2.5} measurements divided by year and station - Part 1 . .	75
4.14 PM ₁₀ and PM _{2.5} Measurements divided by year and station - Part 2 . .	76
4.15 Box plot comparison of PM ₁₀ concentrations between days with dust events and days without dust events across monitoring stations by ARPA.	78
4.16 R2 score vs Number of Features with XGBoost model for PM ₁₀	82
4.17 Mean Shap Value with standard deviation with XGBoost model for PM ₁₀ .	83
4.18 Selected features by RF (Boruta approach) for PM ₁₀	84
4.19 XGB predicted vs observed PM ₁₀ density scatter plot.	86
4.20 Correlation between PM ₁₀ and PM _{2.5} over four stations.	88
4.21 Scatterplot of observed and 10-fold CV predicted PM _{2.5} ($\mu\text{g}/\text{m}^3$) . . .	90
4.22 Correlation between completed PM _{2.5} and PM ₁₀ over all stations. . . .	91
4.23 R2 score vs number of features.	93
4.24 Mean shap value with standard deviation with XGBoost model (stage 4).	94
4.25 Comparison of observed and predicted PM _{2.5} time series for monitoring station of Caldarola using both methodological approaches.	97
4.26 Comparison of observed and predicted PM _{2.5} time series for monitoring station of Cavour using both methodological approaches.	98
4.27 Comparison of observed and predicted PM _{2.5} time series for monitoring station of Bitonto using both methodological approaches.	99
4.28 Comparison of observed and predicted PM _{2.5} time series for monitoring station of Modugno using both methodological approaches.	100
4.29 Estimated trend of daily PM _{2.5} concentrations, 2010-2022.	102
4.30	105
4.31	106
5.1 Box area for PTA-Rome of MAIA mission.	108
5.2 Frequencies distributions of metereological data.	110
5.3 AERONET station for PTA-Rome.	111
5.4 ARPA station in PTA-Rome.	112
5.5 EC time series.	114
5.6 EC box plot.	114
5.7 OC time series.	115
5.8 OC box plot.	115
5.9 SO ₄ time series.	116
5.10 SO ₄ box plot.	117
5.11 NO ₃ time series.	117
5.12 NO ₃ box plot.	118
5.13 Dust time series.	118
5.14 Dust box plot.	119

List of Tables

3.1 AERONET stations	32
3.2 AOD products	33
3.3 Selected meteorological variables for AOD improved task.	34
3.4 Monitoring stations and pollution data	38
3.5 Meteorological variables by WRF ARPA used in the model.	41
3.6 List of LC classes mapped.	43
3.7 Building Use classification based on Urban Atlas Classes.	43
3.8 Urban variables.	44
4.1 Dataset for AOD improved task.	56
4.2 Table of Linear Regression relations	64
4.3 LOSO Cross Validation for AOD Prediction	71
4.4 LOYO Cross Validation for AOD Prediction	73
4.5 10-Folds Cross Validation for AOD Prediction	73
4.6 10-Fold CV comparison between Sentinel-3 and AOD improved	80
4.7 LOSO CV comparison between Sentinel-3 and AOD improved	80
4.8 LOYO CV comparison between Sentinel-3 and AOD improved	81
4.9 Optimized XGB and RF Cross Validation.	85
4.10 PM ₁₀ observed and predicted in the stage 3 concentrations ($\mu\text{g}/\text{m}^3$) (2019-2022)	87
4.11 Categorical variables for ARPA stations.	88
4.12 Evaluation of RF model for PM _{2.5} over four stations.	89
4.13 PM _{2.5} observed and predicted in the stage 3 concentrations ($\mu\text{g}/\text{m}^3$)(2019-2022)	92
4.14 XGB Cross Validation for PM ₂₅ predictions.	93
4.15 XGB evaluation metrix for PM _{2.5} metrix.	95
4.16 Annual PM _{2.5} concentrations [$\mu\text{g}/\text{m}^3$] : ARPA stations vs. spatial map averages (2019-2022)	103
5.1 Data collection for PM _{2.5} species estimations. Credits: Dave Diner and Sina Hasheminassab (JPL-NASA)	109
5.2 Monitoring stations ARPA in PTA-Rome.	113

5.3	Performance metrics for PM _{2.5} components by GRM. Credits: Liu Yang and Jin Zhihao (Emory University)	121
5.4	Comparison of model performance metrics	122

Lista degli Acronimi

WHO World Health Organization

IHME Institute for Health Metric and Evaluation

PM₁₀ Coarse particulate matter

PM_{2.5} Fine particulate matter

OC Organic Carbon

EC Elementar Carbon

SO₄ Sulfate

NO₃ Nitrate

UFP Ultrafine particle

ARPA Regional Environmental Protection Agencie

SNPA National Environmental Protection System

RRQA Rete Regionale di Monitoraggio della Qualità dell'Aria

MODIS Moderate Resolution Imaging Spectroradiometer

AOD Aerosol optical depth

MAIAC Multi-Angle Implementation of Atmospheric Correction

RH Relative Humidity

PBLH Planetary Boundary Layer Height

MLR Multiple Linear Regression

GAM Generalized Additive Models

GWR Geographically Weighted Regression

CTM Chemical Transport Models

MEM Mixed-Effect Models

SVMR Support Vector Machine Regression

MLP Multilayer Perceptron

RF Random Forest

MAIA Multi-Angle Imager for Aerosols

NASA National Aeronautics and Space Administration

ASI Italian Space Agency

JPL Jet Propulsion Laboratory

SCaN Space Communications and Navigation

PTA Primary Target Areas

STA Secondary Target Areas

BC Black Carbon

ASDC Atmospheric Science Data Center

IOC Initial Operational Capability

CVTA Calibration/Validation Target Area

TOO Targets of Opportunity

APEMAIA Assessment of PM Exposure at the intra-urban scale in preparation for the MAIA mission

IIA Institute of Atmospheric Pollution

HReBLOCK High Resolution Ensemble Block Level Of Cities Knowledge

AERONET AErosol RObotic NETwork

CMCC Euro-Mediterranean Centre on Climate Change

ESA European Space Agency

WRF Weather Research and Forecasting

NCAR National Center for Atmospheric Research

UA Urban Atlas

GBoost Gradient Boosting

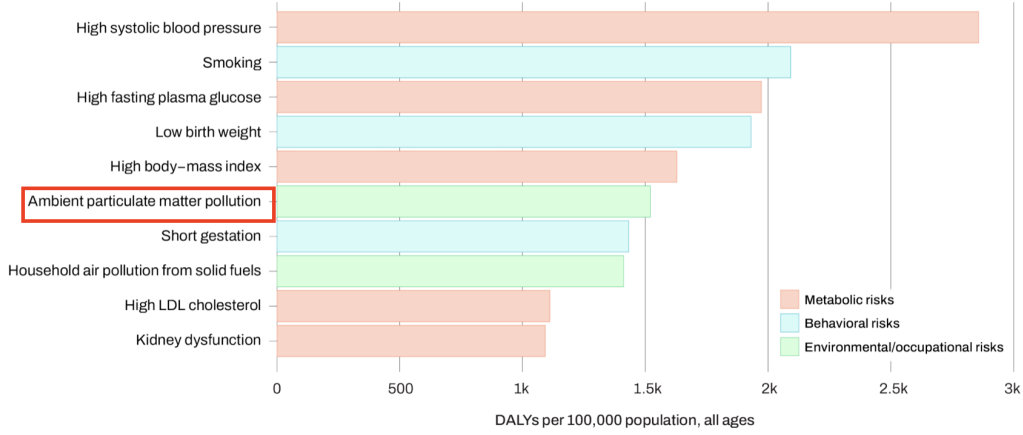
XGB Extra Gradient Boosting

Chapter 1

Introduction

Air pollution represents one of the most significant environmental challenges, with implications that go beyond the purely ecological. Scientific evidence accumulated over the last two decades has led to a radical revision of the understanding of human health impacts, bringing to the update of the World Health Organization (WHO) Global Air Quality Guidelines in September 2021. This update represents a fundamental turning point in the understanding of the phenomenon, highlighting how the harmful effects of air pollution manifest themselves even at concentration levels previously considered safe, also highlighting that, in 2019, 99% of the world's population resided in areas where air quality did not meet the recommended levels by the WHO guidelines [78]. This is further confirmed in 2021 data from the Institute for Health Metric and Evaluation (IHME) presented in the Global Burden of Disease Study [46], where *ambient particulate matter pollution* ranks among the main environmental risk factors [1.1]. Together with domestic solid fuel pollution, in fact, exposure to atmospheric particulate matter represents one of the greatest environmental risks to global public health, with a significant impact quantified at approximately 1500 DALYs (Disability-Adjusted Life Years) per 100000 inhabitants. This epidemiological evidence underlines the urgent need to develop innovative approaches for air quality monitoring and management.

Leading risk factors for disease burden, 2021

**Figure 1.1:** *Leading risk factors for disease burden, 2021, IHME*

The main sources of air pollution globally derive from the combustion of fossil fuels, including coal, diesel, petrol, oil and natural gas [82], mainly in the power generation, heating transport and industrial processes sectors. This issue is amplified by factors such as population growth, urbanization and industrial expansion. Current projections indicate that the world's population, starting at 8.2 billion in 2024, will continue to grow over the next 50 to 60 years to reach a peak of about 10.3 billion in the mid-2080s [75]. The health impacts of air pollution are comparable to other global health risks such as smoking and poor dietary habits [32]. The effects of contaminants vary by type, amount and duration of exposure, with scientific evidence linking air pollution to several serious diseases: lung cancer [40], cardiovascular disease [68] [84], impaired cognitive function and impaired emotional state [27]. This profound health impact is underscored by global estimates from 2019, which attribute *approximately* 4.2 million premature deaths worldwide to air pollution [96]. In addition to direct health effects, air pollution has significant socioeconomic consequences [101]. These include reduced physical activity in highly polluted areas (both during pollution peaks and as a general deterrent to outdoor activities [90]), premature mortality, negative impacts on social and educational outcomes, and contributions to climate change [87]. Research has also shown a significant correlation between air pollution and poverty, suggesting a negative feedback loop that amplifies existing socioeconomic inequalities [83].

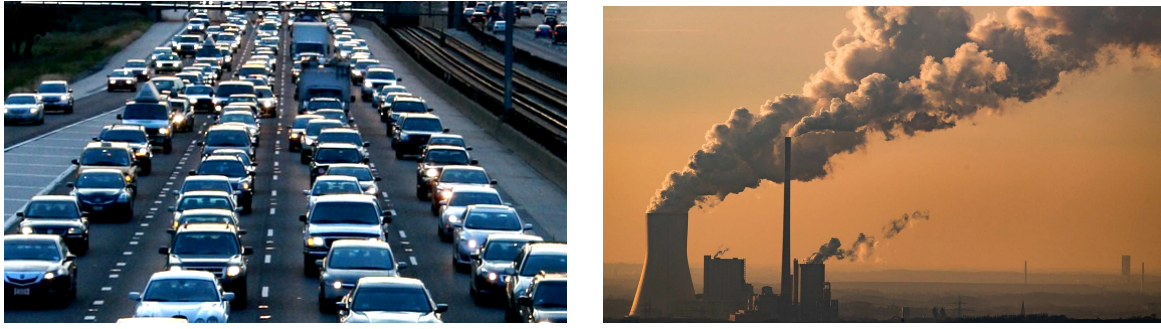


Figure 1.2: *Anthropic pollution.*

1.1 Atmospheric Pollutants

Atmospheric pollutants, particularly Particulate Matter (PM), can be classified according to different criteria. Aerodynamic diameter represents one of the most significant parameters, as it determines both the transport capacity in the atmosphere and the penetration capability into the respiratory system. In the European context, according to Directive 2008/50/EC on air quality [36], implemented in Italy by Legislative Decree 155/2010, and in line with WHO guidelines, two main dimensional categories are identified for particulate matter regulation: **coarse particulate matter** (PM_{10}), with aerodynamic diameter less than $10\ \mu m$, and **fine particulate matter** ($PM_{2.5}$), with aerodynamic diameter less than $2.5\ \mu m$.

The origin of atmospheric particulate matter can be traced to two main mechanisms: direct emission into the atmosphere (primary particulate) and conversion from gaseous precursors (secondary particulate). Gaseous precursors include sulfur dioxide, nitrogen oxides, ammonia, and non-methane volatile organic compounds, released from both anthropogenic and natural sources. Anthropogenic sources, which are highly variable, include combustion of solid fuels (coal, lignite, heavy oil, and biomass), industrial and agricultural activities, road surface erosion due to vehicular traffic, and wear of brakes and tires. Natural sources include volcanoes, sand storms, forest fires, vegetation, and sea spray.

A distinctive characteristic of fine particles is their relatively short persistence in the troposphere, with residence times that can extend from days to weeks and transport distances that can reach hundreds or thousands of kilometers. These residence times are strongly influenced by particle size, meteorological conditions, and removal processes. Smaller particles in the accumulation mode ($0.1 - 1\ \mu m$) can remain airborne longer due to their lower gravitational settling velocity, while removal occurs primarily through wet deposition (washout and rainout) and dry deposition processes. The atmospheric lifetime directly influences the spatial extent of population exposure: particles with longer residence times can be transported thousands of kilometers from

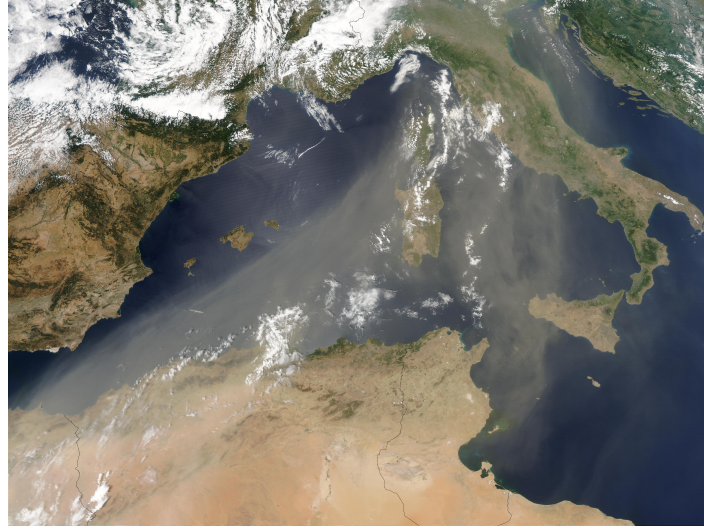


Figure 1.3: *Saharian dust over Italy.*

their sources, affecting populations far from emission points and contributing to the regional and intercontinental transport of air pollution. The concentrations of different PM fractions can undergo significant temporal fluctuations, both on a daily and hourly basis, depending on wind patterns and atmospheric stability [26]. The concentrations of different PM fractions can undergo significant temporal fluctuations, both on a daily and hourly basis, depending on wind patterns and atmospheric stability [49]. The chemical composition of atmospheric particulate matter is dominated by five main components, each with specific characteristics, origins, and impacts: organic carbon, elemental carbon, nitrates, sulfates, and mineral dust [Fig. 1.4].

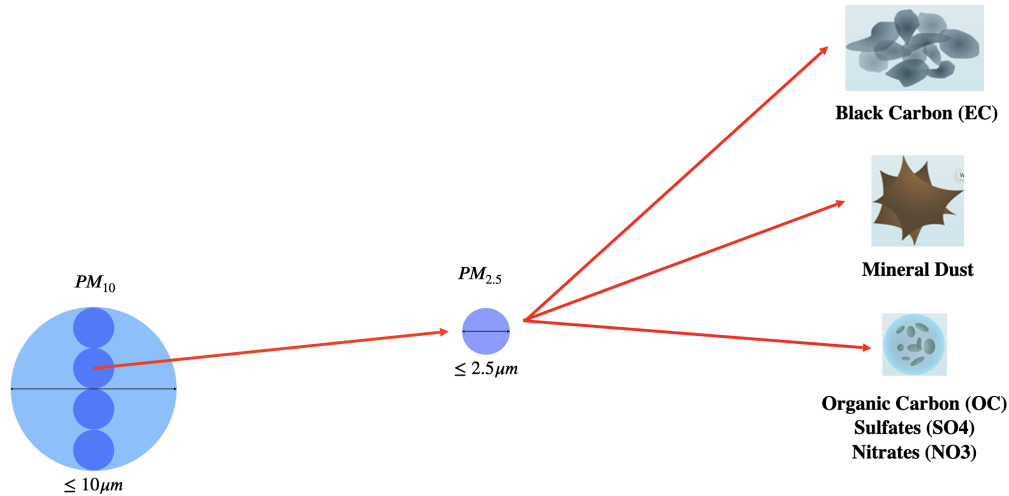


Figure 1.4: *PM species classification*

Among the carbonaceous components, **organic carbon** (OC) represents the fraction

that forms organic species through chemical bonds with other atoms (carbon, hydrogen, oxygen, sulfur, nitrogen, phosphorus, and chlorine). Its presence in particulate matter has a dual origin: it can derive from both primary emissions, coming from anthropogenic and natural sources, and secondary formation processes in the atmosphere through oxidation, condensation, or transformation of pre-existing compounds or gaseous precursors. **Elemental carbon** (EC), also known as black carbon or soot, is composed exclusively of carbon atoms not bound to other elements and is exclusively a primary pollutant, resulting from incomplete combustion processes, regardless of the type of fuel used.

Regarding ionic components, sulfates (SO_4^{2-}) and nitrates (NO_3^-) are both negatively charged ions but with distinctive characteristics. **Sulfates** present a relatively homogeneous spatial and seasonal distribution, to the extent that they are considered indicators of a super-regional pollution background, with both natural origins (volcanic eruptions and sea spray) and anthropogenic ones (mainly combustion in domestic systems and thermoelectric power plants). Nitrates, on the other hand, are secondary pollutants formed in the atmosphere from nitrogen oxides emitted by anthropogenic combustion sources (vehicular traffic, biomass burning, domestic heating, and industrial activities), and represent one of the most significant secondary components of $PM_{2.5}$, especially during the winter period [7]. The mineral component (**dust**), particularly relevant in the coarse fraction of PM_{10} , can have both natural (soil erosion, transport of Saharan dust, marine aerosol) and anthropogenic origin (construction activities, road asphalt wear, industrial activities), often representing the dominant mass fraction of PM_{10} . In the Italian context, the contribution of Saharan dust is of particular importance: the Mediterranean basin, and Italy in particular, is frequently affected by dust transport events from the Sahara desert, which can significantly influence PM_{10} levels especially in the central-southern regions and during the spring and summer seasons [15]. Beyond chemical composition, the size distribution of particulate matter plays a crucial role in determining health impacts. While $PM_{2.5}$ and PM_{10} standards are based on mass concentration ($\mu g/m^3$), this metric may inadequately represent the health risks posed by ultrafine particles (UFP, $< 100nm$ diameter) [54]. Despite contributing negligibly to total particle mass, UFP dominate particle number concentration and exhibit higher toxicity per unit mass due to their ability to penetrate deeper into the respiratory system, larger surface area per unit mass, and capacity to translocate to systemic circulation and organs including the brain. These ultrafine particles, primarily generated by combustion processes, are currently unregulated despite growing evidence of their independent health effects, highlighting a critical gap in current air quality standards that focus exclusively on mass-based metrics [85], [86]. The chemical speciation of particulate matter is of fundamental importance not only for source identification but especially for health impact assessment. Recent epidemiological evidence demonstrates that relying solely on total mass of PM_{10} and $PM_{2.5}$ as indicators of air pollution exposure can substantially underestimate the true health impact, as individual chemical species exhibit differential toxicity independent of their mass contribution [16]. Some components, for example, such as nitrates, elemental

carbon, and ammonium, have shown stronger associations with certain health effects compared to other PM species [17]. Furthermore, the emergence of particle number concentration as a complementary metric to mass concentration reveals that current mass-based standards may overlook the health risks posed by numerous UFP that contribute minimal mass but potentially significant toxicity. Understanding and predicting PM_{10} and $PM_{2.5}$ concentrations remain primary objectives in air quality management, given the established scientific evidence of their impacts on human health. However, the growing awareness that different chemical components of particulate matter can exert differentiated toxicological effects suggests the importance of also developing predictive capabilities for individual chemical species. This evolution toward component-specific and multi-metric approaches (mass, number, and chemical composition) would enable a more comprehensive assessment of air pollution health risks and support the development of targeted emission control strategies that prioritize the reduction of the most toxic components and sources.

This evolution in the approach to monitoring and predicting air pollution would allow not only more accurate assessment of public health risks but also the development of more targeted and effective mitigation strategies, based on reducing the most harmful components and controlling the emission sectors that produce them.

1.2 Monitoring Networks and Satellite Data

Given the evidence of air pollution impacts on human health and the environment, understanding how pollutant levels are monitored is fundamental. In Italy, this monitoring is carried out through a network of measurement stations managed by Regional Environmental Protection Agencies (ARPA) under the coordination of the National Environmental Protection System (SNPA) [48]. European regulations, through Decision 2001/752/EC and the document "Criteria for EUROAIRNET" [34], defined a methodological framework for classifying monitoring stations, aimed at ensuring representativeness and comparability of collected data [28].

The air quality monitoring network RRQA [1] is organized through a dual classification system that distinguishes stations by area type (urban, suburban, rural) and station type (industrial, traffic, background) to comprehensively capture different aspects of air pollution [35]. Traffic stations are strategically positioned near major roads to quantify the direct impact of vehicular emissions, characterized by high concentration gradients over reduced spatial scales, with their placement following precise geometric criteria to ensure representativeness of roadway pollution phenomena. Industrial stations are installed in areas influenced by production plants and industrial facilities to monitor the impact of industrial emissions on surrounding air quality, with their positioning considering emission source characteristics, prevailing wind patterns, and proximity to residential settlements. Background stations are designed to characterize baseline pollution levels not directly influenced by specific individual sources, measuring general air quality conditions across different spatial scales and allowing

discrimination between local contributions and long-range transport phenomena. This dual classification system enables comprehensive monitoring by combining area context with pollution source focus, providing both population exposure assessment and source-specific impact evaluation across different environmental settings, from densely populated metropolitan zones to rural areas with minimal anthropogenic interference.

This hierarchical structure of the monitoring system allows not only quantification of pollution levels in different environment types by providing direct and continuous measurements of air pollutant concentrations with high temporal frequency, but also comparison of measurements with current legislative limits. However, these networks have significant limitations in spatial coverage, providing point information that doesn't always capture the heterogeneity of pollutant distribution despite their high accuracy. To overcome these intrinsic limitations of traditional monitoring systems, the scientific community has increasingly focused on alternative and complementary observation methodologies in recent decades. **Remote sensing** data [39] is particularly relevant, not only for air pollution, but also for meteorological variables, land use, and geophysical parameters that can serve as predictors in pollutant concentration estimates. In this context, satellite remote sensing represents a fundamental tool in air quality monitoring, offering a complementary perspective to traditional ground monitoring networks. Integration between satellite observations and surface measurements proves essential for a complete understanding of air pollution dynamics.

Satellite remote sensing offers extensive and continuous spatial coverage, providing dense datasets with global coverage. Satellite observations, with resolutions ranging from 300 m to 50 km, provide valuable information on areas between monitoring stations, remote regions, oceans, and countries lacking dense monitoring networks. Among the most relevant satellite missions for air pollution monitoring, MODIS (Moderate Resolution Imaging Spectroradiometer) has represented a milestone, providing daily global observations since 2000. While MODIS imagery achieves spatial resolutions from 250 m to 1 km for different spectral bands, it should be noted that the aerosol optical depth (AOD) products derived from MODIS data are available at coarser spatial resolutions, typically 10 km for the standard Dark Target and Deep Blue algorithms [72]. The MAIAC (Multi-Angle Implementation of Atmospheric Correction) algorithm applied to MODIS data has allowed for more accurate estimates of AOD, a key parameter for estimating atmospheric particulate matter. The ESA's Copernicus program, through the Sentinel constellation, has further expanded observation capabilities: Sentinel-2 provides high-resolution images (10-60 m) useful for characterizing land use and surface conditions, while Sentinel-3 offers specific measurements for atmospheric and oceanic monitoring. Of particular interest is Sentinel-5P, launched in 2017, specifically designed for air quality monitoring with measurements of NO_2 , SO_2 , O_3 and other atmospheric pollutants with a spatial resolution of $7 \times 3.5 \text{ km}^2$.

1.3 Predicting PM: a review

The satellite remote sensing, with its ability to provide continuous data in space and time, combined with accurate point measurements from ground monitoring stations, has opened new perspectives in spatial modeling of atmospheric particulate matter. The integration of these diverse data sources, along with meteorological and land use variables, has led to the development of increasingly sophisticated modeling approaches for estimating PM concentrations across vast territorial areas. This leads us to analyze the state of the art methodologies developed for spatial prediction of atmospheric particulate matter, with particular attention to machine learning techniques that have demonstrated superior capabilities in managing complex relationships between the various variables involved.

Aerosol Optical Depth and PM

The use of satellite data for estimating atmospheric particulate matter began in the 1970s, but the real breakthrough occurred in 2003 when Wang and Christopher [95] introduced the use of AOD derived from the NASA MODIS sensor for PM prediction. AOD is a dimensionless physical quantity that quantifies the aerosol present in the atmospheric column in the vertical direction and is used to characterize the degree of atmospheric turbidity. Mathematically, at a given wavelength λ , the atmospheric AOD is represented by:

$$AOD_{\lambda} = \ln\left(\frac{\Phi_{e,\lambda}^i}{\Phi_{e,\lambda}^t}\right) = -\ln(T_{\lambda}) \quad (1.1)$$

where:

- $\Phi_{e,\lambda}^i$ is the spectral radiant flux in wavelength received by the atmosphere;
- $\Phi_{e,\lambda}^t$ is the spectral radiant flux in wavelength λ transmitted by the atmosphere;
- T_{λ} is the spectral transmittance in wavelength of the atmosphere.

AOD, therefore, measures the amount of light lost due to the presence of aerosols on a vertical path through the atmosphere. AOD values can take any positive value, but typically range from 0.01 indicating extremely clear conditions, to much higher values, such as 5.0, indicating extremely turbid atmospheric conditions [51].

Despite a correlation between AOD and PM, this relationship presents considerable complexity. Several meteorological factors significantly influence this relationship by affecting the conversion from columnar AOD to surface PM concentrations. For a given AOD value, variations in relative humidity (RH) can substantially alter surface PM concentrations through changes in aerosol extinction efficiency [38]. High RH increases aerosol extinction efficiency as hygroscopic particles grow through water uptake, meaning that the same AOD value can correspond to lower dry particle mass concentrations

at the surface [41]. Similarly, planetary boundary layer height (PBLH) plays a crucial role: for the same columnar AOD, a lower PBLH concentrates aerosols in a thinner atmospheric layer, resulting in higher surface PM concentrations, while a higher PBLH leads to more diluted surface concentrations. Although PM and AOD have different units, for $PM_{2.5}$ they are related to each other through the following equation

$$AOD = PM_{2.5} PBLH f(RH) \frac{3Q_{ext,dry}}{4\rho r_{eff}} = PM_{2.5} PBLH S \quad (1.2)$$

where $f(RH)$ is the ratio of ambient and dry extinction coefficients, ρ is aerosol mass density (g m^{-3}), $PBLH$ is the boundary layer height, $Q_{ext,dry}$ is the Mie extinction efficiency, and r_{eff} is the particle effective radius. S is the specific extinction efficiency ($\text{m}^2 \text{g}^{-1}$) of the aerosol at ambient RH [51]. Wind speed also plays an important role, altering this relationship by modifying the spatial distribution of aerosols [102]. Finally, geographical and seasonal variability plays an important role, requiring the implementation of locally calibrated models to adequately represent the regional peculiarities of this complex relationship [98].

1.3.1 Prediction methods

In the early years of research, approaches focused mainly on two methodologies: statistical models and deterministic numerical models. One of the most enduring statistical approaches is represented by *Multiple Linear Regression (MLR)*, which sets ground-measured PM as the dependent variable and AOD as the independent variable, integrating various meteorological factors as covariates [62]. Despite variable results, MLR has shown significant limitations in including covariates such as seasonal aerosol variation and land use. Subsequently, *Generalized Additive Models (GAM)* were introduced, which expanded the approach by including both meteorological and land use factors, and *Geographically Weighted Regression (GWR)*, which addresses the spatial variability in the relationships between AOD and $PM_{2.5}$, albeit with slower development.

On the deterministic front, *Chemical Transport Models (CTM)* [56] are based on solving differential equations to simulate the physical and chemical processes of atmospheric pollutants. A relevant aspect of CTM is its ability to operate even in the absence of ground monitoring data, proving effective on a global scale, although limited by the need for detailed emission data and high computational complexity. A significant methodological breakthrough is represented by *Mixed-Effect Models (MEM)*, developed in 2011 by Kloog et al. [50]. MEMs are characterized by their statistical structure, combining fixed effects and random effects to account for both population-level relationships and subject-specific variations. The introduction of MAIAC algorithms to MODIS products has further improved the performance of these models [57]. However, their accuracy strongly depends on the density of ground monitoring stations needed for spatial interpolation.

In the field of atmospheric particulate prediction, the methodological evolution has seen the emergence of increasingly sophisticated machine learning techniques that have

progressively overcome the limitations of traditional statistical models. In recent years, diverse machine learning modeling approaches have been developed to analyze and predict the evolution of PM_{10} and $\text{PM}_{2.5}$ [20], [11]. These approaches have demonstrated a greater capacity to capture non-linear relationships between satellite AOD data and ground-level PM_{10} and $\text{PM}_{2.5}$ concentrations.

The evolution of predictive techniques has led to the adoption of *Support Vector Machine Regression (SVMR)* for PM estimation, offering significant advantages over classical linear models. Introduced by Vapnik [94], SVMR distinguishes itself through its ability to manage non-linear relationships through the principle of structural risk minimization. This technique has demonstrated its superiority over multiple linear regression in processing MODIS AOD data and meteorological variables, thanks to the mapping of data into high-dimensionality feature spaces through kernel functions [21]. However, SVMR shows limitations with voluminous datasets and in parameter calibration, with particular reference to the choice of kernel and regularization coefficient.

The *Multilayer Perceptron (MLP)* architecture, fundamental in the field of artificial neural networks, has found wide application in PM prediction thanks to its flexibility and ability to model highly non-linear relationships between predictive and target variables [76], [69], [55]. MLP models have been implemented for their capacity to effectively integrate different types of data; they can simultaneously process satellite data, meteorological variables, and land use information, creating intermediate representations that capture the complexity of interactions between these factors. The optimal configuration of the MLP architecture remains a challenge, however. The number of hidden layers, the number of neurons per layer, and the activation function significantly influence the model’s performance.

Machine learning methods have recently become an integral part of modern air quality research, offering flexible and automated procedures for PM prediction based on past observations, while uncovering underlying patterns in complex environmental data and handling complex interactions between predictor variables. Among the various available machine learning approaches, *Random Forest (RF)* has proven to be particularly advantageous for PM concentration modeling due to several key features: the availability of easy-to-use open-source libraries designed to efficiently handle large environmental datasets; the simplicity of implementation; remarkable robustness to parameter specifications; and the inherent ability to capture nonlinearity and higher-order interactions between environmental predictors without a priori assumptions about functional relationships [23], [92], [100]. The ensemble nature of Random Forest, which combines numerous decision trees via bootstrap aggregation, provides exceptional stability and significantly reduces the risks of overfitting, while offering an intrinsic assessment of variable importance that improves model interpretability for environmental scientists and policy makers. Random Forest methods have been successfully applied to estimate PM exposure at different spatial scales, with notable applications including [19], who used RF approaches to estimate PM_{10} and $\text{PM}_{2.5}$ concentrations in China during 2005–2016, integrating AOD, meteorological parameters, and land use information as predictors. Similarly, spatiotemporal RF models have been developed for metropoli-

tan NO₂ exposure estimation in Japan [5] and PM forecasting in the United States [44], demonstrating the versatility and effectiveness of this approach across diverse geographic settings and pollution scenarios. The application of these methodologies has yielded significant results in different environmental contexts, each of which presents specific challenges and requirements. In regions with sparse monitoring networks, such as parts of Africa [6], ensemble techniques have successfully enabled PM_{2.5} concentration estimation in remote areas through sophisticated interpolation and extrapolation strategies. The Chinese context, characterized by high levels of pollution and rapid urbanization, has witnessed the development of advanced boosting algorithms integrating satellite observations with weather forecasts, focusing in particular on PM_{2.5} estimation in megacities, where traditional monitoring approaches have significant limitations [81]. At the European level, research efforts have focused on standardizing methodologies and exploiting open source data repositories, facilitating the development of harmonized frameworks for continental-scale air quality assessment and cross-border pollution monitoring initiatives [45].

Gradient Boosting (GBoost) algorithms are another powerful class of ensemble methods that have demonstrated outstanding performance in PM concentration forecasting due to their sequential learning approach. The GBoost methodology operates as an iterative algorithm composed of multiple decision trees that work together to obtain the final predictions, offering the ability to handle both linear and nonlinear regression problems. While in RF, where trees are built independently, gradient boosting builds models sequentially, with each new tree designed to correct the errors made by the previous ensemble, resulting in a powerful predictive framework that can capture complex environmental relationships with high accuracy. *Extreme Gradient Boosting (XGBoost)* has emerged as a particularly effective implementation of gradient boosting for air quality forecasting, representing what researchers describe as “enhanced gradient boosting” due to its increased efficiency and performance. XGBoost provides better control against overfitting thanks to a more regularized model formalization compared to traditional gradient boosting algorithms, while maintaining computational efficiency. Studies conducted in different geographical settings have demonstrated the superiority of XGBoost in PM forecasting tasks, with applications ranging from hourly forecasting of PM_{2.5} concentration in Chinese megacities such as Shanghai and Beijing to comprehensive air quality assessment in Tehran [67], [99], [22].

1.4 Thesis Structure

This thesis is organized into the following chapters:

- Chapter 2: Introduction to the JPL-NASA MAIA mission, a cutting-edge Earth observation initiative dedicated to the study of aerosols and atmospheric particulate matter, aimed at supporting epidemiological research on the relationship between pollution and human health. The chapter includes a brief mention of the preparatory research project APEMAIA, to which this thesis work is linked,

designed to develop and validate methodologies in preparation for the operational phase of the mission,

- Chapter 3: Presentation of the HReBLOCK framework, the methodological core of the research, with a description of its architecture, operating principles, and materials used for the different tasks,
- Chapter 4: Exploration of the practical applications of the framework, with presentation of the results obtained in improving AOD data and spatial predicting atmospheric particulate matter at intra-urban scale (pixel resolution of 300m),
- Chapter 5: Initial case study on the innovative field of PM_{2.5} speciation, with description of materials, methods and preliminary results of this promising research direction that will serve as foundation for future investigations,
- Chapter 6: discussion of the obtained results, their scientific and practical implications, conclusions, and future perspectives.

Chapter 2

Application Area

The need for increasingly sophisticated analytical tools to characterize air quality, together with the need to understand the composition and spatial distribution of air pollutants at the urban scale, has led to the development of new dedicated satellite missions. It is in this context that the MAIA (Multi-Angle Imager for Aerosols) mission fits in, representing a turning point in the remote sensing of atmospheric aerosols. This mission introduces an innovative approach that combines high-resolution multi-angle observation with advanced data processing techniques specifically designed to characterize the composition and distribution of atmospheric particulate matter in urban areas.

2.1 MAIA Mission

The MAIA Mission [12][74] represents a milestone in the history of space exploration, being the first NASA(National Aeronautics and Space Administration) mission whose primary objective is directly related to society's health. This mission also marks the first time that epidemiologists and public health researchers have been directly involved in the development of a satellite mission. Originally selected as a NASA mission, MAIA later evolved into a partnership with the Italian Space Agency (ASI). The project, scheduled to be launched by the end of 2026, now consists of a satellite (PLATiNO-2) provided by ASI and a scientific instrument developed at NASA's Jet Propulsion Laboratory in California (JPL). The mission will integrate data from three main sources: the space observatory, a network of ground-based sensors and atmospheric models. The synergy of data is aimed at understanding the health impacts of solid and liquid particles that contaminate the air

The importance of the MAIA mission lies in its ability to find a connection between environmental data and health records on births, deaths and hospitalizations to answer questions about the health impacts of atmospheric aerosols. These have been linked to a wide range of diseases, such as respiratory diseases, lung cancer, cardiovascular diseases and negative impacts on reproduction and pregnancy outcomes (premature

births, low birth weight, etc.).

The atmospheric particulate matter impact on these topics is strictly related to its size and compositional characteristics. Particles with a diameter of less than 10 micrometers PM_{10} are small enough to be inhaled, potentially causing tissue damage and inflammation in the nose, throat and lungs. Even more critical are particles with a diameter of less than 2.5 micrometers $PM_{2.5}$, which can penetrate deeper into the lungs and be absorbed into the bloodstream, where they can cause more serious health problems.

The composition of these particles varies according to their origin: black carbon comes from the combustion of fossil fuels and biomass, while mineral dust comes from soil and sand. Other particles - such as organic carbon, sulfates and nitrates - are formed through chemical reactions between gases in the atmosphere. The main objective of MAIA is to study whether exposure to these different types of particulate pollution leads to different health impacts.

2.1.1 Instrument

The observatory's scientific component consists of a steerable spectropolarimetric camera capable of acquiring digital images at different angles in the ultraviolet, visible, near-infrared and shortwave infrared regions of the electromagnetic spectrum. This multi-spectral capability will allow the MAIA team to explore the size, geographical distribution, composition and abundance of airborne particles, investigating their correlations with patterns and prevalence of health problems resulting from poor air quality. The mission is a joint effort in which responsibilities are distributed as follows:

- JPL, managed for NASA by Caltech in Pasadena, leads the US component of the project, providing the observatory science instrument and hosting the instrument operations centre
- ASI supplies the PLATiNO-2 satellite, contributes launch services and hosts the mission operations centre
- The NASA Space Communications and Navigation (SCaN) programme provides uplink and downlink services for commands and data
- The NASA Atmospheric Science Data Center provides the computational resources to generate and archive science products

To achieve these scientific goals, the MAIA instrument was designed with cutting-edge technical features that make it unique in the panorama of atmospheric remote sensing instrumentation. Operating at an altitude of 740 kilometers above the earth's surface, the property of the instrument is to discriminate different types of atmospheric particles by analysing their optical properties, i.e. the way the particles reflect or absorb

sunlight. This gives the chance to distinguish between different types of microscopic particles, such as sulfate particles and dust particles of similar size.

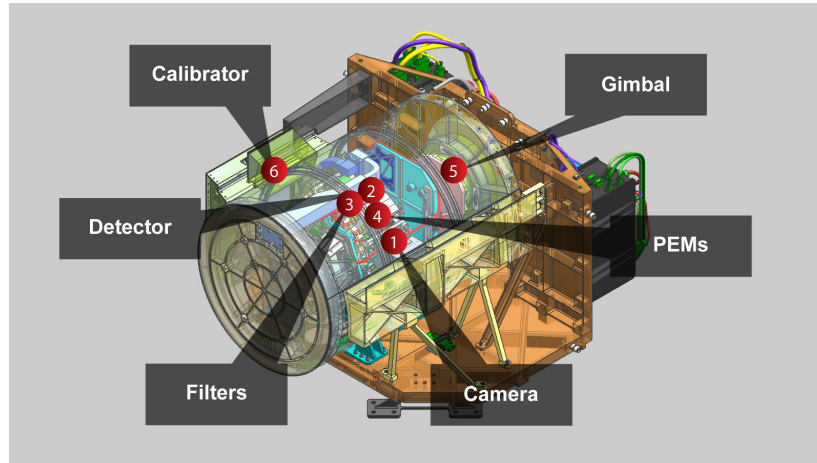


Figure 2.1: *Maia Digital Camera parts*

The architecture of the instrument is based on a specialized **digital camera**, designed with specific characteristics for observing atmospheric particulate matter. The optical system employs a configuration based on polished aluminum mirrors, rather than traditional glass lenses.

The **filtration system** represents one of the most innovative elements of the instrument. MAIA is equipped with 14 specific filters, a multi-spectral configuration that has been designed to exploit the principle of light scattering by atmospheric particles: the spectral bands in the visible range provide information on smaller particles ($PM_{2.5}$ and below), while the short-wave infrared bands are sensitive to larger aerosols, such as dust and volcanic ash. The ultraviolet wavelengths are particularly useful for detecting the absorption of sunlight by particles containing specific minerals and organic matter.

A key element of the MAIA instrument is the **polarization modulation system**, which utilizes photoelastic modulators (PEMs) in combination with quarter wave plates (QWPs). This system is based on the principle that sunlight becomes polarized when scattered by atmospheric particles, and the degree and direction of this polarization depend on the physical characteristics of the particles themselves, such as size and shape. Through the combined use of PEMs, QWPs, and polarizing filters positioned above the camera detectors, MAIA is capable of making accurate measurements of the degree of polarization of incoming light caused by atmospheric particles.

The precision of measurements is guaranteed by an onboard calibration system. The **calibrator** integrates a series of metallic grid polarizers illuminated by sunlight passing through a diffuser screen. Routine observation of the calibrator allows the MAIA team to correct any variations in the camera's sensitivity to polarization. The system also includes a dark target that enables the measurement and subsequent subtraction of the background signal generated by the detectors in the absence of light, thus ensuring the accuracy of the measurements.

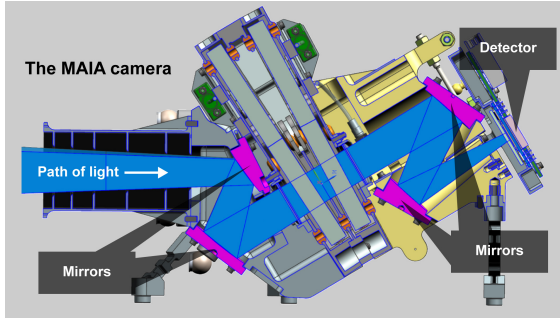
From a data acquisition perspective, the instrument implements a **pushbroom technology**, where the detectors are arranged in single rows rather than rectangular arrays as in conventional cameras. This configuration, combined with the satellite's movement at approximately 25000 km/h, allows for image acquisition through a progressive scanning technique, similar to that used in scanners.

The **motorized gimbal system** allows the camera to rotate 60 degrees both forward and backward. This capability enables implementation of the *step and stare* technique, through which the instrument can observe the same target from multiple angles during the satellite's passage. This multi-angle acquisition mode is crucial for obtaining more detailed information about the size and shape of atmospheric particles, leveraging the principle that the atmosphere appears different when observed from different angles. The gimbal can also orient laterally, allowing observation of urban areas not directly under the satellite's track. The ASI's PLATiNO-2 satellite platform, which hosts the MAIA instrument, provides the necessary power, navigation information, receives uplink commands, and transmits the scientific and engineering data of the instrument to the ground. The satellite will be positioned in a low Earth polar sun-synchronous orbit, passing near the north and south poles in each orbit. This orbital configuration ensures that each target is observed approximately at the same local time during each satellite pass, with a designated equator crossing time in mid-morning and an orbital period of approximately 100 minutes.

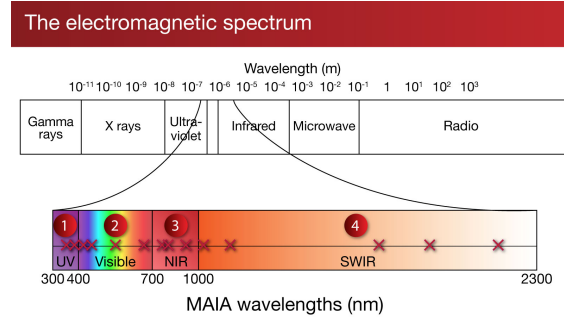
2.1.2 Target Area

The MAIA mission operates across four types of target areas, each with specific objectives and distinctive characteristics. The Primary Target Areas (PTAs) represent the mission's main focus, including large metropolitan areas where MAIA epidemiologists will study the impacts of various types of atmospheric particulate matter on human health. The selection of the 12 PTAs is based on criteria that include:

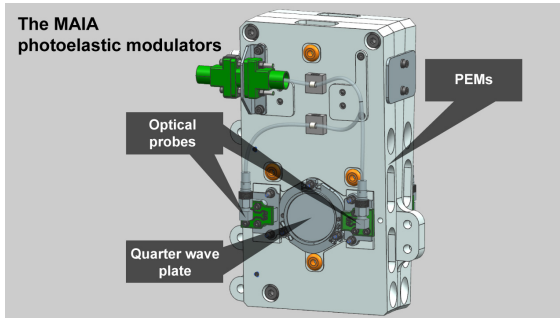
- the size of the population exposed to air pollution;
- the availability of health registries accessible to MAIA epidemiologists;
- the characteristics of air pollution, with particular attention to areas where the quantity and types of particulate matter vary spatially and/or seasonally;
- the presence of ground-based air monitoring networks;
- cloud cover conditions, which could limit satellite observations;
- the geographical location, which must be compatible with the MAIA satellite's orbit.



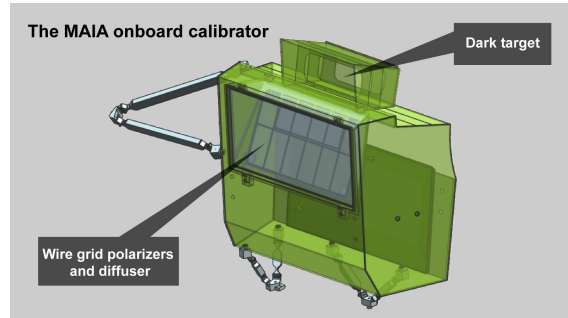
(a) MAIA camera with light path and mirrors



(b) Electromagnetic spectrum and MAIA wavelengths



(c) MAIA polarization modulation system



(d) MAIA onboard calibrator

Figure 2.2: Main components of the MAIA instrument: (a) optical system, (b) operational spectral range, (c) polarization modulation system, (d) calibration system.



Figure 2.3: Target area defined for MAIA Mission

In addition to the PTAs, the mission identifies three further categories of target areas:

- Secondary Target Areas (STAs): areas of additional scientific interest, not limited to air quality research, where Level 1 radiance and polarization images will be generated;
- Calibration/Validation Target Areas (CVTAs): routine observation zones to verify and maintain the accuracy of instrument measurements and data products;
- Targets of Opportunity (TOO): episodic events such as large wildfires and volcanic eruptions that affect air quality, with observations planned based on the occurrence of the event.

2.1.3 Monitoring Network

To ensure accurate conversion of satellite data into PM concentrations near the surface, the MAIA project utilizes a **monitoring network** on the ground. This network integrates data from both existing PM measurement stations, managed by government agencies and research groups, as well as additional instruments strategically positioned in PTAs where surface measurement coverage is limited.

A key element of this network is the SPARTAN (Surface Particulate Matter Network), which uses filter-based AirPhoton samplers operated by Washington University for $\text{PM}_{2.5}$ speciation. The MAIA project has expanded this network through the installation of supplementary samplers. These instruments collect atmospheric PM on Teflon filters, which are subsequently analyzed in laboratories to determine the concentrations of different chemical constituents in the particulate matter. The monitoring network has been further strengthened through the implementation of various types of instrumentation:

- AMOD (Aerosol Mass and Optical Depth) samplers, developed by Colorado State University (CSU) as part of a citizen science project, progressively replaced by the modified version AMOS (Aerosol Mass Outdoor Sampler)
- AethLabs microAeth MA350 monitors, installed in some PTAs to provide black carbon (BC) measurements, calibrated against co-located filter-based instruments to convert BC concentrations to EC values
- Low-cost PurpleAir sensors, particularly used in areas with poor historical $\text{PM}_{2.5}$ monitoring coverage, such as Ethiopia, where they are calibrated through co-location with reference instruments managed by the US State Department in Addis Abeba

	Existing Networks		MAIA Deployed Monitors			
						
Instruments	Various		SS5i- $\text{PM}_{2.5}$ filter-based sampler	Aerosol Mass Outdoor Sampler (AMOS)	microAeth MA350	PA-II-SD
Supplier	Various		AirPhoton (SPARTAN network)	Colorado State University	AethLabs	PurpleAir
Pollutants	Total $\text{PM}_{2.5}/\text{PM}_{10}$	$\text{PM}_{2.5}$ speciation	$\text{PM}_{2.5}$ speciation	$\text{PM}_{2.5}$ speciation	Black carbon	Total $\text{PM}_{2.5}$
Locations	All primary target areas ($\text{PM}_{2.5}$ only in Ethiopia)	Spain, Italy, South Korea, USA	Chile, Ethiopia, India, Israel, Taiwan, South Africa, USA	Ethiopia, India, Taiwan, South Africa, USA	Ethiopia, India USA	Ethiopia

Figure 2.4: MAIA Project Surface Monitoring Network

2.2 Structure and Management of MAIA Data Products

The management and distribution of MAIA mission data products is organized through a well-defined architecture[73]. The main storage and distribution system is implemented at NASA's Atmospheric Science Data Center (ASDC), with a symmetric replica managed by ASI. The products are generated in NetCDF format, ensuring metadata compliance with international standards ISO-3166-1, ISO-8601, and the MENDS protocol (Metadata Evolution for NASA Data Systems).

The MAIA data processing is structured in four main phases:

- **Conversion from Raw Data to Radiance:** Raw telemetry data are converted into radiance (brightness) measurements for each of the instrument's 14 spectral bands. For three of these bands, the degree and orientation of polarization are also calculated. This information constitutes the **Level 1 product** and forms the basis for generating calibrated images;
- **Aerosol Data Processing:** Through specific algorithms, radiance and polarization measurements are used to calculate AOD and other optical and physical characteristics of atmospheric pollution. AOD represents a dimensionless measure of the amount of sunlight blocked by particles in the atmosphere. These parameters are stored in the **Level 2 Aerosol product**.
- **Conversion from Aerosol to PM:** Although the MAIA instrument observes particles at all atmospheric levels, the impact on human health is primarily determined by particulate matter present in the lower layers of the atmosphere. For this reason, a conversion strategy is implemented that combines AOD and other aerosol properties with data from ground pollution monitors and meteorological and environmental information to derive concentrations of different types of PM. These data constitute the **Level 2 PM product**.
- **Generation of Gap-filled Maps:** Since the MAIA instrument cannot observe each PTA daily, a gap-filling procedure is implemented that integrates satellite data with outputs from numerical models and measurements from surface monitoring stations. This process allows for the generation of complete daily maps of average concentrations of different types of PM, which constitute the **Level 4 product**.

The hierarchical structure of data products is organized into different processing levels, each characterized by specific spatial, temporal, and qualitative properties.

2.2.1 Level 1 Products

The Level 1 Georectified Imagery constitutes the first processing level, providing map-projected radiance and polarization data, along with solar and viewing geometry and geographic coordinates. The spatial resolution of 250 m and the bounding box extent of 1000 km x 1000 km for acquisitions in step and stare mode (extended to 1000 km x 3000 km for sweep mode) ensure extensive territorial coverage. The processing timeline includes a latency of approximately 12 weeks from telemetry receipt, with the first public release planned within 6 months of Initial Operational Capability (IOC).

2.2.2 Level 2 Products

Level 2 processing is divided into three distinct subproducts:

- **Cloud Mask Product:** implements discrimination between cloudy and clear areas, with a spatial resolution of 1 km over a 360 km x 480 km bounding box. The temporal sampling frequency is synchronized with target overpasses, with a processing latency of 8 weeks from the completion of Level 1 processing.
- **Aerosol Product:** quantitatively characterizes the optical and physical properties of aerosols, maintaining the same spatial resolution and bounding box as the Cloud Mask Product. The processing methodology follows protocols similar to the previous product.
- **PM Product:** provides quantitative characterization of atmospheric particulate matter concentrations, distinguishing between coarse and fine fractions and analyzing specific components of fine PM. Processing maintains the spatial resolution of previous Level 2 products, implementing 24-hour temporal averages on target overpass days.

2.2.3 Level 4 Products

The Level 4 PM Product represents the highest level of processing, providing spatially and temporally continuous atmospheric particulate concentration fields obtained through sophisticated gap-filling techniques. The spatial resolution of 1 km is maintained over a 360 km x 480 km bounding box, with daily average processing. Public release is planned within 18 months of IOC, with a latency of 8 weeks from the completion of Level 2 processing.

2.2.4 Ancillary Products

The product system is completed by two fundamental ancillary datasets:

- **Geographic Product:** provides detailed land use information with variable resolution (125 m - 1 km) over a 1000 km x 1000 km bounding box, made available in the pre-launch phase.
- **Surface Monitor Product:** integrates PM data from surface measurements at specific sites, with differentiated latencies of 1 week for total PM and 6 months for component analysis.

The territorial implementation of products follows a hierarchical strategy based on target area type. Complete coverage is guaranteed for PTAs and selected STAs, while for the remaining STAs, CVTAs, and TOO, product availability is selective.

The geometric structure of acquisition includes a system of nested bounding boxes centered on target coordinates. The atmospheric CTM, functional for generating meteorological data and PM data gap-filling, operates on a minimum domain of 1080 km x 1140 km. Visualization of bounding boxes is implemented through the Geographic Information Visualization Tool (GIVT), with different extensions based on product level and target area type.

It's important to note that effective coverage for each satellite pass may differ from the nominal extent of bounding boxes, depending on along-track and cross-track acquisition geometry. Oblique acquisitions generate more extensive coverage of the Level 1 bounding box compared to nadir acquisitions. Level 2 and 4 products derive from the intersection of areas observed at different step and stare angles, with the extension determined by the least oblique view.

In accordance with strict health data privacy protocols, although peer-reviewed epidemiological studies conducted by the MAIA scientific team constitute an integral part of the mission's products, maximum confidentiality of associated health records is guaranteed.

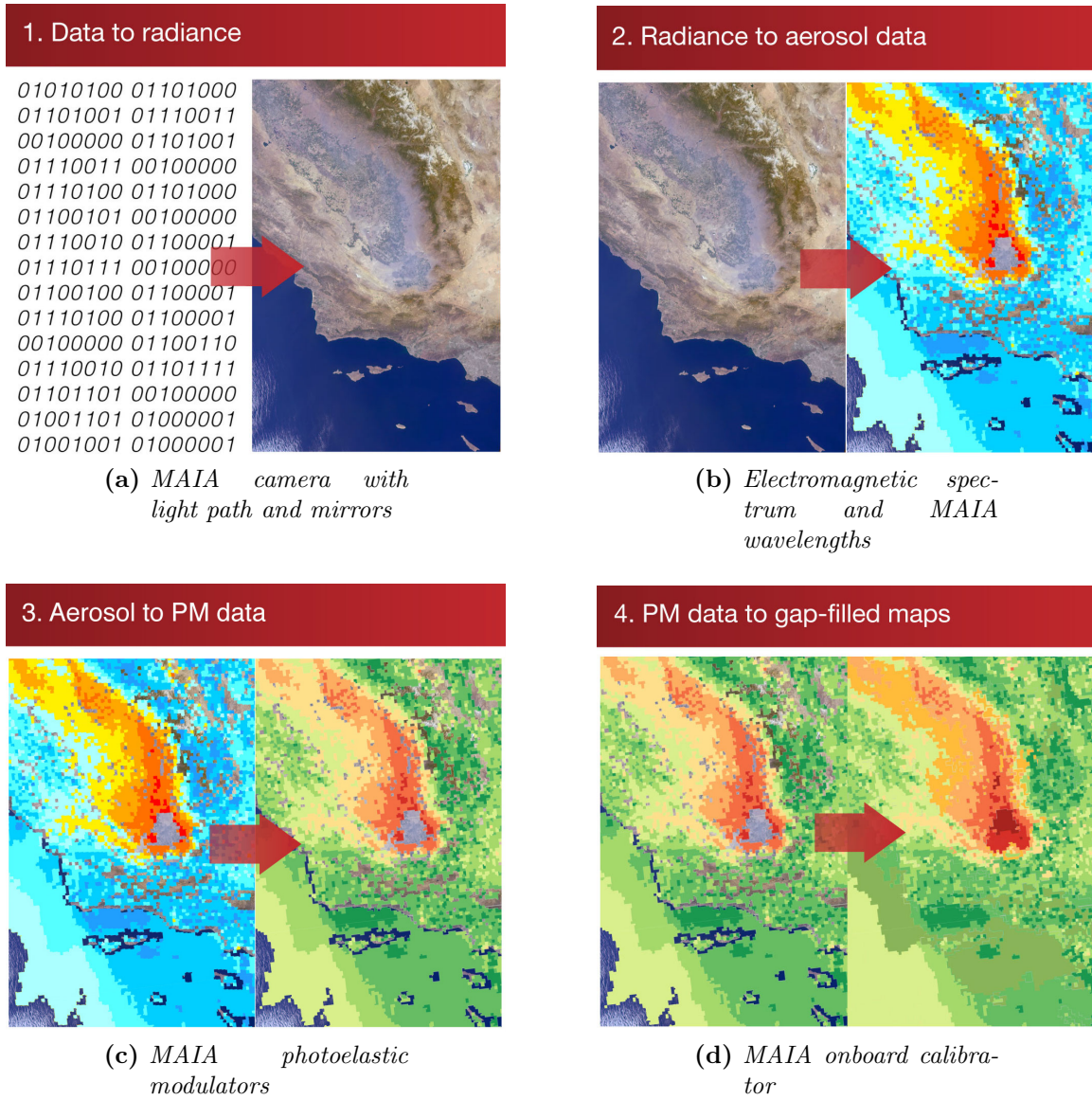


Figure 2.5: MAIA Simulated Data Products.

2.3 APEMAIA project

In preparation for the MAIA mission, the ASI has funded the APEMAIA project (Assessment of PM Exposure at the intra-urban scale in preparation for the MAIA mission), coordinated by the group of Earth Observation of the Institute of Atmospheric Pollution of Bari (IIA-CNR) in collaboration with the Physics Department of the University of Bari and ARPA Puglia. The project is configured as a preparatory initiative aimed at developing and validating methodologies for the characterization of atmospheric particulate matter at an intraurban scale. With particular focus on the metropolitan area of Bari as the main study site, APEMAIA aims to generate high spatial resolution maps of PM_{10} and $PM_{2.5}$ concentrations, along with the quantification of atmospheric pollution exposure indices. This project represents a testbed for the development of methodologies in the field of remote sensing of atmospheric pollution, laying the groundwork for optimal use of future MAIA data at the urban scale.



Figure 2.6: *APEMAIA project logo.*

Chapter 3

Methodology

The work of this research is structured around two main tasks, both aimed at improving the estimation of atmospheric particulate matter. The tasks addressed are:

- The **first task** focuses on improving AOD estimates at the fine spatial scale. Given the importance of AOD as a predictor in the estimation of atmospheric PM, and considering the goal of mapping PM at an intraurban scale with sub-kilometer resolution (thus overcoming the limitations of current applications in the literature), the decision was made to exploit Sentinel-3 satellite data, which offers a resolution of 300 meters. This task involves the development of a machine learning model that integrates meteorological data and several established AOD satellite products to generate improved AOD estimates at 300 meter resolution.
- The **second task** focuses on the prediction of PM_{10} and $PM_{2.5}$ concentrations with a machine learning model that uses as main predictors the improved AOD estimates obtained in the first task, integrated with other meteorological variables and ancillary data. The objective is to obtain high spatial resolution PM concentration maps capable of capturing intraurban variability in air pollution.

For these two interconnected tasks, respectively, the sections [3.2.1](#) and [3.3.3](#) describe the areas of study and the input characteristics considered for each task; while the "Machine Learning technique" section [3.4](#) collects the main notions of the machine learning models employed; and finally the "Applications and results" chapter [4](#) demonstrates the practical implementation and the results of these models.

In addition to these two interconnected objectives, this research has also explored an independent line of investigation during the international research period.

- The **third task** concerns the speciation of particulate matter within the PTA of the MAIA mission. This work focuses specifically on predicting key particulate matter components: *EC*, *OC*, *NO₃*, *SO₄*, and dust. This work represents a preliminary exploration that lays the foundation for future research on the characterization of particulate matter composition using advanced remote sensing

techniques and machine learning approaches. This third task, given its distinct focus and methodology, is presented in a separate dedicated chapter 5 that follows the same structural organization: materials, methods, and results specific to the particulate matter speciation investigation.

Despite their different objectives and implementations, all three tasks share a common methodological approach through the HReBLOCK framework (High Resolution Ensemble Block Level Of Cities Knowledge), which provides a unified structure for data processing and machine learning application. This framework is described in detail in the following section.

3.1 HReBLOCK framework

For all established tasks, the methodological approach of this research is based on the **HReBLOCK** (High Resolution Ensemble Block Level Of Cities Knowledge) framework. As shown in Figure 3.1, this framework is structured into two main phases:

- **Data preparation:** Data preparation: involves the acquisition and integration of satellite images from different geographic areas, with particular attention to the standardization of spatial resolutions based on the reference grid selected for this research work. The acquired satellite data together with the georeferenced variables are subsequently organized into three-dimensional blocks representing the study areas, allowing a complete spatial characterization of the observed phenomenon. The process is completed with the transformation of these blocks into high spatial resolution feature matrices, thus creating a structured and homogeneous representation of the data for subsequent analyses.
- **Learning model implementation:** initially, validation involves the selection of the most relevant features, thus reducing and optimizing the dimensionality of the problem. Selected and parameterized machine learning models are then applied to the extracted dataset. The evaluation and validation of the system includes a daily assessment of the results produced, through the analysis of different metrics that are based on the comparison with reference targets measured by ground monitoring stations (ground truth measurements).

This general architecture of the HReBLOCK framework constitutes the core methodological structure of the entire research work. Figure 3.2 shows the interconnection between Task 1 (AOD improvement) and Task 2 (PM prediction), highlighting the data flow and functional dependency between them.

- **Task 1,** focused on improving AOD (Aerosol Optical Depth) data, utilizing the HRe-BLOCK framework to enhance the quality and accuracy of satellite-derived aerosol optical depth measurements to obtain "improved AOD" products. This process integrates various input sources, including meteorological data from CMCC,

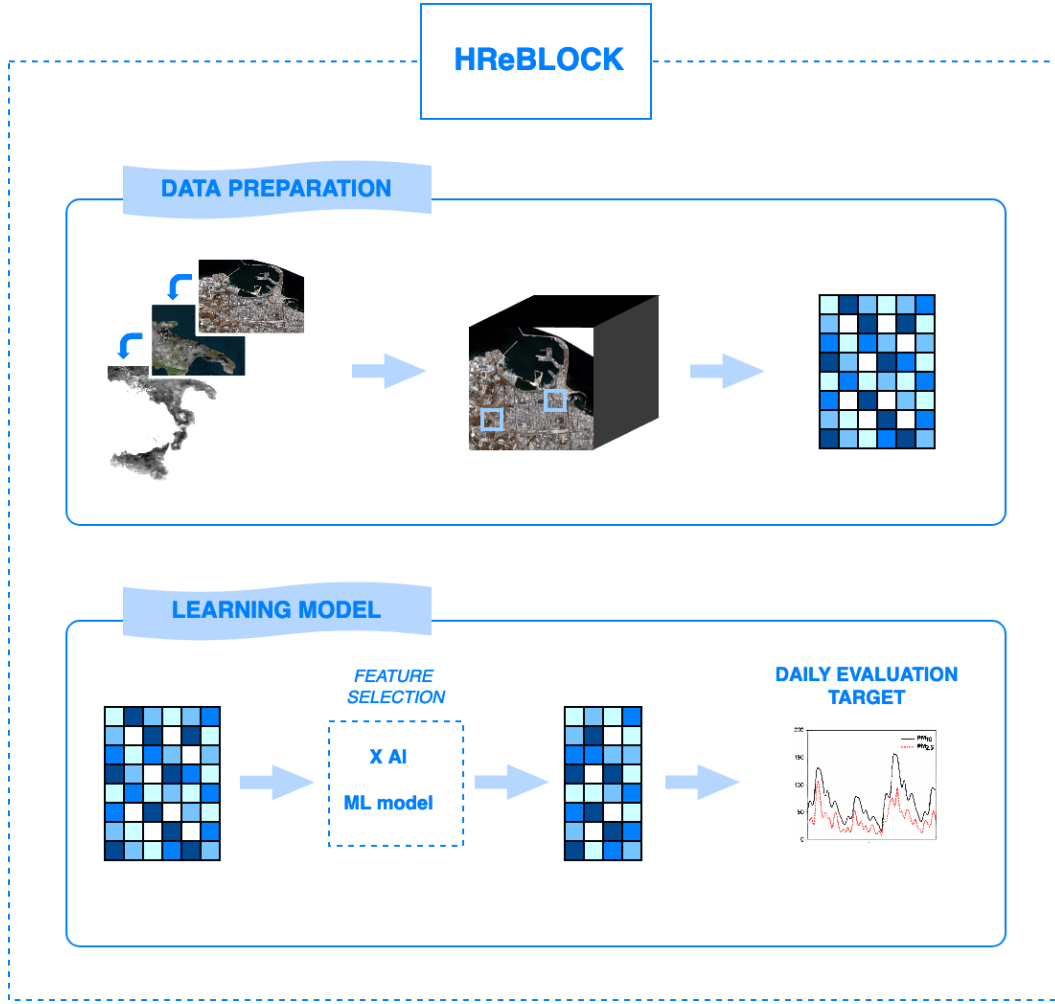


Figure 3.1: *HReBLOCK framework.*

satellite observations from Sentinel-3A and Sentinel-3B sensors (SYN3-A and SYN3-B products), as well as aerosol retrievals from the MAIAC system. The specific characteristics and processing methods for each input dataset will be detailed in the subsequent methodology sections,

- **Task 2**, centered on the prediction of atmospheric PM, builds directly on the results obtained from Task 1, using the improved AOD generated by the previous model as primary input. In addition to the improved AOD, the second model analyzes meteorological data derived from WRF simulations provided by ARPA, vegetation indices, and a set of ancillary and urban variables whose specifications will be shown in subsequent sections.

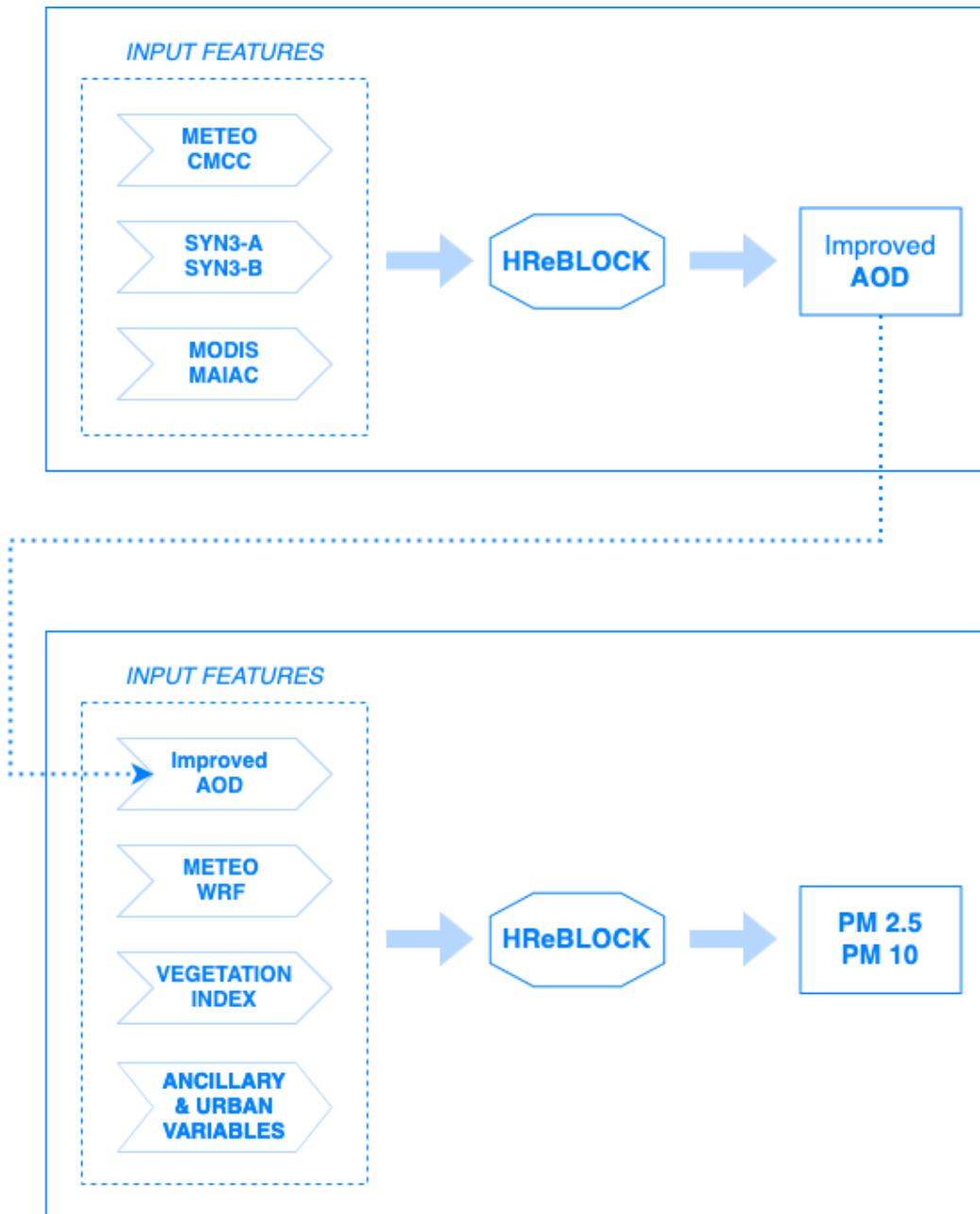


Figure 3.2: Task 1 and task 2 framework.

3.2 Improved AOD methodology

The following subsections detail the materials utilized in the "improved AOD" task, including satellite products, ground-based reference measurements, and meteorological datasets that will be integrated and processed by the HReBLOCK framework.

3.2.1 Materials

The significant correlation between PM and AOD has been widely validated in the scientific literature, confirming the role of AOD as one of the most relevant features in air quality predictive models [37]. This study aims to map PM at intra-urban scale with high spatial resolution (less than one kilometer). To achieve this goal, the research exploits the high resolution provided by the SYNERGY sensors on board the Sentinel-3 satellite, which offers AOD data with a spatial resolution of 300 meters. However, considering that Sentinel-3 represents a relatively new mission, it was considered appropriate to complement these data with AOD products derived from more established satellite missions. The construction of the training dataset is a crucial step of this study. To ensure high-quality training data, points corresponding to the AERONET (AErosol RObotic NETwork) monitoring stations were selected. These stations provide high-precision ground-level AOD measurements, which constitute the ground truth for training and validating predictive models. The dataset construction procedure involved the extraction and temporal synchronization of the variables of interest to the geographic coordinates of the AERONET stations, thus ensuring a reliable basis for the comparison between satellite estimates and ground measurements. The developed methodology involves the implementation of a machine learning model that integrates three fundamental components:

1. the high spatial resolution offered by Sentinel-3 SYNERGY data
2. the increased stability and temporal reliability of the AOD products derived from the NASA MODIS MAIAC satellite mission
3. a selected set of meteorological variables that influence the distribution and concentration of atmospheric aerosols.

Data were collected over 4 years, from January 2019 to December 2022.

AERONET AOD data

AERONET represents a federation of ground-based remote sensing networks dedicated to monitoring atmospheric aerosols. Established and managed by NASA's Goddard Space Flight Center, and operational for over 25 years, AERONET has provided a continuous and easily accessible public database of optical, microphysical, and radiative properties of aerosols, supporting aerosol research, characterization of their properties,

validation of satellite data, and synergy with other databases. The AERONET network is extensively documented on the official website (<https://aeronet.gsfc.nasa.gov/>) and in scientific literature through the works of Holben et al. (1988) and Giles et al. (2019). AERONET's ground-based solar photometers directly measure the attenuation of solar radiation without interference from Earth's surface reflections, providing accurate measurements of AOD with an uncertainty of approximately 0.01-0.02 in the spectral range of 340-1640 nm.

For the purpose of AOD validation, the AERONET dataset was utilized, which is widely acknowledged as the standard for assessing the precision of AOD satellite products. These products are accessible at the 500nm wavelength. It should be noted that AERONET does not possess direct measurements at this wavelength. To facilitate a direct and consistent comparison, the Ångström equation is applied, as demonstrated in this section.

AERONET AOD data are available in three quality levels: level 1 represents the original, unprocessed data, level 1.5 includes data processed through cloud masking algorithms, and level 2 includes data that have undergone both cloud masking and rigorous quality control. In this study, level 2 data spanning the period from 2019 to 2022 were utilized for seven AERONET stations dispersed throughout Italy, as detailed in Table 3.1

STATION	COORDINATE	WAVELENGTH	ANGSTROM EXPONENT
Lecce	(40.33,18.11)	500	500–870_Angstrom_Exponent
Foggia	(41.46,15.50)	500	500–870_Angstrom_Exponent
Potenza	(40.60,15.72)	500	500–870_Angstrom_Exponent
Roma La Sapienza	(41.90,12.51)	500	500–870_Angstrom_Exponent
Roma Tor Vergata	(41.83, 12.64)	500	500–870_Angstrom_Exponent
Ispra	(45.80,8.62)	490	440–675_Angstrom_Exponent
Modena	(44.63,10.94)	500	500–870_Angstrom_Exponent

Table 3.1: *AERONET stations*

AOD satellite products

Through a literature review, the satellite products listed in Table 3.2 were selected for the measurement of AOD. The selection focused in particular on the Sentinel-3 products (S3A and S3B SY 2 SYN), relatively new on the market and characterized by a high spatial resolution of 300 m, in combination with the well-established MODIS MAIAC product (MCD19A2). In this study, it was deliberately chosen to use direct Earth observation products rather than reanalysis products such as MERRA-2 or CAMS.

PLATFORM PRODUCT				PROC. LEVEL	SPATIAL RESOLU- TION	TEMPORAL RESOLU- TION	ACQUISTI ON TIME	BAND
Sentinel3	S3A	SY	2	Level 2	300m	5 images per week	8-10 UTC	550 nm
Sentinel3	S3B	SY	2	Level 2	300m	5 images per week	8-10 UTC	550nm
Modis	MCD19A2			Level 2	1km	daily	9-11 UTC	550nm

Table 3.2: *AOD products*

A significant advancement in satellite-based aerosol monitoring is the MAIAC algorithm, which processes time series measurements acquired by the Aqua and Terra sensors aboard the MODIS satellite platform. This algorithm produces AOD data at a spatial resolution of 1 km, representing a substantial improvement over previous products and enabling high-resolution mapping of $PM_{2.5}$, particularly valuable for urban areas ([66]). This study employs the MCD19A2 Version 6 MAIAC data product for $PM_{2.5}$ estimation. The MCD19A2 product is a Level 2 dataset with daily availability, collected between 9:00-11:00 UTC. It provides AOD measurements at the 550 nm wavelength band, ensuring compatibility and comparability with other aerosol products.

The products S3A SY_2_SYN and S3B SY_2_SYN from the Sentinel-3 mission, part of the European Space Agency (ESA) Copernicus program, represent one of the most recent datasets for monitoring atmospheric aerosols. These products are generated through the synergistic combination of data from two instruments: OLCI (Ocean and Land Colour Instrument) and SLSTR (Sea and Land Surface Temperature Radiometer). The SY_2_SYN product provides surface directional reflectance and aerosol parameters over land at 300-meter spatial resolution, derived from OLCI's native sampling distance of 0.3 km for terrestrial observations and SLSTR's 500-meter resolution in the reflectance channels. This high spatial resolution makes the product particularly suitable for regional and local-scale atmospheric studies, offering detailed aerosol monitoring capabilities, although it may exhibit greater variability in AOD estimates

compared to spatially aggregated products that are processed at coarser resolutions for improved signal-to-noise ratios.

Meteorological data

The dataset has been enriched with high-resolution meteorological data from the Euro-Mediterranean Centre on Climate Change (CMCC). These data represent a downscaling of the ERA-5 dataset, with a refined spatial resolution of 2.2 km - a substantial improvement over the native resolution of ERA-5. For the construction of the training dataset, the integration with satellite data and AERONET measurements was done by extracting the values of meteorological variables corresponding to the geographical coordinates of the AERONET stations included in the study. Meteorological parameters significantly influence the distribution and properties of atmospheric aerosols. The height of the planetary boundary layer, for example, determines the volume within which aerosols disperse, while precipitation, humidity and wind influence the removal processes, hygroscopic properties and transport mechanisms of aerosols, respectively. Table 3.3 provides the complete list of meteorological variables integrated in the dataset and their main characteristics. CMCC meteorological data are available with hourly frequency for each day of the study period and have been downloaded from the CMCC official website in NetCDF format (<https://dds.cmcc.it>).

VARIABLE	DESCRIPTION	UNIT	SOURCE	RESOLUTION
U_{10M}	u-wind component at 10m	m/s^{-1}	CMCC	2.2km
U_{10M}	u-wind component at 10m	m/s^{-1}	CMCC	2.2km
MODV	Wind Intensity	m/s^{-1}	Computed	2.2km
T_2M	Temperature at 2m	K	CMCC	2.2km
TD_{2M}	Dew Point at 2m	K	CMCC	2.2km
TOT_prec	Total Precipitation	$\frac{kg}{m^2}$	CMCC	2.2km
PMSL	Pression Mean Sea Level	Pa	CMCC	2.2km
BLH	Boundary Layer Heigh	m	ERA-5	9km
RH	Relative Humidity	%	Computed	2.2km
DEM	Elevation	m	AERONET	-

Table 3.3: *Selected meteorological variables for AOD improved task.*

3.3 PM prediction methodology

This section presents the materials and the distinct steps for PM prediction. The research approach is structured in multiple cascading phases designed to address data availability challenges while maximizing the quantity of available data and prediction accuracy across various PM metrics and monitoring stations.

3.3.1 Methodological Framework

The spatial prediction phases consist of several interconnected phases specifically developed for effective modeling of PM data each using HRBLOCK [3.1](#) as a framework:

- **Stage 1:** PM_{10} Prediction Model Development

In the initial stage, the author develops a machine learning model for $PM_{2.5}$ spatial prediction using data collected from nine monitoring stations. This foundational model establishes the basis for subsequent prediction tasks and serves as the primary predictive framework for $PM_{2.5}$ concentrations.

- **Stage 2:** Missing PM_{10} Data Imputation

The second stage addresses temporal gaps in the PM_{10} time series. Leveraging the trained model from Stage 1, the researchers impute PM_{10} values for days where measurements are unavailable (identified by placeholder values of -999 in the dataset). This critical step ensures temporal continuity in the PM_{10} data series, which is essential for robust subsequent modeling tasks.

- **Stage 3:** $PM_{2.5}$ Prediction for Stations with Incomplete Data

Five monitoring stations in the network completely lack $PM_{2.5}$ measurements. To address this limitation, the researchers develop a specialized model that utilizes the now-complete $PM_{2.5}$ data as the primary predictor for $PM_{2.5}$ concentrations. This approach enables the extension of $PM_{2.5}$ predictions to locations where only PM_{10} data is available, significantly expanding the spatial coverage of $PM_{2.5}$ estimates.

- **Stage 4:** Comprehensive PM_{25} Model

The final stage incorporates data from all nine stations to construct a robust training set for PM_{25} prediction. This comprehensive model is then specifically applied to predict PM_{25} concentrations in Bari, providing high-resolution particulate matter estimates for this key urban area of interest.

3.3.2 Study Area

The study area for this research is a central portion of the Metropolitan City of Bari, including the capital Bari and the neighboring municipalities of Modugno and Bitonto,

together with some surrounding areas. This territory, which we can define as the "Metropolitan Area of Central Bari", is located in the central-eastern part of Puglia, in the South-East of Italy. Geographically, the area forms a compact territorial "box" that includes important urban centers and peri-urban areas. The territory covers a surface of approximately 450 square kilometers and includes the Adriatic coastal strip corresponding to the city of Bari and its immediate hinterland. This area includes Bari as the main urban center, and Modugno and Bitonto as important satellite centers with significant production and commercial realities. The selected area includes a population of approximately 450,000 inhabitants, with the highest population density concentrated in the municipality of Bari.

The metropolitan area of central Bari is characterized by a typical Mediterranean climate, with specific microclimatic influences due to the presence of the sea and the urban conformation. Summers are hot and dry, often accompanied by high humidity levels in the areas closest to the coast. Winters are mild, with possible short colder periods caused by continental air irruptions. Precipitation is moderate and uneven throughout the year, concentrated mainly in the autumn and winter periods. During the year, episodes of Sirocco occur frequently, a hot wind coming from the south-east that can carry Saharan dust, significantly influencing the levels of atmospheric particulate matter.

Vehicle traffic is one of the main sources, particularly intense along the connecting arteries between residential areas, such as the State Road 96 that connects Bari to Modugno and Bitonto, the Bari ring road (SS16) and the main urban roads of the capital. The port area of Bari, located in the north-eastern part of the city, is a significant source of emissions, deriving from both maritime activities and related logistics operations. The most significant industrial area is located in the territory of Modugno, where numerous production activities of various kinds are concentrated, from the mechanical industry to the food, chemical and building materials industries. Other minor industrial areas are present on the outskirts of Bari, in particular in the Bari-Mungivacca area, and in the industrial area of Bitonto. The picture is completed by emissions deriving from commercial and service activities, spread throughout the territory, and the occasional influence of natural sources such as Saharan dust and marine aerosols.

3.3.3 Materials

For this study, a comprehensive dataset was compiled to develop and validate the PM_{10} and $PM_{2.5}$ forecast model. Four different types of data were used: improved AOD derived from the results of Task 1 [4.1](#); meteorological variables from the WRF model; PM concentrations measured at ground-based air monitoring stations; and ancillary data related to urban characteristics. The datasets cover a four-year period, from January 2019 to December 2022. More details about each dataset are described below.

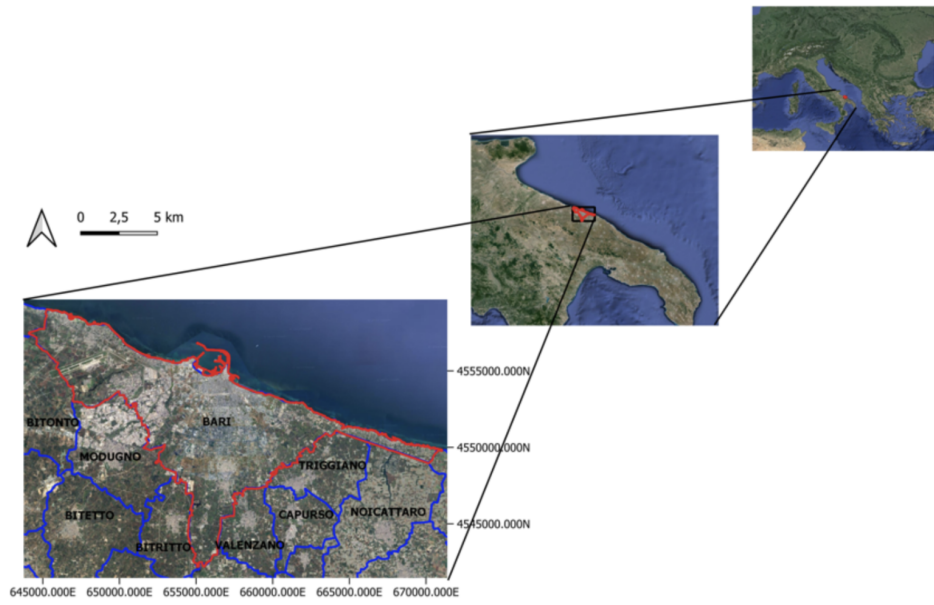


Figure 3.3: *Bari study area for Task 1.*

Monitoring Data for PM

The atmospheric particulate data used in this study were provided by ARPA of the Puglia Region [8]. ARPA Puglia monitors air quality through fixed stations of the RRQA, mobile laboratory campaigns, and additional sampling instruments [9]. The agency evaluates and forecasts air quality across the entire regional territory using pollutant dispersion simulation models. This activity is part of a broader environmental monitoring and control system, which also includes verifying compliance with prescriptions and emission limit values from industrial plants defined during authorization. The decision to use monitoring data from fixed ground stations is motivated by their high reliability and compliance with European measurement standards, defined within the Assessment Program and the Regional Air Quality Plan (PRQA).

The monitoring focused on the central Metropolitan City of Bari area described in 3.3.2, covering a four-year period from 2019 to 2022. The monitoring network selected for this study includes nine stations for measuring PM_{10} and four stations for $PM_{2.5}$, strategically distributed throughout the Bari area as shown in image 3.4 and 3.5. Table 3.4 reports in detail the characteristics of each station, specifying the type of pollutant monitored, the area type (urban or suburban), and the classification of the station based on the main pollution source (traffic, industrial, or background). The stations of Caldarola, Cavour, Kennedy, and Bitonto monitor both PM_{10} and $PM_{2.5}$, while CUS, Carbonara, and the three Modugno stations (EN02, EN03, EN04) record only PM_{10} levels. This distribution allows for analysis of both urban areas subject to intense vehicular traffic (Caldarola, Cavour, CUS), suburban areas influenced by

STATION	POLLUTION	TYPE AREA	TYPE STATION
Caldarola	PM ₁₀ , PM _{2.5}	Urban	Traffic
Cavour	PM ₁₀ , PM _{2.5}	Urban	Traffic
Kennedy	PM ₁₀ , PM _{2.5}	Urban	Background
Bitonto	PM ₁₀ , PM _{2.5}	Suburban	Industrial
CUS	PM ₁₀	Suburban	Traffic
Carbonara	PM ₁₀	Suburban	Background
Modugno_EN02	PM ₁₀	Suburban	Industrial
Modugno_EN03	PM ₁₀	Urban	Industrial
Modugno_EN04	PM ₁₀	Suburban	Industrial

Table 3.4: *Monitoring stations and pollution data*

industrial activities (Bitonto and the Modugno stations), as well as urban and suburban background areas (Kennedy and Carbonara) that represent the background pollution levels in the metropolitan area.

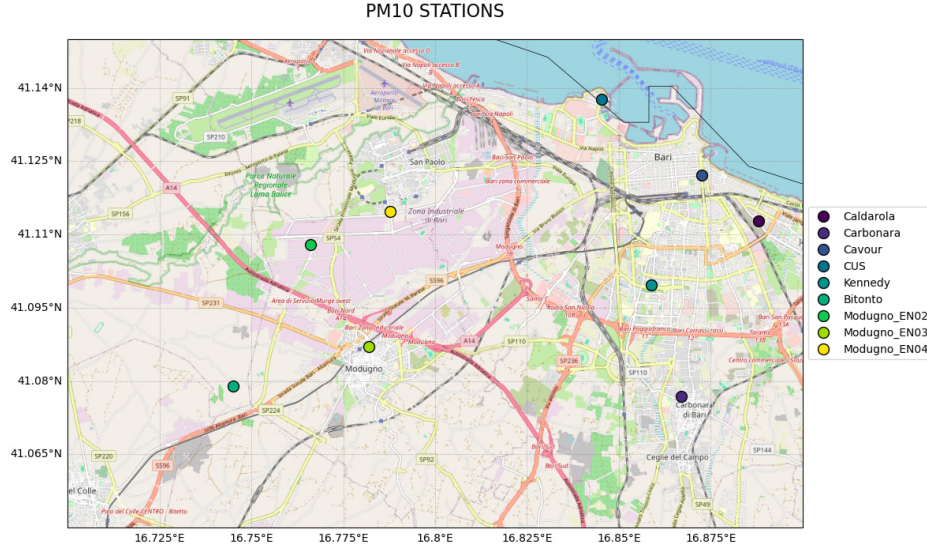


Figure 3.4: *PM₁₀ measuring stations.*

The data collected consist of *daily average* concentrations of PM₁₀ and PM_{2.5}, expressed in $\eta g/m^3$. The continuous hourly measurements are processed by ARPA Puglia to provide daily average values, which represent the reference dataset for this

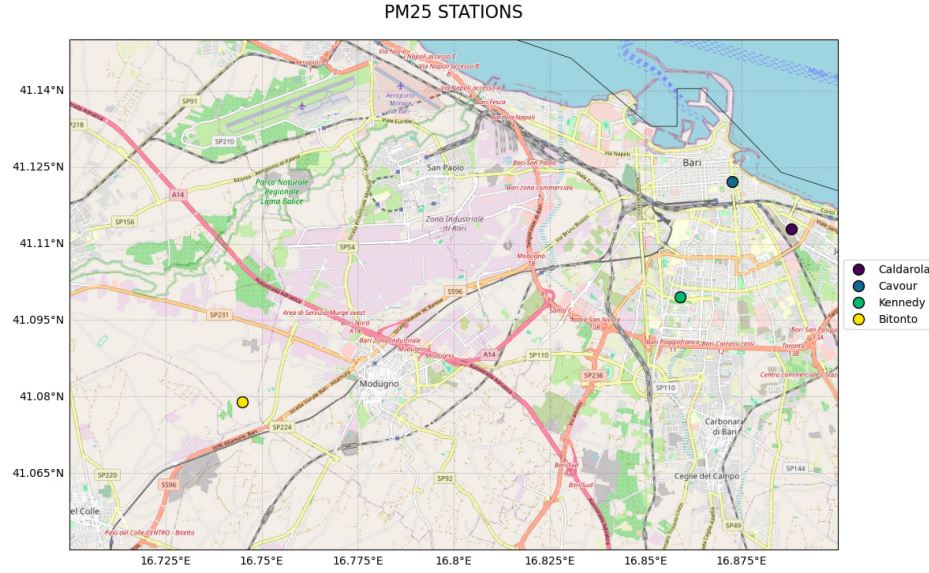


Figure 3.5: $PM_{2.5}$ measuring stations.

study. These stations provide essential reference measurements both for the validation of predictive models and for training the machine learning algorithms used in this study.

Metereological features

In this study, several meteorological variables that influence both atmospheric particulate concentration and AOD measurement quality were considered. These variables were selected based on their known influence on the processes of formation, dispersion, and removal of atmospheric particulate matter. The meteorological data used were provided by ARPA Puglia and derived from the WRF model. WRF [88] is a state-of-the-art mesoscale numerical weather prediction system designed for both atmospheric research and operational forecasting applications. As described in the official documentation, the model was developed through collaboration between the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (represented by the National Centers for Environmental Prediction and the Earth System Research Laboratory), the U.S. Air Force, the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration (FAA). In the context of this study, ARPA Puglia implemented the WRF model using real atmospheric conditions (derived from observations and analyses) for the Metropolitan City of Bari area, thus providing meteorological data at hourly temporal resolution and spatial resolution of $4km$ that accurately represent local conditions. The choice to use WRF model outputs instead of direct measurements from meteorological stations is motivated by the need for continuous and homogeneous spatial coverage of the entire study area, overcoming the representativeness limitations of individual measurement

stations.

Table 3.5 summarizes the meteorological variables derived from the WRF model as potential predictive features for PM, selected based on literature review ([103], [63], [100]). Temperature significantly influences secondary particulate formation processes. Surface temperature, in particular, has proven to be an effective predictor for PM concentrations in scientific literature, given its direct influence on the lower atmospheric layers. Relative humidity modifies the physical properties of particulate matter through hygroscopic growth processes, increasing particle mass in high humidity conditions. This parameter is particularly relevant for PM_{2.5}, whose particles have greater capacity to absorb water compared to PM₁₀. Wind speed and direction are fundamental for understanding pollutant transport and dispersion phenomena, while atmospheric pressure and air stability can determine conditions favorable to particulate accumulation [80]. Precipitation (RAIN, RAINNC) represents one of the main mechanisms for removing particulate from the atmosphere through wet deposition [79]. The planetary boundary layer height (PBLH) represents the height to which ground-emitted pollutants can disperse vertically; low values of this parameter, typical of nighttime or winter conditions, tend to be associated with higher concentrations of pollutants in the lower atmospheric layers [30]. This variable is particularly informative for predictive models, as it synthesizes various aspects of the vertical structure of the atmosphere into a single parameter [59], [89], [58]. Surface emissivity (EMISS) influences the radiation balance and, consequently, photochemical processes and atmospheric stability. This variable, while having less direct relationships with PM concentrations, could contribute to characterizing overall environmental conditions, as radiation variables as NETRAD and TOTRAD [29]. The dataset of the available ARPA variables also provides the list of days in which the dust thresholds were exceeded. In Puglia, in fact, precisely in the Foggia area, there is the ARPA station of Ciuffreda, which serves as a reference for the Bari area for monitoring long-distance dust transport events. The days were coded in binary variables (0/1).

Ancillary features

To integrate features with information on the environmental and anthropic context, a set of auxiliary urban variables was considered as potential predictors for the PM estimation model. The Land Cover variables derive from a GEOBIA (Geographic Object-Based Image Analysis, [18]) approach applied to multitemporal Sentinel-2 images acquired during 2022. The methodology followed a process structured in two distinct phases [91]: first, segmentation was performed using the multi-resolution algorithm implemented in the eCognition 9.0 environment [93], then classification was carried out using the Random Forest algorithm.

The resulting land cover classes, listed in detail in Table 3.6, were then aggregated into three main macro-categories (vegetated areas, artificial surfaces, and water bodies) to simplify their use and because they were considered sufficiently representative of the

VARIABLE	DESCRIPTION	UNIT	SOURCE
U	u-wind component at 10m	m/s^{-1}	WRF
V	v-wind component at 10m	m/s^{-1}	WRF
W	w-wind component at 10m	m/s^{-1}	WRF
MODV	Wind intensity	m/s^{-1}	Computed
T	Temperature at 2m	K	WRF
TSK	Surface Skin Temperature	K	WRF
PBLH	Planetary Boundary Layer Height	m	WRF
RH	Relativity humidity	%	WRF
P	Perturbation pressure	Pa	WRF
REL	Terrain Height	m	WRF
PHI	Geopotential	m^2/s^2	WRF
TOTRAD	Total Solar Radiation	W/m^2	WRF
NETRAD	Net Radiation	W/m^2	WRF
RAINC	Accumulated Convective Precipitation	mm	WRF
RAINNC	Non-Convective Precipitation	mm	WRF
ALBEDO	Albedo	%	WRF
QVAPOR	Water Vapor Mixing Ratio	kg/kg	WRF
SMOISS	Soil Moisture	m^3/m^3	WRF
dust	dust exceedance (0/1)	-	ARPA station

Table 3.5: Meteorological variables by WRF ARPA used in the model.

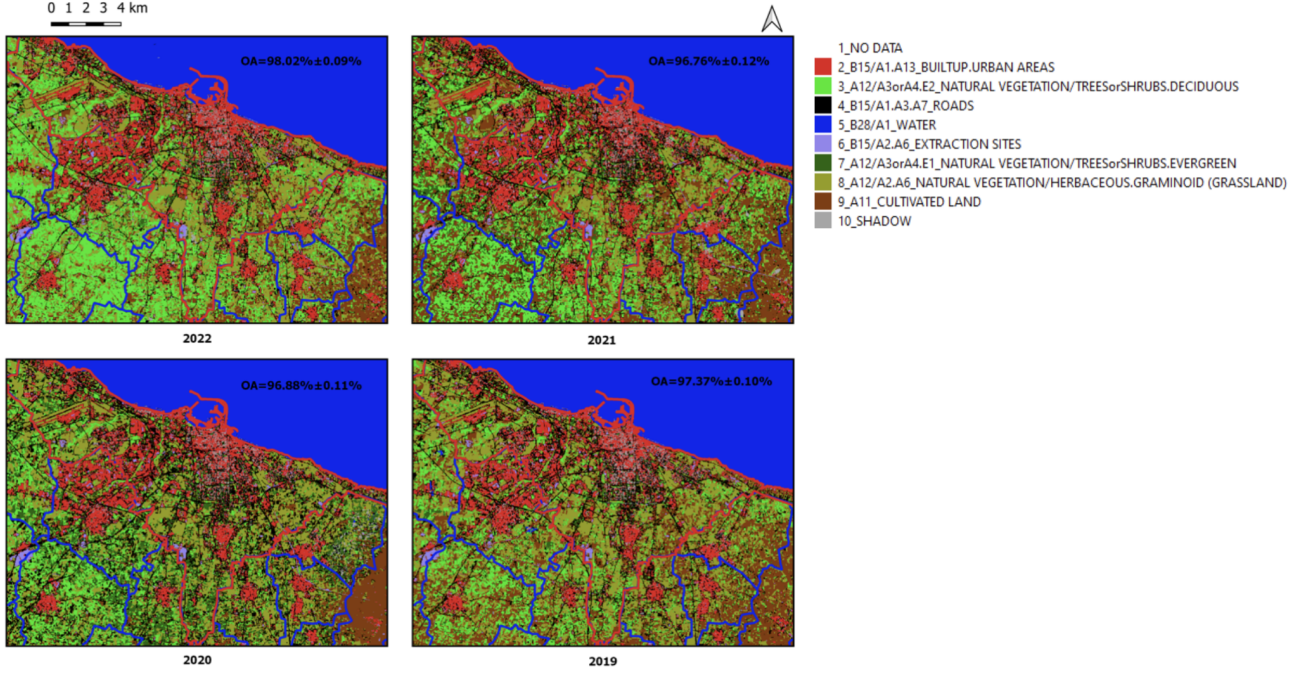


Figure 3.6: *Land Cover classification over Bari area.*

territorial context. These macro-categories are represented respectively by the variables VEGETATED AREAS, ARTIFICIAL SURFACES, and WATER, which provide information on the percentage of coverage of each type at a spatial resolution of 10m.

Regarding spatial predictors related to demographic density, the POP grid variable has been included, which represents the average population density. This data comes from the permanent census survey of the National Institute of Statistics (ISTAT) [47], which to date has published only one update of the census data for the year 2021. Therefore, this update was used as input for the dasymetric method implemented in the QGIS plugin [4] developed within the European Horizon 2020 projects SMURBS and e-SHAPE, thus producing the population density map updated to 2021 with a resolution of 100×100 meters.

The urban land use classes were derived from the Copernicus Urban Atlas (UA) [25] service available for 2018. The UA classes were aggregated based on building use typology, following the associations shown in Table 3.7, which led to the definition of five classes (percentage of road area, residential, other classes, rural, and commercial-industrial) presented with a 100 m resolution.

Regarding indicators related to vehicular traffic, indirect information derived from the road network was used, with data acquired from the Open Street Map platform [77]. In particular, two main characteristics were calculated for each element of the regular grid with a resolution of 300 meters: the total length of road segments present within the cell, and the number of road intersections (crossroads) located in each cell. LiDAR height data were combined with building layers of Ebucco [33] to obtain a

Aggregated LC	LC Code	LC Description
VEGETATED AREAS	A12/A3orA4.E2	Natural Vegetation/Trees or Shrubs
	A12/A3orA4.E1	Natural Vegetation/Trees or Shrubs Evergreen
	A12/A2.A6	Natural Vegetation/Herbaceous Graminoid
	A11	Cultivated Land
ARTIFICIAL SURFACES	B15/A1.A13	Built up.urban areas
	B15/A1.A3.A7	Roads
	-	Shadows
WATER	B28/A1	Water
BARE SOIL	B15/A2.A6	Extraction Sites

Table 3.6: *List of LC classes mapped.*

Name	Building Use	Urban Atlas Classes
Buclass_1	Residential	1111; 11210; 11220; 11230; 11240; 13400
Buclass_4	Rural	21000; 220000; 23000; 24000; 32000; 330000
Buclass_5	Industrial, Commercial and Leisure	12100; 14100; 14200
Buclass_3	Roads, Railways, Port and Airport	12210; 12220; 12230; 12300; 12400
Buclass_2	Other	All remaining classes

Table 3.7: *Building Use classification based on Urban Atlas Classes.*

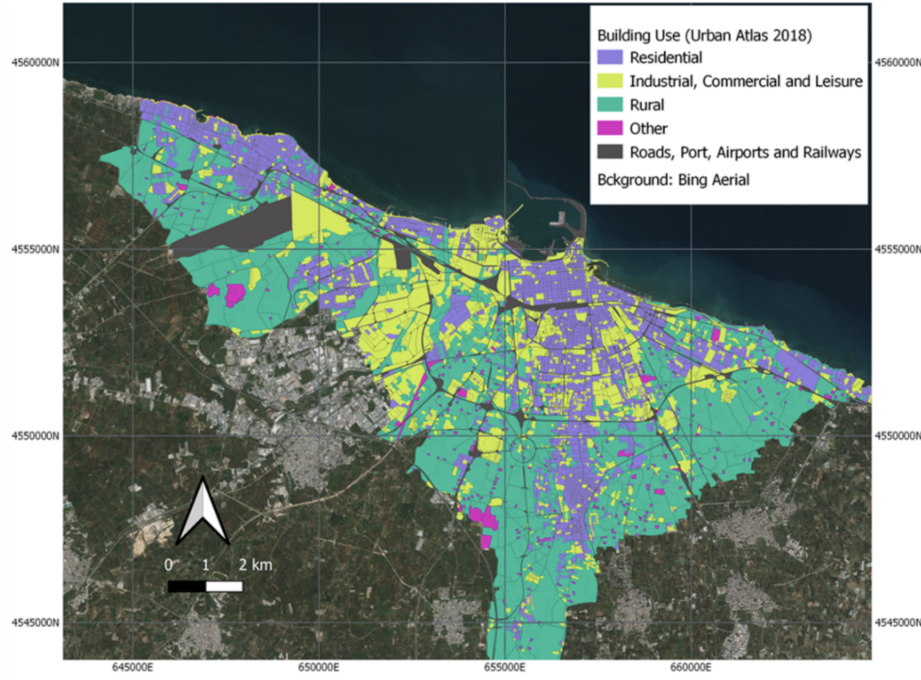


Figure 3.7: *Building use classification over Bari area.*

representative 3-dimensional model of buildings of the city.

Name	Description	Source
HHmean	Mean Building Height	Ebucco
POP grid	Mean Population Density	Open Street Map
Length	Mean Streets Length	Open Street Map
Count	Number of street nodes	Open Street Map
IDWIDW	Inverse Distance from ARPA stations	computed
SeaHubDist	Minimum Sea Distance	computed

Table 3.8: *Urban variables.*

To capture the influence of vegetation on atmospheric particulate matter dynamics, several vegetation indices were calculated and considered from Sentinel-2 A/B time series [3.8](#). Vegetation can significantly influence PM concentrations through dry deposition processes, air filtration, and modification of local circulation patterns [97](#). The indices considered include: NDVI (Normalized Difference Vegetation Index, an indicator of vegetation health and density, correlated with particulate absorption capacity), MSAVI (Modified Soil Adjusted Vegetation Index, which improves analysis

in urban and peri-urban areas where vegetation is intermixed with artificial surfaces), MCARI2 (Modified Chlorophyll Absorption Ratio Index, potentially correlated with plants' ability to resist pollution stress), NBR and NBR2 (Normalized Burn Ratio, useful for identifying burned areas that can be sources of particulate emissions), NDWI (Normalized Difference Water Index, relevant for understanding vegetation moisture that influences particulate resuspension), and BSI (Bare Soil Index, which identifies exposed soils potentially subject to wind erosion and therefore sources of natural particulate matter). Despite the partial redundancy among these indices, they were all considered in the feature selection phase for the PM predictive model, in order to identify which aspects of vegetation have greater influence on air pollution dynamics. The complete stack of Sentinel-2 images was downloaded through the WEkEO API [24], and interpolation was performed using Gaussian Process Regression.

3.4 Machine Learning technique

3.5 Ensemble Methods

In the context of predictive methods, approaches based on decision trees represent a class of algorithms that operate through stratification or segmentation of the predictive space into elementary regions. The principle of these methods lies in their ability to recursively subdivide the input variable space into increasingly homogeneous subsets with respect to the target variable. To make a prediction on a new observation, these methods typically use the mean (in the case of regression) or the mode (in the case of classification) of the target values of training observations belonging to the identified region.

Despite their interpretative simplicity and ability to capture non-linear structures in the data, individual decision trees rarely achieve competitive performance compared to other more sophisticated supervised learning approaches. It is precisely to overcome this limitation that tree-based ensemble learning methods have been developed, among which Random Forest represents one of the most significant advancements.

3.5.1 Random Forest

Mathematical Formulation

Random forest, proposed by Breiman in 2001 [13], is a non-parametric supervised learning algorithm that combines the principles of bagging with randomness in the construction of decision trees. The method operates according to the "divide and conquer" principle: it builds multiple randomized decision trees using subsamples of the data, then aggregates their predictions through averaging (in the case of regression) or majority voting (in the case of classification). Its popularity is due to its ability to effectively handle large datasets, its robustness against overfitting, and its capacity to operate

successfully in spaces characterized by a high number of variables even relative to the number of observations.

Given an observed input vector $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$, the aim is to predict the response $Y \in \mathbb{R}$ by estimating the regression function

$$m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

We therefore assume that we have a training set

$$\mathcal{D}_n = ((\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n))$$

of independent and identically distributed variables. The objective is to use the dataset $\mathcal{D}_n$ to construct an estimate $m_n : \mathcal{X} \rightarrow \mathbb{R}$ of the function m . In this regard, we will say that the estimated regression function m_n is consistent with respect to the mean squared error if $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \rightarrow 0$ for $n \rightarrow \infty$.

A random forest is a predictor consisting of a collection of M random regression trees. Given the j -th tree of the family, the predicted value at the query point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$, where $\Theta_1, \dots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ and independent of $\mathcal{D}_n$. In practice, the variable Θ is used to resample the training set before growing the individual trees and to select the subsequent directions for splitting.

In mathematical terms, the estimate of the j -th tree takes the form:

$$m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) = \sum_{i \in \mathcal{D}_n^*(\Theta_j)} \frac{\mathbb{1}_{\mathbf{x}_i \in A_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)} Y_i}{N_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)},$$

where $\mathcal{D}_n^*(\Theta_j)$ is the set of data points selected before the tree construction, $A_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ is the cell containing $\mathbf{x}$, and $N_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ is the number of (preselected) points that fall in the cell

At this point, the trees are combined to form the estimate of the (finite) forest:

$$m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^M m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n). \quad (3.1)$$

Since M can be chosen arbitrarily large (limited only by available computational resources), it makes sense, from a modeling perspective, to let M tend to infinity and consider, instead of (3.1), the estimate of the (infinite) forest

$$m_{\infty,n}(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_{\Theta}[m_n(\mathbf{x}; \Theta, \mathcal{D}_n)].$$

Algorithm

In its original formulation ([13]), each node of a single tree is associated with a hyperrectangular cell. The roots of the tree are in turn $\mathcal{X}$ and, at each step of the tree construction, a node (or equivalently its corresponding cell) is divided into two parts.

All terminal nodes (also called leaves) form, together, a partition of $\mathcal{X}$. The algorithm is described below:

Algorithm 1 Predicted value of Breiman's random forest at point $\mathbf{x}$

Input: Training dataset $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1, \dots, n\}$,
number of variables to select $\mathbf{mtry} \in \{1, \dots, p\}$,
minimum node size $\mathbf{nodesize} \in \{1, \dots, a_n\}$, and $\mathbf{x} \in \mathcal{X}$

Output: Random forest prediction at point $\mathbf{x}$

- 1: **for** $j = 1, \dots, M$ **do**
- 2: Select a_n points, with (or without) replacement, uniformly in $\mathcal{D}_n$. In the following steps, only these a_n observations will be used.
- 3: Set $\mathcal{P} = (\mathcal{X})$ as a list containing the cell associated with the root of the tree.
- 4: Set $\mathcal{P}_{\text{final}} = \emptyset$ as an empty list.
- 5: **while** $\mathcal{P} \neq \emptyset$ **do**
- 6: Let A be the first element of $\mathcal{P}$.
- 7: **if** A contains fewer than $\mathbf{nodesize}$ points or if all $\mathbf{X}_i \in A$ are equal **then**
- 8: Remove the cell A from the list $\mathcal{P}$.
- 9: $\mathcal{P}_{\text{final}} \leftarrow \text{Concatenate}(\mathcal{P}_{\text{final}}, A)$.
- 10: **else**
- 11: Select uniformly, without replacement, a subset $\mathcal{M}_{\text{try}} \subset \{1, \dots, p\}$ of cardinality $\mathbf{mtry}$.
- 12: Select the best split in A by optimizing the CART splitting criterion along the coordinates in $\mathcal{M}_{\text{try}}$ (see text below for details).
- 13: Split the cell A according to the best split. Call A_S and A_D the two resulting cells.
- 14: Remove the cell A from the list $\mathcal{P}$.
- 15: $\mathcal{P} \leftarrow \text{Concatenate}(\mathcal{P}, A_S, A_D)$.
- 16: **end if**
- 17: **end while**
- 18: Calculate the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ as the average of the Y_i that fall in the cell of $\mathbf{x}$ in the partition $\mathcal{P}_{\text{final}}$.
- 19: **end for**
- 20: Calculate the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$ at the query point $\mathbf{x}$ according to equation (3.1).

To conclude, we describe the functioning of the CART split criterion. For simplicity, we consider a tree without resampling, which uses the entire original dataset $\mathcal{D}_n$ for its construction. Let A be a generic cell and denote with $N_n(A)$ the number of points that fall in A . A cut in A is a pair (j, z) , where j is a value (dimension) from $\{1, \dots, p\}$ and z is the position of the cut along the j -th coordinate, within the limits of A . Let $\mathcal{C}_A$ be the set of all possible cuts in A . Then, with the notation $\mathbf{X}_i = (X_i^{(1)}, \dots, X_i^{(p)})$, for every $(j, z) \in \mathcal{C}_A$, the CART split criterion takes the form:

$$\begin{aligned}
L_{\text{reg},n}(j, z) = & \frac{1}{N_n(A)} \sum_{i=1}^n (Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\
& - \frac{1}{N_n(A)} \sum_{i=1}^n \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{X_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{X_i^{(j)} \geq z} \right)^2 \mathbb{1}_{\mathbf{X}_i \in A},
\end{aligned} \tag{3.2}$$

where $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$, and $\bar{Y}_A$ (respectively, $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) is the mean of Y_i such that $\mathbf{X}_i$ belongs to A (respectively, A_L , A_R), with the convention that the mean is equal to 0 when no point $\mathbf{X}_i$ belongs to A (respectively, A_L , A_R). For each cell A , the best cut (j_n^*, z_n^*) is selected by maximizing $L_{\text{reg},n}(j, z)$ over $\mathcal{M}_{\text{try}}$ and $\mathcal{C}_A$; that is,

$$(j_n^*, z_n^*) \in \underset{j \in \mathcal{M}_{\text{try}}, (j,z) \in \mathcal{C}_A}{\operatorname{argmax}} L_{\text{reg},n}(j, z). \tag{3.3}$$

To remove some of the ties in the argmax, the best cut is always performed at the midpoint between two consecutive data points. Finally, note that the optimization program extends effortlessly to the case of resampling, optimizing over the a_n preselected points instead of the original dataset $\mathcal{D}_n$.

Therefore, at each node of each tree, the algorithm uniformly randomly chooses `mtry` coordinates in $\{1, \dots, p\}$, evaluates the criterion (3.2) along all possible cutting thresholds according to the directions in `Mtry`, and returns the best one. The quality measure (3.2) is the criterion used in the CART algorithm [14]. This criterion calculates the (renormalized) difference between the empirical variance in the node before and after performing a cut. There are three essential differences between CART and a tree in Breiman's forest [13]:

- In Breiman's forest, the criterion (3.2) is evaluated on a randomly selected subset `Mtry` of coordinates, rather than on the entire set $\{1, \dots, p\}$.
- The individual trees are not pruned, and the final cells contain no more than `nodesize` observations (unless all points in the node have the same $\mathcal{X}_i$).
- Each tree is built on a subset of a_n examples drawn from the initial sample, not on the entire sample $\mathcal{D}_n$; only these a_n observations are used to calculate the estimate. When $a_n = n$ (and resampling is performed with replacement), the algorithm operates in *bootstrap* mode, while $a_n < n$ corresponds to *subsampling* (with or without replacement).

3.5.2 Gradient Boosting

Gradient Boosting is a machine learning ensemble technique that combines multiple simple predictive models (called **weak learners**, typically decision trees) to create a single strong predictive model. It belongs to a family of boosting algorithms that build

models sequentially, where each new model attempts to correct errors made by previous ones. The core idea of gradient boosting is to optimize a differentiable loss function by iteratively adding weak models that follow the anti-gradient of the loss function. This process resembles gradient descent in optimization problems but operates in a functional space rather than a parametric one.

The Gradient Boosting Algorithm

Given a training dataset $D = \{\mathbf{x}_i, y_i\}_1^N$, the objective of gradient boosting is to find an approximation $\hat{F}(\mathbf{x})$ of the function $F^*(\mathbf{x})$ that maps instances $\mathbf{x}$ to their output values y by minimizing the expected value of a given loss function $L(y, F(\mathbf{x}))$. The approximation of $F^*(\mathbf{x})$ is then constructed iteratively as a weighted sum of functions

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h_m(\mathbf{x})$$

where ρ_m is the weight of the m -th function $h_m(\mathbf{x})$. These functions are the ensemble models (such as decision trees, for example). The approximation is built iteratively. Initially, a constant approximation of $F^*(\mathbf{x})$ is obtained as

$$F_0(\mathbf{x}) = \arg \min_{\alpha} \sum_{i=1}^N L(y_i, \alpha)$$

The subsequent models will minimize

$$(\rho_m, h_m(\mathbf{x})) = \arg \min_{\rho, h} \sum_{i=1}^N L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i))$$

However, rather than solving the optimization problem directly, each h_m can be viewed as a greedy step in a gradient descent optimization problem for F^* . Therefore, each model h_m is trained on a new dataset $D = \{\mathbf{x}_i, r_{mi}\}_{i=1}^N$, where the pseudo-residuals r_{mi} are calculated as

$$r_{mi} = \left[\frac{\partial L(y_i, F(\mathbf{x}))}{\partial F(\mathbf{x})} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$$

The value of ρ_m is subsequently calculated by solving a line search optimization problem.

Regularization in Gradient Boosting

The algorithm can suffer from overfitting if the iterative process is not properly regularized. To prevent this, various regularization techniques can be used:

1. **Shrinkage:** Multiply each weak learner's contribution by a learning rate ν (typically 0.1 or smaller):

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \rho_m h_m(x) \tag{3.4}$$

2. **Subsampling:** Train each weak learner on a random subset of the training data
3. **Tree constraints:** Limit the depth, minimum samples per leaf, etc. of decision trees

3.5.3 XGBoost

XGBoost is a gradient boosting algorithm that builds an ensemble of decision trees sequentially, with each new tree correcting errors made by previous trees. Unlike standard gradient boosting, XGBoost employs a modified loss function that includes regularization terms to control tree complexity, preventing overfitting while maintaining predictive power:

$$L_{xgb} = \sum_{i=1}^N L(y_i, F(\mathbf{x}_i)) + \sum_{m=1}^M \Omega(h_m)$$

having

$$\Omega(h) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

where T is the number of leaves of the tree and w are the output scores of the leaves. XGBoost incorporates regularization directly into its loss function to control tree complexity. The gamma (γ) parameter defines the minimum loss reduction required for node splitting, with higher values creating simpler trees. Additional complexity controls include shrinkage (reducing step size in the additive model), maximum tree depth, and randomization techniques (subsampling observations and features at tree/node levels). These controls reduce overfitting and decrease training time. XGBoost also can be extended to custom loss functions by defining a function loss functions by allowing users to specify gradient and hessian calculations through the "objective" parameter. Moreover, XGBoost introduces a specialized algorithm that efficiently handles sparse data. This sparsity can arise from missing entries or the presence of many zero in the data. XGBoost is able to automatically remove these entries from the computation of the gain for split candidates. XGBoost doesn't simply discard or impute them. Instead, each tree learns the optimal default child branch for these entries.

3.6 Feature Selection

Feature selection is process that involves identifying and selecting a subset of the most relevant features (variables, predictors) from the original feature set. The primary goal is to remove redundant or irrelevant features while retaining those that contribute most significantly to the predictive power of a model. Feature selection offers several important benefits: it reduces model complexity, mitigates the risk of overfitting, improves model performance, enhances interpretability, and reduces computational costs.

By focusing on only the most informative features, models become more efficient and often more accurate, particularly when dealing with high-dimensional datasets where the "curse of dimensionality" can significantly impact performance. There are three main approaches to feature selection: filter methods (which evaluate features independently of the model using statistical measures), wrapper methods (which evaluate subsets of features using the model itself), and embedded methods (which perform feature selection as part of the model training process) [31]. Each approach offers different trade-offs between computational efficiency, model performance, and feature dependencies consideration. Next, some of the most common methods used in this work will be introduced.

3.6.1 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a wrapper method for feature selection. It systematically eliminates features by recursively training the model, ranking features by importance, and removing the least important ones.

In the context of Recursive Feature Elimination with Random Forest, the importance of features is defined through two methods:

- Mean Decrease in Impurity (MDI) - Measures how much each feature decreases impurity (Gini impurity for classification or variance reduction for regression) across all trees in the forest. Features that result in larger decreases in impurity are considered more important.
- Mean Decrease in Accuracy (MDA) - Also called permutation importance, this method randomly shuffles the values of each feature and measures the resulting decrease in model accuracy. A significant decrease indicates the feature is important.

During the RFE process, these importance scores determine which features to eliminate in each iteration based on their importance. The Random Forest model is then retrained on the reduced feature set, and importances are recalculated. This recursive process continues until reaching the desired number of features.

The importance scores provide a meaningful ranking because they capture both the direct effect of each feature and some of its interactions with other features, making Random Forest RFE particularly effective for complex datasets with non-linear relationships.

3.6.2 Boruta

Boruta is a feature selection algorithm designed to find all relevant variables in a dataset. The algorithm works as a wrapper around Random Forest classification and iteratively removes features proven by statistical tests to be less relevant than random probes [53].

The key principle behind Boruta is comparing the importance of real features with artificially created random attributes called **shadow features**. This approach helps determine which variables contain truly meaningful information versus those that might appear important due to random fluctuations. Unlike many feature selection methods that focus on finding minimal-optimal subsets (the smallest set of features giving optimal classification), Boruta addresses the all-relevant problem identifying all attributes that are relevant to the classification under any circumstances. The algorithm works by extending the dataset with shadow attributes created by shuffling the original features, running Random Forest classification, and then comparing each feature's importance score with the highest-scoring shadow attribute. Features performing significantly better than the best shadow attribute are considered important, while those performing worse are removed. This process repeats iteratively until all features are either confirmed or rejected, or the algorithm reaches a predefined maximum number of runs. To deal with importance score fluctuations, Boruta uses multiple Random Forest runs and statistical testing. Studies have shown that Boruta provides unbiased and stable selection of important features, effectively capturing both direct relationships and complex interactions between attributes.

3.6.3 Shap values

SHAP (SHapley Additive exPlanations) [64] values represent a unified approach to interpret model predictions. This method assigns each feature an importance value for a particular prediction based on game theory concepts. The key contributions of SHAP include:

1. The identification of a new class of additive feature importance measures that unifies existing explainability methods
2. Theoretical results proving there is a unique solution with desirable properties

SHAP values attribute to each feature the change in the expected model prediction when conditioning on that feature (by keeping fixed the other features). They explain how to get from the base value (the expected prediction without knowing any features) to the current model output.

The SHAP framework is built on three properties:

- *Local accuracy*: The sum of feature attributions must equal the model prediction
- *Missingness*: Features with zero values should have no impact
- *Consistency*: If a model changes so that a feature's contribution increases, its attribution shouldn't decrease

We give now the following definition

Definition 1 *Additive feature attribution methods have an explanation model that is a linear function of binary variables:*

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i, \quad (3.5)$$

where $z' \in \{0, 1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

Methods with explanation models matching Definition 1 attribute an effect ϕ_i to each feature, and summing the effects of all feature attributions approximates the output $f(x)$ of the original model.

Given this definition and the three previous properties, there is only one possible explanation model g follows Definition 1 and satisfies the properties:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \quad (3.6)$$

where $|z'|$ is the number of non-zero entries in z' , and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x' .

The theorem follows from combined cooperative game theory results, where the values ϕ_i are known as Shapley values.

Experimental studies have shown that SHAP values better align with human intuition and more effectively discriminate between model output classes than competing methods like LIME and DeepLIFT [64]. The idea behind this approach is choosing it for feature selection instead of explainability purpose, by choosing the features that better explain a model on the training set.

Chapter 4

Results

In the following sections, the main results of the research are presented:

- Section 4.1 illustrates the application of the HReBLOCK framework to the 'Improved Aod' task, highlighting how the approach improves the correlation of 300 m AOD data with AERONET data.
- Section 4.2 explores the results of the four methodological stages developed for the prediction of PM_{10} and $PM_{2.5}$, culminating in the generation of annual maps at intra-urban scale that reveal the detailed spatial distribution of pollutants in the study area.

4.1 Task 1: AOD Improved

In this exploratory phase, a comprehensive set of potential features was considered, encompassing diverse natures. The subsequent exploratory analysis will facilitate the identification of variables that exert a tangible influence on the quality of AOD estimates, while those that demonstrate negligible impact can be excluded from the final model. This feature selection process will allow for the optimization of the machine learning model, enhancing both the performance and interpretability of the results. The initial features integrated into the Task 1 dataset are outlined in Table 4.1. This dataset was constructed by integrating various data sources from 2019 to 2022, as detailed in the "Data Collection" section. In addition to the various AOD satellite products, the dataset was enriched with meteorological and temporal variables to capture potential seasonal and cyclical patterns. Validation was performed using measurements from seven AERONET distributed stations over the Italian territory, which provide high-precision reference data.

	Features	Name
AOD Products	S3A SY 2 SYN	SYN3_A
	S3B SY 2 SYN	SYN3_B
	MODIS MAIAC	Optical_Depth_055
Metereological variables	Temperature at 2m	T_2M
	Dewpoint Temperature at 2m	TD_2M
	Mean Pressure at sea level	PMSL
	v-wind component at 10m	V_10M
	u-wind component at 10m	U_10M
	wind speed	MODV
	Total Precipitation	TOT_PREC
	Boundary Layer Heigh	BLH
	Relative Humidity	RH
Spatio temporal variables	Day of the year	DOY
	Day in the week	week_day
	Cosine of month	cos_month
	Sine of month	sin_month
	Month in the year	month
	Elevation	DEM

Table 4.1: *Dataset for AOD improved task.*

Note that the variables *wind_speed*, *month_cos*, and *month_sin* are computationally derived as:

$$MODV = \sqrt{U_{10}M^2 + V_{10}M^2} \quad (4.1)$$

$$\sin_month = \sin\left(\frac{2\pi \times month}{12}\right) \quad (4.2)$$

$$\cos_month = \cos\left(\frac{2\pi \times month}{12}\right) \quad (4.3)$$

where u and v are the wind components and $month \in [1, 12]$ represents the calendar month.

4.1.1 Data preparation

This section discusses in detail the data preparation methodology for the meteorological and AOD datasets. All data processing was performed with the Python programming language, using several specialized libraries. In particular, data manipulation and analysis were conducted using pandas (McKinney, 2010 [70]) and NumPy (Harris et al., 2020 [42]). For the management and processing of the NetCDF meteorological files, the xarray library (Hoyer & Hamman, 2017 [43]) was used due to its powerful capabilities to handle multidimensional arrays with labeled coordinates. For visualization, the matplotlib and seaborn libraries were used. Statistical analyses were performed with scipy and scikit-learn, which also provided the machine learning implementations used in the subsequent modeling steps.

Metereological data preparation

As previously mentioned in the data collection section, meteorological variables were incorporated into the dataset due to their established relationship with AOD. These variables play a crucial role in both improving satellite AOD data and predicting particulate matter concentrations. The meteorological data were obtained from the CMCC downscaling of ERA5 reanalysis. These data offer a high spatial resolution of 2.2 km and hourly temporal resolution, making them suitable for integration with satellite-derived AOD measurements. The data were downloaded from the CMCC data portal and extracted based on the spatial coordinates of AERONET stations. Only the Boundary Layer Height was downloaded directly from ERA-5 with a resolution of 9km. To ensure temporal alignment with satellite overpass times, morning averages (07:00-12:00 local time) were calculated for each meteorological parameter. This time window was selected to encompass the overpass times of all satellites considered in this study, ensuring the meteorological conditions accurately represented those during AOD acquisitions. In addition to the standard meteorological variables provided by the CMCC dataset, two important parameters were derived:

- Relative Humidity (RH): Calculated using the Magnus-Tetens approximation formula, which relates saturation vapor pressure to temperature. The calculation

uses air temperature and dew point temperature to determine the relative humidity percentage, with values constrained between 0 and 100%. This calculation is particularly important as relative humidity significantly affects aerosol hygroscopic growth and consequently influences both AOD measurements and particulate matter concentrations.

- **Wind Speed (MODV):** Computed as the vector magnitude of the wind components by taking the square root of the sum of squared zonal and meridional wind components at 10 meters height. Wind speed provides critical information about the potential for horizontal transport of aerosols.

The preprocessed meteorological dataset was then merged with the AOD observations based on temporal and spatial coincidence, providing a comprehensive set of variables for both the AOD improvement task and the particulate matter prediction models. To provide a comprehensive understanding of the weather conditions during the study period and in the region, the frequency distributions of each weather variable were analyzed. These distributions provide insights into the predominant weather conditions during satellite overflights and help identify potential biases or anomalies in the data set. Histograms were generated to visualize the statistical properties of the meteorological features, providing a context for the spatial and temporal heterogeneity observed in AOD measurements and particulate concentrations.

AOD data preparation

As described earlier in the data collection section, this study uses AOD data from several satellite products: Sentinel-3 Synergy Land A and B from ESA’s Copernicus programme and NASA’s MODIS MAIAC. Each product offers different temporal and spatial resolution characteristics. The Sentinel-3 Synergy products provide observations with about 5 passes per week and a high spatial resolution of 300 metres, providing a good basis for working towards intra-urban resolution of PM concentrations. The MODIS MAIAC algorithm, on the other hand, provides daily AOD data with a spatial resolution of 1 kilometre, providing consistent temporal coverage. In order to validate the model, data from AERONET stations over the Italian territory were chosen as a reference.

AERONET AOD data preparation To prepare the AERONET AOD data and to ensure the high quality and reliability of the data used as ground truth reference, only Level 2 data were selected, representing the highest quality standard available in the AERONET network, as they have already undergone calibration, cloud screening and manual quality control. To ensure temporal consistency with satellite passes, the 7 am-1 pm UTC time window was used for averaging the data. For almost all stations, the AOD value at $\lambda = 500nm$ was available, so the AOD at $550nm$ was calculated using

Frequency Distributions of Meteorological Variables

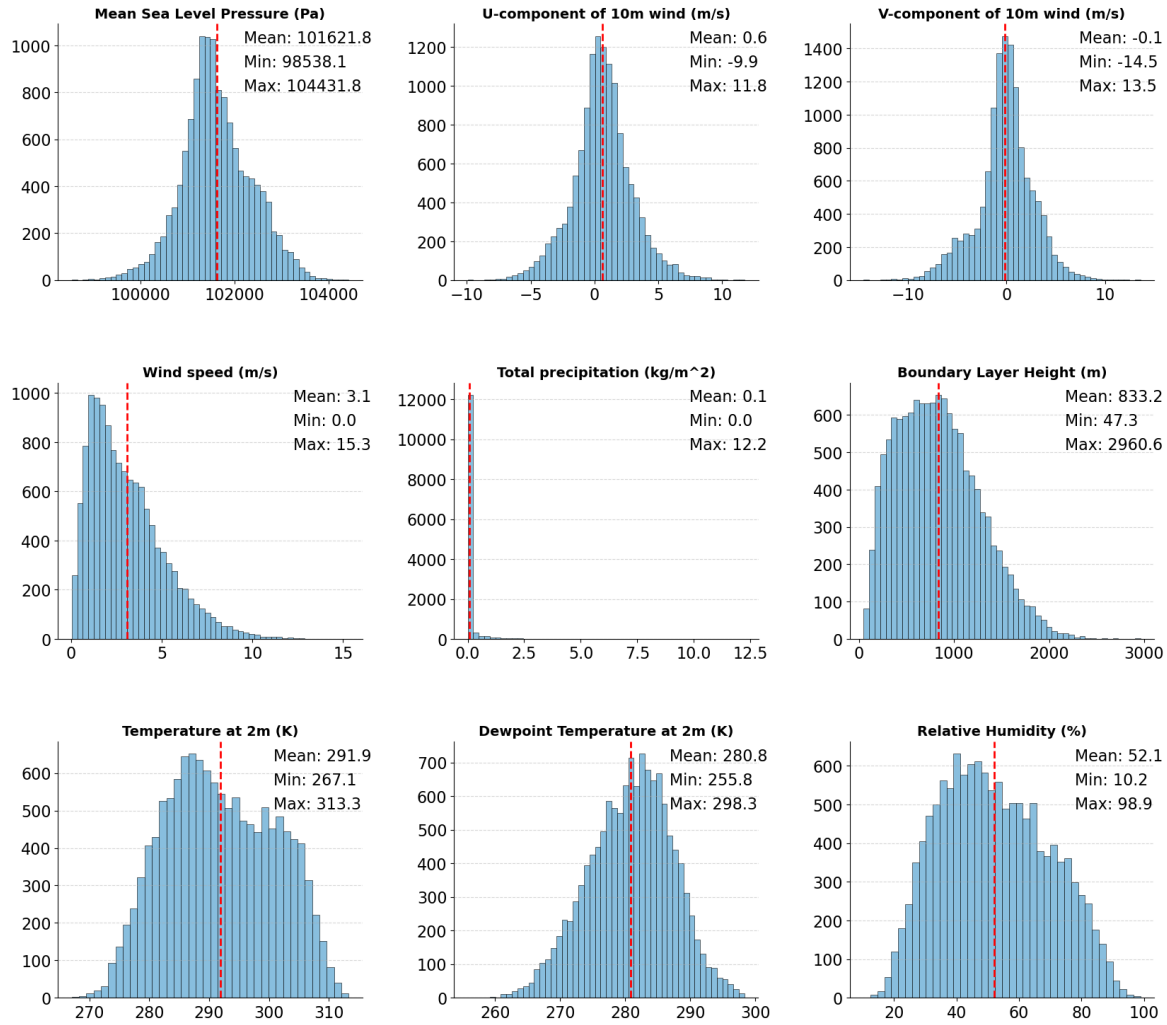


Figure 4.1: Frequency Distributions of Metereological Variables

the `500 - 870_Angstrom_Exponent` value reported in the downloaded datasets:

$$AOD_{550nm} = AOD_{500nm} \left(\frac{550nm}{500nm} \right)^{-\alpha} \quad (4.4)$$

The only exception concerns the Ispra station, where the AOD value at $490nm$ was used and the Ångström exponent was considered to be in the range $440 - 675nm(440 - 675_Angstrom_Exponent)$, and the conversion formula was adjusted accordingly to obtain the AOD at $550nm$. This conversion was necessary to standardise the AERONET measurements to the $550nm$ wavelength commonly used in satellite AOD products, thus ensuring a direct and accurate comparison between ground-based and satellite data. To ensure the integrity of the data, an exploratory analysis of the collected AERONET data was performed as a secondary validation step. This process identified and removed extreme AOD values that deviated significantly from the typical distribution as potential outliers. These outlier AOD measurements are likely indicative of uncorrected instrumental errors rather than actual atmospheric conditions.

MAIAC AOD data preparation For the MODIS MAIAC AOD product, research focused on extracting and validating the parameter “Optical_Depth_055,” which represents the optical depth of aerosols in the green band (550 nm). This spectral band is particularly important for air quality applications because of its sensitivity to fine particulate matter. To ensure the highest data quality, according to the manual of MAIAC AOD product [65], a rigorous filter based on MAIAC’s Quality Assessment (QA) flags was applied, selecting only pixels with a QA value of “0000” (Best Quality). This classification corresponds to the combination of two critical quality criteria: ‘QA.CloudMask = Clear’ and ‘QA.Adjacencymask = Clear’, which effectively eliminate observations contaminated by clouds or affected by adjacency effects from nearby features. The daily temporal resolution of the MAIAC algorithm allowed the construction of a consistent time series of AOD measurements throughout the study area. To ensure spatial consistency of the extracted data, a 3×3 pixel window centered on the location of each AERONET station was considered. The mean value of AOD within this window was calculated for each timestamp. Data points were excluded from the time series when the standard deviation within the window exceeded 0.5 or when fewer than 3 valid (non-NaN) pixels were available in the window [10]. This spatial filtering approach helped minimize the impact of local anomalies and ensure the reliability of the data for subsequent analyses. This spatial filtering approach helped minimize the impact of local anomalies and ensure data reliability. Figure 4.2 presents the correlation analysis between MAIAC’s AOD detections and the corresponding AERONET measurements of 0.75 based on 2561 coincident observations. This result is consistent with the established accuracy range of MAIAC AOD retrievals reported in literature [60]. The original MAIAC aerosol algorithm validation by Lyapustin et al [66] demonstrated variable performance across different geographical regions, with correlation coefficients ranging from 0.33 to 0.99 depending on site characteristics and environmental

conditions. The obtained correlation of 0.75 falls well within this established range, particularly consistent with the intermediate-to-good performance levels observed for sites with mixed surface and atmospheric conditions.

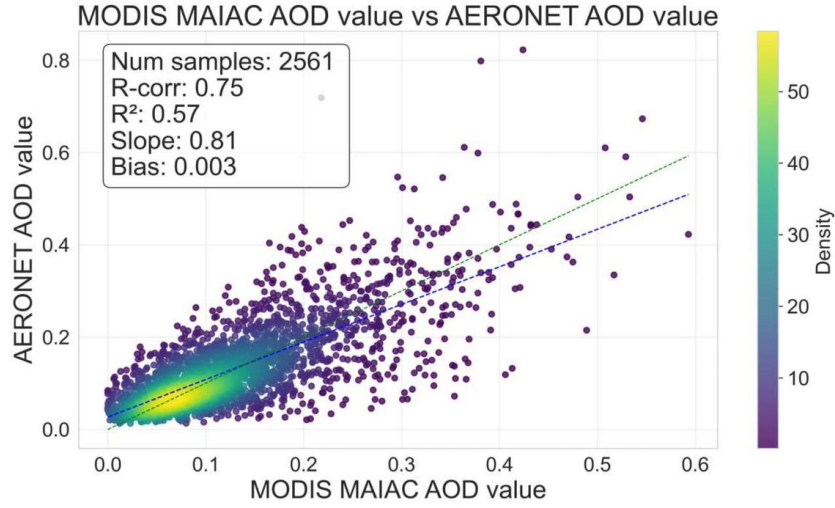


Figure 4.2: *Correlation between MODIS MAIAC AOD (550 nm) and AERONET ground-based measurements.*

SENTINEL-3 SYNERGY data preprocessing In the context of preprocessing AOD data, we evaluated how to handle the overlapping products (synergy-syn) from Sentinel-3 Synergy A and B by examining their correlation with AERONET reference data. As illustrated in figure 4.3, both products showed a Pearson correlation coefficient with AERONET of approximately 0.49, based on 2550 and 2088 observations, respectively.

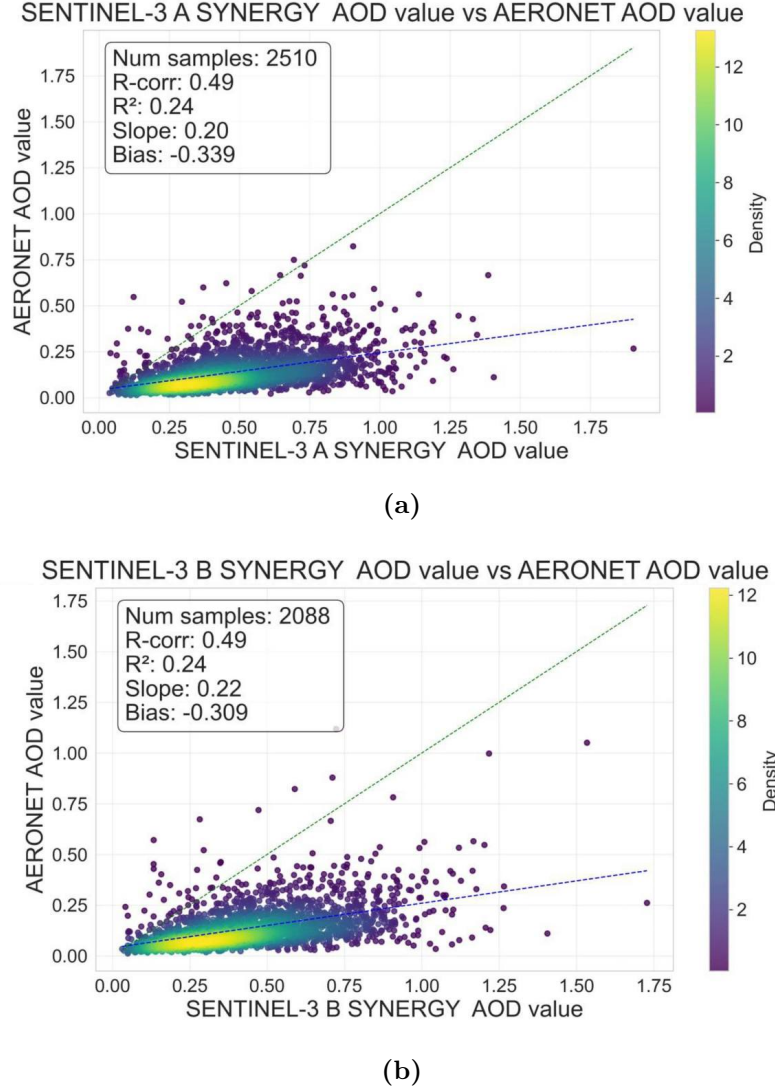


Figure 4.3: *Correlation analysis between satellite Sentinel-3 retrievals and AERONET ground-based measurements for (A) Synergy A and (b) Synergy B*

Starting from 2020, the Sentinel-3 dataset has made available both ST (Short Time Critical) and NT (Non Time Critical) products, with distinct characteristics and applications. ST products are designed for operational applications with rapid availability within 1-2 days, offering intermediate quality and using auxiliary data available with short delays, making them suitable for time-sensitive operational monitoring and atmospheric/marine forecasting services. In contrast, NT products are specifically developed for scientific and climatological applications, featuring delayed availability that can require several days but compensated by more rigorous calibration and validation processes. These products incorporate enhanced precision calculations including improved orbit determinations and refined atmospheric corrections. In cases where both types were available for the same acquisition, NT products were systematically chosen, thus prioritizing data quality and reliability over timeliness for the scientific objectives of this study.

The Sentinel-3 Synergy data were downloaded applying a series of quality filters to ensure measurement reliability. A spatial consistency check was implemented to further validate the AOD data quality. For each ground station location, we extracted AOD values using a 3 x 3 pixel window centered on the station coordinates. Within each window, the mean AOD value and its standard deviation was calculated. If the standard deviation exceeded 0.02, the measurement was classified as unreliable and subsequently removed from the dataset [71]. The final AOD value assigned to each location was determined by averaging only the reliable measurements within the window that passed this spatial consistency criterion. This filtering procedure inevitably created gaps in the datasets. To optimize temporal and spatial coverage, the possibility of merging data from Synergy A and B sensors into a single composite variable was evaluated. Supporting this strategy, the correlation between the two Sentinel-3 products was analyzed, which revealed a significant correlation coefficient of 0.79 across a set of 1100 coincident observations, as shown in the figure 4.4.

Based on these results, we implemented a data fusion approach structured in two phases:

- On days with overlapping coverage, where both sensors (Synergy A and B) provide valid measurements, the arithmetic mean of values was calculated:

$$AOD_{SYN3} = \frac{(AOD_{SYN3A} + AOD_{SYN3B})}{2} \quad (4.5)$$

- For points where AOD is measured by only one platform, cross-platform adjustment was applied using station-specific linear regression coefficients, provided that adequate correlation existed between Sentinel-3A and Sentinel-3B measurements ($R - corr > 0.7$, $p < 0.05$). This approach ensures harmonization between sensors while maintaining data quality standards. For stations failing to meet these correlation criteria, no cross-platform adjustment was applied, and original sensor values were retained to avoid introducing artificial errors. All regression line are reported in Table 4.2.

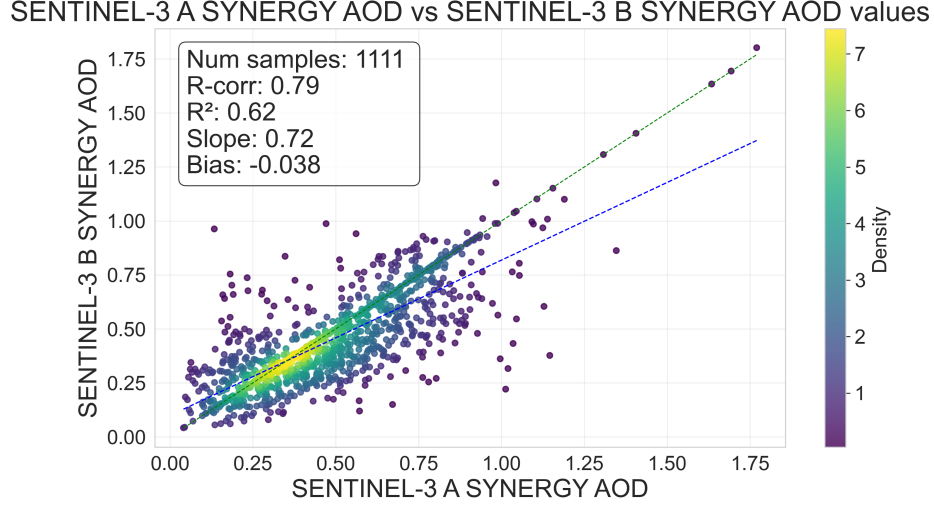


Figure 4.4: *Correlation between Sentinel-3 A Synergy and Sentinel-3 B Synergy.*

SITE	SYN3_A to SYN3_B	SYN3_B to SYN3_A
ISPRA	$AOD_B = 0.84AOD_A - 0.02$	$AOD_A = 0.67AOD_B + 0.15$
POTENZA	$AOD_B = 0.99AOD_A - 0.002$	$AOD_A = 0.96AOD_B + 0.01$
FOGGIA	$AOD_B = 0.7AOD_A + 0.13$	$AOD_A = 0.77AOD_B + 0.15$
MODENA	$AOD_B = 0.52AOD_A + 0.11$	$AOD_A = 0.81AOD_B + 0.20$
TOR VERGATA	$AOD_B = 0.44AOD_A + 0.17$	$AOD : A = 0.76AOD_B + 0.16$
SAPIENZA	$AOD_B = 0.46AOD_A + 0.29$	$AOD_A = 0.86AOD_B + 0.08$
LECCE	$AOD_B = 0.68AOD_A + 0.13$	$AOD_A = AOD_B + 0.01$

Table 4.2: *Table of Linear Regression relations*

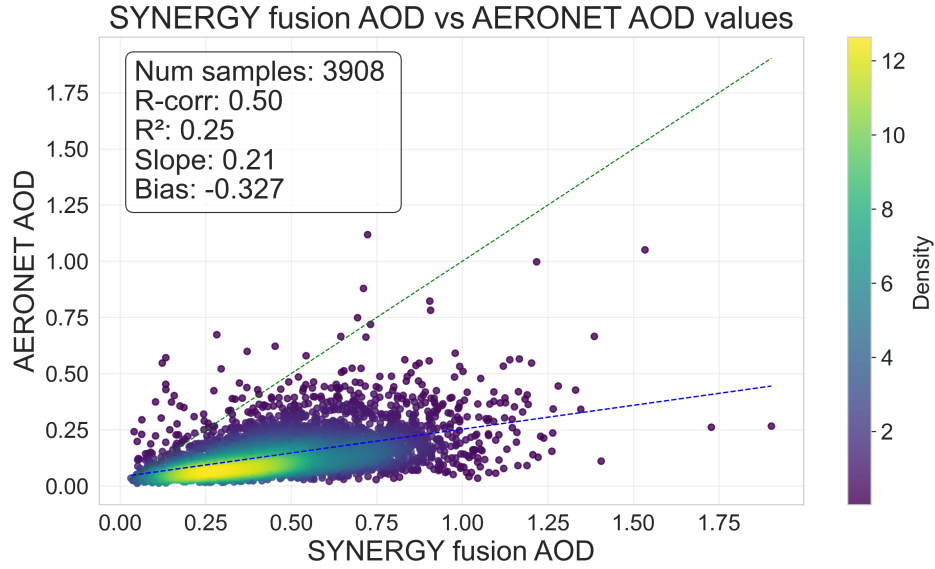


Figure 4.5: *Correlation between Sentinel-3 Synergy fusion AOD and AERONET ground-based measurements.*

The relatively low correlation coefficient ($R=0.50$) observed for Sentinel-3 Synergy A and B AOD compared to AERONET measurements is consistent with known limitations of the high-resolution (300m) Sentinel-3 Synergy land product documented in the literature. Lipponen et al. (2022) explicitly addressed the accuracy limitations of the 300m resolution Sentinel-3 Level-2 Synergy land product, noting that "the drawback in the Synergy land product aerosol parameters is their relatively poor accuracy" compared to the dedicated 4.5km aerosol product. [61].

This methodology allowed us to generate a more complete time series [4.5], optimizing the use of available information from both sensors and significantly improving the spatial and temporal coverage of AOD data for subsequent analyses.

4.1.2 Data exploration

With regard to the availability of data, it is imperative to emphasize the substantial discrepancy between meteorological variables and the AOD and AERONET products. The meteorological variables present complete and valid data for all days of the study period, due to their hourly temporal resolution that ensures continuous coverage. Conversely, the satellite AOD products exhibit a more constrained coverage, attributable to the implementation of quality and cloud mask flags. These flags are imperative for ensuring the reliability of measurements, as they automatically exclude observations made under suboptimal conditions, such as in the presence of clouds or other factors that would compromise the accuracy of the data. Similarly, the data from AERONET exhibit discontinuities in temporal coverage due to the methodological decision to consider exclusively level 2 data.



Figure 4.6: Average monthly AOD data products availability.

The first full dataframe has 10.230 rows, but after removing missing values (NaN), there are only 1,595 rows left. As shown in Figure 4.6, the average monthly availability of AOD data from SYN3, MODIS MAIAC and AERONET varies a lot among the seven monitoring stations during 2019-2022. The graph shows a clear pattern over the year, with more data in the summer than in the winter. In addition, Figure 4.7 presents the time series of AOD products (SYN3, MODIS MAIAC, and AERONET) acquired at the "Roma la Sapienza" station during the period 2019-2022. The analysis reveals a clear seasonal pattern characterized by greater data availability during summer months compared to winter months, typical of observation conditions in the Mediterranean area. The time series highlight the presence of well-defined seasonal cycles, with maximum AOD values systematically occurring during summer months and minimum values in winter periods. This behavior is consistent with aerosol climatology in the Mediterranean basin, where seasonality is influenced by several factors: [15] the intensification of Saharan dust transport during summer, increased photochemical formation of secondary aerosols due to higher temperatures and solar radiation, reduced precipitation that limits wet deposition processes, and the frequency of stagnant meteorological conditions that favor particulate matter accumulation. However, the most significant aspect emerging from the analysis is represented by the substantial differences observed between the three AOD products. The SYN3 dataset (light blue line) presents notably higher values, frequently exceeding 1.0 and reaching peaks up to 1.4, with considerably higher temporal variability and presence of outliers. These results are consistent with findings reported by Lipponen et al. [60], who demonstrated

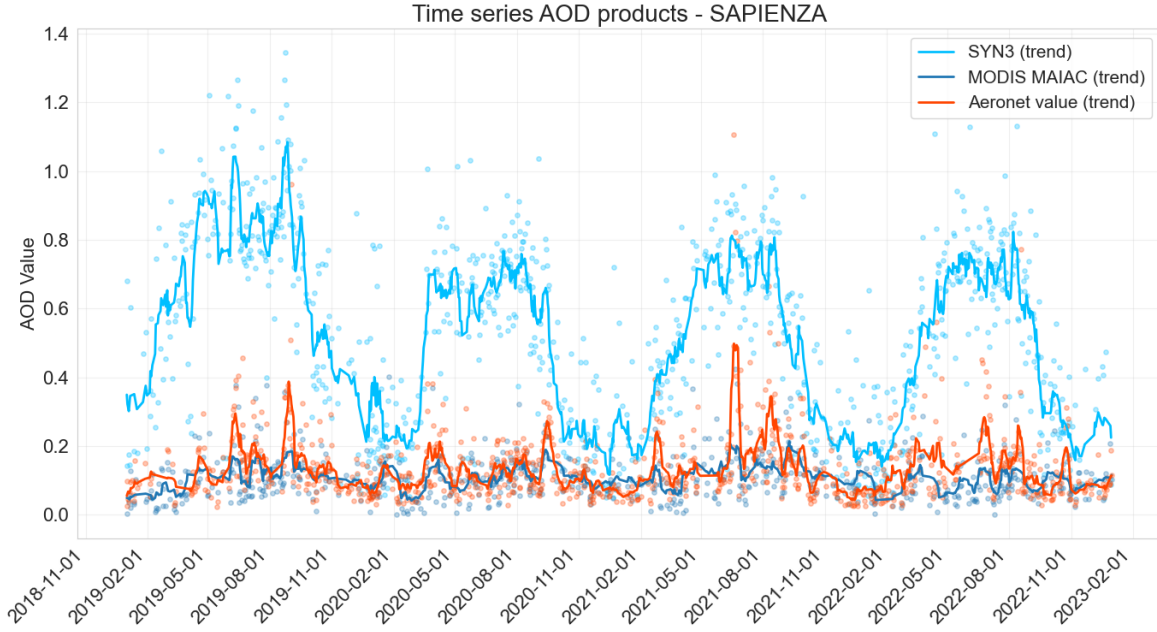


Figure 4.7: Time series of AOD products (SYN3, MODIS MAIAC, and AERONET) at the "Roma la Sapienza" station (2019-2022)

that the original 300m Synergy land product showed systematic overestimation of AOD values and limited correlation with AERONET measurements, making it necessary to develop corrective approaches based on deep learning to achieve acceptable accuracy. The MODIS MAIAC product (dark blue line) provides intermediate estimates, generally contained within 0.4-0.5, showing more gradual and consistent temporal variations with AERONET (orange line) that representing ground-based reference measurements.

To better understand the relationships between the variables, we performed an exploratory correlation analysis using Pearson correlation coefficients between the ground-truth AERONET AOD measurements and all available predictor variables. The correlation matrix presented in Figure 4.8 helps interpret the dependencies between AOD and environmental parameters, which will influence the feature selection strategy discussed in the next section 4.1.3. Interestingly, the temperature-related variables (T_2M: $r=0.37$, TD_2M: $r=0.36$) show moderate positive correlations with AOD, suggesting that thermal conditions influence aerosol concentrations, probably through processes such as secondary aerosol formation and hygroscopic growth.

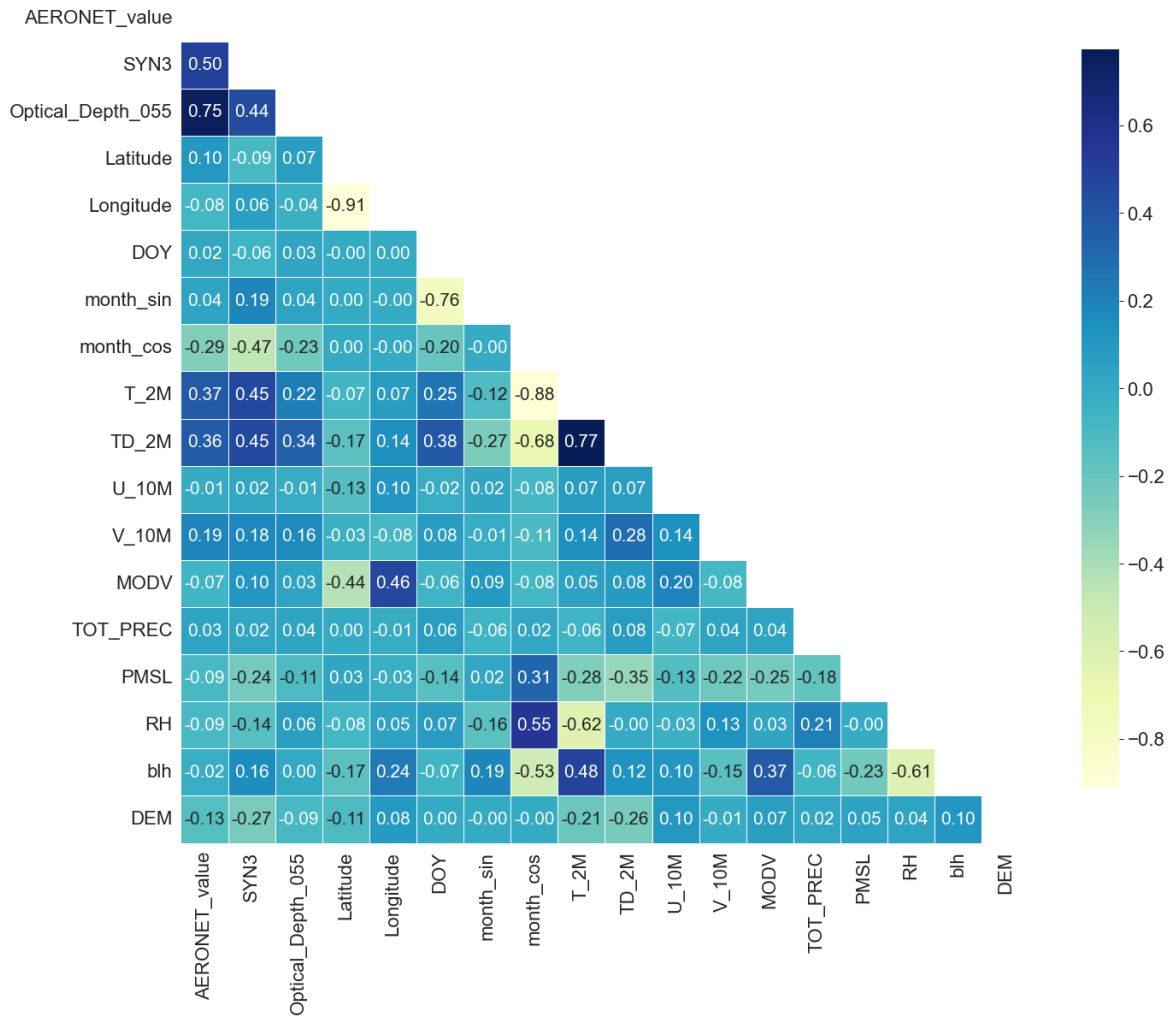


Figure 4.8: *Correlation analysis in AOD improvement task*

4.1.3 Machine Learning Approaches and Performance

In the context of improving satellite AOD data, several machine learning approaches were implemented and evaluated. The methodology followed a structured path involving the separate implementation of two main models, followed by model-specific feature selection procedures, culminating in an ensemble approach that exploits the strengths of both models.

Random Forest approach

The Random Forest algorithm (subsection 3.5.1) was implemented using the Python scikit-learn library with the *RandomForestRegressor* function, setting the number of trees to 500 to ensure robust model performance. A systematic approach to feature selection was then adopted to obtain a simpler and more interpretable model while maintaining or improving predictive performance. Recursive feature elimination (RFE, subsection 3.6.1) was applied as the primary feature selection method. This technique iteratively removes the least important features based on their contribution to the model, measured by the feature importance metric inherent to Random Forest algorithms. Figure 2 illustrates the relative importance of each feature computed by the model, providing a quantitative basis for the selection process.

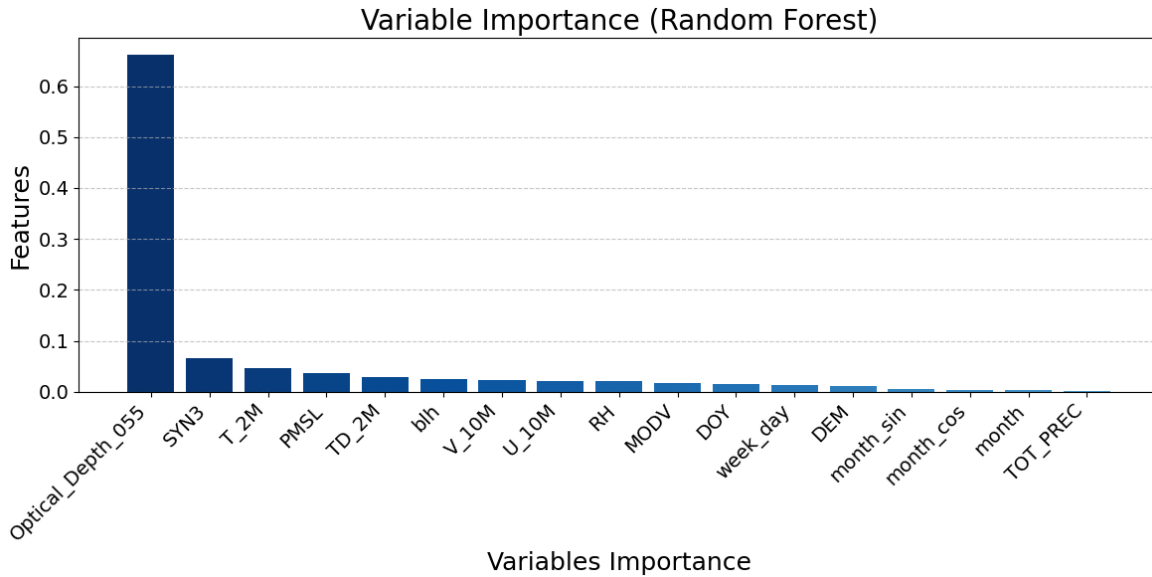


Figure 4.9: *Random Forest Variable importance.*

To validate the feature selection results, the BORUTA algorithm (subsection 3.6.2) was used as a complementary wrapper method. BORUTA uses shadow features (randomly permuted copies of the original features) to establish a statistical threshold for feature significance. The results of this analysis are presented in Figure 4.10. In particular, a substantial convergence was observed between the features identified as

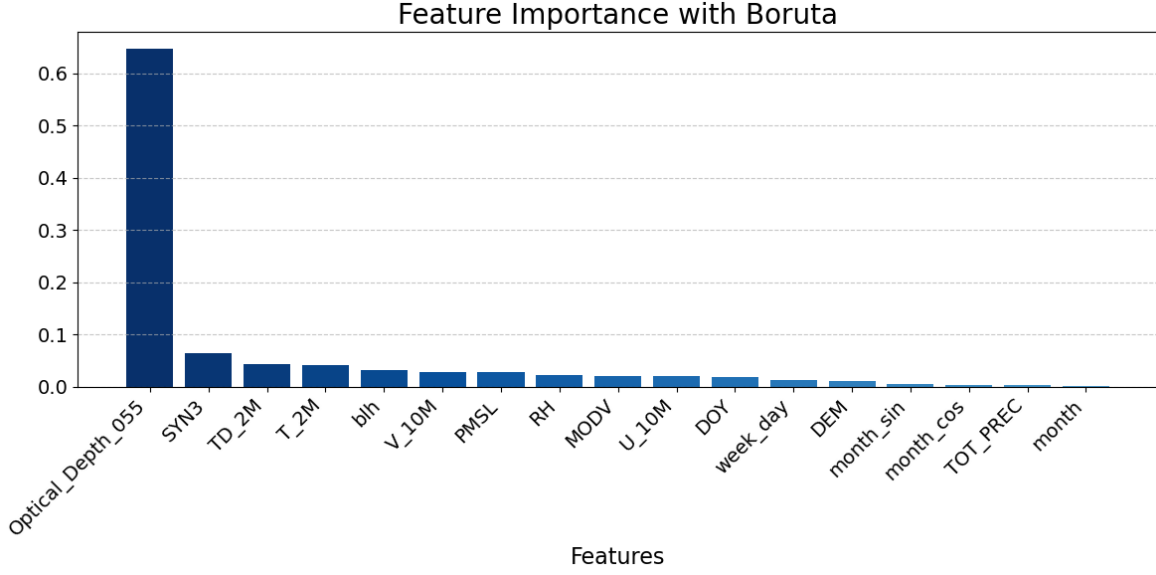


Figure 4.10: *Boruta variables importance.*

significant by the RFE and BORUTA algorithms, which increases confidence in the robustness of the feature selection process.

The convergence between the RFE and BORUTA algorithms led to the selection of the following optimal feature set:

$$y = RF(\text{Optical_Depth_055}, \text{SYN3}, \text{T_2M}, \text{T2_2M}, \text{PMSL}, \text{blh}, \text{V_10M}, \text{U_10M}, \text{MODV}, \text{RH}, \text{DOY})$$

The full validation results of this feature selection approach, including performance metrics on the training and test datasets, are presented in subsection [4.1.3](#).

Multiple Linear Regression Approach

In parallel with the Random Forest approach, a multiple linear regression model was developed to exploit its interpretability and computational efficiency. This model established direct relationships between AERONET measurements, satellite AOD products, and meteorological variables. Feature selection was implemented through a recursive feature elimination (RFE, subsection [3.6.1](#)) process, which iteratively removed variables with the lowest Pearson coefficient (see correlation matrix [4.8](#)) while monitoring model performance. This methodical dimensionality reduction was continued until performance optimization was achieved without unnecessary complexity. Full cross-validation results are detailed in the following subsection [4.1.3](#).

With $M = \{V_10M, T_2M, TD_2M, RH\}$ and $\alpha_j \in \mathcal{R}$ the final model can be represented by:

$$y = \alpha_0 + \alpha_1(\text{Optical_Depth_055}) + \alpha_2 \text{SYN3} + \alpha_3(\text{month_cos}) + \sum_{i \in M} \alpha_i i$$

Ensemble Model

A weighted ensemble model was implemented to leverage the strengths of both approaches by combining predictions from the optimized Random Forest and multiple linear regression models. The weighting scheme used the reciprocal of the root mean square error (RMSE) of each constituent model, thus assigning greater influence to the most accurate model. This methodology automatically adjusts the contribution of each model based on its predictive performance shown in different validation stages, resulting in a model that leverages the complementary strengths of both algorithmic approaches.

Validation and performance evaluation

To rigorously evaluate the performance of each methodology, a comprehensive validation framework consisting of three separate cross-validation strategies was implemented using several complementary metrics: coefficient of determination (R^2), root mean square error (RMSE), slope, bias and Pearson's correlation coefficient.

The first two convalidation are: Leave-One-Station-Out (LOSO) and Leave-One-Year-Out (LOYO). LOSO is a procedure that involves the iterative exclusion of each station from the training dataset, using it instead as a validation set. This approach evaluates the model's ability to generalize to new geographic locations not in the training set, simulating a realistic application scenario in areas without monitoring stations. LOYO similarly, each year is iteratively excluded from the training set and used for validation. This strategy tests the temporal stability of the model by testing its predictive ability under temporal conditions other than the training set.

MODEL	R^2	RMSE	Slope	Bias
RF	0.63	0.056	1.04	-0.005
LR	0.64	0.055	1.03	-0.002
Ensemble	0.66	0.054	1.05	-0.006

Table 4.3: *Leave One Station Out Cross Validation for AOD Prediction*

Tables 4.3 and 4.4 respectively show the results of the metrics, which are the averages calculated on the performance of each station or each year. Meanwhile, in the scatter plots (Figure 4.11), each point represents a data point predicted according to the leave-one-out principle: every element has been "predicted" by a model trained on a dataset that did not include that element in the training set.

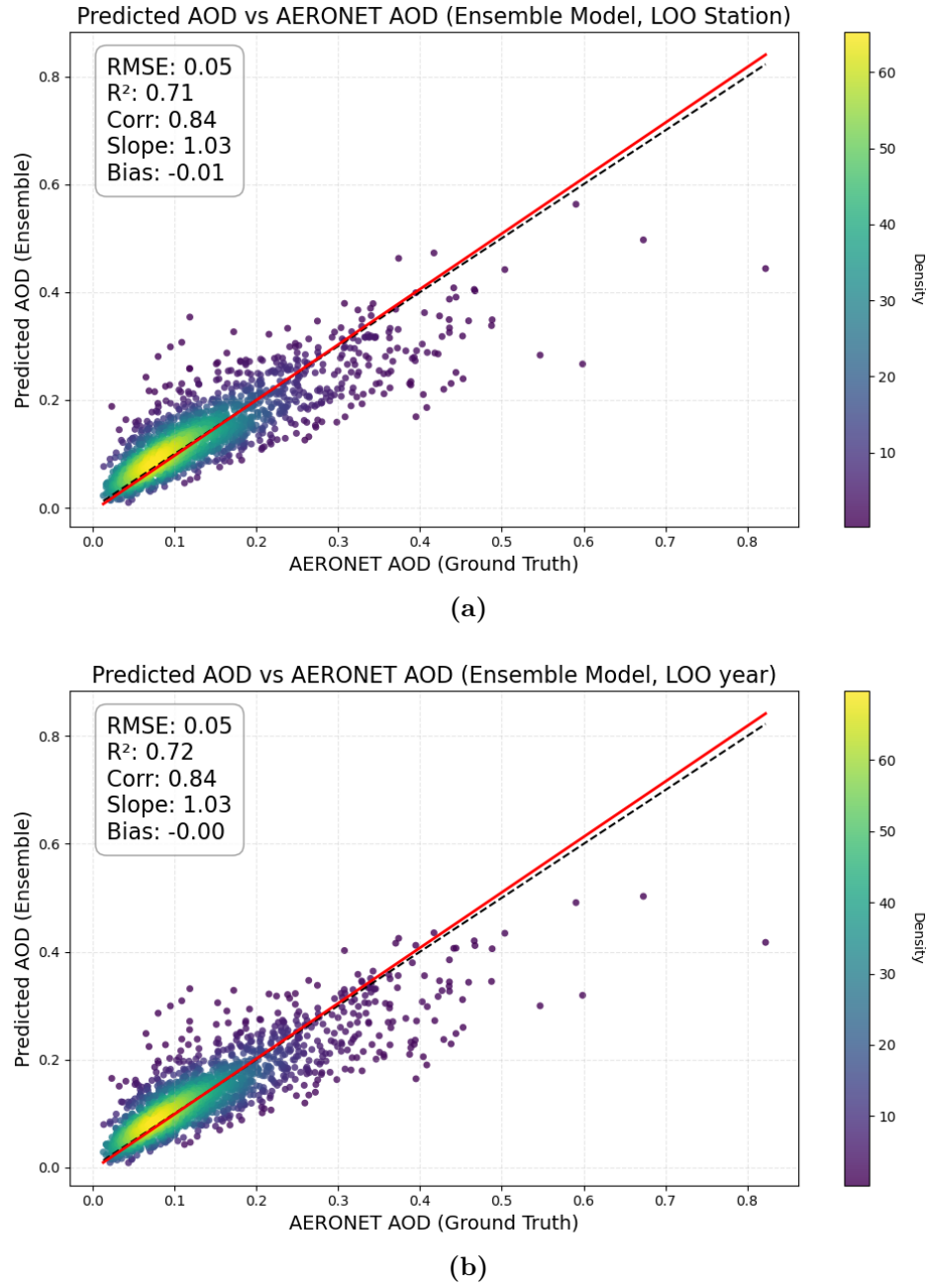


Figure 4.11: Panel (a) shows comparison between predicted AOD and ground truth AERONET AOD using the Ensemble Model with Leave-One-Station-Out validation and panel (b) using Leave-One-Year-Out validation.

MODEL	R^2	RMSE	Slope	Bias
RF	0.70	0.050	0.71	0.001
LR	0.69	0.051	0.98	0.001
Ensemble	0.71	0.049	1.02	-0.004

Table 4.4: *Leave One Year Out Cross Validation for AOD Prediction*

The third validation was performed by K-fold Cross-Validation with $k = 10$: the dataset was randomly divided into 10 partitions of equal size, iteratively using 9 partitions for training and 1 for validation. This approach provides an assessment of the overall robustness of the models while minimizing the effects of sampling variability. Similarly, in Tab 4.5 there are the average results of the metrics while in Figure 4.12 the scatterplot related to the ensemble model is depicted.

MODEL	R^2	RMSE	Slope	Bias
RF	0.73	0.048	0.73	-0.001
LR	0.70	0.051	1.03	-0.001
Ensemble	0.73	0.047	1.03	-0.004

Table 4.5: *10-Folds Cross Validation for AOD Prediction*

Based on the comprehensive analysis of the results of the three validation strategies, all models implemented (Random Forest, multiple linear regression, and ensemble) demonstrated good predictive abilities with generally comparable performance. However, the ensemble approach stood out, albeit marginally, in each validation strategy, recording slightly higher R^2 values and a corresponding reduction in RMSE compared to the individual models. This improvement, although limited, confirms the effectiveness of the weighted combination in capitalizing on the complementary strengths of both approaches, optimizing predictive performance in different spatiotemporal contexts. However, the relative homogeneity of performance suggests that each model, even individually, is a good option for application.

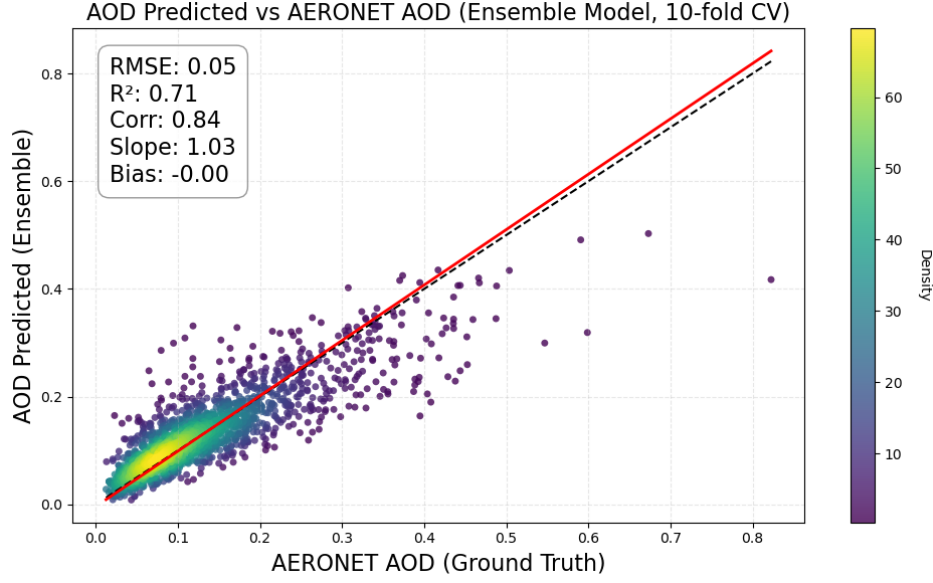


Figure 4.12: *AOD Predicted vs AERONET AOD with a 10-fold CV and an Ensemble Model*

4.2 Task 2: PM prediction

4.2.1 Data preparation

To reconstruct the spatial patterns of PM, different types of predictors were employed, classified into four main categories: meteorological data, satellite data, geographic data and temporal data. A systematic approach was developed using a 300-meter resolution square grid covering the entire study area. The WGS-32 [52] reference system was adopted for this operation. All collected raster satellite variables were projected into this coordinate reference system (CRS) and their values were extracted within the grid cells using different aggregation functions. This process allowed the construction of a multidimensional grid in which, for each 300-meter pixel, the values of all the selected features were associated, creating a coherent multivariate space. Although the multidimensional grid covers the entire study area with a high spatial resolution, the actual PM concentration values are available exclusively at the ARPA monitoring stations. Consequently, the dataset used for training and validating the Machine Learning model consisted exclusively of pixels corresponding to these stations.

PM data preparation

The data related to the concentration of atmospheric particulate matter (PM_{10} and $PM_{2.5}$) were collected from a total of 9 monitoring stations distributed across the territory, covering a period of four consecutive years (2019-2022), as explained in 3.3.3. During the initial cleaning phase, invalid values were identified and removed, repre-

sented in the original datasets by the value -999, which indicated data unavailability.

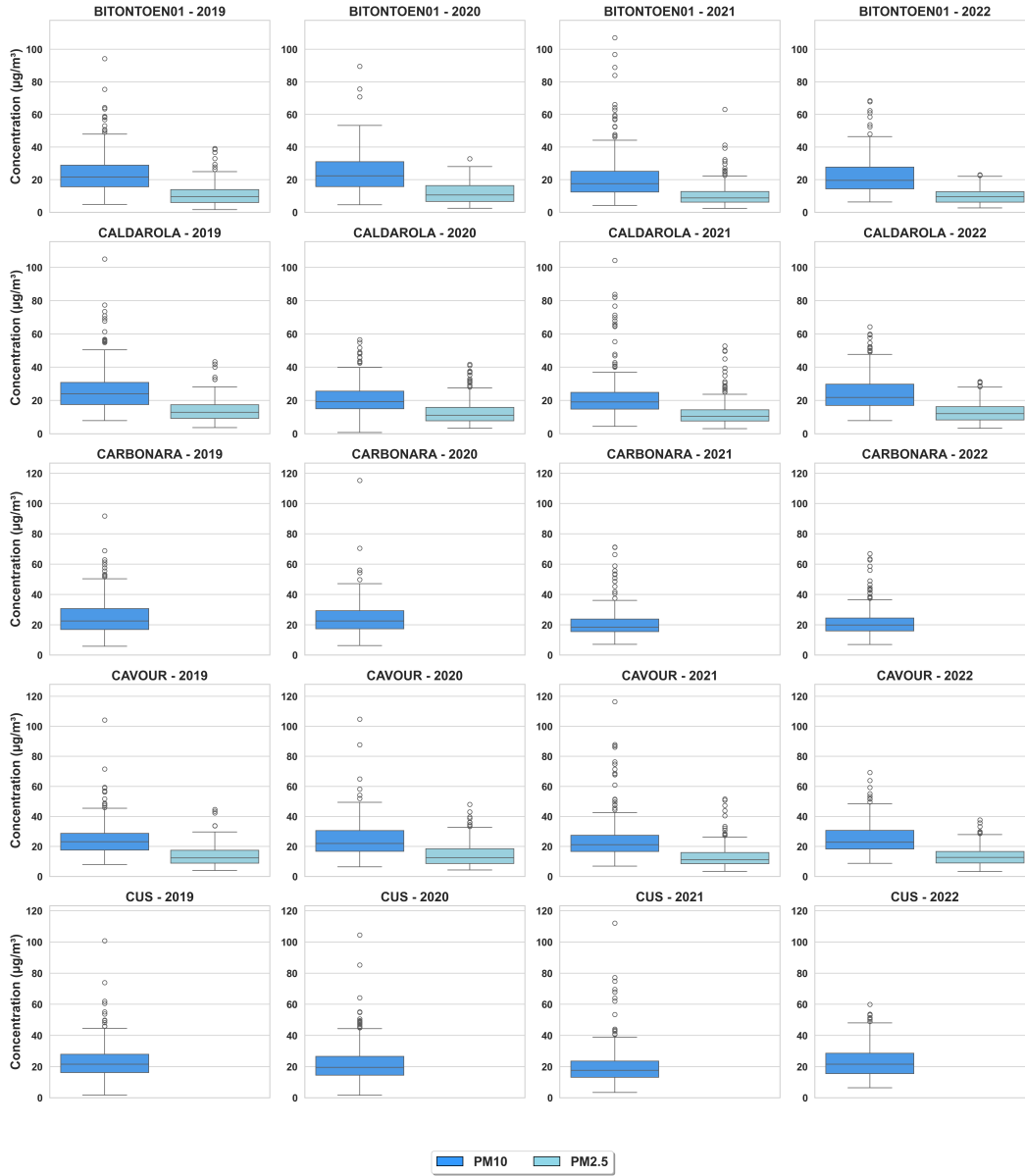


Figure 4.13: PM_{10} and $PM_{2.5}$ measurements divided by year and station
- Part 1

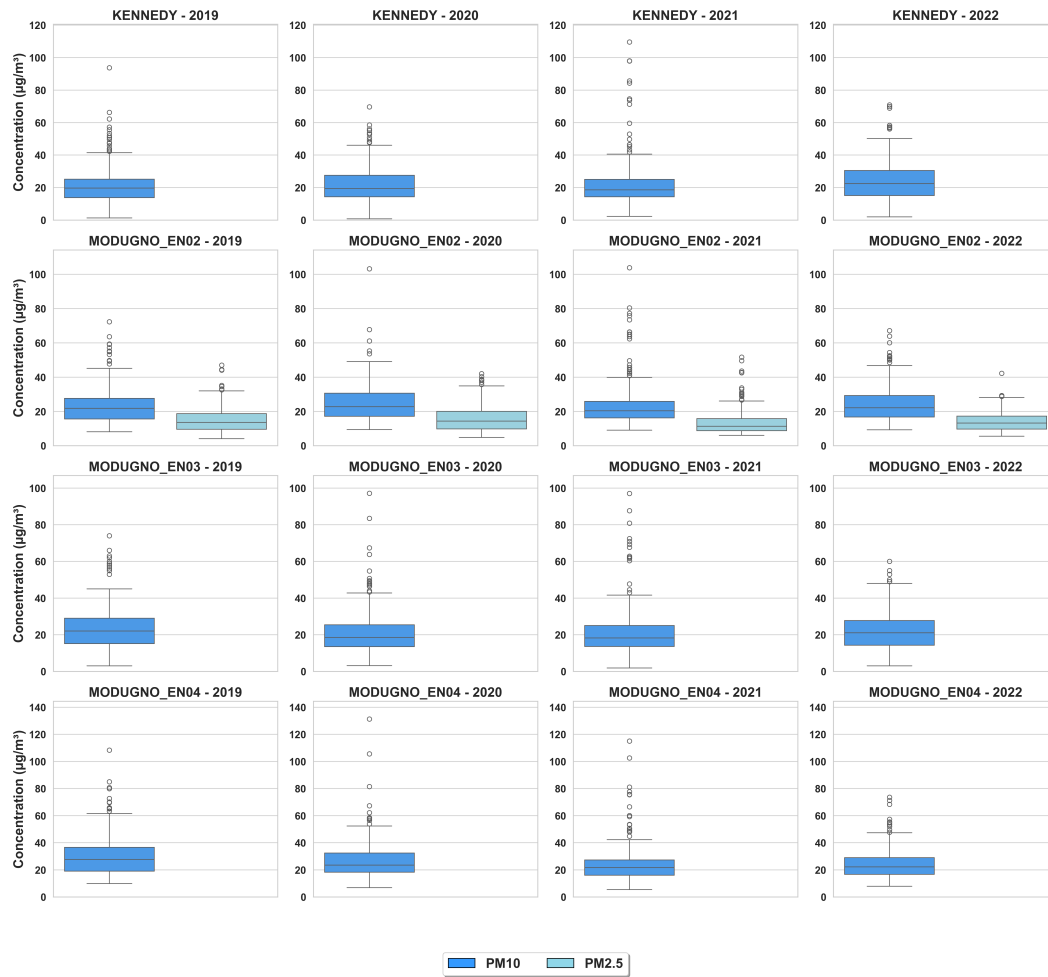


Figure 4.14: PM_{10} and $PM_{2.5}$ Measurements divided by year and station
- Part 2

This operation allows the exclusion from the analysis of periods with sensor malfunctions or interruptions in data collection. Figures 4.13 and 4.14 present box plots showing the distribution of PM_{10} (in blue) and $PM_{2.5}$ (in light blue) concentrations divided by year and monitoring station. To enrich the dataset with information on Saharan dust transport events, a binary variable (0/1) was used to indicate the days that had elapsed since dust thresholds were exceeded. These data were acquired from the ARPA station in Ciuffreda, located in the Foggia area, which serves as a reference for the Bari area for monitoring long-distance dust transport events.

Figure 4.15 shows the comparison of PM_{10} concentrations between days with dust events (labeled "Dust") and days without dust events (labeled "No Dust") at all monitoring stations. The box plots clearly show a substantial increase in PM_{10} concentrations during dust events at all stations. This pattern is consistent with Saharan dust being a significant component of PM_{10} , contributing to elevated levels during transport events.

Metereologic data preparation

Meteorological data were acquired from the Weather Research and Forecasting (WRF) model in NetCDF format. Since these data were available with hourly resolution, aggregation was performed by calculating daily averages to align them with the temporal scale of PM values. For some three-dimensional variables, such as the vertical component of wind, an average along the z-axis was calculated. This simplification was necessary due to the high correlation between measurements taken at different atmospheric heights, which would have introduced redundancy in the Machine Learning model. Furthermore, since the meteorological data from the WRF model originally had a spatial resolution of 4 km, *nearest neighbor resampling* was applied with a resolution of 300 meters. This approach associates each 300m×300m grid cell with the value of the 4km×4km pixel in which it falls, thus preserving the original pixel values without introducing artificial interpolated values.

4.2.2 Machine Learning Approaches and Performance

Stage 1 and 2: PM_{10} Prediction Model

In developing the PM_{10} forecasting model, three different machine learning algorithms were systematically evaluated: Random Forest (RF), Gradient Boosting (GBoost) and XGBoost. This comparative approach allowed to identify the most effective algorithm for the specific characteristics of the PM_{10} forecasting task.

Initial model comparison The initial evaluation framework incorporated several cross-validation strategies to ensure a comprehensive performance evaluation under different data partitioning scenarios:

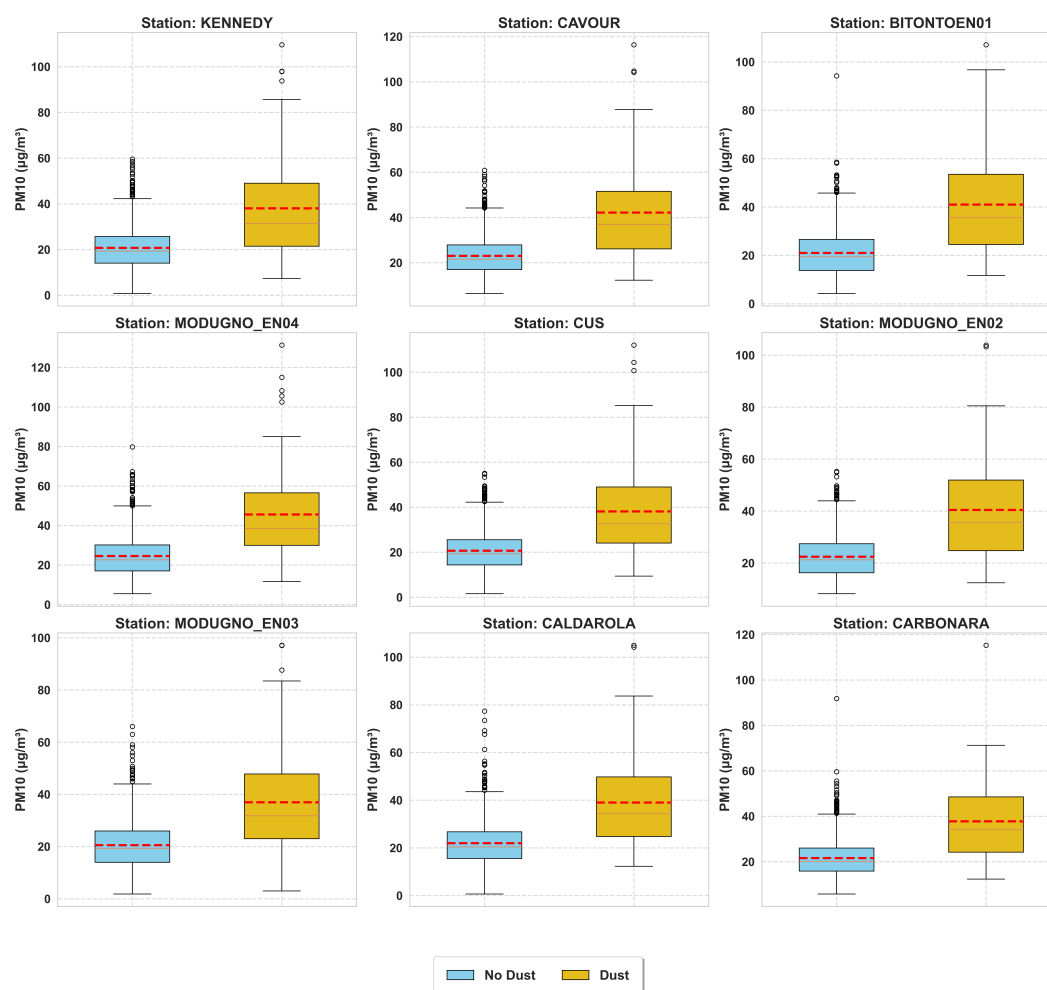


Figure 4.15: *Box plot comparison of PM_{10} concentrations between days with dust events and days without dust events across monitoring stations by ARPA.*

- k-fold cross-validation (k=10): the dataset was systematically split into five equal subsets, each of which sequentially served as the validation set, while the remaining four contributed to the training corpus. This balanced approach provides an assessment of the model performance across the entire spectrum of the dataset.
- Leave-One-Station-Out (LOSO) cross-validation: In this spatially oriented validation approach, data from individual monitoring stations were sequentially excluded from the training set and used exclusively for validation. This methodology assesses the spatial generalization capabilities of the model.
- Leave-One-Year-Out (LOYO) cross-validation: Following a similar principle, but applied to the temporal dimension, each year of data was systematically excluded from training and reserved for validation, assessing its temporal generalization capabilities.

Each algorithm was evaluated in two distinct configurations to assess the differential impact of AOD data sources on prediction accuracy:

- the complete set of features described in SECTION plus the standard AOD derived from Sentinel-3 SYNERGY satellite observations,
- an identical feature set, with the only difference being the substitution of standard AOD with the AOD improved (developed in 4.1)

The results of the different cross-validation strategies are presented in Tables 4.6, 4.7 and 4.8. Each table is organized in two main sections: the top half shows the results obtained using the standard AOD derived from Sentinel-3 SYNERGY observations, while the bottom half presents the results obtained with the improved AOD developed in Task 1 (section 4.1). For each configuration, the tables report the R^2 , RMSE and MAE values for all the implemented machine learning models. From the comparative analysis of these results, a constant improvement emerges with the use of the improved AOD. In particular, it can be observed that, for all the tested models and in all cross-validation strategies, the use of the improved AOD leads to a slight but systematic increase in the coefficient of determination (R^2), accompanied by a corresponding reduction in the predictive errors measured as RMSE and MAE. This result confirms how the AOD estimation influences the more accurate prediction of particulate matter.

Among the tested models, XGBoost (XGB) and Random Forest (RF) have demonstrated the most promising results in both configurations, as shown by their higher R^2 values and lower error metrics. Given their superior performance, these two models were selected for further optimization. XGB and RF achieved comparable results across all cross-validation strategies, with slight variations in their relative performance. Specifically, XGB showed slightly higher R^2 values in the 10-fold cross-validation and Leave-One-Year-Out (LOYO), while RF demonstrated superior performance in the Leave-One-Station-Out (LOSO) validation scenario.

Metrix	XGB	GBoost	RF
R^2	0.76 ± 0.03	0.73 ± 0.02	0.73 ± 0.02
RMSE	5.40 ± 0.32	5.65 ± 0.27	5.56 ± 0.27
MAE	3.73 ± 0.11	4.03 ± 0.10	3.91 ± 0.12
Slope	0.78 ± 0.02	0.70 ± 0.02	0.70 ± 0.02
Bias	0.13 ± 0.21	0.04 ± 0.25	0.18 ± 0.31
R^2	0.77 ± 0.01	0.75 ± 0.02	0.74 ± 0.01
RMSE	5.16 ± 0.31	5.39 ± 0.39	5.43 ± 0.28
MAE	3.67 ± 0.16	3.89 ± 0.18	3.89 ± 0.15
Slope	0.79 ± 0.04	0.72 ± 0.05	0.70 ± 0.03
Bias	0.03 ± 0.28	0.03 ± 0.31	0.19 ± 0.28

Table 4.6: 10-Fold Cross Validation comparison between Sentinel-3 and AOD improved

	XGB	GBoost	RF
R^2	0.55 ± 0.23	0.51 ± 0.20	0.60 ± 0.16
RMSE	6.40 ± 1.63	6.71 ± 1.67	5.93 ± 1.49
MAE	4.69 ± 1.34	5.09 ± 1.47	4.33 ± 1.18
Slope	0.66 ± 0.15	0.61 ± 0.09	0.63 ± 0.12
Bias	1.03 ± 1.79	1.82 ± 2.31	0.90 ± 1.88
R^2	0.57 ± 0.13	0.52 ± 0.33	0.62 ± 0.14
RMSE	6.10 ± 1.19	6.62 ± 2.45	5.70 ± 1.38
MAE	4.63 ± 1.02	5.19 ± 2.30	4.23 ± 1.09
Slope	0.69 ± 0.13	0.63 ± 0.11	0.64 ± 0.11
Bias	1.57 ± 1.74	2.25 ± 3.15	0.86 ± 1.86

Table 4.7: Leave one Station Out Cross Validation comparison between Sentinel-3 and AOD improved

	XGB	GBoost	RF
R^2	0.25 ± 0.12	0.26 ± 0.17	0.25 ± 0.17
RMSE	8.14 ± 0.80	7.91 ± 1.34	8.05 ± 1.07
MAE	5.85 ± 0.76	5.90 ± 1.31	5.74 ± 0.81
Slope	0.41 ± 0.24	0.40 ± 0.19	0.38 ± 0.22
Bias	0.76 ± 1.62	1.39 ± 2.27	0.50 ± 0.93
R^2	0.32 ± 0.11	0.31 ± 0.11	0.27 ± 0.17
RMSE	7.47 ± 0.99	7.49 ± 0.91	7.94 ± 0.81
MAE	5.22 ± 0.46	5.23 ± 0.58	5.52 ± 0.68
Slope	0.41 ± 0.13	0.40 ± 0.08	0.35 ± 0.11
Bias	-0.03 ± 1.13	0.26 ± 1.07	0.61 ± 0.96

Table 4.8: *Leave one Year Out Cross Validation comparison between Sentinel-3 and AOD improved*

The optimization process was structured in two sequential phases:

- a feature selection procedure was implemented to identify the most relevant variables for PM prediction, potentially reducing model complexity while maintaining or improving predictive power,
- a systematic hyperparameter optimization was performed to identify the optimal configuration for each algorithm.

Models optimization To identify the most relevant variables for PM prediction using the XGBoost model, SHAP (SHapley Additive exPlanations, subsection 3.6.3) values were calculated for each feature. SHAP values provide a consistent and theoretically valid approach to quantify feature importance, measuring the contribution of each feature to the prediction for each instance. Specifically, the SHAP implementation in TreeExplainer, optimized for tree-based models, was utilized. The magnitude of SHAP values served as an indicator of feature importance, allowing variables to be ranked based on their overall impact on the model's predictions. To determine the optimal number of features for XGBoost, an iterative selection process was implemented. Starting from the complete set of features, models were progressively built with a decreasing number of features, always selecting those with the highest SHAP magnitude values. For each subset of features, model performance was evaluated using a robust 10-fold cross-validation approach to obtain reliable estimates of R^2 , RMSE, and MAE.

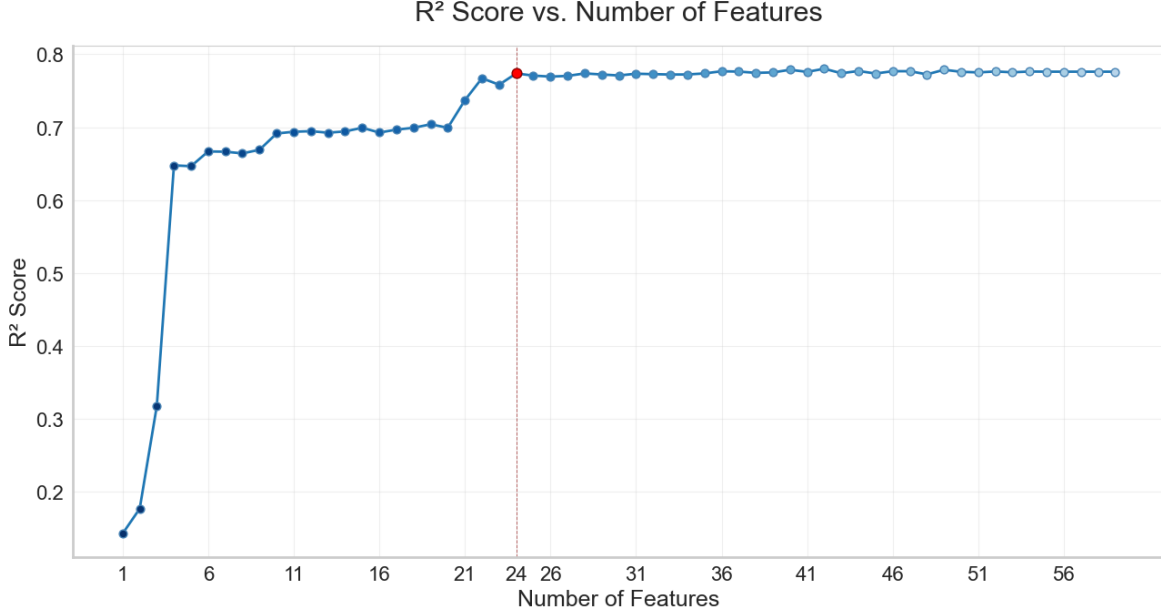


Figure 4.16: R^2 score vs Number of Features with XGBoost model for PM_{10}

This cross-validation strategy ensured that the feature selection process was not influenced by peculiarities in the data partitioning. The optimal feature subset was selected based on the configuration that maximized R^2 and minimized RMSE, thus achieving a balance between model accuracy and complexity.

The analysis illustrated in Figure 4.16 shows the relationship between the number of features used and the coefficient of determination R^2 . The results highlight an optimal point corresponding to 24 features, highlighted in red in the graph. Up to this threshold value, there is a rapid increase in R^2 , which rises from approximately 0.15 with a single feature to almost 0.78 with 24 features. Beyond this critical point, the curve stabilizes, with marginal gains of less than 0.01 in R^2 even when doubling the number of variables. This behavior identifies an optimal balance between computational complexity and predictive power of the model, allowing for significant dimensional reduction without compromising the accuracy of predictions. Figure 4.17 illustrates the relative contribution of the 28 most influential features, classified according to the absolute mean SHAP values. The horizontal bars represent the impact of each variable, while the overlaid black lines indicate the standard deviation, providing information on the consistency of their contribution. The analysis reveals that four variables clearly dominate the predictive capacity of the model: the planetary boundary layer height (PBLH) with a mean SHAP value greater than 2.0, the improved AOD (AOD_impr) derived from the first phase of the research with a value of approximately 1.75, the atmospheric dust concentration (dust) around 1.5, and the temporal variable of the Julian day (julian_day) with a value close to 1.25. These key variables also present

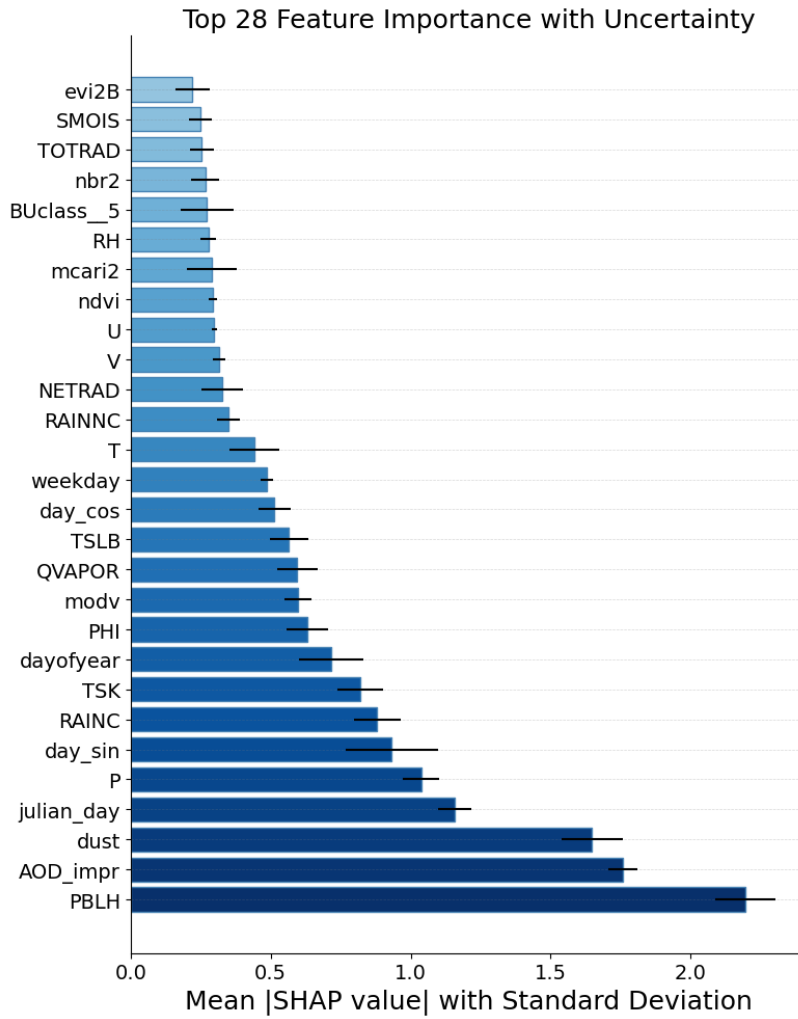


Figure 4.17: Mean Shap Value with standard deviation with XGBoost model for PM_{10} .

contained standard deviations, confirming their consistent influence across different predictive scenarios.

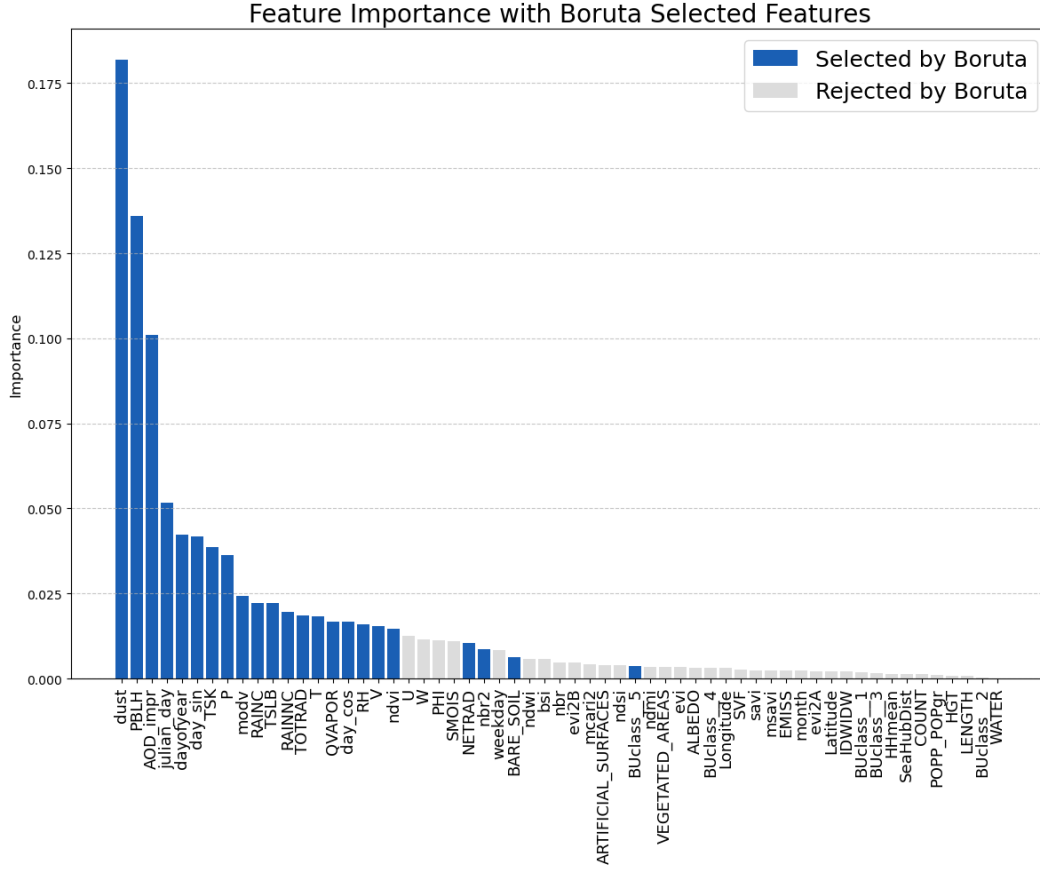


Figure 4.18: Selected features by RF (Boruta approach) for PM_{10} .

For the Random Forest model, a different feature selection approach was employed. The Boruta algorithm (subsection 3.6.2) was used as the main feature selection method. Boruta is a wrapper algorithm specifically designed as an extension of Random Forest, making it particularly suitable for this model. It determines the statistical confidence of each feature by comparing the importance of original attributes with the importance of their randomly permuted copies (shadow features), leveraging the same intrinsic variable importance measure of Random Forest. Figure 4.18 shows the results of the Boruta analysis, highlighting the relative importance of all considered variables. The blue bars represent the 23 features identified as statistically significant by the algorithm, while the gray bars indicate variables discarded because they were not sufficiently relevant for PM prediction. Similar to what emerged from the SHAP analysis for XGBoost, in this case as well, the variables with greater importance turn out to be atmospheric dust concentration (dust), planetary boundary layer height (PBLH), and improved AOD (AOD_impr). These are followed by temporal variables such as Julian day (julian_day), as well as meteorological parameters such as surface temperature (TSK) and atmospheric pressure (P).

After feature selection, the hyperparameters of both models were optimized using Grid Search. For XGBoost, the following combinations were explored:

- Number of estimators (*n_estimators*): 600 and 1000, with 1000 being optimal
- Learning rate (*learning_rate*): 0.07, 0.08, 0.09, with 0.07 being optimal
- Maximum tree depth (*max_depth*): 5, 6, 7 and 10, with 6 being optimal
- Training sample fraction (subsample): 0.7, 0.8, 0.9, with 0.9 being optimal
- Column fraction per tree (*colsample_bytree*): 0.7, 0.8, 0.9, with 0.8 being optimal

For Random Forest, the Grid Search examined:

- Number of trees (*n_estimators*): 600 and 1000, with 1000 being optimal
- Maximum tree depth (*max_depth*): 4, 6, 8, with 6 being optimal
- Minimum number of samples to split a node (*min_samples_split*): 2, 4, 10, with 4 being optimal

Both processes used parallelization (*n_jobs=-1*) to optimize computational resources. XGBoost completed the Grid Search in significantly less time compared to Random Forest, a notable advantage for applying the model to extensive spatial areas.

Model	Metrix	10-FOLDS	LOSO	LOYO
XGB	R^2	0.81 ± 0.01	0.63 ± 0.15	0.36 ± 0.09
	RMSE	4.90 ± 0.31	5.64 ± 1.59	7.19 ± 0.92
	MAE	3.43 ± 0.13	4.12 ± 1.22	5.00 ± 0.50
	Slope	0.78 ± 0.04	0.69 ± 0.13	0.40 ± 0.12
	Bias	0.07 ± 0.21	0.42 ± 1.79	0.22 ± 0.83
RF	R^2	0.75 ± 0.01	0.62 ± 0.15	0.25 ± 0.18
	RMSE	5.40 ± 0.26	5.67 ± 1.47	8.01 ± 0.91
	MAE	3.89 ± 0.13	4.13 ± 1.13	5.59 ± 0.79
	Slope	0.71 ± 0.03	0.67 ± 0.15	0.36 ± 0.12
	Bias	0.22 ± 0.28	0.69 ± 1.84	0.68 ± 1.09

Table 4.9: *Optimized XGB and RF Cross Validation.*

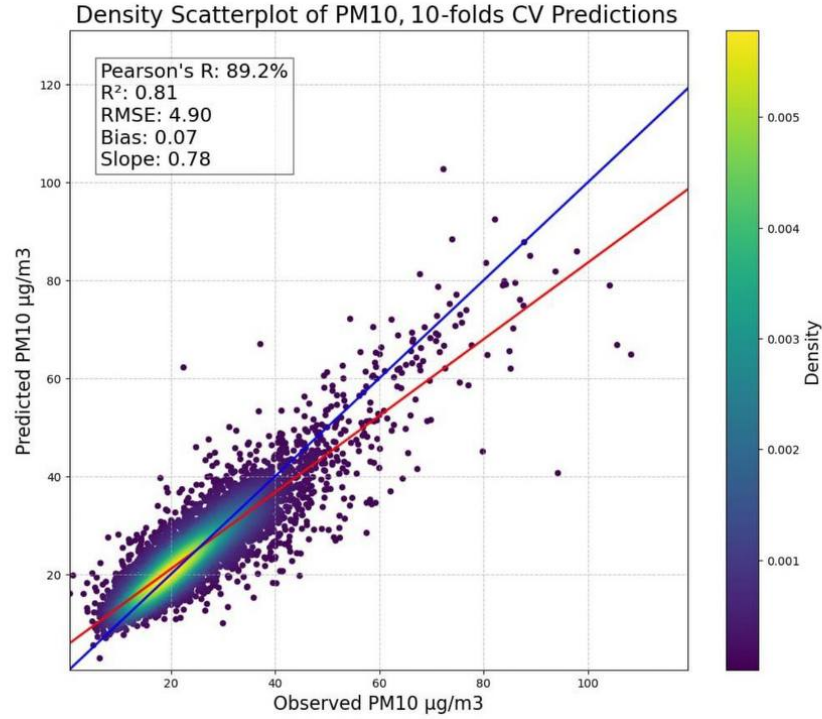


Figure 4.19: *XGB predicted vs observed PM_{10} density scatter plot.*

The models with optimal configurations were evaluated and compared through three cross validations, whose metrics are presented in table 4.9. These metrics represent the average performance across test sets in k-fold, leave-one-station-out, and leave-one-year-out cross-validation strategies, providing a comprehensive assessment of model generalizability across different spatial and temporal data partitions. XGBoost demonstrated superior performance in terms of both predictive accuracy and computational efficiency and will therefore be used as the reference model for the next phase of the analysis.

As described in Section 4.2.1, some PM_{10} values (and similarly $PM_{2.5}$) were encoded by ARPA with the value -999 to indicate their unavailability. The dataset used contained 6576 total observations, of which 264 had $PM_{10}=-999$ and were therefore excluded from the initial analysis. The second stage of the methodology involves applying the developed XGBoost model to these observations, in order to replace the missing values with PM predictions. This approach allows for significantly increasing the number of observations available for the subsequent stage.

As shown in Table 4.10, the model maintains the temporal trends and variability of PM_{10} concentrations, while expanding data coverage by approximately 5% per year, effectively filling gaps in the monitoring network.

Target	Year	Samples	Mean	Std	Percentiles				
					5	25	50	75	95
Observed									
PM10	2019	3130	24.40	11.69	10.40	16.47	22.40	29.60	45.71
PM10	2020	3126	23.39	11.23	10.00	15.70	21.20	29.12	42.77
PM10	2021	3171	21.87	12.38	9.30	14.78	19.22	25.31	40.87
PM10	2022	3163	23.28	10.11	10.26	16.00	21.49	28.56	41.51
Predicted									
PM10	2019	3285	24.57	11.56	10.50	16.78	22.81	29.75	45.60
PM10	2020	3294	23.58	11.32	10.13	15.90	21.42	29.30	42.89
PM10	2021	3285	22.12	12.52	9.37	14.90	19.47	25.60	41.60
PM10	2022	3285	23.33	10.03	10.30	16.20	21.60	28.70	41.40

Table 4.10: PM_{10} observed and predicted in the stage 3 concentrations ($\mu g/m^3$) (2019-2022)

Stage 3: $PM_{2.5}$ Prediction for Incomplete Stations

The objective of Stage 3 is to expand the availability of $PM_{2.5}$ data using a predictive model mainly based on daily PM_{10} concentrations from monitoring stations that do not directly measure $PM_{2.5}$. This approach was necessitated by the limited spatial coverage of stations measuring $PM_{2.5}$ compared to those measuring PM_{10} . The preliminary analysis showed a strong correlation between PM_{10} and $PM_{2.5}$ (Pearson's $R = 0.82$, 4.20), supporting the hypothesis that PM_{10} can be used as an effective predictor of $PM_{2.5}$.

For this stage, a Random Forest model was defined and structured as follows:

$$PM_{2.5} \sim \{PM_{10}, ST, month, weekday, day_sin, day_cos, latitude, longitude, IDWIDW, SeaHubDist\}$$

where "month" and "weekday" define respectively the month and day of the week; "sin_day" and "cos_day" are cyclical transformations of the day of the year that capture annual seasonality through sinusoidal functions, allowing the model to recognize temporal cyclicity without discontinuities between the end and beginning of the year:

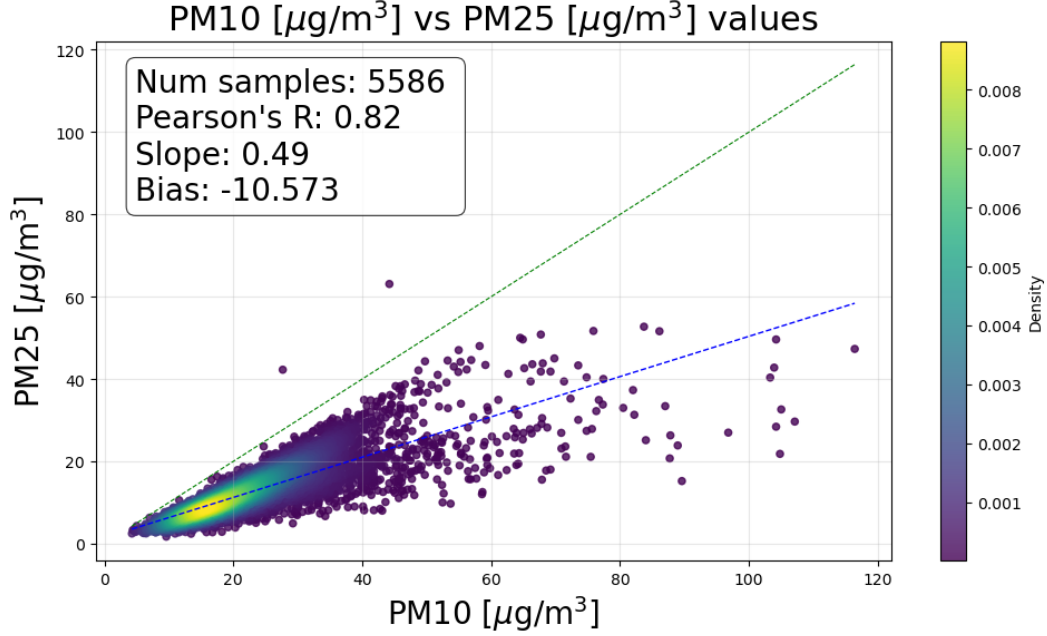


Figure 4.20: Correlation between PM_{10} and $PM_{2.5}$ over four stations.

$$\sin_day = \sin\left(\frac{2\pi \times DOY}{12}\right) \quad (4.6)$$

$$\cos_day = \cos\left(\frac{2\pi \times DOY}{12}\right) \quad (4.7)$$

where DOY is the day of the year, "lat" and "lon" represent the geographical coordinates of monitoring station, "IDWIDW" represents the sum of inverse distances from other ARPA stations for station, providing a measure of monitoring density in the area; "SeaHubDist" indicates the minimum distance of station from the sea, constituting an important factor for capturing the influence of coastal areas on particulate composition. "ST" is a categorical variable that classifies monitoring ARPA station as "background", "traffic", or "industrial", Tab. 4.11.

Station Type	Category
Background	0
Traffic	1
Industrial	2

Table 4.11: Categorical variables for ARPA stations.

The variables were selected a priori with the aim of capturing general spatial trends

(geographical coordinates), seasonal patterns (month), and weekly variations (day of the week) in PM concentrations. After a preliminary analysis for parameter optimization, we selected $num.trees = 1000$ and $mtry = 4$ for the RF model. A distinctive feature of this stage was the use of a limited training approach: the model was trained using data from only four monitoring stations that provided measurements of both PM_{10} and $PM_{2.5}$.

The model evaluation was carried out through two distinct approaches:

- Assessment of the correlation between observed $PM_{2.5}$ concentrations and those predicted in *out-of-bag* (*OOB*) samples, which are data not included in the bootstrap samples at each forest iteration, and therefore do not contribute to the definition of each decision tree.
- Application of a 10-fold cross-validation approach on monitoring stations, randomly dividing the total set of four stations into ten groups, applying the model on nine groups ("training set") and predicting it on the tenth group ("testing set"), finally verifying the correlation between observed $PM_{2.5}$ concentrations and those predicted at the excluded stations.

CV	R2	RMSE	Bias	Slope
OOB	0.89	2.32	0.006	0.87
10-FOLDS CV	0.88	2.33	0.002	0.87

Table 4.12: *Evaluation of RF model for $PM_{2.5}$ over four stations.*

After validating the model robustness through both out-of-bag evaluation and station-based cross-validation, we proceeded to train the final model on the complete dataset of the four stations that simultaneously measured both PM_{10} and $PM_{2.5}$. Prior to model training, we performed data cleaning by removing all invalid measurements (marked as -999) for both $PM_{2.5}$ and PM_{10} parameters to ensure data quality. Once the final model was trained, we applied it to the remaining five stations that only measured PM_{10} to infer the missing $PM_{2.5}$ values. This approach allowed us to complete the $PM_{2.5}$ dataset across all monitoring stations in the network, significantly enhancing the spatial coverage of fine particulate matter data. Figure 4.22 illustrates the correlation between the PM_{10} data completed in Stage 2 and the $PM_{2.5}$ data completed in this stage. The analysis confirms that the statistical metrics remain consistent with the validation results obtained earlier, demonstrating the reliability of our approach.

Table 4.13 compares the observed and predicted PM_{25} concentrations from 2019 to 2022. The results show good correspondence between measured and estimated values. The statistics (mean, standard deviation, and percentiles) of the predicted data are very similar to those of the observed data for all years. The difference between annual

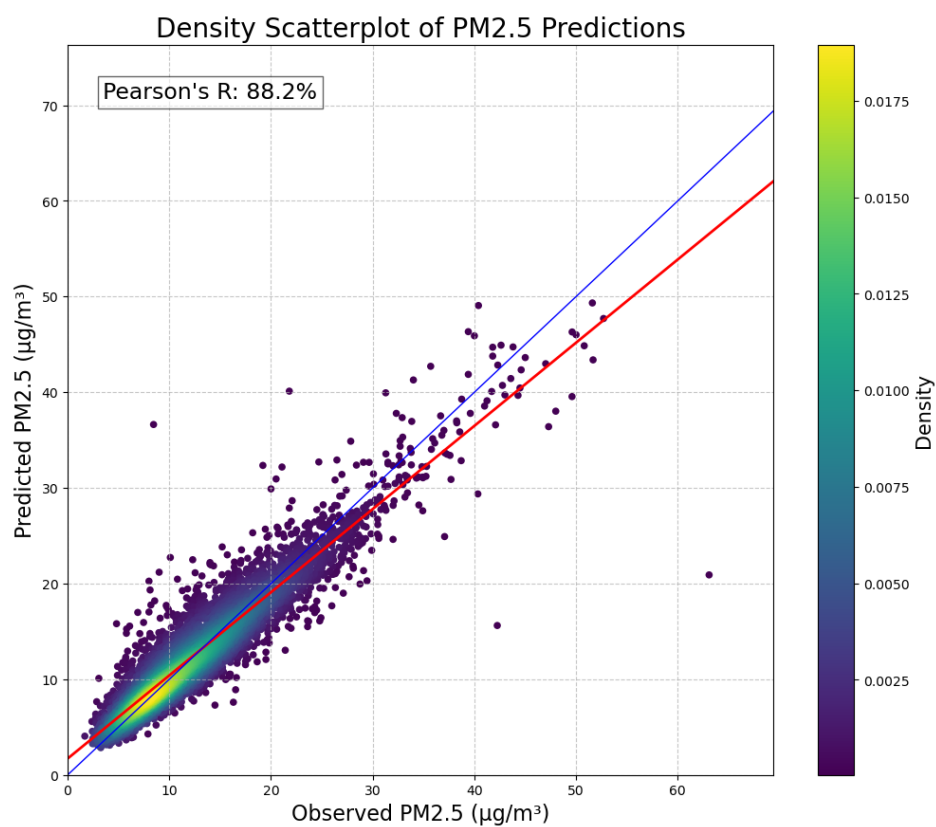


Figure 4.21: Scatterplot of observed and 10-fold CV predicted $PM_{2.5}$ ($\mu\text{g}/\text{m}^3$)

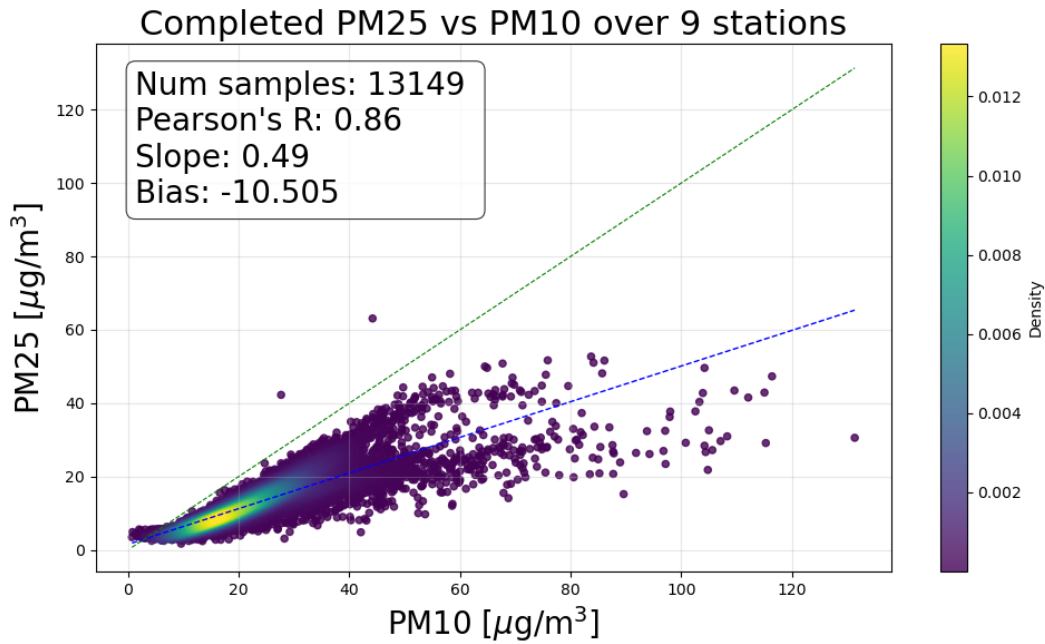


Figure 4.22: *Correlation between completed $PM_{2.5}$ and PM_{10} over all stations.*

means is minimal, always less than $0.2 \mu\text{g}/\text{m}^3$, with a maximum deviation of 0.75% in 2019 (13.43 vs $13.33 \mu\text{g}/\text{m}^3$). The distribution of values across the various percentiles is also well aligned between the two datasets. The model has significantly expanded data availability, increasing from approximately 1400 observed samples to over 3280 predicted values per year, more than doubling the coverage of the monitoring network.

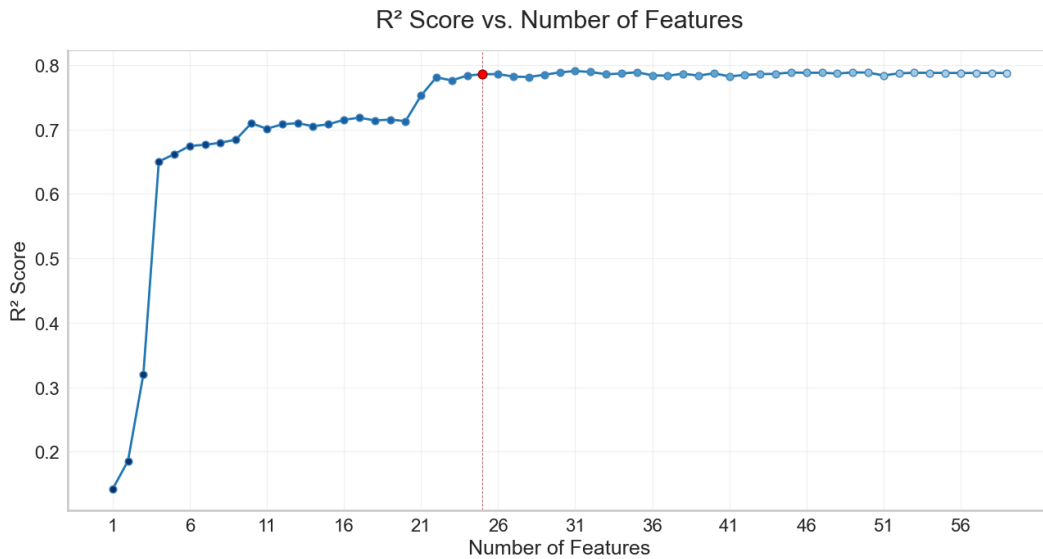
Target	Year	Samples	Mean	Std	Percentiles				
					5	25	50	75	95
Observed									
PM25	2019	1372	13.33	6.72	4.82	8.46	12.17	17.10	25.03
PM25	2020	1400	13.84	7.33	5.10	8.17	12.30	17.82	27.98
PM25	2021	1427	12.07	6.74	4.80	7.80	10.50	14.70	23.72
PM25	2022	1416	12.63	5.62	5.10	8.40	11.89	15.71	23.16
Predicted									
PM25	2019	3285	13.43	6.49	4.73	8.54	12.62	17.40	24.27
PM25	2020	3294	13.65	7.10	5.03	8.24	12.07	17.45	27.16
PM25	2021	3285	12.01	6.50	5.00	7.83	10.38	14.62	23.84
PM25	2022	3285	12.48	5.44	5.25	8.21	11.78	15.70	22.50

Table 4.13: $PM_{2.5}$ observed and predicted in the stage 3 concentrations ($\mu g/m^3$)(2019-2022)

Stage 4: Comprehensive PM_{2.5} Prediction Model

In this phase, a complete model for predicting PM_{2.5} is developed using the full dataset of 6300 observations obtained from previous phases. Based on prior considerations, the eXtreme Gradient Boosting (XGBoost) algorithm was used as the primary modeling technique. Initially, the model incorporated the entire spectrum of features described in the data preparation section (4.2.1), including meteorological variables, temporal factors, AOD measurements, geographical coordinates, and land use parameters. Before proceeding with optimization, an evaluation of the initial model using three different cross-validation strategies was conducted: as in previous models, for this complete model, an evaluation was conducted through the three cross-validation strategies already described: K-Folds (k=10), Leave-One-Station-Out (LOSO), and Leave-One-Year-Out (LOYO), the results of which are found in the table:

XGB	10-FOLDS	LOSO	LOYO
R^2	0.77 ± 0.03	0.57 ± 0.29	0.42 ± 0.09
RMSE	3.05 ± 0.25	3.86 ± 1.15	4.83 ± 0.46
Slope	0.79 ± 0.03	0.70 ± 0.12	0.44 ± 0.05
Bias	0.00 ± 0.20	0.99 ± 1.75	0.61 ± 1.34

Table 4.14: XGB Cross Validation for PM_{2.5} predictions.**Figure 4.23:** R^2 score vs number of features.

Similar to what was described in Stage 1 [4.2.2](#), we apply feature selection based on variable importance given by SHAP values. As shown in Figure [4.24](#), the increase in R^2 as a function of the number of features used highlights an optimal point corresponding to 25 features (indicated by the red dot), after which the addition of further variables does not produce significant improvements in the predictive performance of the model. Figure [4.23](#) illustrates the relative importance of the 25 main features according to SHAP values, confirming that PBLH (Planetary Boundary Layer Height) and AOD_impr (the AOD data improved in the previous phase) are the variables with the greatest predictive impact.

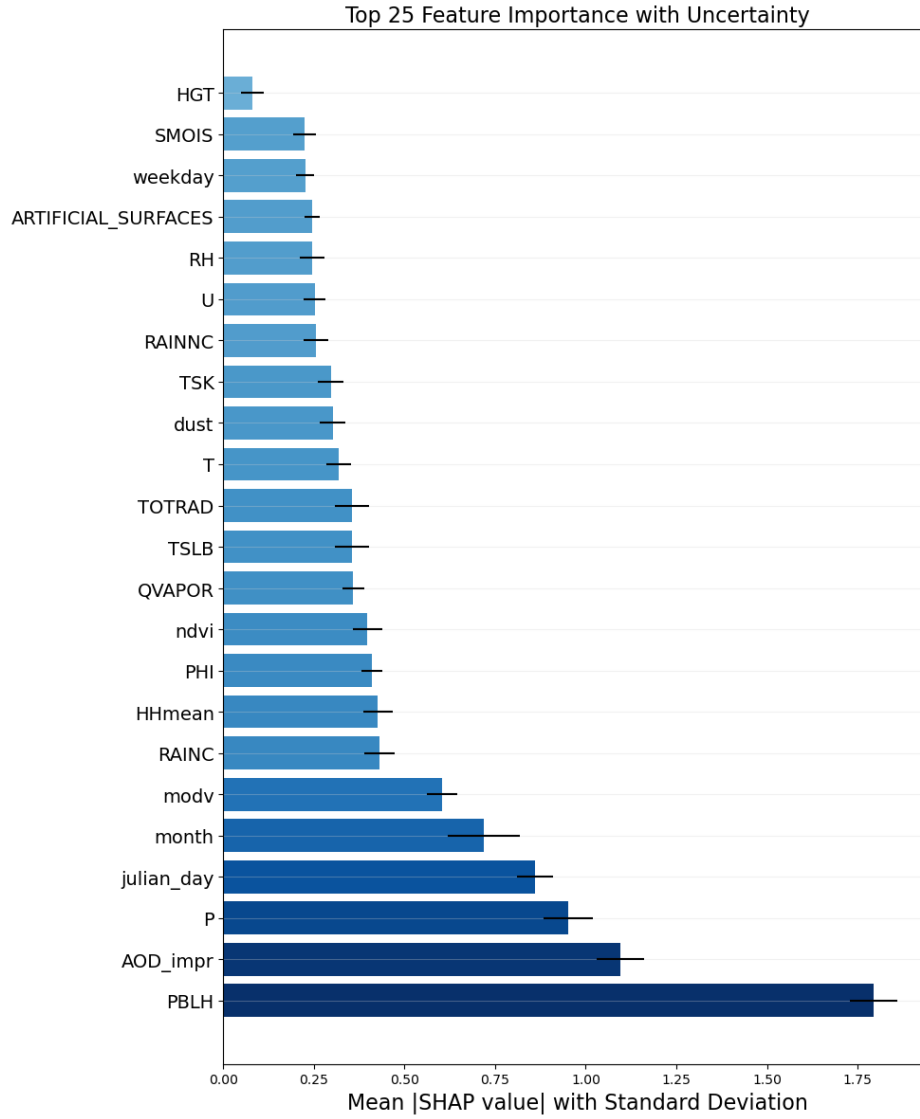


Figure 4.24: Mean shap value with standard deviation with XGBoost model (stage 4).

Figure 4.23 illustrates the relative importance of the 28 main features according to SHAP values, confirming that PBLH (Planetary Boundary Layer Height) and AOD impr (the AOD data improved in the previous phase) are the variables with the greatest predictive impact. The 25 selected features include meteorological parameters (such as temperature, relative humidity, and pressure), temporal variables (julian day), vegetation indicators (NDVI), and factors related to soil and dust presence. Notably, urban morphology features such as building height (HHmean) and artificial surfaces percentage (ARTIFICIAL_SURFACES) appear among the predictive variables, highlighting the influence of urban structure on air quality patterns. The inclusion of these anthropogenic factors alongside natural variables demonstrates the complex interplay between built environment characteristics and atmospheric particulate matter concentrations. For the tuning of parameters the following combinations were explored:

- Number of estimators (*n_estimators*): 600 and 1000, with 1000 being optimal
- Learning rate (*learning_rate*): 0.07, 0.08, 0.09, with 0.07 being optimal
- Maximum tree depth (*max_depth*): 5, 6, 7 and 10, with 6 being optimal
- Training sample fraction (subsample): 0.7, 0.8, 0.9, with 0.8 being optimal
- Column fraction per tree (*colsample_bytree*): 0.7, 0.8, 0.9, with 0.7 being optimal

Table 4.15 presents the results of the XGBoost model evaluation for $PM_{2.5}$ prediction through three different cross-validation strategies. The reported values represent performance metrics calculated by averaging across all predictions of observations that served as the test set during the validation process. Overall, the model shows good predictive capability at the spatial level. However, in temporal validation, the performance is more limited ($R^2 = 0.45$), with a higher RMSE and a significantly lower regression line slope (0.44), indicating a reduced ability of the model to extrapolate predictions to time periods different from those used in training.

Metrix	Overall	Spatial	Temporal
R^2	0.80 ± 0.04	0.69 ± 0.15	0.45 ± 0.07
RMSE	2.90 ± 0.04	3.30 ± 0.70	4.51 ± 0.07
Slope	0.79 ± 0.02	0.73 ± 0.12	0.44 ± 0.07
Bias	0.05 ± 0.20	0.15 ± 0.8	0.27 ± 1.05

Table 4.15: XGB evaluation metrix for $PM_{2.5}$ metrix.

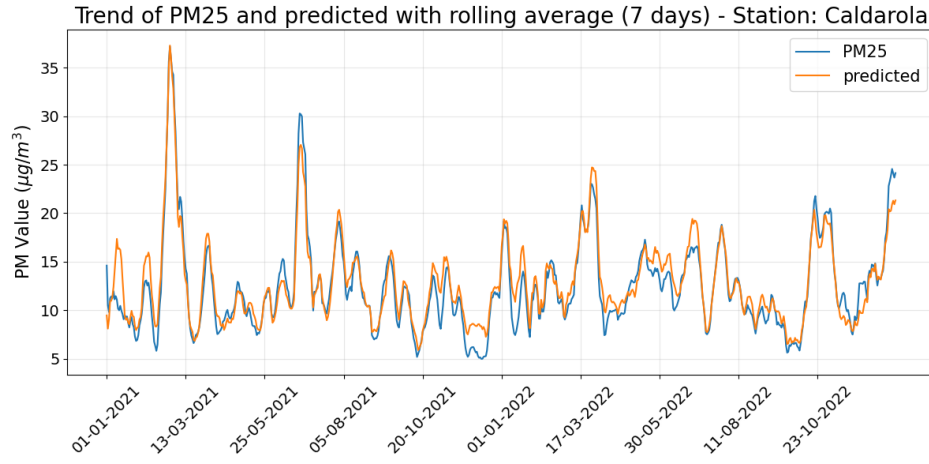
To further validate the effectiveness of the multi-phase approach in forecasting $PM_{2.5}$ concentrations, a comparison with a more direct methodological approach was

conducted. The multi-phase methodology, as previously described, involves a two-step process in which $\text{PM}_{2.5}$ concentrations are initially estimated for stations natively measuring only PM_{10} , creating a complete dataset that is subsequently used to train predictive models for PM_{10} . This approach takes advantage of the wide spatial coverage and increased data density resulting from the data completion procedure.

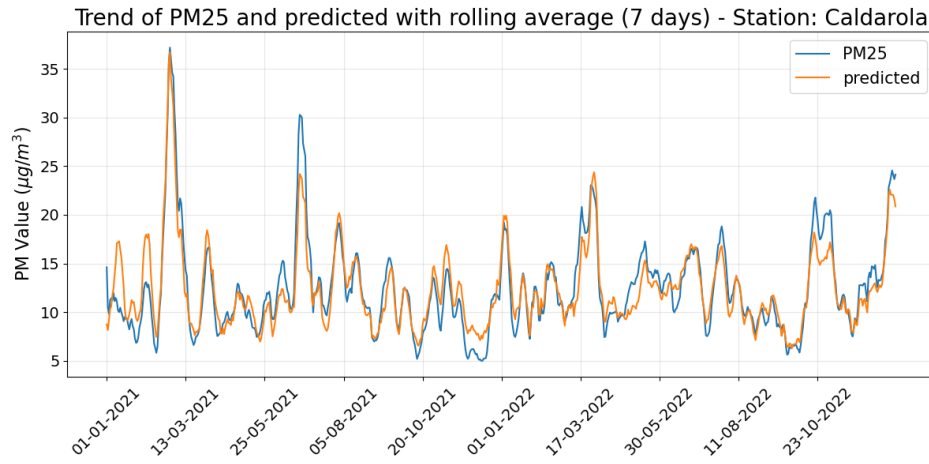
In contrast, the direct approach uses only data from the four monitoring stations natively measuring $\text{PM}_{2.5}$ concentrations (Cavour, Caldarola, Bitonto EN01 and Modugno EN02) without any data completion procedure. This basic methodology represents the conventional approach typically employed when working with limited observational data, providing a benchmark to assess the added value of the multi-phase framework.

The comparative analysis was structured using a leave-one-station-out cross-validation approach, where each of the four $\text{PM}_{2.5}$ monitoring stations was used individually as a test case for the years 2021 and 2022. For each station serving as a test case, the predictive models were trained using data from all other monitoring stations in the network, ensuring complete independence between the training and testing datasets. This methodology follows a time split of 80%-20%, with the first part of the data used for training and the second part reserved for validation.

The following figures present a visual comparison between the observed and forecasted $\text{PM}_{2.5}$ time series for each monitoring station during the period 2021-2022. To improve the clarity of the comparison and reduce the influence of short-term noise, a 7-day moving average was applied to both the observed and forecasted values. For each station, the results obtained with both the multistage and direct approaches are presented one below the other.

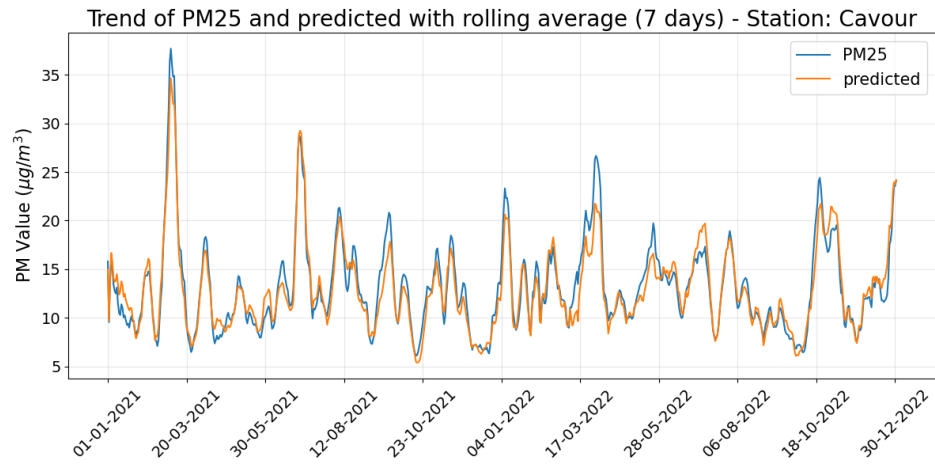


(a) Caldarola $\text{PM}_{2.5}$ time series with multi stage model.

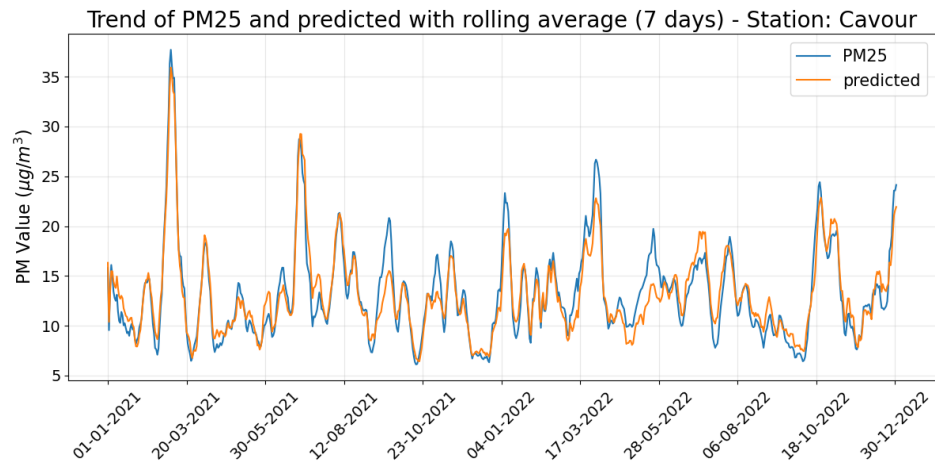


(b) Caldarola $\text{PM}_{2.5}$ time series with simple model.

Figure 4.25: Comparison of observed and predicted $\text{PM}_{2.5}$ time series for monitoring station of Caldarola using both methodological approaches.



(a) Cavour $\text{PM}_{2.5}$ time series with multi stage model.



(b) Cavour $\text{PM}_{2.5}$ time series with simple model.

Figure 4.26: Comparison of observed and predicted $\text{PM}_{2.5}$ time series for monitoring station of Cavour using both methodological approaches.

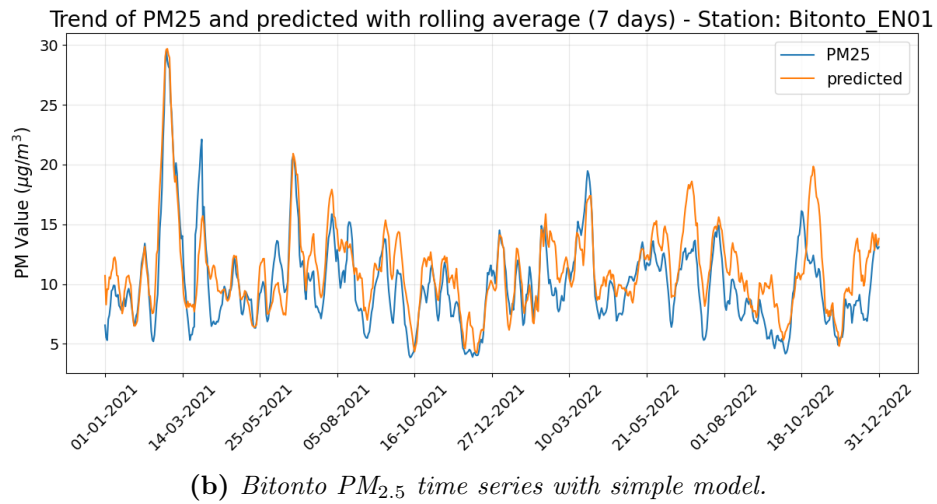
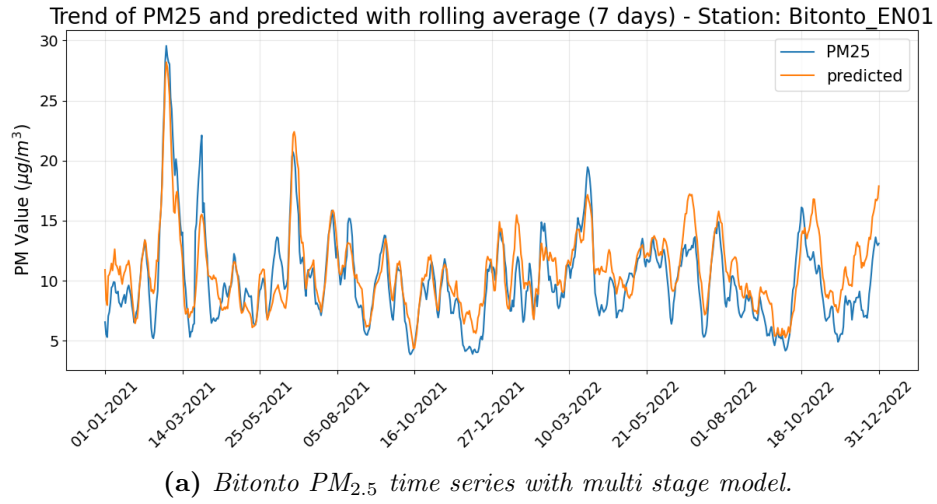


Figure 4.27: Comparison of observed and predicted $PM_{2.5}$ time series for monitoring station of Bitonto using both methodological approaches.

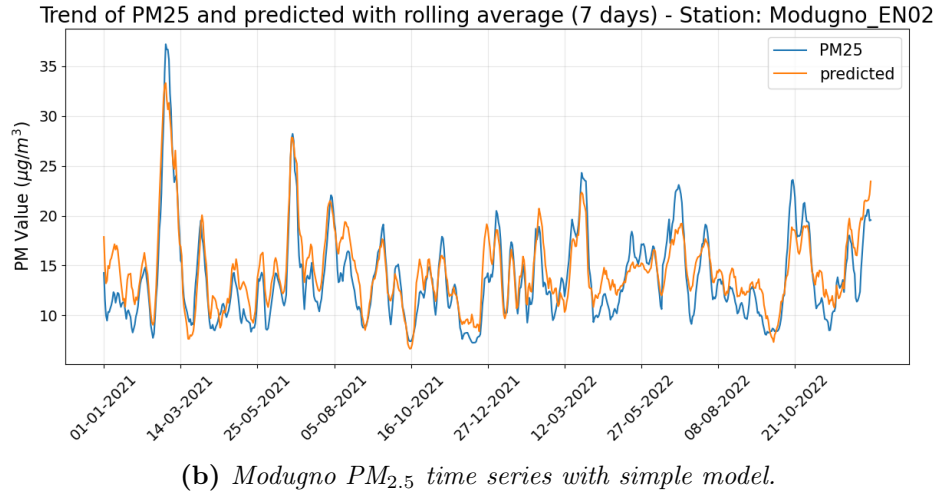
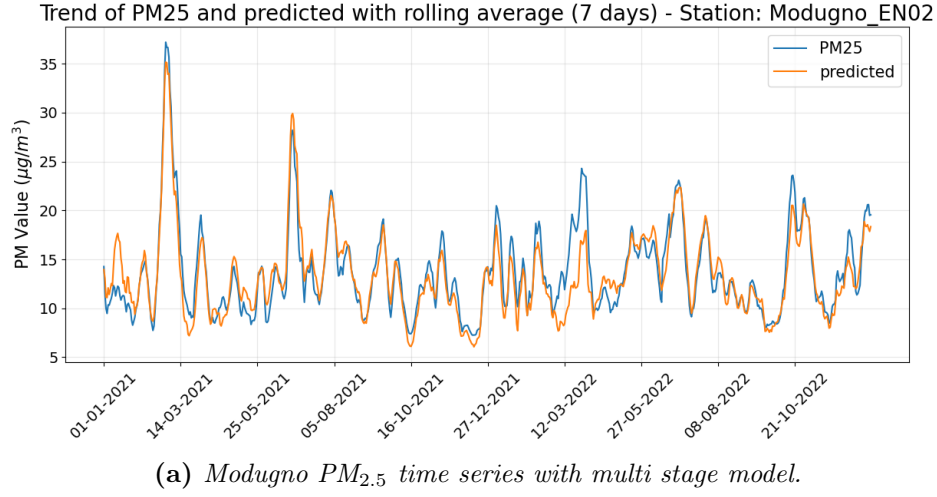


Figure 4.28: Comparison of observed and predicted $PM_{2.5}$ time series for monitoring station of Modugno using both methodological approaches.

The multi-stage model demonstrates slightly superior performance in capturing peak $\text{PM}_{2.5}$ concentrations, particularly evident at the Modugno_EN02 station (fig. 4.28a), where it accurately reproduces the February 2021 maximum (approximately $37 \mu\text{g}/\text{m}^3$), which the simple model (fig. 4.28b) underestimates. For Bitonto_EN01, for example, the multi-stage approach (fig. 4.27a) shows an improvement in avoiding the systematic overestimation observed in the simple model 4.27b during the summer-autumn period of 2022. The results from the Caldarola (fig. 4.25a) station further illustrate the good capability of the multi-stage model to capture rapid concentration variations, especially during spring 2022. Even at the Cavour station (fig. 4.26a), where differences are more subtle, the multi-stage approach provides slightly more accurate predictions during transition periods. Overall, these observations suggest that the additional complexity of the multi-phase methodology translates into tangible, albeit not drastic, improvements in predictive accuracy, particularly during pollution episodes and periods of significant concentration variability.

4.3 Particulate Matter 2.5 maps

The maps illustrate the annual average concentration of $\text{PM}_{2.5}$ ($\mu\text{g}/\text{m}^3$) for the Bari area from 2019 to 2022, calculated on the 300 m spatial resolution grid on which the feature data cube was built and on which the predictive model saved in .json format was applied. A critical aspect of the methodology concerned the management of missing data for AOD, a fundamental satellite parameter for $\text{PM}_{2.5}$ estimation but characterized by numerous gaps. The missing values are estimated by averaging the Inverse Distance Weighting (IDW) and the Nearest Neighbours (NN) interpolation methods. IDW is a widely used spatial interpolation technique that estimates the value of a variable z at an unsampled location x_0 through a weighted average of values at sampled points x_i as shown in Equation 4.8

$$z(x_0) = \frac{\sum_{i=1}^n w_i z(x_i)}{\sum_{i=1}^n w_i} \quad (4.8)$$

The underlying assumption of the IDW method is that points closer to the location of the prediction have a greater influence than points further away. Therefore, the weight assigned to each sample point is inversely proportional to its distance from the prediction location. Furthermore, these distances can be raised to a power parameter $\beta > 0$ that controls the rate of decline of influence with distance, as shown in Equation 4.9

$$w_i = d^{-\beta}(x_0, x_i) \quad (4.9)$$

The nearest neighbor interpolation method is another technique in which the value of a variable z at an unsampled location x_0 is determined using the average of the k nearest points to that location, as shown in equation 4.10

$$z(x_0) = \frac{\sum_{i=1}^k z(x_i)}{\sum_{i=1}^k 1} \quad (4.10)$$

In this case, k is chosen to be equal to the rounded square root of the number of sampled cells.

The trained .json model was applied to the entire time series of daily data cubes, generating daily $PM_{2.5}$ concentration maps with full spatial coverage of the study area. The annual average $PM_{2.5}$ concentrations were then computed by temporally aggregating these daily predictions on a pixel-by-pixel basis, producing the four annual mean concentration maps spanning the 2019-2022 period. The granularity of the maps ("pixel effect") derives from the 4 km resolution of the WRF model used among the inputs, which influences the ability to distinguish finer variations in the spatial distribution of the pollutant. All maps were visualized using the same color scale ($7-16 \mu g/m^3$) to enable direct visual comparison of concentration levels and temporal trends across the four-year period.

The maps (figure 4.30, 4.31) of annual mean $PM_{2.5}$ concentrations in the Bari metropolitan area reveal consistent spatial trends throughout the examined period (2019-2022), characterized by two particular hotspots, represented by the darker red: the city center of Bari (closer to the coast) and the industrial area of Modugno, the first characterized by a high density of vehicular traffic and the second by a high density of industrial activities, both reflecting persistent emission sources and local topographic features that influence pollutant dispersion. The similar spatial distribution patterns observed in all four years demonstrate the stability of emission sources and dispersion mechanisms in the region.

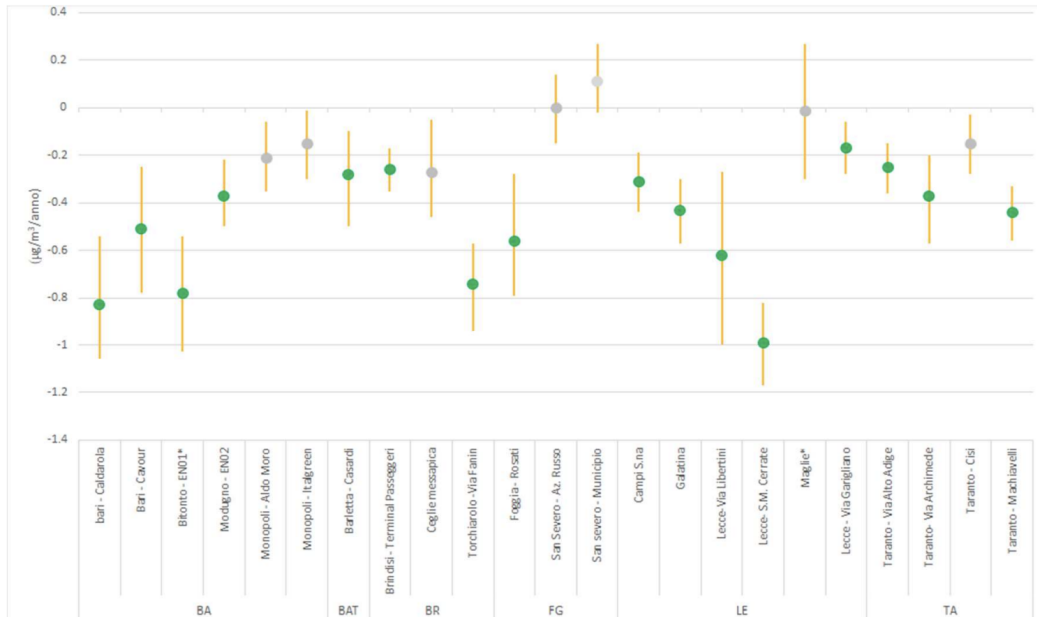


Figure 4.29: *Estimated trend of daily $PM_{2.5}$ concentrations, 2010-2022.*

ARPA Puglia has conducted a comprehensive trend analysis of $PM_{2.5}$ concentrations for the period 2010-2022 using the Theil-Sen method, a robust non-parametric statistical approach suitable for detecting monotonic trends in environmental time series [3]. As illustrated in Figure 4.29 of the ARPA report, the trend analysis reveals a heterogeneous pattern across monitoring stations, with circle markers indicating the trend magnitude, color coding representing statistical significance (green = statistically significant decrease; red = statistically significant increase; gray = non-significant trend), and yellow bars indicating the 95% confidence intervals. The analysis demonstrates that $PM_{2.5}$ exhibits a general statistical improvement trend for the majority of monitoring stations across the region, with a median regional reduction of $-0.16 \mu g/m^3/year$. Examining the specific stations relevant to this study, the Bari metropolitan area shows particularly pronounced decreasing trends: Bari-Caldarola exhibits a statistically significant decrease of $-0.8 \mu g/m^3/year$ (indicated by the green marker), while Modugno EN02 demonstrates an even more substantial reduction of $-0.9 \mu g/m^3/year$.

	2019	2020	2021	2022
Annual mean ARPA Stations	13.33 ± 6.71	13.84 ± 7.33	12.07 ± 6.73	12.63 ± 5.61
Annual Spatial Mean	11.6 ± 1.29	11.8 ± 1.19	10.48 ± 1.16	10.75 ± 1.34

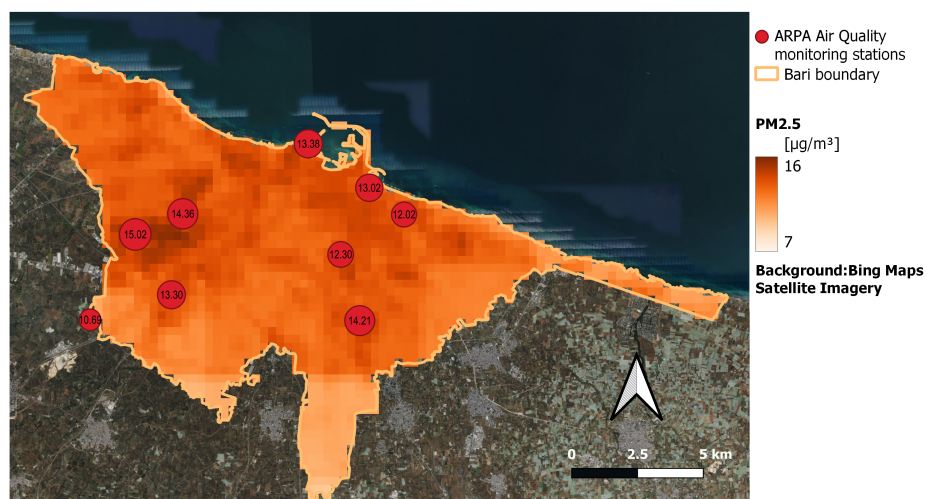
Table 4.16: *Annual $PM_{2.5}$ concentrations [$\mu g/m^3$] : ARPA stations vs. spatial map averages (2019-2022)*

Table 4.16 reports the annual mean $PM_{2.5}$ concentrations calculated using two different approaches for the years 2019-2022. The first row shows the annual means derived from the ARPA monitoring stations, obtained by averaging the daily measurements of all available stations for each year. The second row presents the spatial means calculated from the generated maps, calculated by averaging all pixel values within the study area for each corresponding year. Although these two metrics are not directly comparable, they provide information on the annual variations of $PM_{2.5}$ concentrations. Consistent trends between the two approaches can help validate temporal trends and assess the reliability of the mapping methodology in capturing interannual variability in air pollution levels.

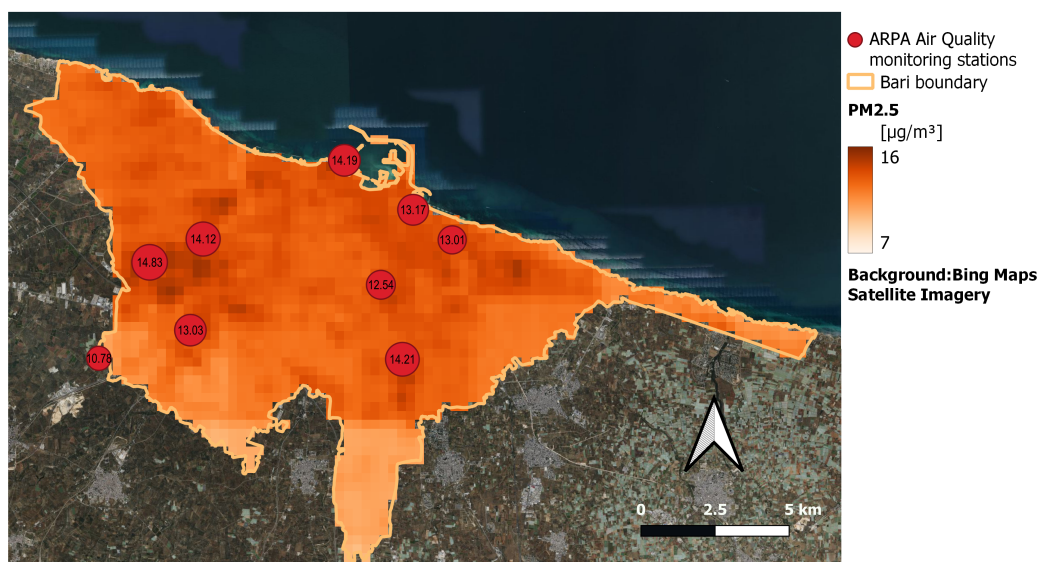
When examining the shorter temporal window of this study (2019-2022), the annual mean concentrations from ARPA monitoring data reveal distinct inter-annual patterns. The Bari province data shows relatively stable conditions between 2019 ($13.33 \mu g/m^3$) and 2020 ($13.84 \mu g/m^3$), representing only a marginal increase of $0.51 \mu g/m^3$. However, the most evident decrease in both the ARPA station data and the satellite-derived

mapping sequence occurs from 2020 to 2021, where Bari stations recorded a substantial reduction from 13.84 to 12.07 $\mu\text{g}/\text{m}^3$ (a decrease of 1.77 $\mu\text{g}/\text{m}^3$). This temporal pattern is corroborated by the satellite-derived annual maps, where pixel-averaged values across the study domain decreased from 11.8 to 10.48 $\mu\text{g}/\text{m}^3$ (a reduction of 1.32 $\mu\text{g}/\text{m}^3$), providing independent confirmation of the ground-based observations. The meteorological explanation of this 2020-2021 transition is documented in the ARPA report "Valutazione integrata della Qualità dell'Aria in Puglia Anno 2021" [2], which provides an analysis of the atmospheric conditions. The report states: "On an annual basis, a slight but significant increase in the average wind speed and a decrease in the percentage of calm winds were observed for the Foggia and Bari-Idrografico stations. Ultimately, the greater windiness observed in 2021 compared to 2020 at the Bari-Idrografico and Foggia stations, combined with the reduction of calm winds (particularly significant in Bari) and the increase in the number of rain events (especially in winter), could have determined meteorological conditions in these areas more favorable to the dilution and removal of pollutants", where Bari-Idrografico is the meteorological station representative of the Bari area.

The importance of this meteorological analysis goes beyond the simple explanation of the trend, as it provides a validation of the variable selection process of the machine learning model developed in 4.2.2. The main meteorological parameters identified by ARPA as drivers of the concentration reduction in the period 2020-2021 – wind speed, calm frequency and precipitation pattern – are found within the predictor variables selected through the feature importance analysis within the machine learning framework. This convergence between the statistical analysis of ground measurements and the satellite modeling approach confirms that the model successfully captures the dominant physical processes governing the spatial distribution and temporal variability of PM_{2.5}, thus increasing the reliability of satellite-derived concentration estimates and their applicability for air quality assessment and policy support.

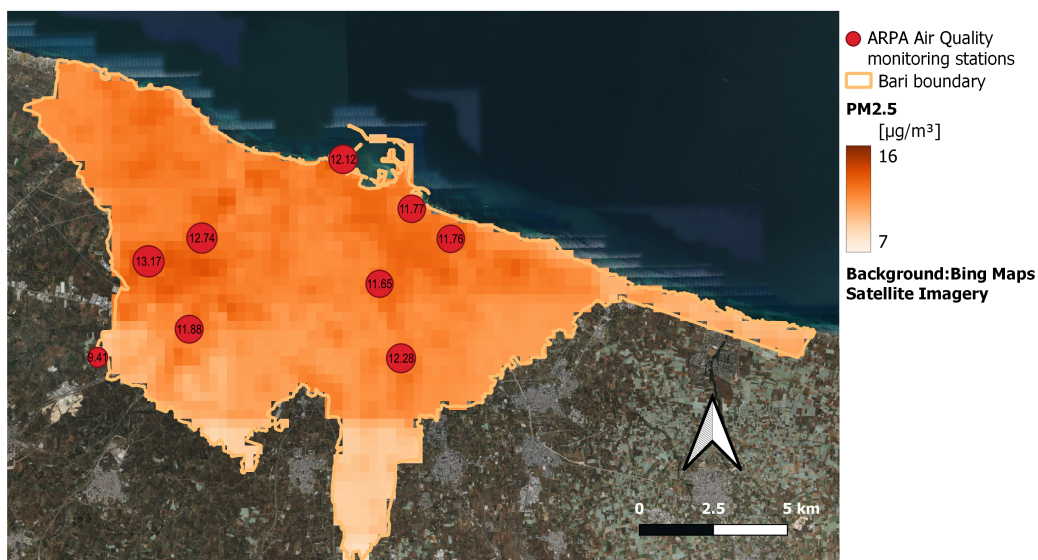


(a) Annual average $PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$) in 2019 over Bari metropolitan area.

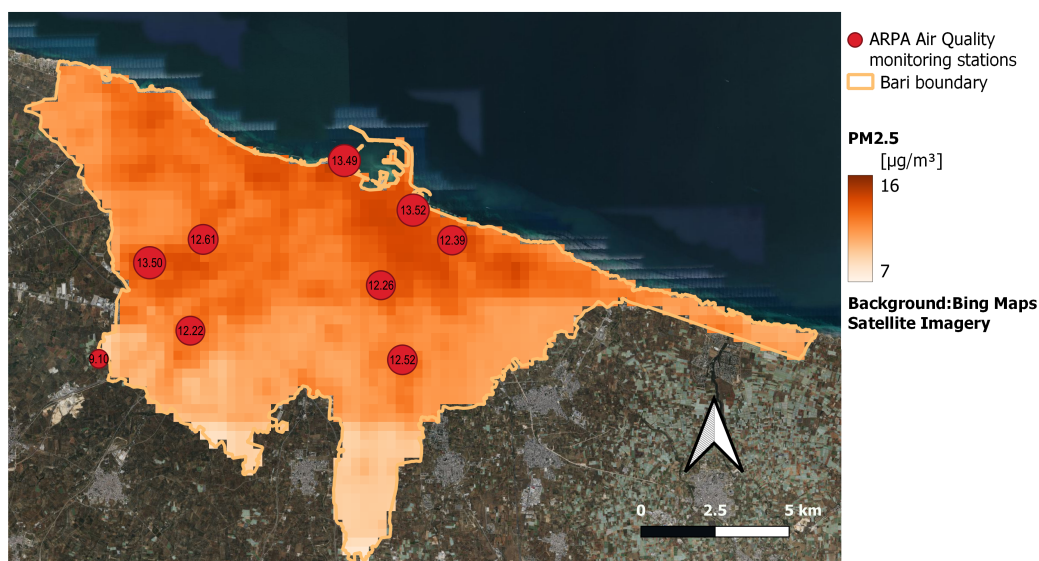


(b) Annual average $PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$) in 2020 over Bari metropolitan area.

Figure 4.30



(a) Annual average $PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$) in 2021 over Bari metropolitan area.



(b) Annual average $PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$) in 2022 over Bari metropolitan area.

Figure 4.31

Chapter 5

PM species estimations

This section presents the preliminary results of PM_{2.5} speciation estimates, work developed during the research period abroad in collaboration with the scientific group of the NASA MAIA mission. The results presented here represent the methodological foundations for future work on the chemical characterization of atmospheric particulate matter through the use of machine learning, opening new perspectives for understanding the distribution and impact of different pollutant components on human health and the environment.

5.1 Study Area

The study area focuses on the Primary Target Area (PTA-Rome, fig 5.1) of the MAIA mission in Italy, which comprises a 352 x 420 km area that includes the regions Lazio, Toscana, Umbria, Marche, and Emilia Romagna. In particular, two regions have distinct pollution characteristics:

- **Lazio Region:** This southern portion of the PTA includes Rome, one of the most densely populated urban centers in Italy, with approximately 2.8 million inhabitants. Rome represents an important case study due to its complex urban pollution patterns, mainly due to traffic emissions, residential heating and occasional transport of dust from the Sahara.
- **Emilia Romagna Region:** The northern portion of the PTA includes the southern part of the Po Valley, one of the most notoriously polluted regions in Europe due to its geographical conformation (surrounded by mountains on three sides), its population density, intensive agricultural activities and industrial development.

This PTA has been selected as one of the priority areas of the MAIA mission for surface monitoring data collection and epidemiological studies, with a satellite observation frequency of 3-4 times per week. The contrasting characteristics of pollution between the urbanized area of Rome and the industrial agricultural region of the Po Valley

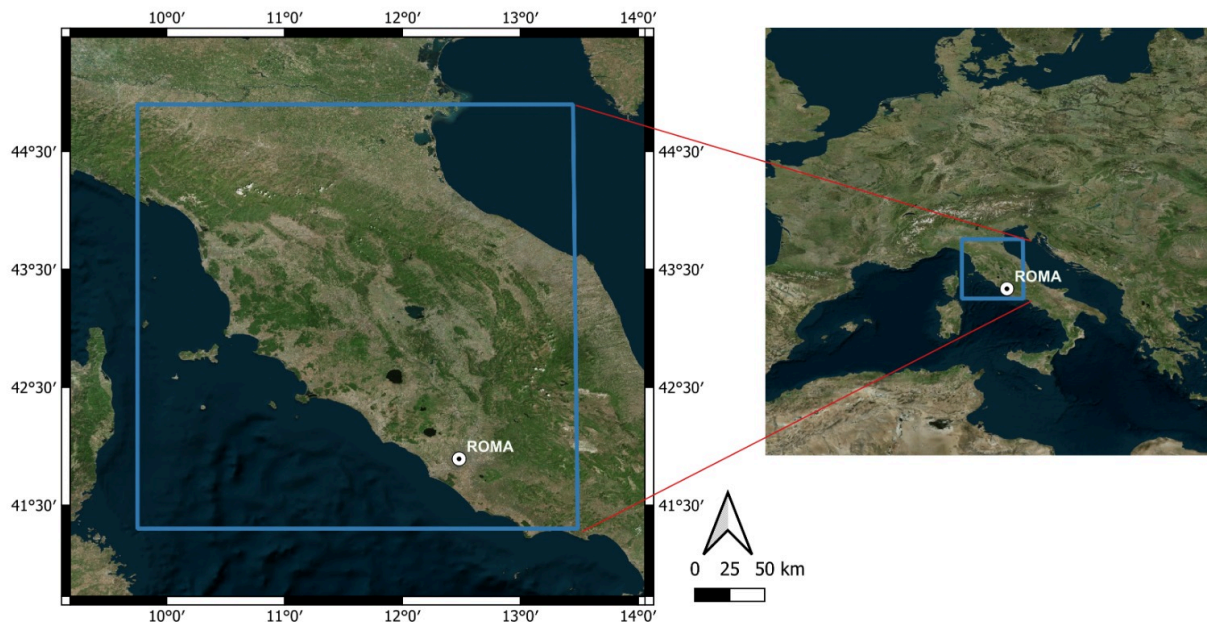


Figure 5.1: Box area for PTA-Rome of MAIA mission.

constitute an excellent test bed for PM speciation models predicting EC , OC , NO_3 , SO_4 and dust components.

5.2 Dataset

A comprehensive dataset has been assembled in collaboration with the regional environmental protection agencies ARPA Lazio and ARPA Emilia-Romagna to support the development of machine learning models for PM component forecasting. This integrated dataset combines ground-based monitoring measurements of PM chemical components (EC , OC , NO_3 , SO_4 and dust), meteorological parameters from monitoring stations and satellite-derived aerosol optical depth values. This dataset, covering the period between 01 September 2022 and 30 August 2023, provides a solid basis for the development and validation of PM speciation models within the Italian primary target area of the MAIA mission. The specific characteristics of each data source and their processing methodologies will be detailed in the following sections.

Variable	Description	Unit
U	u-wind component at 10m	m/s
V	v-wind component at 10m	m/s
MODV	surface wind	m/s
PBLH	Planetary Boundary Layer Height	m
RH	Relative Humidity	%
AOD	Mean Aerosol Optical Depth	-
EC	Elementar Carbon	$\mu g/m^3$
OC	Organic Carbon	$\mu g/m^3$
SO_4	Sulfate	$\mu g/m^3$
NO_3	Nitrate	$\mu g/m^3$
dust	saharian dust	$\mu g/m^3$

Table 5.1: *Data collection for $PM_{2.5}$ species estimations.*
Credits: Dave Diner and Sina Hasheminassab (JPL-NASA)

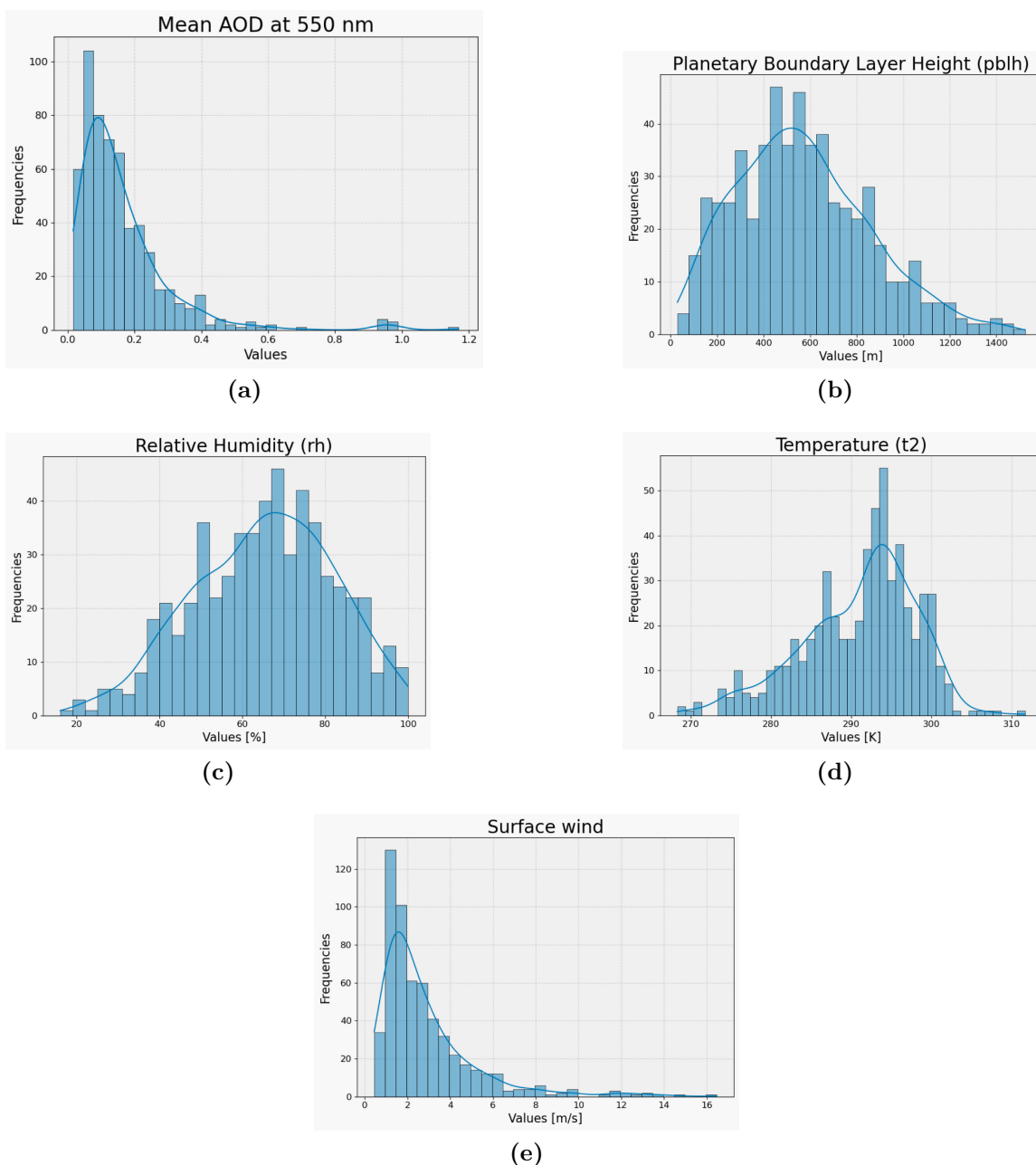


Figure 5.2: *Frequencies distributions of meteorological data.*

The meteorological data incorporated in this study include key atmospheric parameters known to influence the formation, dispersion, and deposition of particulate matter components. These variables, including temperature, relative humidity, precipitation, wind speed, wind direction, and boundary layer height, were obtained from the monitoring stations' weather sensors and supplemented with data from nearby meteorological stations when necessary. The inclusion of these parameters is crucial for capturing the atmospheric conditions that govern PM speciation and transport

processes.

Aerosol optical depth values used in this study were obtained directly from ground-based AERONET stations (Figure 5.3). Four AERONET stations were used: Roma_La_Sapienza_1, AAOT_1 (Adriatic Sea), ISPRA_1, and NAPOLI_CeSMA_1. For each ARPA monitoring station, AOD values were associated with the closest AERONET station to ensure optimal spatial correlation and data quality.

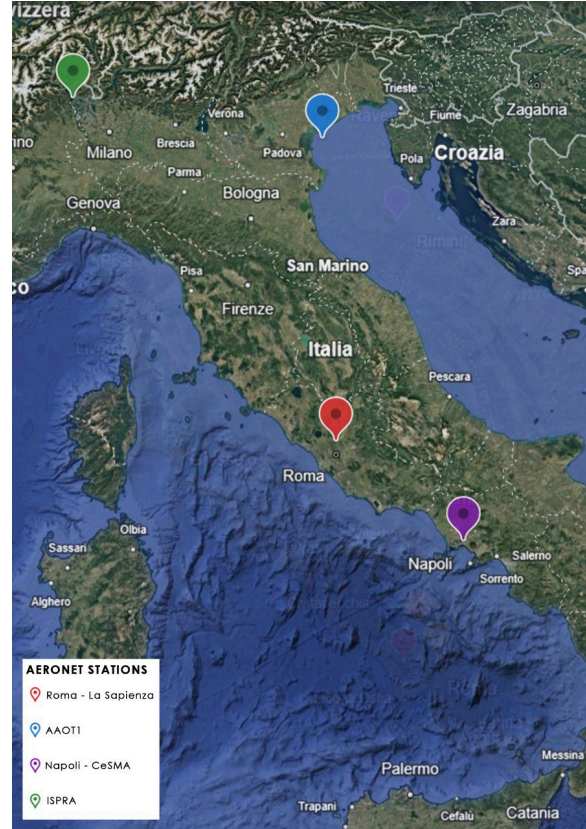


Figure 5.3: *AERONET station for PTA-Rome.*

5.2.1 PM Species data

The PM speciation data used in this study were collected from seven monitoring stations strategically distributed in the PTA-Rome (5.4), which includes both Lazio and Emilia-Romagna. In Emilia-Romagna, measurements were obtained from four monitoring stations managed by Arpa, positioned to represent the different typologies of areas within the region. Three urban sites were included (tab. 5.2): Bologna (via Gobetti) representing the central metropolitan area, Parma (Cittadella) covering the western and innermost part of the Po Valley, and Rimini (Marecchia) representing the eastern coastal area. Additionally, Molinella (BO) was included as a representative rural site, providing an important contrast with the urban locations. In Lazio, three



Figure 5.4: ARPA station in PTA-Rome.

monitoring stations managed by ARPA Lazio were selected: Corso Francia, classified as an urban traffic site with significant influence on vehicular pollution; Villa Ada, an urban background site less directly impacted by local emission sources; and Castel di Guido, a suburban traffic site that represents the transition between urban and rural environments. All stations provided daily mean concentrations (in $\mu g/m^3$) of the five target components of PM.

STATION	ARPA	TYPE AREA	TYPE STATION
Corso Francia	ARPA-Lazio	Urban	Traffic
Villa Ada	ARPA-Lazio	Urban	Background
Castel di Guido	ARPA-Lazio	Suburban	Traffic
Bo-Gobbetti	ARPA-Emilia	Urban	Traffic
PR-Cittadella	ARPA-Emilia	Urban	Background
BO-Molinella	ARPA-Emilia	Rural	Background
RN-Marecchia	ARPA-Emilia	Urban	Background

Table 5.2: *Monitoring stations ARPA in PTA-Rome.*

The distinctive characteristics of each monitoring station and the spatial-temporal variations in PM speciation across different urban and rural environments are illustrated in the following figures, which highlight the unique pollution profiles and seasonal patterns observed throughout the study period.

Figure 5.6 presents a comparative analysis of Elemental Carbon (EC) concentrations at all seven monitoring stations. The boxplot clearly illustrates significant spatial variations in EC levels, with Corso Francia in Rome showing significantly higher concentrations (median of approximately $2.0 \mu\text{g}/\text{m}^3$) than other sites. This high level is consistent with its classification as an urban traffic site directly affected by vehicular emissions, which represent a primary source of EC. Villa Ada, despite being an urban context site in Rome, shows the second highest EC levels (median of approximately $0.8 \mu\text{g}/\text{m}^3$), highlighting the pervasive nature of EC pollution in the urban environment of Rome. In contrast, Molinella, in a rural area, and Castel di Guido, in a suburban area, show the lowest EC concentrations, demonstrating their distance from the main traffic emission sources.

Figure 5.5 presents the temporal evolution of EC concentrations during the study period (September 2022 to August 2023), showing daily measurements along with 7- and 30-day moving averages. The time series reveals a pronounced seasonal pattern with significantly higher EC concentrations during the winter months (December 2022 to February 2023), when concentrations frequently exceeded $1.5 \mu\text{g}/\text{m}^3$ and peaked above $3.5 \mu\text{g}/\text{m}^3$ in early January. This seasonal variation can be attributed to a combination of increased residential heating emissions during winter, less favorable meteorological conditions for pollutant dispersion (lower mixing height and more frequent temperature inversions), and a potentially longer atmospheric residence time of EC particles under colder conditions. The summer months (June to August 2023) show consistently lower EC levels, typically below $0.5 \mu\text{g}/\text{m}^3$, demonstrating the strong

seasonal influence on this pollutant component.

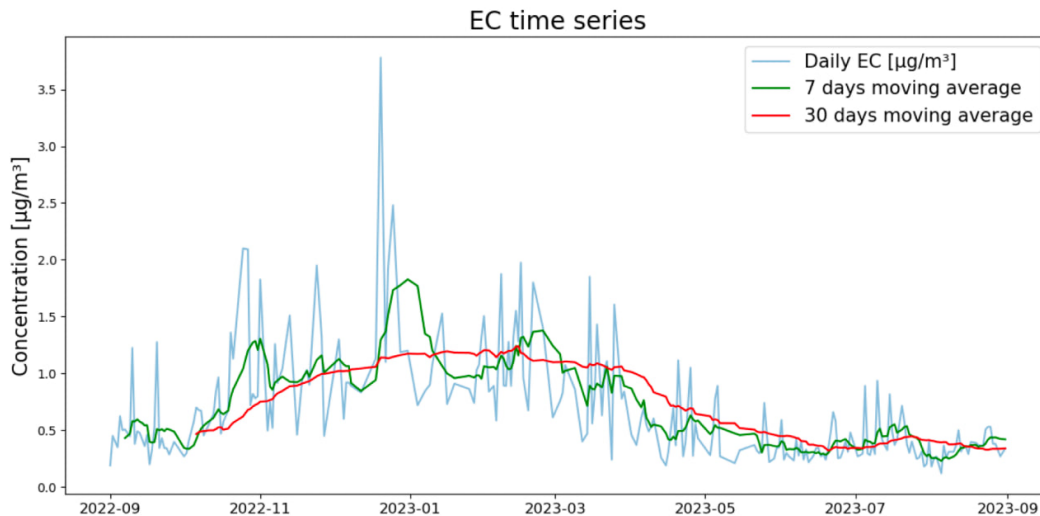


Figure 5.5: *EC time series.*

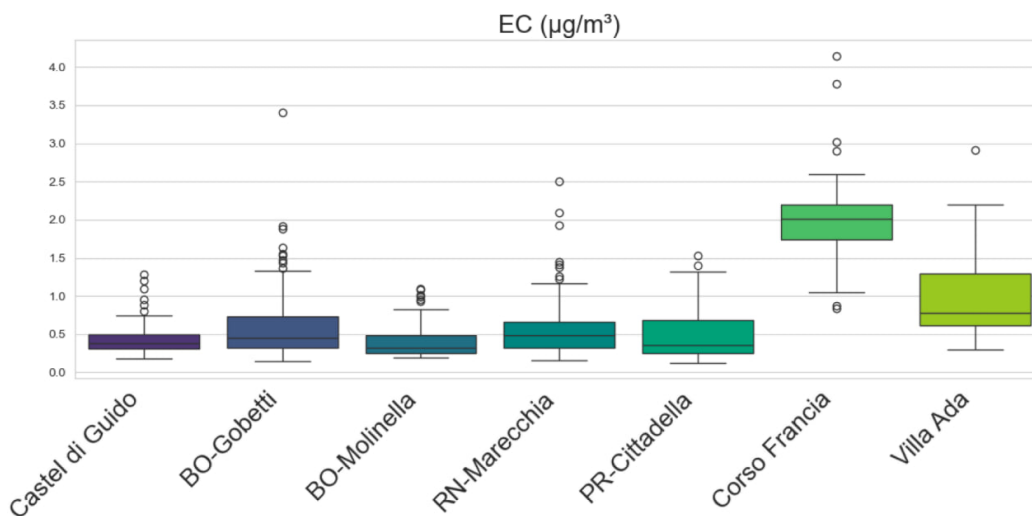


Figure 5.6: *EC box plot.*

Organic Carbon (OC) analysis reveals both spatial and temporal trends that provide further information on the composition of PM in the study area. Spatially, Corso Francia shows once again the highest concentrations (median of about $5.0 \mu\text{g}/\text{m}^3$), closely followed by PR-Cittadella. Unlike EC, where rural sites showed significantly lower values, OC concentrations showed less marked variations between urban and rural contexts, suggesting the influence of both anthropogenic and biogenic sources.

The temporal profile of OC follows a seasonal pattern similar to that of EC, but with significantly higher absolute concentrations, typically 4-5 times higher than EC levels.

Winter months show peak concentrations above $12 \mu\text{g}/\text{m}^3$ in December 2022, while summer levels stabilize around $3 \mu\text{g}/\text{m}^3$. The 30-day moving average reveals a more gradual decline from winter to summer than the EC, potentially indicating different mechanisms of OC formation and removal. This could be attributed to secondary organic aerosol formation processes that continue to contribute to OC levels even when primary emissions decline in the warmer months.

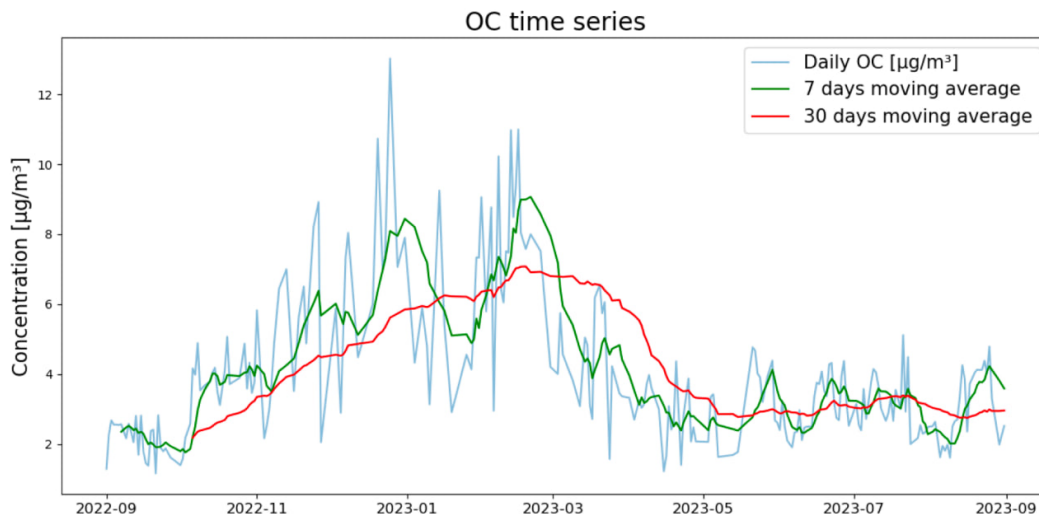


Figure 5.7: *OC time series.*

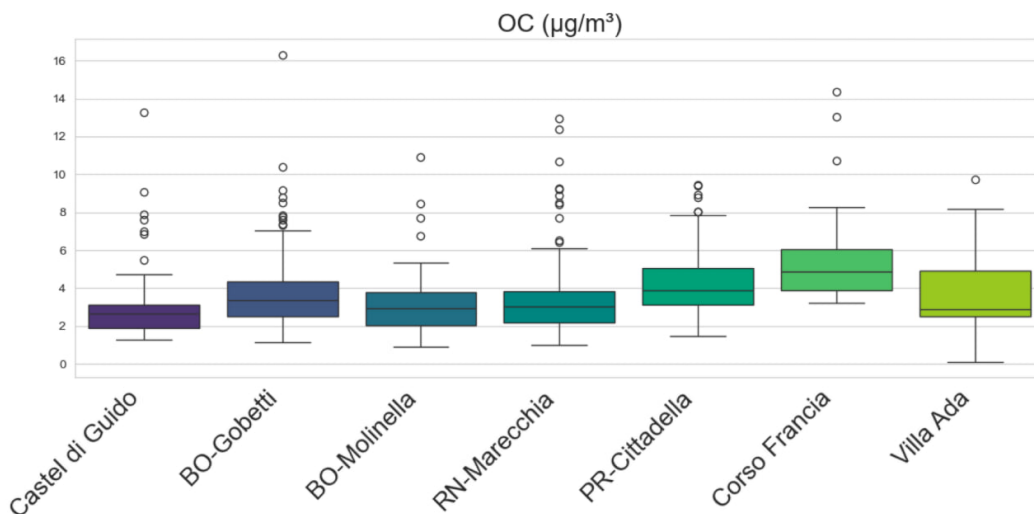


Figure 5.8: *OC box plot.*

Sulfate (SO_4) concentrations show a markedly different pattern than carbon components. The spatial distribution among monitoring stations shows relatively uniform levels, with Corso Francia presenting concentrations only marginally higher than other

sites. This spatial homogeneity suggests that sulfate represents more of a regional background pollution rather than being strongly influenced by local emission sources.

The temporal profile of SO_4 reveals a unique seasonal pattern, characterized by multiple peaks throughout the year rather than a single winter maximum. These episodic peaks likely correspond to specific meteorological conditions that favor secondary sulfate formation through photochemical processes, such as periods of high solar radiation, temperature, and humidity that favor the oxidation of sulfur dioxide to sulfate.

The relatively constant background concentrations at rural and urban sites, combined with the episodic nature of concentration increases, support the interpretation that sulfate represents a regional-scale pollutant with significant contributions from secondary atmospheric formation rather than direct emissions. This characteristic makes sulfates an important indicator of background air quality conditions affecting the entire study area, regardless of local emission intensities.

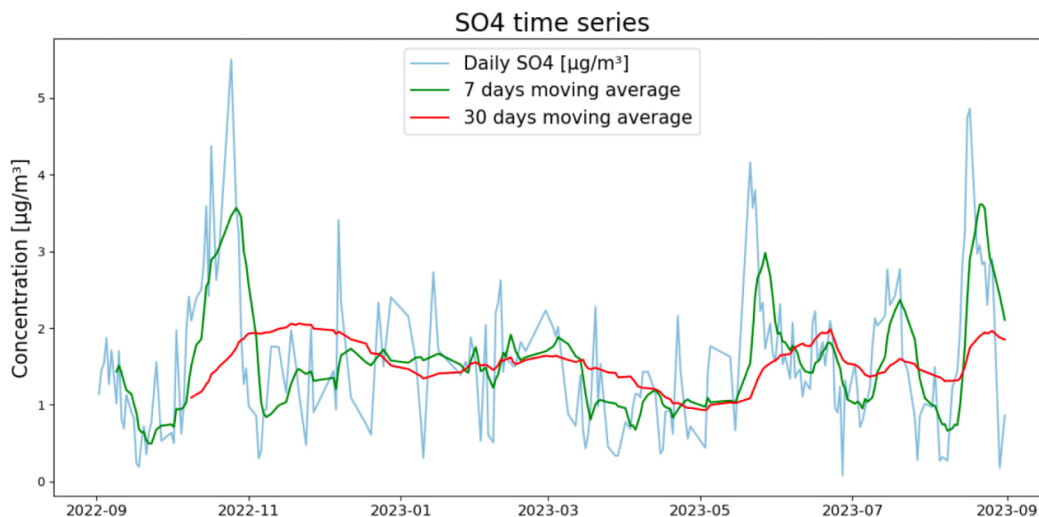


Figure 5.9: SO_4 time series.

The time series of NO_3 shows a marked seasonal pattern, with nitrates virtually disappearing during the summer months (May to September) when concentrations approach zero, in sharp contrast to winter values that often exceed $10 \mu g/m^3$ and reach peaks above $20 \mu g/m^3$ in November and January. This extreme seasonality reflects the temperature-dependent gas-particle partitioning behavior of nitrate: ammonium nitrate readily dissociates into gaseous ammonia and nitric acid at higher temperatures.

The spatial concentration gradient, higher at the Po Valley sites and lower in Rome, indicates the influence of specific regional factors, such as ammonia emissions from agricultural activities (abundant in the Po Valley), combined with the meteorological conditions of the basin that favor nitrate formation during winter. The negligible summer concentrations at all sites underline the fundamental role of temperature in

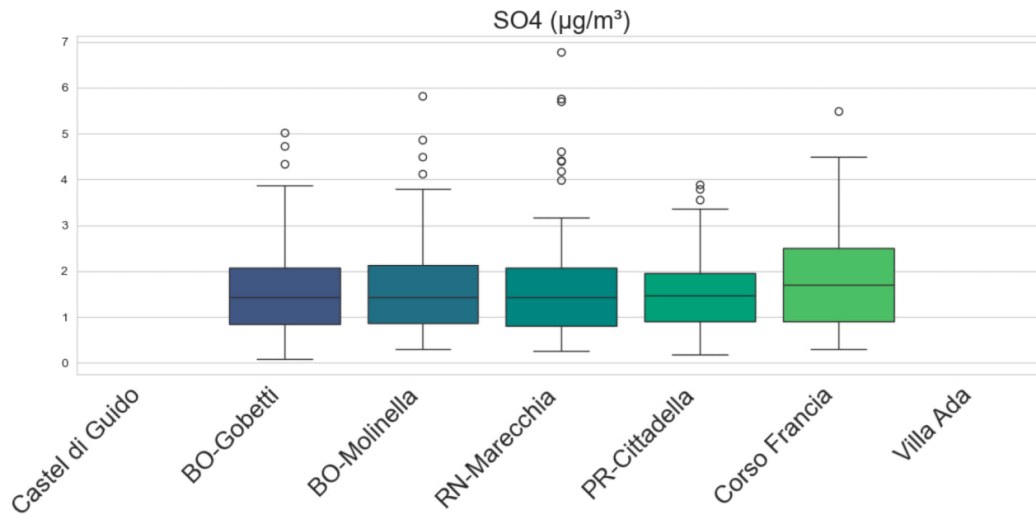


Figure 5.10: SO_4 box plot.

controlling the nitrate contribution to PM, making it a highly seasonal component unlike the more persistent EC and OC.

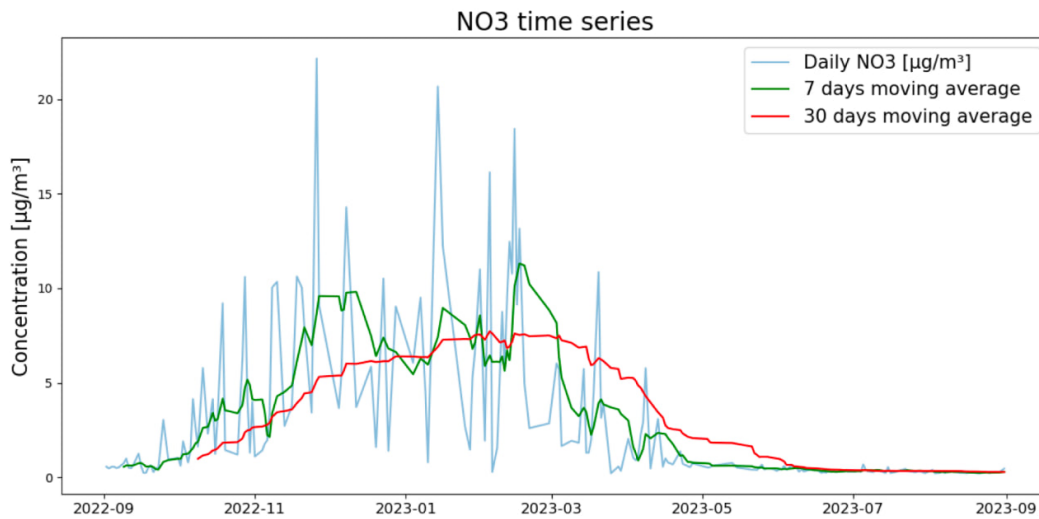


Figure 5.11: NO_3 time series.

Mineral dust exhibits a distinctly different pattern from all other PM components, characterized by sharp and isolated concentration peaks rather than constant seasonal patterns. This episodic pattern reflects the primary origin of dust from sporadic events such as Saharan dust transport, construction activities and resuspension of road dust during drought periods.

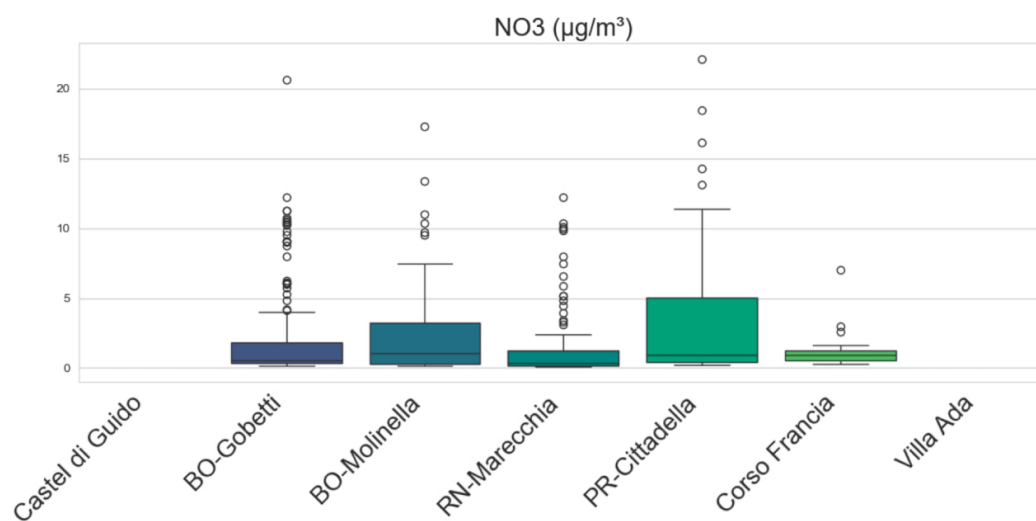


Figure 5.12: NO_3 box plot.

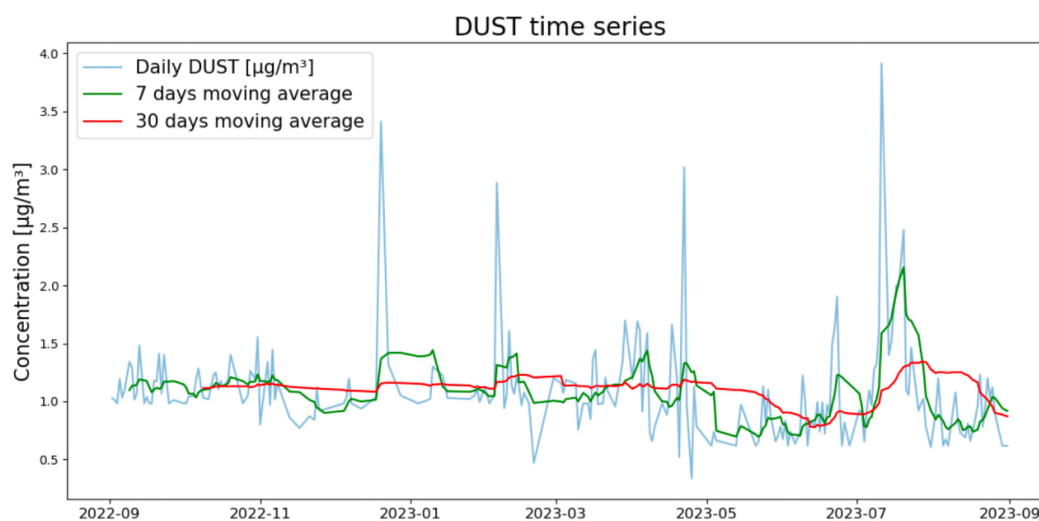
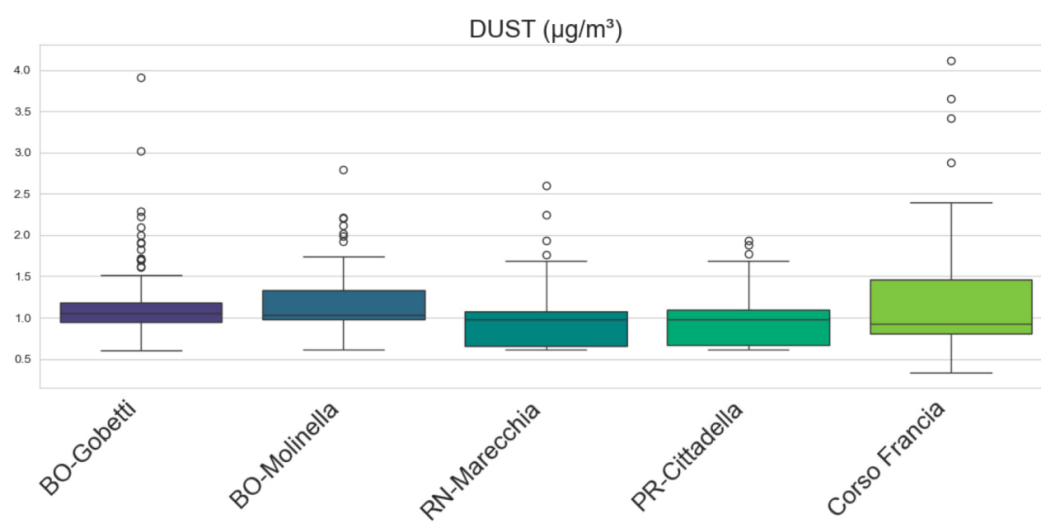


Figure 5.13: Dust time series.

**Figure 5.14:** *Dust box plot.*

5.3 Random Forest and GRM model comparison

In the context of the MAIA mission, the model used for estimating atmospheric particulate matter concentrations is the **GRM (Geostatistical Regression Model)**. This model represents a geostatistical regression approach that establishes empirical relationships between aerosol parameters detected by satellite and ground-level atmospheric PM concentrations. The GRM is based on a main equation that includes several types of predictors, including AOD derived from satellite observations, geospatial variables such as elevation and land use, and spatiotemporal predictors such as meteorological parameters and concentrations modeled by the CTM. The mathematical structure of the model allows capturing spatial and temporal variabilities in the relationships between AOD and PM through coefficients with random components. The mathematical structure of the model allows capturing spatial and temporal variabilities in the relationships between AOD and PM through coefficients with random components. The general mathematical equation that describes the GRM is represented by equation:

$$\begin{aligned}
 PM_{component,s,t} = & \left(\alpha_0 + \alpha_{w(t)} + \nu_1 \alpha_{s,m(t)} \right) \\
 & + \left(\beta_0 + \beta_{w(t)} + \nu_2 \beta_{s,m(t)} + \nu_3 \beta_1 \frac{[PM_{component,s,t}]_{CTM}}{[AOD_{s,t}]_{CTM}} \right) AOD_{s,t} \\
 & + \sum_{i=1}^{N_L} \nu_{4,i} \gamma_i L_{i,s} \\
 & + \sum_{j=1}^{N_M} \nu_{5,j} \phi_j M_{j,s,t} \\
 & + \varepsilon_{s,t}
 \end{aligned}$$

where the term $(\alpha_0 + \alpha_{w(t)} + \nu_1 \alpha_{s,m(t)})$ represents a spatiotemporal additive bias not explained by AOD or the other geospatial and spatiotemporal inputs, the term $(\beta_0 + \beta_{w(t)} + \nu_2 \beta_{s,m(t)} + \nu_3 \beta_1 \frac{[PM_{component,s,t}]_{CTM}}{[AOD_{s,t}]_{CTM}})$ governs spatiotemporal associations between total AOD and $PM_{component,s,t}$ that may be impacted by various factors such as PM composition and atmospheric conditions, $\sum_{i=1}^{N_L} \nu_{4,i} \gamma_i L_{i,s}$ allows a selected set of geospatial predictors to be incorporated in the GRM, the term $\sum_{j=1}^{N_M} \nu_{5,j} \phi_j M_{j,s,t}$ allows a selected set of spatiotemporal predictors and the final term in $\varepsilon_{s,t}$ denotes the error term and represents the residual difference between the measured and predicted PM concentrations.

Table [5.3](#) shows the results of 10-fold cross-validation for the GRM model using the basic predictive variables: relative humidity (rh), planetary boundary layer height (pblh), mean AOD at 550nm (mean_AOD_550), surface wind (wind_surface), and temperature at 2 meters (t2).

Component	R^2	RMSE	Bias	Slope
EC	0.67	0.32	0.02	0.97
OC	0.61	1.36	0.03	0.99
SO ₄	0.55	0.71	0.02	0.99
NO ₃	0.59	2.26	0.03	0.98
Dust	0.01	2.49	0.09	0.93

Table 5.3: *Performance metrics for PM_{2.5} components by GRM. Credits: Liu Yang and Jin Zhihao (Emory University)*

To evaluate alternative modeling approaches, the Random Forest algorithm was tested in two phases. First, Random Forest was implemented using the same predictors (5.1) employed in the GRM implementation to establish a direct baseline comparison between methodologies. This allowed for an assessment of whether Random Forest’s ensemble learning approach could capture the same patterns as GRM’s geostatistical approach. In the second phase, the predictor set was enhanced by incorporating additional spatial and temporal variables, including latitude, longitude, and Julian day. These coordinates were added to explicitly capture geographic and seasonal patterns potentially not fully represented in the original predictor set. The inclusion of these spatiotemporal variables attempts to replicate the role of GRM coefficients influenced by spatiotemporal random effects. This approach leverages Random Forest’s intrinsic capability to model complex non-linear relationships without requiring explicit specification of the spatial and temporal correlation structures integrated into the GRM formulation.

Table 5.4 presents a detailed comparison of the performance of the three implemented models (RF-base, RF-Enhanced, and GRM) in predicting different PM species: Elemental Carbon (EC), Organic Carbon (OC), Sulfate (SO₄), Nitrate (NO₃), and Dust. For each model-target combination, four evaluation metrics are reported: R^2 , RMSE, SLOPE, and BIAS.

The systematic improvement in performance in the transition from RF-base to RF-Enhanced demonstrates the crucial importance of spatial and temporal components in the distribution of atmospheric chemical species. This result validates the hypothesis that spatiotemporal random effects incorporated in the GRM model can be approximated, at least partially, through the explicit inclusion of geographic and temporal coordinates in a machine learning framework. It is notable how the RF-Enhanced model achieves, and sometimes exceeds, the performance of the GRM despite the contrasting methodological approaches. While GRM explicitly incorporates a spatiotemporal correlation structure through regression coefficients influenced by random effects, Random Forest autonomously identifies these relationships through recursive partitioning of the predictor space. This data-driven learning offers a significant advantage in

TARGET	Model	R2	RMSE	SLOPE	BIAS
EC	RF-base	0.14	0.47	0.18	0.00
	RF-Enhanced	0.68	0.27	0.69	-0.00
	GRM	0.67	0.32	0.97	0.02
OC	RF-base	0.16	1.89	0.29	0.01
	RF-Enhanced	0.53	1.33	0.58	-0.01
	GRM	0.61	1.36	0.99	0.03
SO ₄	RF-base	0.22	0.90	0.30	0.01
	RF-Enhanced	0.46	0.72	0.47	-0.01
	GRM	0.55	0.71	0.99	0.02
NO ₃	RF-base	0.29	2.79	0.38	0.02
	RF-Enhanced	0.58	2.10	0.62	0.03
	GRM	0.59	2.26	0.98	0.03
Dust	RF-base	-1.96	1.82	0.12	0.02
	RF-Enhanced	-1.71	1.74	0.20	0.03
	GRM	0.01	2.49	0.93	0.09

Table 5.4: Comparison of model performance metrics

complex contexts with non-stationary or difficult-to-parameterize spatiotemporal relationships. The SLOPE values, however, reveal a limitation of Random Forest models: the tendency toward regression to the mean, with values consistently less than 1, which leads to underestimating high concentrations and overestimating low ones. The GRM, instead, maintains nearly optimal SLOPE values for most species, better preserving the scale of concentrations. This difference stems from the nature of the algorithms: GRM is based on a statistical framework that preserves the scale of dependent variables, while Random Forest uses averages of local models that compress the range of predicted values. A relevant aspect that emerged from the analysis concerns the variability of performance across different chemical species. The results suggest that some PM components appear to be more easily modeled than others, regardless of the methodological approach used. This difference might be connected to the various chemical-physical dynamics that characterize the formation, transformation, and transport of various atmospheric species. Primary components, such as EC, tend to show better performance compared to secondary components or those influenced by long-range transport, such as dust. The latter, in particular, presents difficulties for all tested models, probably due to its episodic nature, suggesting the opportunity to consider additional predictors related to specific transport events or particular meteorological conditions. It is important to note that negative values of the coefficient of determination R^2 indicate that the Random Forest model performs worse than a baseline model that constantly predicts the mean of the observed values. The R^2 coefficient is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5.1)$$

where:

y_i are the observed values $\hat{y}_i$ are the values predicted by the model $\bar{y}$ is the mean of the observed values n is the number of observations

When $R^2 < 0$, it means that the sum of squared residuals of the model ($\sum_{i=1}^n (y_i - \hat{y}_i)^2$) is greater than the total sum of squares ($\sum_{i=1}^n (y_i - \bar{y})^2$), indicating that the model provides less accurate predictions compared to the simple sample mean.

The complementarity observed between the geostatistical and machine learning approaches suggests possible developments for hybrid methodologies. The GRM distinguishes itself in preserving the scale of concentrations and incorporating the spatiotemporal correlation structure, while Random Forest shows greater flexibility in identifying complex and non-linear relationships between predictors and chemical species concentrations. An approach that combines the strengths of both methodologies could offer an interesting direction for future research.

Chapter 6

Discussion and Conclusions

This research addressed the challenge of fine particulate matter (PM_{2.5}) prediction in urban environments, focusing on the city of Bari as a study area. The research process was structured around two interconnected objectives: (1) improving the resolution and accuracy of satellite-derived Aerosol Optical Depth (AOD) data and (2) developing a robust PM_{2.5} prediction model at an intra-urban scale (300 m resolution). Recognizing the importance of AOD for PM prediction, as established in the literature, the approach, detailed in the section 4.1, has leveraged the HReBLOCK framework 3.1 to combine Sentinel-3 SYNERGY's high spatial resolution (300 m) with MODIS MAIAC's robust retrieval algorithm (1 km), obtaining an improved AOD product that was used as input for the subsequent PM prediction activity (4.2). As demonstrated in section 4.2.2, this improved AOD product offered modest but consistent improvements in predictive capabilities compared to using only Sentinel-3 data. These improvements were consistent across all three machine learning models tested and in all cross-validation configurations applied. The approach maintained Sentinel-3's high spatial resolution of 300 m while incorporating aspects of MAIAC's algorithmic robustness, creating a balanced product more suitable for fine-scale urban analysis than either satellite product used individually.

The PM prediction methodology, also based on the HReBLOCK framework, was implemented through four sequential phases, as described in Section (4.2): (1) PM₁₀ prediction at monitoring stations, (2) filling in missing PM₁₀ values, (3) estimation of PM_{2.5} for stations with incomplete data, and (4) development of the final PM_{2.5} prediction model for the entire study area (4.2.2).

Although increasing the spatial resolution to this level introduced some challenges, the model maintained performance parameters that remain comparable to those reported in previous studies at coarser resolutions, as demonstrated by the validation results in table 4.14. The spatial trend of PM_{2.5} concentration demonstrated consistency with the ARPA validated data reported in the annual reports during the four-year study period, highlighting the robustness and stability of the developed modelling approach. The maps clearly reveal two critical areas with higher concentrations: Bari's city center, characterized by intense vehicular traffic, and the industrial zone of Modugno. This

spatial distribution is particularly significant as it accurately reflects the main pollution sources in the urban area, confirming the model’s ability to realistically capture pollution patterns at an intra-urban scale.

Although the absolute values between spatial means and station-based measurements are not directly comparable due to the different methodologies, the concordance in temporal trends provides a good validation of the ability of our model to capture interannual variability and supports the reliability of our methodological approach for monitoring long-term air quality patterns.

SHAP analysis provided crucial information on the main factors determining PM_{2.5} occurrence in the study area. Planetary boundary layer height (PBLH) and AOD enhancement emerged as the most influential predictors, confirming the importance of atmospheric dynamics and aerosol loading. Urban morphology characteristics, including building height and percentage of artificial surface, ranked among the significant predictors, demonstrating the critical role of urban structure in high spatial resolution air quality models.

The evaluation of predicted spatial patterns represents a key challenge in urban air quality modeling, due to the limited spatial coverage of ground-based monitoring networks. Although our approach found consistency in temporal trends through comparison with ARPA station measurements, demonstrating consistent interannual variations, a comprehensive spatial validation remains limited by the sparse distribution of monitoring stations in the urban domain.

To overcome this limitation, future validation strategies could employ two complementary approaches. First, spatial validation could be conducted using validated ARPA maps with a 4 km resolution over the territory of Bari, spatially aggregating our high-resolution forecasts to match the reference resolution, allowing for a systematic pixel-by-pixel comparison at the regional scale. Second, validation by domain experts could provide a qualitative assessment of predicted spatial patterns with respect to local knowledge of pollution sources and urban dynamics.

Among the methodological limitations of this work is the difficulty in integrating multi-source data with different spatial resolutions. The 4 km resolution of the WRF meteorological variables, processed using nearest neighbor interpolation, creates a visible “staircase” effect in some areas, particularly evident in the peripheral areas of the maps. This choice was made to preserve the original accuracy of the meteorological data and avoid introducing interpolation errors. However, future improvements could address this limitation through two strategies: the adoption of higher resolution meteorological datasets or the implementation of advanced resampling techniques to harmonize all inputs to a 300 m resolution.

Despite these visual limitations, the model maintains a good fine-scale (300 m) discrimination capability in the most critical areas, effectively identifying concentration gradients between adjacent zones with different land use characteristics or traffic intensity. Given the proven importance of traffic sources at intra-urban scale, future model improvements should prioritize the integration of high-resolution traffic data to better capture local emission patterns and improve the accuracy of fine-scale forecasts.

6.0.1 Applicability and Future Perspectives

The PM_{2.5} predictive model developed in this research, although focused on the metropolitan area of Bari, opens important avenues for investigating model transferability and scalability across different urban contexts. Within the APEMAIA project framework, future research directions will involve developing individual predictive models for other Italian urban areas, including Taranto and Rome, each tailored to their specific urban morphologies, emission sources, and meteorological conditions.

A key research question concerns the transferability of these models across different cities: whether a model trained on Bari data can effectively predict PM_{2.5} concentrations in Taranto or Rome, and vice versa. This cross-city validation will provide crucial insights into the generalizability of feature importance patterns and model parameters across diverse urban environments.

Additionally, a comparative analysis will be conducted to evaluate the performance differences between city-specific models and a unified multi-city model trained on combined datasets from all study areas. This approach will help determine whether local urban characteristics require dedicated modeling strategies or whether a more generalized model can capture the common underlying processes governing PM_{2.5} dynamics across different Italian urban contexts. Such analysis will inform the development of scalable air quality prediction frameworks applicable to broader geographical domains while maintaining the fine-scale resolution necessary for urban planning applications.

Task 3 (Se.5) of this research, initiated in collaboration with JPL (Jet Propulsion Laboratory), focused on estimating PM_{2.5} species, has laid the groundwork, albeit with a still simplified model, to go beyond PM_{2.5} mass concentration and achieve a more detailed particulate speciation. The same high-resolution methodology used for modeling PM₁₀ and PM_{2.5} can be applied to model specific PM components (such as sulfates, nitrates, carbonaceous particulate, and organic carbon) once more systematic measurements of these components' concentration at ground level become available.

For One Health objectives, studying the effects of pollution on human health is crucial. While the harmful effects of PM on health have already been demonstrated, what has not yet been thoroughly explored is the relative danger associated with each PM species. This is where the importance of speciation becomes evident - understanding which specific components contribute most significantly to health impacts would allow for more targeted and effective interventions for public health protection.

The upcoming MAIA (Multi-Angle Imager for Aerosols) mission represents an important opportunity to enhance the methodology presented in this thesis. MAIA's anticipated products include aerosol measurements that could integrate perfectly with our improved AOD framework. Furthermore, MAIA will provide specific information

on particulate composition, which aligns perfectly with the particulate speciation principles established in Task 3 of this research. This synergy could, in addition to demonstrating the potential of this new satellite, significantly advance our understanding not only of particulate concentration but also of its chemical composition at fine spatial scales, potentially overcoming current limitations in the availability of ground-based speciation data.

A natural evolution of the current model would be its extension to a forecasting framework. By incorporating predictive meteorological variables and temporal dynamics, the model could transition from a retrospective analysis tool to a forward-looking decision support system. Such predictive capability would substantially increase the model's utility for environmental protection agencies like ARPA and local administrations, enabling proactive rather than reactive air quality management.

Furthermore, the high-resolution $PM_{2.5}$ maps generated through this approach provide an excellent foundation for developing exposure assessment maps. By overlaying population density data with pollution concentration patterns, it becomes possible to identify areas where large numbers of people are exposed to high pollution levels. This is particularly valuable for protecting vulnerable communities and aligns with Sustainable Development Goal indicators related to air quality and public health.

The methodological framework developed in this research, combining satellite data fusion, multi-stage modeling, and machine learning approaches, has potential applications beyond air quality monitoring. Similar approaches could be adapted to other environmental parameters where high spatial resolution is required but direct measurements are sparse, such as urban heat island effects.

In summary, future research directions should focus on:

- Address the "stepped" effect of meteorological variables through more advanced interpolation techniques
- Verify the model's transferability in different urban contexts to establish boundary conditions for its applicability
- Integrate data from the future MAIA mission to improve both high-resolution AOD products and PM speciation
- Develop temporal forecasting capability by integrating meteorological prediction models
- Create population exposure metrics by combining high-resolution PM maps with demographic data

- Initiate epidemiological studies using these high-resolution maps to more accurately assess the association between long-term PM_{2.5} exposure and health effects, overcoming limitations of previous studies based on coarser resolution exposure estimates

These future developments would strengthen the practical impact of this research, transforming it from a methodological advance into an operational tool to improve urban air quality management and, ultimately, public health outcomes.

Bibliography

- [1] Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente Puglia. Rete regionale di monitoraggio della qualità dell'aria, 2023. URL https://www.arpa.puglia.it/pagina2806_rete-regionale-di-monitoraggio-della-qualit-dellaria.html. Accessed: 2023-04-25.
- [2] Lorenzo Angiuli, Paolo Rosario Dambruso, Fiorella Mazzone, Livia Trizio, Angela Morabito, Francesca Intini, Ilenia Schipa, and Annalisa Tanzarella. Valutazione integrata della qualità dell'aria in puglia-anno 2021. Report tecnico, Agenzia Regionale per la Protezione Ambientale-ARPA PUGLIA, 2022. URL https://www.arpa.puglia.it/pagina2873_report-annuali-e-mensili-qualit-dellaria-rrqa.html.
- [3] Lorenzo Angiuli, Paolo Rosario Dambruso, Fiorella Mazzone, Livia Trizio, Angela Morabito, Francesca Intini, Ilenia Schipa, and Annalisa Tanzarella. Valutazione integrata della qualità dell'aria in puglia-anno 2022. Report tecnico, Agenzia Regionale per la Protezione Ambientale-ARPA PUGLIA, 2023. URL https://www.arpa.puglia.it/pagina2873_report-annuali-e-mensili-qualit-dellaria-rrqa.html.
- [4] Mariella Aquilino, Maria Adamo, Palma Blonda, Angela Barbanente, and Cristina Tarantino. Improvement of a dasymetric method for implementing sustainable development goal 11 indicators at an intra-urban scale. *Remote Sensing*, 13(14):2835, 2021.
- [5] Shin Araki, Masayuki Shima, and Kouhei Yamamoto. Spatiotemporal land use random forest model for estimating metropolitan no2 exposure in japan. *Science of The Total Environment*, 634:1269–1277, 2018.
- [6] Oluwaseyi Olalekan Arowosegbe, Martin Röösl, Nino Künzli, Apolline Saucy, Temitope C Adebayo-Ojo, Joel Schwartz, Moses Kebalepile, Mohamed Fareed Jeebhay, Mohamed Aqiel Dalvie, and Kees De Hoogh. Ensemble averaging using remote sensing data to model spatiotemporal pm10 concentrations in sparsely monitored south africa. *Environmental Pollution*, 310:119883, 2022.

- [7] *Specie chimiche nel particolato (PM_{2.5}) in atmosfera*. ARPA - Agenzia regionale per la protezione dell'ambiente, 2024.
- [8] ARPA Puglia. Agenzia regionale per la prevenzione e la protezione dell'ambiente, 2023. URL <https://www.arpa.puglia.it/>. Accessed: 2025-04-19.
- [9] ARPA Puglia RRQA. Agenzia regionale per la prevenzione e la protezione dell'ambiente, 2023. URL <https://dati.arpa.puglia.it/qaria>. Accessed: 2023-04-19.
- [10] Hossein Bagheri. A machine learning-based framework for high resolution mapping of pm_{2.5} in tehran, iran, using maiaac aod data. *Advances in space Research*, 69(9):3333–3349, 2022.
- [11] Fabio Biancofiore, Marcella Busilacchio, Marco Verdecchia, Barbara Tomassetti, Eleonora Aruffo, Sebastiano Bianco, Sinibaldo Di Tommaso, Carlo Colangeli, Gianluigi Rosatelli, and Piero Di Carlo. Recursive neural network model for analysis and forecast of pm₁₀ and pm_{2.5}. *Atmospheric Pollution Research*, 8(4):652–659, 2017.
- [12] Stacey W Boland, David J Diner, John C Pearson, and Kevin A Burke. Nasa's multi-angle imager for aerosols (maia) earth venture instrument investigation. In *AGU Fall Meeting Abstracts*, volume 2018, pages GH41C–1443, 2018.
- [13] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [14] Leo Breiman, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, New York, 1984. ISBN 9781315139470. doi: 10.1201/9781315139470.
- [15] F. Calastrini, F. Guarnieri, S. Becagli, C. Busillo, M. Chiari, U. Dayan, F. Lucarelli, S. Nava, M. Pasqui, R. Traversi, R. Udisti, and G. Zipoli. Desert dust outbreaks over mediterranean basin: A modeling, observational, and synoptic analysis approach. *Desert Dust Properties, Modelling, and Monitoring. Advances in Meteorology, Hindawi*, 2012.
- [16] Ekta Chaudhary, Franciosalgeo George, Aswathi Saji, Sagnik Dey, Santu Ghosh, Tinku Thomas, Anura V Kurpad, Sumit Sharma, Nimish Singh, Shivang Agarwal, et al. Cumulative effect of pm_{2.5} components is larger than the effect of pm_{2.5} mass on child health in india. *Nature Communications*, 14(1):6955, 2023.
- [17] Ekta Chaudhary, Franciosalgeo George, Aswathi Saji, Sagnik Dey, Santu Ghosh, Tinku Thomas, Anura V Kurpad, Sumit Sharma, Nimish Singh, Shivang Agarwal, et al. Cumulative effect of pm_{2.5} components is larger than the effect of pm_{2.5} mass on child health in india. *Nature Communications*, 14(1):6955, 2023.

- [18] Gang Chen, Qihao Weng, Geoffrey J Hay, and Yinan He. Geographic object-based image analysis (geobia): Emerging trends and future opportunities. *GI-Science & Remote Sensing*, 55(2):159–182, 2018.
- [19] Gongbo Chen, Shanshan Li, Luke D Knibbs, Nicholas AS Hamm, Wei Cao, Tiantian Li, Jianping Guo, Hongyan Ren, Michael J Abramson, and Yuming Guo. A machine learning method to estimate pm2. 5 concentrations across china with remote sensing, meteorological and land use information. *Science of the Total Environment*, 636:52–60, 2018.
- [20] Jie Chen, Kees de Hoogh, John Gulliver, Barbara Hoffmann, Ole Hertel, Matthias Ketzel, Mariska Bauwelinck, Aaron Van Donkelaar, Ulla A Hvidtfeldt, Klea Katsouyanni, et al. A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide. *Environment international*, 130:104934, 2019.
- [21] Nengcheng Chen, Meijuan Yang, Wenying Du, and Min Huang. Pm2. 5 estimation and spatial-temporal pattern analysis based on the modified support vector regression model and the 1 km resolution maiac aod in hubei, china. *ISPRS International Journal of Geo-Information*, 10(1):31, 2021.
- [22] Zhao-Yue Chen, Tian-Hao Zhang, Rong Zhang, Zhong-Min Zhu, Jun Yang, Ping-Yan Chen, Chun-Quan Ou, and Yuming Guo. Extreme gradient boosting model to estimate pm2. 5 concentrations with missing-filled satellite data in china. *Atmospheric Environment*, 202:180–189, 2019.
- [23] Bahram Choubin, Mahsa Abdolshahnejad, Ehsan Moradi, Xavier Querol, Amir Mosavi, Shahaboddin Shamshirband, and Pedram Ghamisi. Spatial hazard assessment of the pm10 using machine learning models in barcelona, spain. *Science of The Total Environment*, 701:134474, 2020.
- [24] Copernicus and EUMETSAT. Wekeo - copernicus data space ecosystem, 2025. URL <https://wekeo.copernicus.eu/>. EU’s Copernicus Data and Information Access Service (DIAS).
- [25] Copernicus Land Monitoring Service. Urban atlas, 2023. URL <https://land.copernicus.eu/en/products/urban-atlas>. European Environment Agency.
- [26] P Crippa, RC Sullivan, A Thota, and SC Pryor. Sensitivity of simulated aerosol properties over eastern north america to wrf-chem parameterizations. *Journal of Geophysical Research: Atmospheres*, 124(6):3365–3383, 2019.
- [27] J. M. Delgado-Saborit, V. Guercio, A. M. Gowers, G. Shaddick, N. C. Fox, and S. Love. A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population. *Science of The Total Environment*, 2021.

- [28] Eriberto de'Munari, Ivo Allegrini, Nadia Bardizza, Nadia Carfagno, Natascia Di Carlo, Alessandra Gaeta, Guido Lanzani, Marco Malaguti, Giovanna Marson, Chiara Melegari, et al. Linee guida per la predisposizione delle reti di monitoraggio della qualità dell'aria in Italia. *APAT CTN-ACE*, 2004.
- [29] Qian Di, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M Benjamin Sabath, Christine Choirat, Petros Koutrakis, Alexei Lyapustin, et al. An ensemble-based model of pm_{2.5} concentration across the contiguous united states with high spatiotemporal resolution. *Environment international*, 130:104909, 2019.
- [30] Hiep Nguyen Duc, Md Mahmudur Rahman, Toan Trieu, Merched Azzi, Matthew Riley, Thomas Koh, Shaohua Liu, Kasun Bandara, Vishall Krishnan, Yujing Yang, et al. Study of planetary boundary layer, air pollution, air quality models and aerosol transport using ceilometers in new south wales (nsw), australia. *Atmosphere*, 13(2):176, 2022.
- [31] Dimitrios Effrosynidis and Avi Arampatzis. An evaluation of feature selection methods for environmental data. *Ecological Informatics*, 61:101224, 2021.
- [32] W. H. Organization et al. Who global air quality guidelines: particulate matter (pm_{2.5} and pm₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Technical report, World Health Organization, 2021.
- [33] EUBUCCO Project. Eubucco - european building and urban cartography collection, 2025. URL <https://eubucco.com/>. European building footprint and urban information dataset.
- [34] European Environment Agency. Criteria for euroairnet: The EEA air quality monitoring and information network, 1999. URL <https://www.eea.europa.eu/en/analysis/publications/tec12>. Accessed: 2023-04-25.
- [35] European Environment Agency. Monitoring station classifications and criteria for including them in the eea's assessment products. <https://www.eea.europa.eu/en/topics/in-depth/air-pollution/monitoring-station-classifications-and-criteria>, 2024. Accessed: March 2025.
- [36] European Parliament and Council. 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, 2008. URL <http://data.europa.eu/eli/dir/2008/50/oj>. OJ L 152, 11.6.2008.
- [37] Jilin Gu, Yiwei Wang, Ji Ma, Yaoqi Lu, Shaohua Wang, and Xueming Li. An estimation method for pm_{2.5} based on aerosol optical depth obtained from

- remote sensing image processing and meteorological factors. *Remote Sensing*, 14(7):1617, 2022.
- [38] Jianping Guo, Feng Xia, Yong Zhang, Huan Liu, Jing Li, Mengyun Lou, Jing He, Yan Yan, Fu Wang, Min Min, et al. Impact of diurnal variability and meteorological factors on the pm_{2.5}-aod relationship: Implications for pm_{2.5} remote sensing. *Environmental Pollution*, 221:94–104, 2017.
- [39] Miaocai Guo. Application of remote sensing technology in macro-ecological environment monitoring. *Remote Sens*, 9:26, 2020.
- [40] Y. Guo, H. Zeng, R. Zheng, S. Li, A. G. Barnett, S. Zhang, X. Zou, R. Huxley, W. Chen, and G. Williams. The association between lung cancer incidence and ambient air pollution in china: A spatiotemporal analysis. *Environmental Research*, 2016.
- [41] Pawan Gupta and Sundar A Christopher. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. a neural network approach. *Journal of Geophysical Research: Atmospheres*, 114(D20), 2009.
- [42] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825): 357–362, 2020.
- [43] Stephan Hoyer and Joe Hamman. xarray: Nd labeled arrays and datasets in python. *Journal of Open Research Software*, 5(1):10–10, 2017.
- [44] Xuefei Hu, Jessica H Belle, Xia Meng, Avani Wildani, Lance A Waller, Matthew J Strickland, and Yang Liu. Estimating pm_{2.5} concentrations in the conterminous united states using the random forest approach. *Environmental science & technology*, 51(12):6936–6944, 2017.
- [45] S Ibrahim, M Landa, O Pešek, L Brodský, and L Halounová. Machine learning-based approach using open data to estimate pm_{2.5} over europe, remote sens., 14, 3392, 2022.
- [46] *Global Burden of Disease*. Institute for Health Metrics and Evaluation, 2021.
- [47] ISTAT. Censimento della popolazione e delle abitazioni - risultati, 2021. URL <https://www.istat.it/statistiche-per-temi/censimenti/popolazione-e-abitazioni/risultati/>. Istituto Nazionale di Statistica.
- [48] Istituto Superiore per la Protezione e la Ricerca Ambientale. Sistema nazionale per la protezione dell’ambiente, 2023. URL <https://www.isprambiente.gov.it/it/sistema-nazionale-protezione-ambiente>. Accessed: 2023-04-25.

- [49] Shamin Kabir Ki-Hyun Kim, Ehsanul Kabir. A review on the human health impact of airborne particulate matter. *Environment International*, 2014.
- [50] Itai Kloog, Petros Koutrakis, Brent A Coull, Hyung Joo Lee, and Joel Schwartz. Assessing temporally and spatially resolved pm_{2.5} exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric environment*, 45(35):6267–6275, 2011.
- [51] RBA Koelemeijer, CD Homan, and J Matthijsen. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over europe. *Atmospheric Environment*, 40(27):5304–5315, 2006.
- [52] Muneendra Kumar. World geodetic system 1984: A modern and accurate global reference frame. *Marine Geodesy*, 12(2):117–126, 1988.
- [53] Miron B Kursa and Witold R Rudnicki. Feature selection with the boruta package. *Journal of statistical software*, 36:1–13, 2010.
- [54] Hyouk-Soo Kwon, Min Hyung Ryu, and Christopher Carlsten. Ultrafine particles: unique physicochemical properties relevant to health and disease. *Experimental & molecular medicine*, 52(3):318–328, 2020.
- [55] F. Sánchez Lasheras, P. J. García Nieto, E. García Gonzalo, L. Bonavera, and F. J. de Cos Juez. Evolution and forecasting of pm₁₀ concentration at the port of gijon (spain). *Scientific Reports*, 10, 2020. doi: 10.1038/s41598-020-68636-5.
- [56] HJ Lee, Y Liu, BA Coull, J Schwartz, and P Koutrakis. A novel calibration approach of modis aod data to predict pm_{2.5} concentrations. *Atmospheric Chemistry and Physics*, 11(15):7991–8002, 2011.
- [57] Mihye Lee, Itai Kloog, Alexandra Chudnovsky, Alexei Lyapustin, Yujie Wang, Steven Melly, Brent Coull, Petros Koutrakis, and Joel Schwartz. Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the southeastern us 2003–2011. *Journal of exposure science & environmental epidemiology*, 26(4):377–384, 2016.
- [58] Danny M Leung, Amos PK Tai, Loretta J Mickley, Jonathan M Moch, Aaron Van Donkelaar, Lu Shen, and Randall V Martin. Synoptic meteorological modes of variability for fine particulate matter (pm_{2.5}) air quality in major metropolitan regions of china. *Atmospheric Chemistry and Physics*, 18(9):6733–6748, 2018.
- [59] Jing Li, Eric Garshick, Jaime E Hart, Longxiang Li, Liuhua Shi, Ali Al-Hemoud, Shaodan Huang, and Petros Koutrakis. Estimation of ambient pm_{2.5} in iraq and kuwait from 2001 to 2018 using machine learning and remote sensing. *Environment International*, 151:106445, 2021.

- [60] Antti Lipponen, Ville Kolehmainen, Pekka Kolmonen, Antti Kukkurainen, Tero Mielonen, Neus Sabater, Larisa Sogacheva, Timo H Virtanen, and Antti Arola. Model enforced post-process correction of satellite aerosol retrievals. *Atmospheric Measurement Techniques Discussions*, 2020:1–17, 2020.
- [61] Antti Lipponen, Jaakko Reinval, Arttu Väisänen, Henri Taskinen, Timo Lähivaara, Larisa Sogacheva, Pekka Kolmonen, Kari Lehtinen, Antti Arola, and Ville Kolehmainen. Deep-learning-based post-process correction of the aerosol parameters in the high-resolution sentinel-3 level-2 synergy product. *Atmospheric Measurement Techniques*, 15(4):895–914, 2022.
- [62] Yang Liu, Jeremy A Sarnat, Vasu Kilaru, Daniel J Jacob, and Petros Koutrakis. Estimating ground-level pm_{2.5} in the eastern united states using satellite remote sensing. *Environmental science & technology*, 39(9):3269–3278, 2005.
- [63] Ying Liu, Chunyuan Li, Dongren Liu, Yulei Tang, Barnabas C Seyler, Zihang Zhou, Xia Hu, Fumo Yang, and Yu Zhan. Deriving hourly full-coverage pm_{2.5} concentrations across china’s sichuan basin by fusing multisource satellite retrievals: A machine-learning approach. *Atmospheric Environment*, 271:118930, 2022.
- [64] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [65] A. Lyapustin and Y. Wang. *MODIS Multi-Angle Implementation of Atmospheric Correction (MAIAC) Data User’s Guide*. NASA, Greenbelt, MD, USA, 2018.
- [66] Alexei Lyapustin, Y Wang, I Laszlo, R Kahn, Sergey Korkin, L Remer, R Levy, and JS Reid. Multiangle implementation of atmospheric correction (maiac): 2. aerosol algorithm. *Journal of Geophysical Research: Atmospheres*, 116(D3), 2011.
- [67] Jinghui Ma, Zhongqi Yu, Yuanhao Qu, Jianming Xu, and Yu Cao. Application of the xgboost machine learning method in pm_{2.5} prediction: A case study of shanghai. *Aerosol and Air Quality Research*, 20(1):128–138, 2020.
- [68] Y. Ma, H. Zhang, Y. Zhao, J. Zhou, S. Yang, X. Zheng, and S. Wang. Short-term effects of air pollution on daily hospital admissions for cardiovascular diseases in western china. *Environmental Science and Pollution Research*, 2017.
- [69] I. McKendry. Evaluation of artificial neural networks for fine particulate pollution (pm₁₀ and pm_{2.5}) forecasting. *Journal of the Air & Waste Management Association*, 52:1096 – 1101, 2002. doi: 10.1080/10473289.2002.10470836.
- [70] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

- [71] Ali Mirzaei, Hossein Bagheri, and Mehran Sattari. Data level and decision level fusion of satellite multi-sensor aod retrievals for improving pm_{2.5} estimations, a study on tehran. *Earth Science Informatics*, 16(1):753–771, 2023.
- [72] NASA Goddard Space Flight Center. MODIS Atmosphere L2 Aerosol Product. <https://modis.gsfc.nasa.gov/data/dataproduct/mod04.php>, 2024. 04/2025.
- [73] NASA Jet Propulsion Laboratory. Maia data and applications, 2023. URL <https://maia.jpl.nasa.gov/resources/data-and-applications/>. Accessed: 2023-04-25.
- [74] NASA Jet Propulsion Laboratory. Maia mission, 2023. URL <https://maia.jpl.nasa.gov/mission/>. Accessed: 2023-04-25.
- [75] United Nations. World population prospects 2024: Summary of results. *Statistical Papers - United Nations, Population and Vital Statistics Report*, 2024.
- [76] P. G. Nieto, F. S. Lasheras, E. García-Gonzalo, and F. J. D. C. Juez. Pm₁₀ concentration forecasting in the metropolitan area of oviedo (northern spain) using models based on svm, mlp, varma and arima: A case study. *The Science of the total environment*, 621:753–761, 2018. doi: 10.1016/j.scitotenv.2017.11.291.
- [77] OpenStreetMap Contributors. Map features - openstreetmap wiki, 2025. URL https://wiki.openstreetmap.org/wiki/Map_features. OpenStreetMap Wiki.
- [78] W. H. Organization. Ambient (outdoor) air pollution. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)., 24 October 2024.
- [79] Wei Ouyang, Bobo Guo, Guanqing Cai, Qing Li, Sheng Han, Bin Liu, and Xingang Liu. The washing effect of precipitation on particulate matter and the pollution dynamics of rainwater in downtown beijing. *Science of the Total Environment*, 505:306–314, 2015.
- [80] James F Pankow, John ME Storey, and Hiroyasu Yamasaki. Effects of relative humidity on gas/particle partitioning of semivolatile organic compounds to urban particulate matter. *Environmental science & technology*, 27(10):2220–2226, 1993.
- [81] Seohui Park, Miae Kim, and Jungho Im. Estimation of ground-level pm₁₀ and pm_{2.5} concentrations using boosting-based machine learning from satellite and numerical weather prediction data. *Korean Journal of Remote Sensing*, 37(2): 321–335, 2021.
- [82] F. Perera. Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: Solutions exist. *International Journal of Environmental Research and Public Health*, 2017.

- [83] J. Rentschler and N. Leonova. Global air pollution exposure and poverty. *Nature Communications*, 2023.
- [84] H. Saygin, Y. Mercan, and F. Yorulmaz. The association between air pollution parameters and emergency department visits and hospitalizations due to cardiovascular and respiratory diseases: A time-series analysis. *International Archives of Occupational and Environmental Health*, 2021.
- [85] Dean E Schraufnagel. The health effects of ultrafine particles. *Experimental & molecular medicine*, 52(3):311–317, 2020.
- [86] Maximilian Schwarz, Alexandra Schneider, Josef Cyrus, Susanne Bastian, Susanne Breitner, and Annette Peters. Impact of ambient ultrafine particles on cause-specific mortality in three german cities. *American Journal of Respiratory and Critical Care Medicine*, 207(10):1334–1344, 2023.
- [87] R. Sivarethinamohan, S. Sujatha, S. Priya, Sankaran, A. Gafoor, , and Z. Rahman. Impact of air pollution in health and socio-economic aspects: Review on future approach. *International Conference on Newer Trends and Innovation in Mechanical Engineering: Materials Science*, 2021.
- [88] William C Skamarock, Joseph B Klemp, Jimy Dudhia, David O Gill, Dale M Barker, Michael G Duda, Xiang-Yu Huang, Wei Wang, Jordan G Powers, et al. A description of the advanced research wrf version 3. *NCAR technical note*, 475 (125):10–5065, 2008.
- [89] Amos PK Tai, Loretta J Mickley, and Daniel J Jacob. Correlations between fine particulate matter (pm_{2.5}) and meteorological variables in the united states: Implications for the sensitivity of pm_{2.5} to climate change. *Atmospheric environment*, 44(32):3976–3984, 2010.
- [90] M. Tainio, Z. Jovanovic Andersen, M. J. Nieuwenhuijsen, L. Hu, A. de Nazelle, R. An, L. M. Garcia, S. Goenka, B. Zapata-Diomed, F. Bull, and T. H. d. Sà. Air pollution, physical activity and health: A mapping review of the evidence. *Environment International*, 2021.
- [91] Cristina Tarantino, Marica De Lucia, Luciana Zollo, Mariagiovanna Dell’Aglia, Maria Adamo, and Rocco Labadessa. Combination of geobia and data-driven approach for grassland habitat mapping in the alta murgia national park. *Science of Remote Sensing*, 2025.
- [92] Abdulwaheed Tella, Abdul-Lateef Balogun, Naheem Adebisi, and Samsuri Abdullah. Spatial assessment of pm₁₀ hotspots using random forest, k-nearest neighbour and naïve bayes. *Atmospheric Pollution Research*, 12(10):101202, 2021.

- [93] Trimble. *eCognition Developer 9.5 Reference Book*. Trimble Inc., 2019. URL <https://geospatial.trimble.com/en/products/software/trimble-ecognition>.
- [94] Vladimir Vapnik, Steven Golowich, and Alex Smola. Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems*, 9, 1996.
- [95] Jun Wang and Sundar A Christopher. Intercomparison between satellite-derived aerosol optical thickness and pm_{2.5} mass: Implications for air quality studies. *Geophysical research letters*, 30(21), 2003.
- [96] World Health Organization. Ambient (outdoor) air quality and health, 2023. URL [https://www.who.int/en/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health). Accessed: 2023-04-25.
- [97] Chih-Da Wu, Yu-Cheng Chen, Wen-Chi Pan, Yu-Ting Zeng, Mu-Jean Chen, Yue Leon Guo, and Shih-Chun Candice Lung. Land-use regression with long-term satellite-based greenness index and culture-specific sources to model pm_{2.5} spatial-temporal variability. *Environmental pollution*, 224:148–157, 2017.
- [98] Qianqian Yang, Qiangqiang Yuan, Linwei Yue, Tongwen Li, Huanfeng Shen, and Liangpei Zhang. The relationships between pm_{2.5} and aerosol optical depth (aod) in mainland china: About and behind the spatio-temporal variations. *Environmental Pollution*, 248:526–535, 2019.
- [99] Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, and Somayeh Talebiesfandarani. Pm_{2.5} prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, 10(7):373, 2019.
- [100] Luo Zhang, Peng Liu, Lei Zhao, Guizhou Wang, Wangfeng Zhang, and Jianbo Liu. Air quality predictions with a semi-supervised bidirectional lstm neural network. *Atmospheric Pollution Research*, 12(1):328–339, 2021.
- [101] Z. Zhang, G. Zhang, and B. Su. The spatial impacts of air pollution and socio-economic status on public health: Empirical evidence from china. *Socio-Economic Planning Sciences*, 2021.
- [102] Caiwang Zheng, Chuanfeng Zhao, Yannian Zhu, Yang Wang, Xiaoqin Shi, Xiaolin Wu, Tianmeng Chen, Fang Wu, and Yanmei Qiu. Analysis of influential factors for the relationship between pm_{2.5} and aod in beijing. *Atmospheric Chemistry and Physics*, 17(21):13473–13489, 2017.
- [103] Qingping Zhou, Haiyan Jiang, Jianzhou Wang, and Jianling Zhou. A hybrid model for pm_{2.5} forecasting based on ensemble empirical mode decomposition

and a general regression neural network. *Science of the Total Environment*, 496: 264–274, 2014.

