

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Assessment of PM Exposure at intra-urban scale in preparation of MAIA mission (APEMAIA) Project

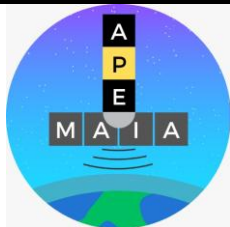
WP2103 – HR PM2.5 and PM10 extraction

DEL009 – Description of the second system module (WP2) for the estimation of PM2.5 and PM10 at high resolution (SW006) and analysis of relations between urban heat islands and intra-urban scale pollution

Contract n. 2023-39-HH.0

CUP n. F83D23000210005

ASI CONTRACT N°	2023-39-HH.0		
PROJECT NAME	Assessment of PM Exposure at intra-urban scale in preparation of MAIA mission (APEMAIA)	CONTRACTOR	National Research Council – Institute of Atmospheric Pollution Research
PROJECT MANAGER	Maria Adamo (CNR-ITIAm)	DATE SUBMITTED	N/A
AUTHOR	M. Adamo, M. Aquilino, Marica De Lucia, Cristina Tarantino (CNR-ITIAm) A. Monaco, A. Fania (DIF) A. Morabito (ARPA Puglia)	CONTRACTOR'S REFERENCE	Issue:01 Rev.:01 Date:25/09/26
PROJECT BEGIN DATE	26/09/2023	PROJECT END DATE	25/09/2026



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

EXECUTIVE SUMMARY			
This document is the release of DEL009 – Description of the second system module (WP2) for the estimation of PM2.5 and PM10 at high resolution (SW006) and analysis of relations between urban heat islands and intra-urban scale pollution – of the Assessment of PM Exposure at an intra-urban scale in preparation for the MAIA mission (APEMAIA) project.			
NAME OF ASI PROJECT MANAGER	Name Surname (PM)		
	Name Surname (RUP)		
	Name Surname (Project Scientist)		
STAR CODE		No OF VOLUMES:	



Project:
APEMAIA

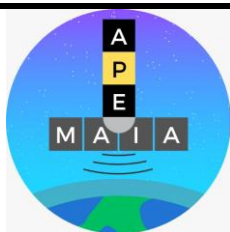
Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

TABLE OF CONTENTS

1	<i>AIM OF THE DOCUMENT</i>	10
1.1	Applicable and reference documents	11
1.2	Applicable documents	11
1.3	Reference documents	11
1.4	Acronyms and abbreviations	11
2	<i>Introduction to the methodological approaches</i>	12
2.1	Study area	15
2.2	PMs Data	15
3	<i>Regional scale approach</i>	24
3.1	Training data	24
3.2	Database of predictors available at regional scale	24
3.3	Method	26
3.3.1	Inverse Distance Weighting	26
3.3.2	Spatial Resample of WRF Data	27
3.3.3	Machine Learning model	28
3.3.4	Explainable Artificial Intelligence	28
3.4	Results	29
4	<i>Intra-urban scale approach</i>	35
4.1	Database of predictors	35
4.1.1	Aerosol Optical Depth (AOD) data	35
4.1.2	Meteorological parameters	36
4.1.3	Vegetation Index	37
4.1.4	Other spatiotemporal data	38
4.1.5	Spatiotemporal aggregation at grid level	43
4.2	ML models	44
4.2.1	Model implementation and validation	45
4.3	Final results	46
4.3.1	Interpretation of the validation results	47
4.3.2	Model performance	49
4.3.3	Predictor relevance	53



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

4.3.4	Model interpretability and SHAP analysis.....	56
4.3.5	Spatial predictions: annual, seasonal and monthly mean maps.....	57
4.3.6	Preliminary investigation of a multi-city approach	62

List of Table

Table 1: Acronyms and abbreviations.....	11
Table 2-Monitoring stations and pollution data in Taranto	17
Table 3-List of all variables collected, with source, spatial and temporal resolution.	25
Table 4-List of all metereological variables collected, with source, spatial and temporal resolution.	37
Table 5- Urban variables used in the intra-urban model.....	39
Table 6-Land Cover Variables used in the intra-urban model	41
Table 7-Urban Land Use Variables used in the intra-urban model	42
Table 8-Type,source and aggregation of predictors used in intra-urban model.....	44
Table 9-PM25 LOSO metrics by station in Taranto.....	52
Table 10-Leave-One-Station-Out (LOSO) cross-validation for PM10, XGB model multi-city vs site-specific approach	63
Table 11-Leave-One-Station-Out (LOSO) cross-validation for PM25, XGB model multi-city vs site-specific approach	64

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

List of figures

Figure 1-Methodological steps for PMs retrieval	14
Figure 2-APEMAIA study areas including metropolitan area of Rome PTA and Apulian sites of Bari and Taranto.....	15
Figure 3-Spatial distribution of Bari monitoring stations	16
Figure 4-Spatial distribution of Taranto monitoring stations	17
Figure 5-Time series of the hours in which PM measurements best approximate the daily averages of PM10 in 2019. Bari Carbonara station at top, ModugnoEN02 in the middle and Modugno EN04 at bottom	19
Figure 6-Comparison between PM10 measured at 8 a.m. and daily average value in Bari Carbonara station.	20
Figure 7-PM ₁₀ and PM _{2.5} measurements divided by station in Bari	21
Figure 8-PM ₁₀ and PM _{2.5} measurements divided by station in Taranto	22
Figure 9-PM ₁₀ and PM _{2.5} boxplot grouped by station	23
Figure 10-Spatial distribution of ARPA Puglia station.....	24
Figure 11-Left panel (panel A) AOD values extracted on the grid (the missing values are attributed to cloud covering and orbit of satellite), on the right panel (panel B) the same image in which NA values are replaced using the Inverse Distance Interpolation meth.....	27
Figure 12-Left panel (A) Surface Emissivity predicted by the WRF model (4 km resolution), on the right panel (B) the resampled image to 300 m with bi-cubic interpolation method.	28
Figure 13-Mean PM10 concentrations predicted by the XGBoost model over the Apulia region at 300 m spatial resolution. Daily predictions were averaged over the 2019–2022 study period. Concentrations are expressed in $\mu\text{g}/\text{m}^3$	30
Figure 14-SHAP summary plot for the XGBoost PM10 model. SHAP values quantify the contribution of each predictor to individual PM10 estimates. Positive values increase the predicted concentration, whereas negative values decrease it. Feature colours represent the	31
Figure 15-Mean PM2.5 concentrations predicted by the XGBoost model over the Apulia region at 300 m spatial resolution. Daily predictions were averaged over the 2019–2022 study period. Concentrations are expressed in $\mu\text{g}/\text{m}^3$	32
Figure 16-SHAP summary plot for the XGBoost PM2.5 model. The plot shows the magnitude and direction of the contribution of each predictor to the estimated PM2.5 concentration. Positive SHAP values correspond to an increase in the model output, whereas negative va.....	33
Figure 17-Interpolated NDVI on the Bari stations	38
Figure 18-Traffic indicators from OpenStreetmap	39

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Figure 19-Taranto, a) population per census zone, b) population density map 100x100m	40
Figure 20-Bari, LC mapping	41
Figure 21-Building use classification over Baria area	42
Figure 22-Modelling workflow for the intra-urban PMs model	45
Figure 23-Configuration of the four final products: selected model, predictor set and number of monitoring stations available for training and validation.	47
Figure -Permutation importance of the predictors of the final XGBoost models for Taranto: PM10 (left) and PM2.5 (right).	54
Figure -Shap values of predictors in the XGBoost PM10 and PM25 model (Bari).....	56
Figure 26-SHAP values of predictors in the XGBoost PM10 and PM25 model (Taranto)	57
Figure -Bari, map of annual average PM10 levels.	Errore. Il segnalibro non è definito.
Figure -Bari, map of annual average PM25 levels	59
Figure 29-Taranto, map of annual average PM10 levels.	60
Figure 30-Taranto, map of annual average PM25 levels.	61
Figure 31-Bari, map of seasonal average PM10 levels (2019).....	62
Figure 32-Leave-One-Station-Out (LOSO) cross-validation for PM25, XGB model multi-city vs site-specific approach	Errore. Il segnalibro non è definito.
Figure -Shap summary plot for the multi-city XGB model for PM10 and PM25	65

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

1 AIM OF THE DOCUMENT

This document represents the preliminary version D009 Description of the second system module (WP2) for the estimation of PM_{2.5} and PM₁₀ at high resolution (SW006) and analysis of relations between urban heat islands and intra-urban scale pollution for the **Assessment of PM Exposure** at an intra-urban scale in preparation for the **MAIA** mission (APEMAIA) Project.

The APEMAIA project's activities are carried out by a consortium led by the Institute of Atmospheric Pollution Research of the National Research Council of Italy (CNR-IRA), acting as the prime contractor, and supported by the Apulian Regional Agency for Environmental Protection (ARPA Puglia) and the Physics Department of the University of Bari (DIF-UNIBA).

Recent World Health Organization reports highlight that air pollution is a significant health issue affecting both current and future populations. Epidemiologists refer to premature deaths cardiovascular and respiratory diseases. This is especially true in cities, where population concentrate and where an estimated 68% of the world's population - about 6.7 billion people - will reside by 2050.

The United Nations 2030 Agenda for Sustainable Development Goals (SDGs) includes specific targets to reduce air pollution and indicators. The latter provides quantitative metrics to assess air quality and its impact on human health. Among these, indicator 11.6.2, which measures annual mean levels of fine particulate matter is most relevant for urban environments.

However, current PM mapping services offered by monitoring agencies have limitations. We found in previous projects that the available 4km resolution PM maps provided by ARPA Puglia monitoring service don't allow local policymakers to extract useful information about areas with the highest exposure within cities. This underscores the need for precise, high-resolution mapping of concentrations within a city that effectively captures variations in residential, industrial, suburban, and port areas. Thus, the goal of APEMAIA project is to improve the spatiotemporal estimation of PM.

Our approach involves extensive use of Machine Learning techniques with the aim to model daily intra-urban PM₁₀ and PM_{2.5}. PMs ground measurements from ARPA Puglia monitoring stations were used as references to train and validate the model because they are the most accurate source of PM measurement available for the study area. AOD was considered as the main proxy of PM spatial variability, but it was used in combination with a wide list of other predictors.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

1.1 Applicable and reference documents

This document incorporates provisions from other dated or undated documents. These references are cited at appropriate places in the text and are listed hereafter. For undated references, the latest edition of the publication referred applies.

In case of conflict between two or more applicable documents, the higher document will prevail.

1.2 Applicable documents

[AD 1] ASI Contract No. 2023-39-HH.0

[AD 2] APEMAIA DEL004 – User Requirement Document -USD (preliminary version)

[AD 3] APEMAIA DEL006 – Dissemination and Exploitation Plan

[AD 4] APEMAIA DEL008 – Description of the first system module (WP2) for the extraction of high-resolution AOD maps (SW004); description of the Data-driven and Knowledge-Driven procedures for the extraction of LCLU maps in urban areas (SW005) (preliminary version)

1.3 Reference documents

[RD 1] WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.

[RD 2] Stafoggia et al. Estimation of daily PM₁₀ and PM_{2.5} concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environmental International (2019).

1.4 Acronyms and abbreviations

Errore. L'origine riferimento non è stata trovata. provides the acronym list.

Table 1: Acronyms and abbreviations

Acronym	Description
AOD	Aerosol Optical Depth
AI	Artificial Intelligence
APEMAIA	Assessment of PM Exposure at intra-urban scale in preparation of MAIA mission
ARPA Puglia	Apulian Regional Agency for Environmental Protection

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

2 Introduction to the methodological approaches

The goal of APEMAIA project is to improve the spatiotemporal estimation of PM. The proposed approach involves extensive use of Machine Learning Deep Learning techniques with the aim to model daily intra-urban PM₁₀ and PM_{2.5}.

Traditionally, PM concentration mapping relies on the use of Chemical Transport Models (CTMs) and statistical analysis which relate PM to other variables. All these model work at a coarse scale due to the strong dependence from the meteorological variables available at low resolution. Satellite data, on the contrary, employed in a ML model aim at providing a continuous monitoring of PM concentration with a daily frequency with high spatial coverage, whereas ground station measurements can only represent the concentration of relatively small areas. Dataset from different sources were considered: the Aerosol Optical Depth (AOD) collected by satellite, meteorological data provided by the WRF model, urban ancillary variables, as Particulate Matter (PM₁₀ and PM_{2.5}) measured at ground-based air monitoring stations. The datasets were collected over four years, from January 2019 to December 2022 AOD has been used as a proxy to model PM spatial variability with a daily frequency and high spatial coverage, but ML model requires additional features to be performed.

The implementation steps have been translated into open-source code, enabling the automated generation of training datasets for ML algorithms and the production of PM concentration maps. Regarding the choice of programming language, all code is currently based on Python (versions 3.8 and above) or Bash scripting. Automated workflows that combine different parts of the code are packaged as Docker containers. The use of Docker stems from the need and opportunity to decouple the software from a specific architecture. Subject to client approval, the Docker containers are shared on a public hub and on the project's official GitHub page: <https://github.com/APEMAIAproject/src>.

The architecture of the PM retrieval software is organised into several modules:

0_Preprocessing — Handles the initial data preparation, cleaning and transformation to ensure the input datasets are in a suitable format for further processing.

1_AOD_extraction — Focuses on extracting Aerosol Optical Depth (AOD) data produced according to the procedure described in [AD 4].

2_WRFOUT_extraction — Processes output from Weather Research and Forecasting (WRF) model simulations to extract relevant meteorological and air quality variables for analysis.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

3_DATABASE_building — Combines data from various sources (AOD, WRF, etc.) to construct a comprehensive database that serves as the foundation for machine learning and statistical analyses.

4_ML_models — Implements machine learning models for predicting air quality parameters, analysing trends and exploring exposure levels. The trained model is serialised in JSON format for operational use.

5_Mapping — Applies the serialised model to daily predictor data cubes stacked on the target grid, generating daily PM concentration maps. These are aggregated temporally to produce annual mean concentration maps, which constitute the final product.

In the figure below these steps are schematized.

Regarding the preliminary steps, several scripts were developed in the 2_WRFOUT_extraction and 3_DATABASE_building modules aimed at:

- Extracting meteorological and air quality variables from WRF model NetCDF files for specific time ranges and locations.
- Ensuring that only the relevant NetCDF files are processed by matching them to the dates found in input .txt files.
- Merging datasets, integrating WRF output variables with other data sources (e.g. AOD, geographical data) using a modular and efficient approach.
- Preparing an updated and enriched dataset for 3_DATABASE_building, facilitating the generation of daily data products for subsequent machine learning analyses.

A detailed description of the individual processing steps — the construction of the training dataset, the configuration and validation of the machine learning model, and the temporal aggregation criteria adopted for the annual maps — is provided in the following paragraphs.

Part of the code examples have been shared on the APEMAIA project's GitHub profile for the benefit of the scientific community; additional modules can be made available upon request.

Regarding such latter aspect, we developed two different approaches:

- **Regional scale approach** aimed at implementing a method for rapid mapping large region with easily accessible list of predictors in order to facilitate reproducibility in other areas of study;
- **Intra-urban scale approach** that makes use of small-scale predictors in order to improving PM estimates at regional scale, accordingly to workflow proposed in [RD 2].

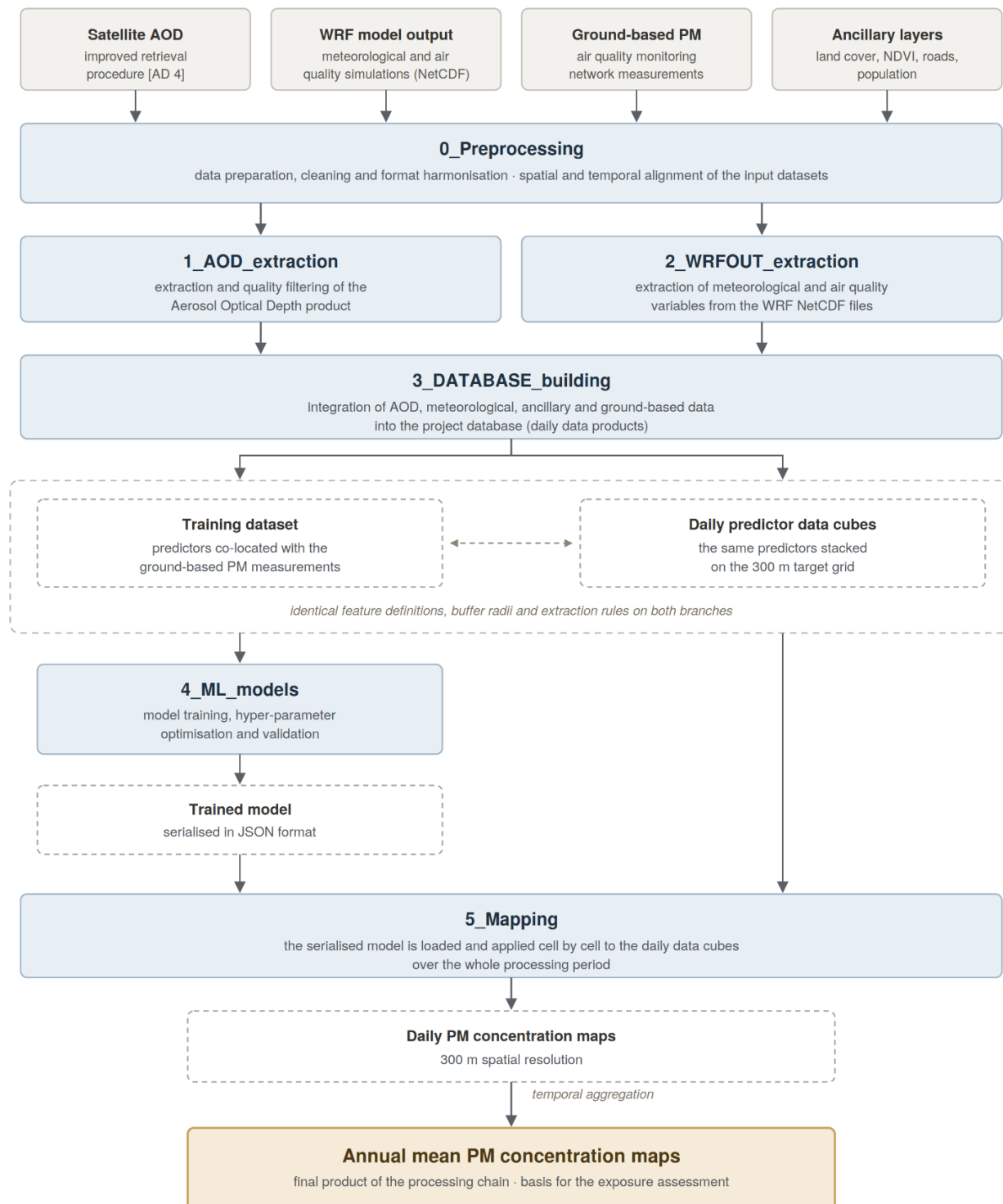


Figure 1-Methodological steps for PMs retrieval



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

2.1 Study area

APEMAIA project focuses on primary and secondary target areas identified in support of NASA JPL's MAIA mission. Bari, although a secondary target area, has been selected as the calibration/validation site for the APEMAIA project. This decision is driven by the availability of comprehensive information about its territory collected through previous projects, but the study includes also an industrial site as Taranto city in Apulia region.



Figure 2-APEMAIA study areas including metropolitan area of Rome PTA and Apulian sites of Bari and Taranto.

2.2 PMs Data

PM data used in this study were provided by the Apulia Regional Agency for the Prevention and Protection of the Environment (ARPA). ARPA Apulia monitors air quality through fixed stations of the Regional network for Air Quality monitoring (RRQA), mobile laboratory campaigns, and additional sampling instruments. 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 network selected for this study includes several stations for PM10 and PM2.5, strategically distributed



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

throughout the Bari and Taranto area as shown in the two images. Tables below details the characteristics of each station, specifying the type of pollutant monitored, the type of area (urban or suburban), and the classification of the station based on the main source of pollution (traffic, industrial, or background). The data collected consists of daily avarages concentrations of PM10 and PM2.5



Figure 3-Spatial distribution of Bari monitoring stations

Station	Pollution	Type Station
Caldarola	PM ₁₀ , PM _{2.5}	Traffic
Cavour	PM ₁₀ , PM _{2.5}	Traffic
Kennedy	PM ₁₀ , PM _{2.5}	Background
Bitonto	PM ₁₀ , PM _{2.5}	Industrial
CUS	PM ₁₀	Traffic
Carbonara	PM ₁₀	Background
Modugno_EN02	PM ₁₀	Industrial
Modugno_EN03	PM ₁₀	Industrial
Modugno_EN04	PM ₁₀	Industrial

Table 2-Monitoring stations and pollution data in Bari



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66



Figure 4-Spatial distribution of Taranto monitoring stations

Station	Pollution	Type Station
CISI	PM ₁₀ , PM _{2.5}	Industrial
Tamburi	PM ₁₀ , PM _{2.5}	Industrial
Archimede	PM ₁₀ , PM _{2.5}	Industrial
Macchiavelli	PM ₁₀ , PM _{2.5}	Industrial
San Vito	PM ₁₀	Background
Statte	PM ₁₀	Industrial
Talsano	PM ₁₀	Background
Ponte Wind	PM ₁₀	Industrial

PM ground measurements from ARPA Puglia monitoring stations were used as references to train and validate the model because they are the most accurate source of PM measurement available for the study area. We obtained daily data on 24-hour mean PM₁₀ and PM_{2.5} concentrations over the period 2019–2022 from all the available monitoring sites in Bari and Taranto.

Notably, the Bitonto EN01 station (within area of Bari) is managed by the private sector. ARPA has responsibility to provide all the validated PM data required for the project. Hourly acquisition data for air quality parameters are also available on the webpage https://www.arpa.puglia.it/pagina3139_dati-storici-qualit-dellaria.html. In general, the stations

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

provide only daily average data for PM₁₀ and PM_{2.5}. In Bari there are three stations that collect data at an hourly or bi-hourly frequency: Carbonara, Modugno EN02, and Modugno EN04. Analyzing the time series of PM measurements from the three stations in Bari with hourly or bi-hourly data in 2021 revealed that the values best approximating the daily averages fall within the time window of interest for satellite observations: 8 a.m., 11 a.m., and 10 a.m. for Bari Carbonara, Modugno EN02, and Modugno EN04, Figure 5-Time series of the hours in which PM measurements best approximate the daily averages of PM₁₀ in 2019. Bari Carbonara station at top, ModugnoEN02 in the middle and Modugno EN04 at bottom.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

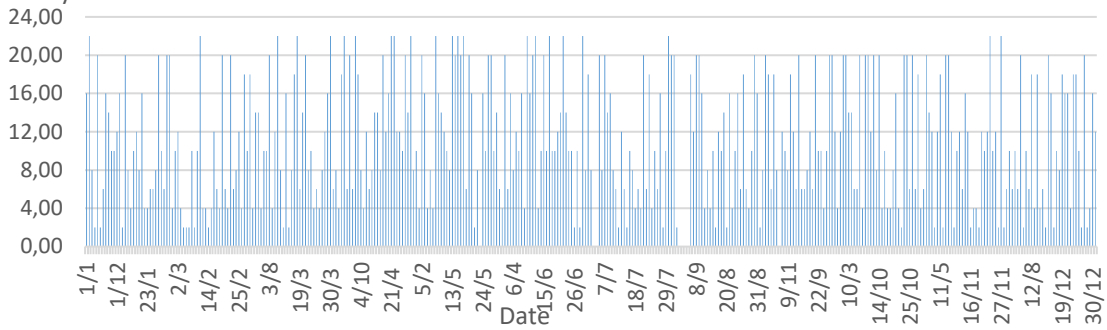
Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

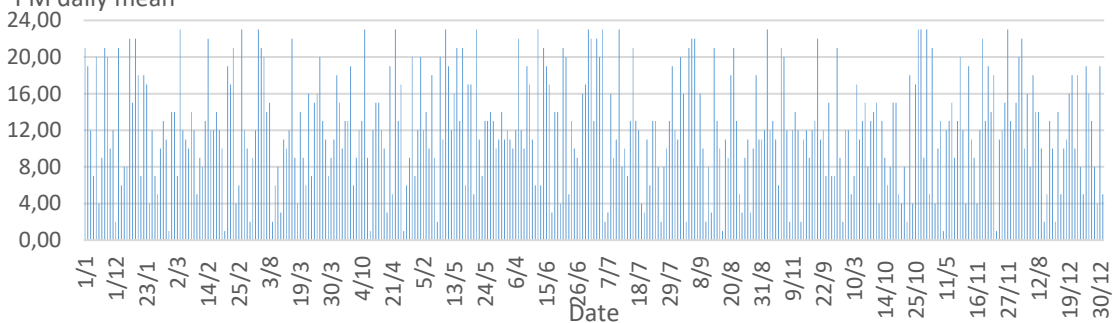
The hour in which
PM measurements
approximate the
PM daily mean

Bari Carbonara station



The hour in which
PM measurements
approximate the
PM daily mean

Modugno EN02



The hour in which
PM measurements
approximate the
PM daily mean

Modugno EN04

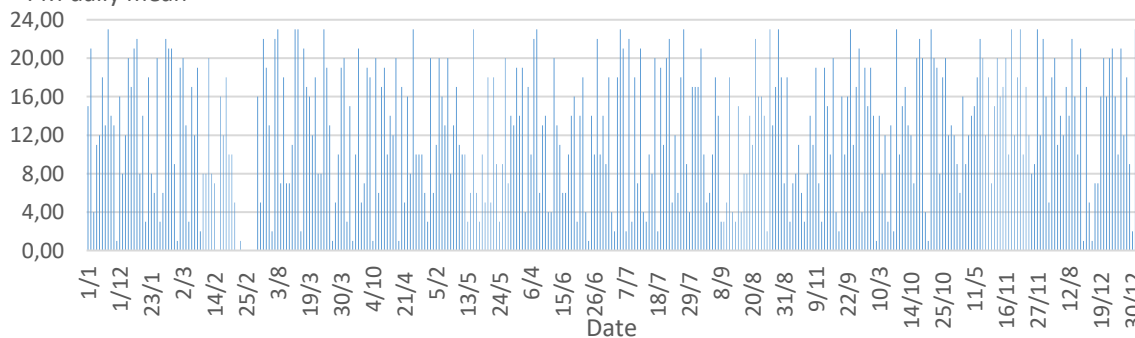


Figure 5-Time series of the hours in which PM measurements best approximate the daily averages of PM10 in 2019. Bari Carbonara station at top, ModugnoEN02 in the middle and Modugno EN04 at bottom



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

At the only station located in an urban area, Bari Carbonara, the best agreement between the daily average and the PM measured in the morning can be observed

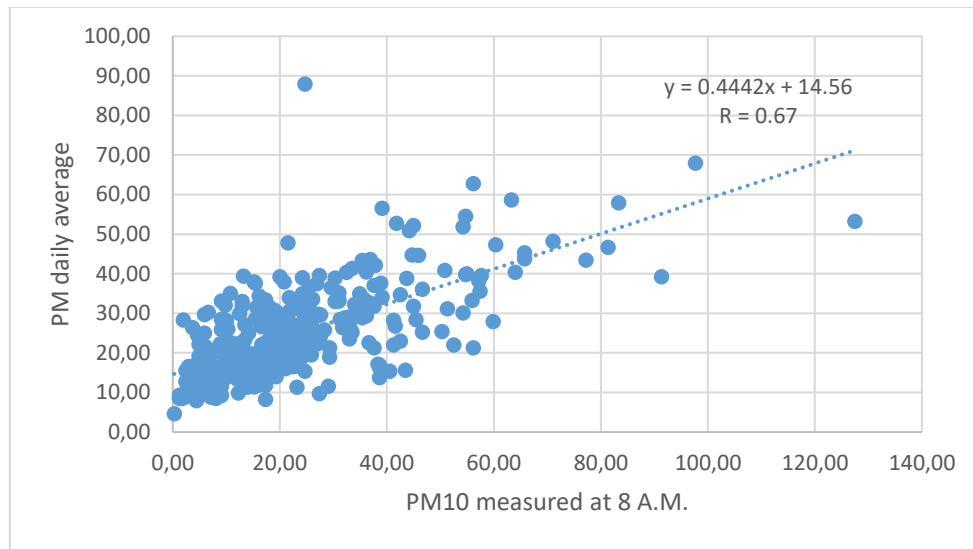


Figure 6-Comparison between PM10 measured at 8 a.m. and daily average value in Bari Carbonara station.

For the APEMAIA project, we work exclusively with the daily average measurements provided by ARPA after statistical validation.

In the following figures, for both PM₁₀ and PM_{2.5} we present frequency histograms of these daily averages and box plots indicating the variance of values recorded at each station and across different years for both Taranto and Bari.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

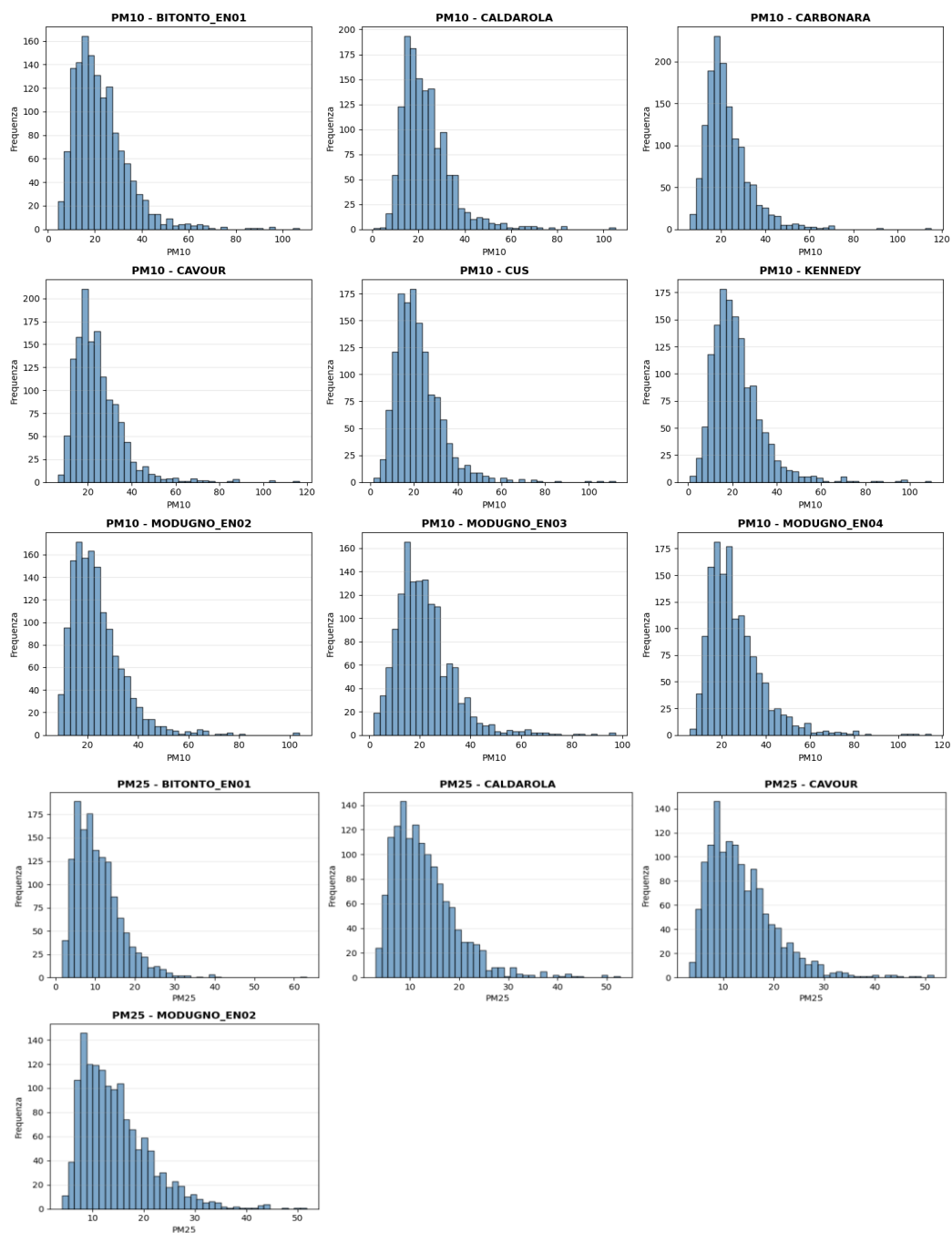


Figure 7- PM_{10} and $PM_{2.5}$ measurements divided by station in Bari



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

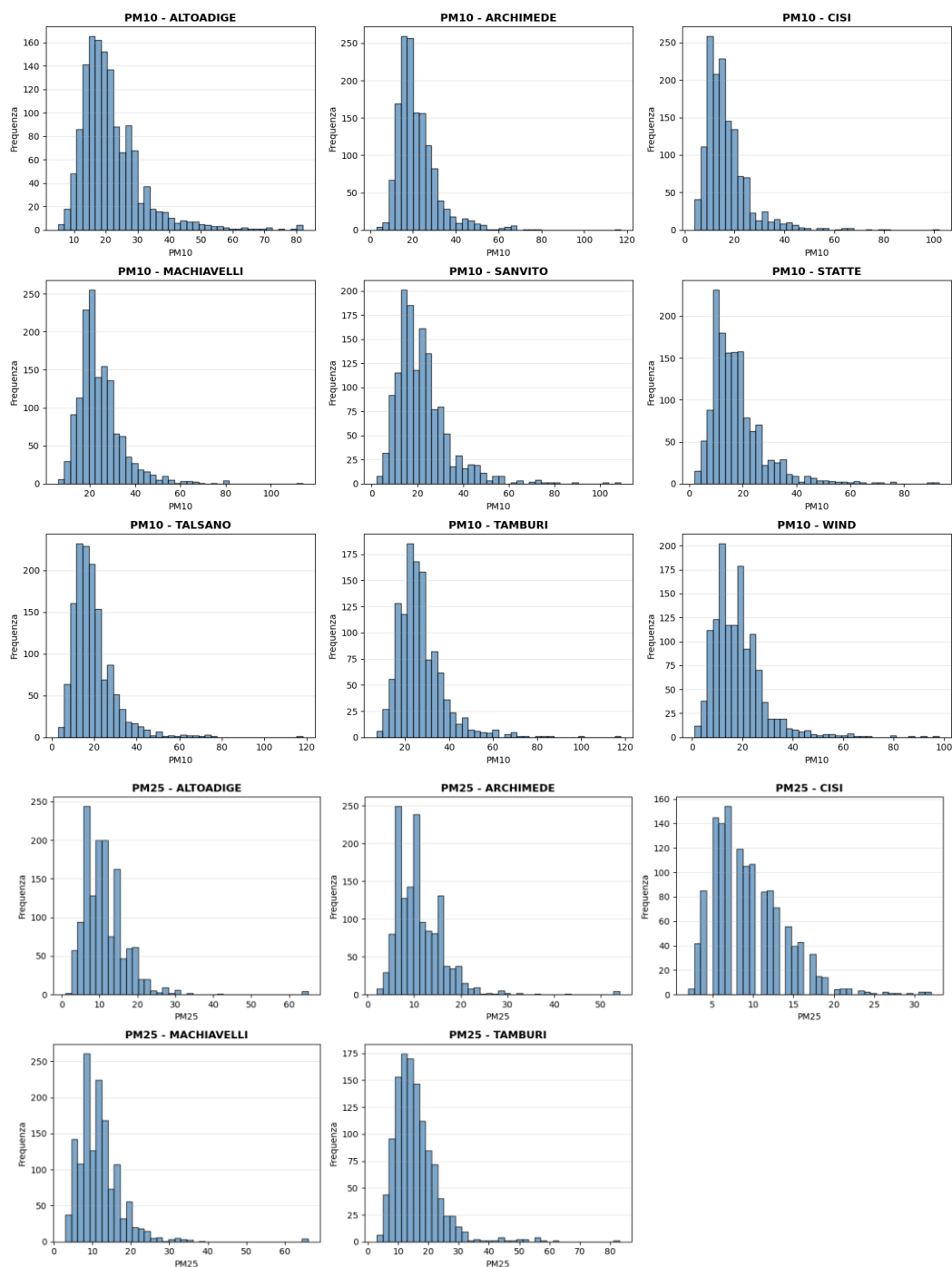


Figure 8- PM_{10} and $PM_{2.5}$ measurements divided by station in Taranto



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

The boxplots display the annual PM₁₀ concentration distributions across monitoring stations in Taranto and Bari. Each box represents the interquartile range (IQR), with the median shown as a horizontal line and outliers plotted individually. Notably, the Tamburi station (Taranto) and the Modugno stations (Bari) exhibit persistently elevated PM₁₀ values compared to other monitoring sites, reflecting local emission sources or unfavourable atmospheric dispersion conditions at these locations.

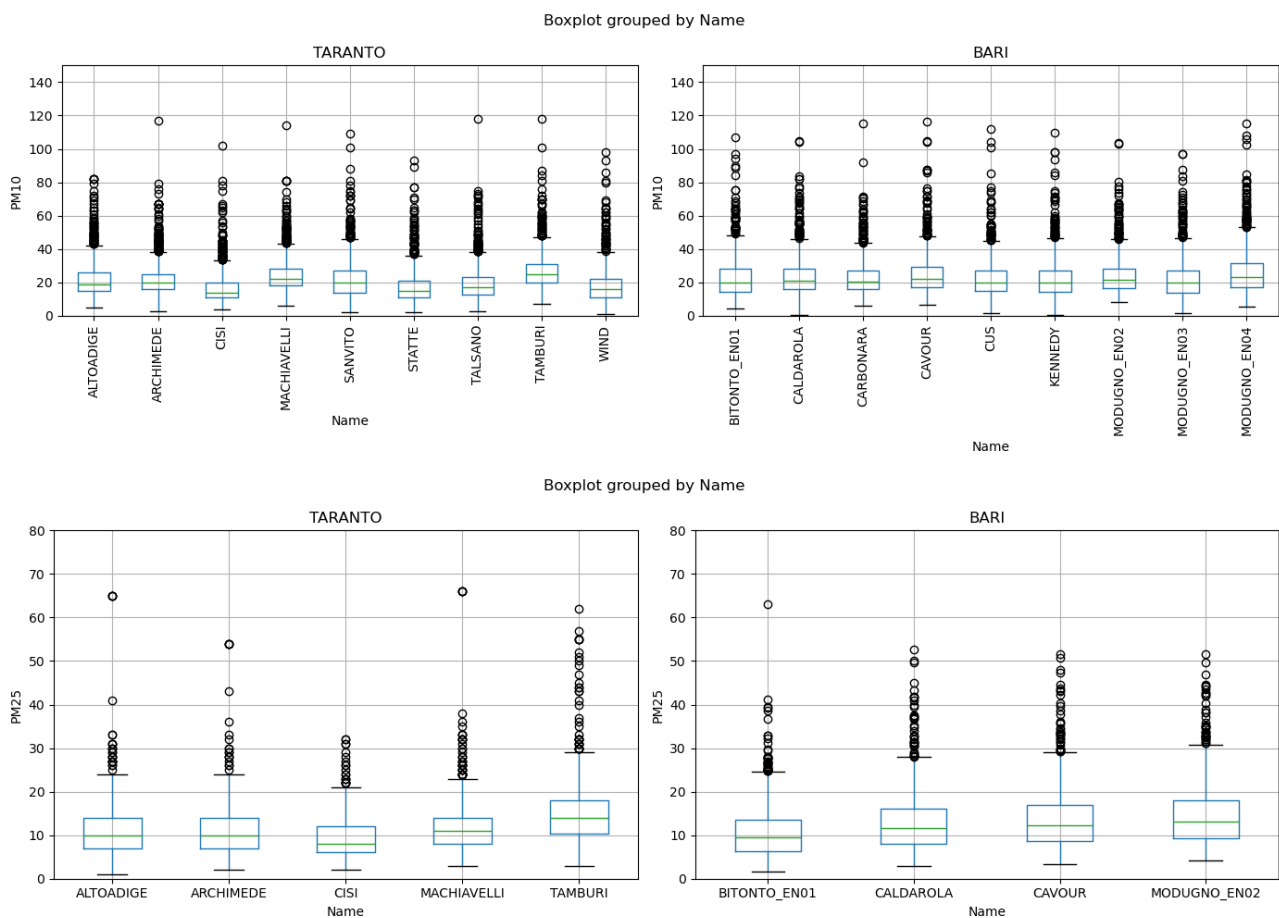


Figure 9-PM₁₀ and PM_{2.5} boxplot grouped by station



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

3 Regional scale approach

3.1 Training data

As already described in section 3.1, we will use ARPA control units as label to fit our machine Learning model. Specifically, in Figure 10 is shown the spatial distribution of these station (53 for PM10, 27 for PM2.5). The measures are provided into $\mu\text{g}/\text{m}^3$. We collected data from 2019 to 2022 and we made a daily average.

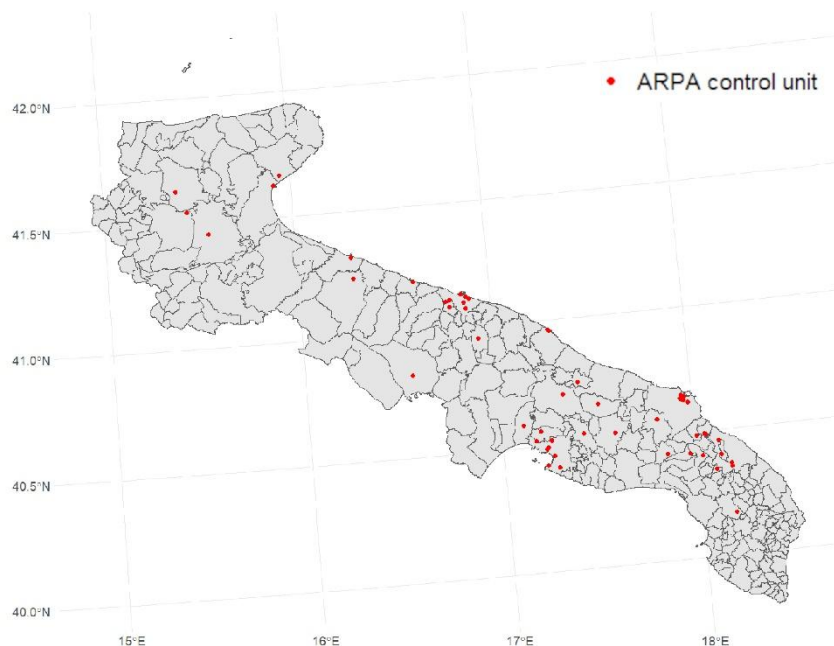


Figure 10-Spatial distribution of ARPA Puglia station.

3.2 Database of predictors available at regional scale

To be able to reconstruct the complex patterns of PMs, we employed different types of predictors, which are listed in table 3. They are divided in Meteorological data, satellite data, geographical data and temporal data.

Name	Product	Spatial resolution	Temporal Resolution
U_wind_component	WRF	4 km	Hourly
V_wind_component	WRF	4 km	Hourly
Wind_speed	Computed	4 km	Hourly
Emissivity	WRF	4 km	Hourly



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

Rain_convective	WRF	4 km	Hourly
Rain_not_convective	WRF	4 km	Hourly
Vertical_wind	WRF	4 km	Hourly
Specific_humidity_2m	WRF	4 km	Hourly
Surface_temperature	WRF	4 km	Hourly
Cloud_fraction	WRF	4 km	Hourly
Surface_pressure	WRF	4 km	Hourly
AOD	Computed (section 4)	300 m	Daily
NO2_sat	Sentinel 5-P	5.5 km x 3.5 km	Daily
O3_sat	Sentinel 5-P	5.5 km x 3.5 km	Daily
Sin_month	Computed		Daily
Cos_month	Computed		Daily
Sin_week	Computed		Daily
Cos_week	Computed		Daily
Population	WorldPop	100 m	
Roads_tot	OpenStreetMap		
Buildings_tot	OpenStreetMap		
Corine	CORINE	100 m	
Continuous_urban_fabric	CORINE	100 m	
Discontinuous_urban_fabric	CORINE	100 m	
Cell_id	Computed	300 m	
Date	Acquired		Daily

Table 3-List of all variables collected, with source, spatial and temporal resolution.

In our analysis, we created a square grid over the Apulia region with a 300-meter resolution. The reference system used is WGS 32. All the collected satellite variables were projected into our reference system, and the values were extracted within the grid cells using various aggregation functions. The AOD data used was computed, as described in section 4.

We used the output from the Weather Research Forecasting (WRF) model in the Puglia region at a scale of 4 km. This dataset was provided by ARPA Puglia. Specifically, we aggregated the hourly data into daily averages to match the temporal scale of the satellite data. Additionally, for some variables, such as the vertical component of wind, we averaged along the z-axis. This was done because these variables across different height layers are highly correlated, making them redundant for a Machine learning model. Finally, a spatial resampling to 300 meters was performed using cubic interpolation.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Subsequently, these Raster-format data were projected into the coordinate reference system (CRS) of our grid, and the values were extracted accordingly.

We used the tropospheric NO₂ and tropospheric O₃ variables provided by the Sentinel-5P dataset. The images were downloaded from the Google Earth Engine platform, with a daily temporal resolution. These variables represent the surface density of these pollutants (expressed in mol/m²), from ground level up to the troposphere. Since there were missing values due to orbit, we employed an Inverse Distance Weighting (IDW) to fill these gaps.

From the OSM site we downloaded a point shapefile containing the locations of all buildings and a line shapefile containing all roads, both pertaining to the Apulian region. From the first file, we calculated the total number of buildings within each grid cell, while from the second file, we computed the total length of roads intersecting each cell.

For greater completeness of the model, we provided information regarding the urban effect. Specifically, we used the dataset provided by WorldPop. The dataset, in GeoTIFF format, contains an estimate of the population across the Italian territory, with a 100-meter grid resolution. We extracted values using our reference grid, applying the sum function as the aggregation method at 300 meters.

Furthermore, we used the CORINE dataset in two ways: we reduced the number of land cover classes from 45 to 5 (grouped into the following macro-categories: 1-Artificial surfaces, 2-Agricultural areas, 3-Forest, 4-Wetlands, and 5-Water bodies). Additionally, we calculated the percentage of site occupation for industrial areas per pixel, to detect patterns related to pollutant emissions.

In order to give temporal information about seasonality and possible week-end effect we encoded the date into sin and cosine of a circle obtained projecting the months and days.

Cell id and Date are features used only to merge and aggregate data.

3.3 Method

3.3.1 Inverse Distance Weighting

Spatial interpolation is a statistical method used to estimate the values of a variable at unsampled locations based on the values observed at known points \cite{lam1983spatial}. There are several techniques, such as Nearest Neighbours, Inverse Distance Weighting (IDW), Kriging, and spline interpolation, which are commonly applied with different assumptions about the spatial correlation between data points. In this analysis, we filled the missing values of the input data by making an



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

average of the Nearest Neighbours (NN) and Inverse Distance Weighting (IDW) interpolations. This is essential for the Sentinel-5P and AOD data, which contain large gaps due to orbital constraints or cloud cover. An example of this procedure can be seen in Figure 11.

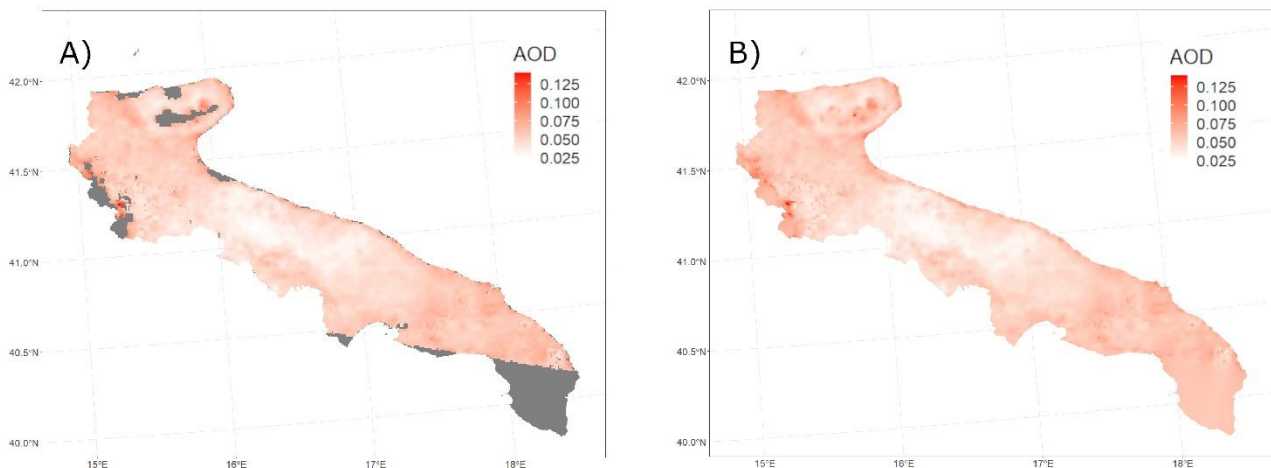


Figure 11-Left panel (panel A) AOD values extracted on the grid (the missing values are attributed to cloud covering and orbit of satellite), on the right panel (panel B) the same image in which NA values are replaced using the Inverse Distance Interpolation meth

3.3.2 Spatial Resample of WRF Data

In our analysis, we needed to convert the WRF data structure into our prediction grid, enhancing the resolution of the model to a 300-meter resolution. The first step consisted of a spatial interpolation of the original raster image. The most common techniques used in the interpolation of the discrete image are the Nearest Neighbour Interpolation, the Bi-Linear Interpolation, Cubic B-Splines and the Bi-Cubic Interpolation. Among these methods we chose to use the Bi-Cubic Interpolation method, which is known to produce smoother results, especially compared to the Bi-Linear Interpolation.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

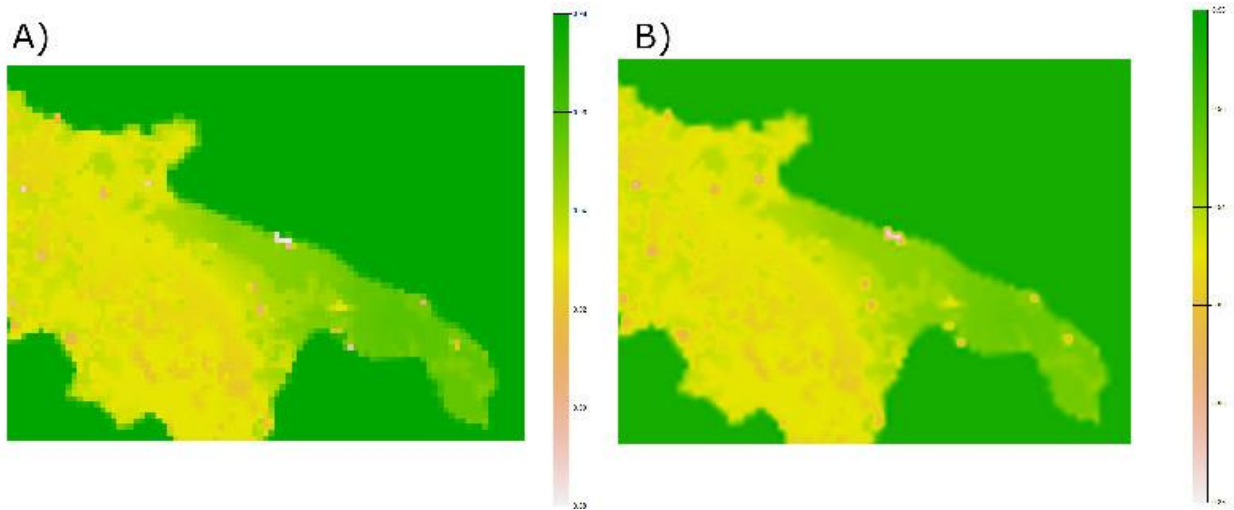


Figure 12-Left panel (A) Surface Emissivity predicted by the WRF model (4 km resolution), on the right panel (B) the resampled image to 300 m with bi-cubic interpolation method.

3.3.3 Machine Learning model

After the preprocessing and harmonization of all predictors on the common 300 m grid, an XGBoost regression model was developed to estimate daily ground-level PM₁₀ and PM_{2.5} concentrations. Ground measurements from ARPA Puglia monitoring stations were used as target values, while satellite-derived, meteorological, land-use, anthropogenic and temporal variables were used as predictors.

XGBoost is a gradient-boosting decision tree algorithm in which successive trees are trained to progressively reduce the residual error of the ensemble. This approach is particularly suitable for the integration of heterogeneous environmental predictors, as it can capture non-linear relationships and interactions among variables without requiring an explicit functional form. Model performance was evaluated using repeated 5-fold cross-validation and, additionally, a Leave-One-Year-Out validation strategy to assess temporal generalization.

3.3.4 Explainable Artificial Intelligence

To improve model interpretability, SHAP (SHapley Additive exPlanations) was applied to the trained XGBoost models. SHAP values quantify the contribution of each predictor to individual model predictions, indicating both the magnitude and the direction of its effect on the estimated pollutant

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

concentration. Positive SHAP values indicate that a predictor increases the model output, whereas negative values indicate a reduction.

The SHAP analysis was used not only to rank predictor relevance, but also to verify whether the relationships learned by the model were consistent with known atmospheric processes, such as aerosol loading, dispersion by wind, wet scavenging and the influence of urban and industrial land-use characteristics.

3.4 Results

After the preprocessing and data integration steps, XGBoost models were trained to estimate daily ground-level PM₁₀ and PM_{2.5} concentrations over the entire Apulia region at a spatial resolution of 300 m. The models integrate satellite-derived information, WRF meteorological fields, land-use and anthropogenic variables, temporal predictors, and ARPA Puglia ground measurements used as reference data.

Model performance was assessed through repeated 5-fold cross-validation. For PM₁₀, the XGBoost model achieved an R^2 of 0.64, an RMSE of 6.26 $\mu\text{g}/\text{m}^3$ and an MAE of 4.45 $\mu\text{g}/\text{m}^3$. For PM_{2.5}, the corresponding values were $R^2 = 0.67$, RMSE = 3.74 $\mu\text{g}/\text{m}^3$ and MAE = 2.66 $\mu\text{g}/\text{m}^3$. These results indicate that the non-linear integration of satellite, meteorological, geographical and temporal information can reproduce a substantial fraction of the observed daily variability of particulate matter concentrations.

In addition to repeated random cross-validation, a Leave-One-Year-Out (LOYO) strategy was applied to assess the temporal transferability of the models. In this more demanding validation setting, an entire year was excluded from model training and subsequently used for testing. At daily resolution, R^2 decreased to 0.24 for PM₁₀ and 0.30 for PM_{2.5}, reflecting the difficulty of reproducing interannual variability and episodic particulate matter events that are not represented in the training period.

Once trained, the models were applied to the complete regional grid to generate spatially continuous daily PM concentration fields. The 300 m output grid makes it possible to estimate particulate matter concentrations beyond the discrete locations covered by the monitoring network and to represent local-scale spatial gradients associated with urban, industrial and coastal environments. The daily predictions can subsequently be aggregated to obtain monthly, seasonal or annual exposure maps.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

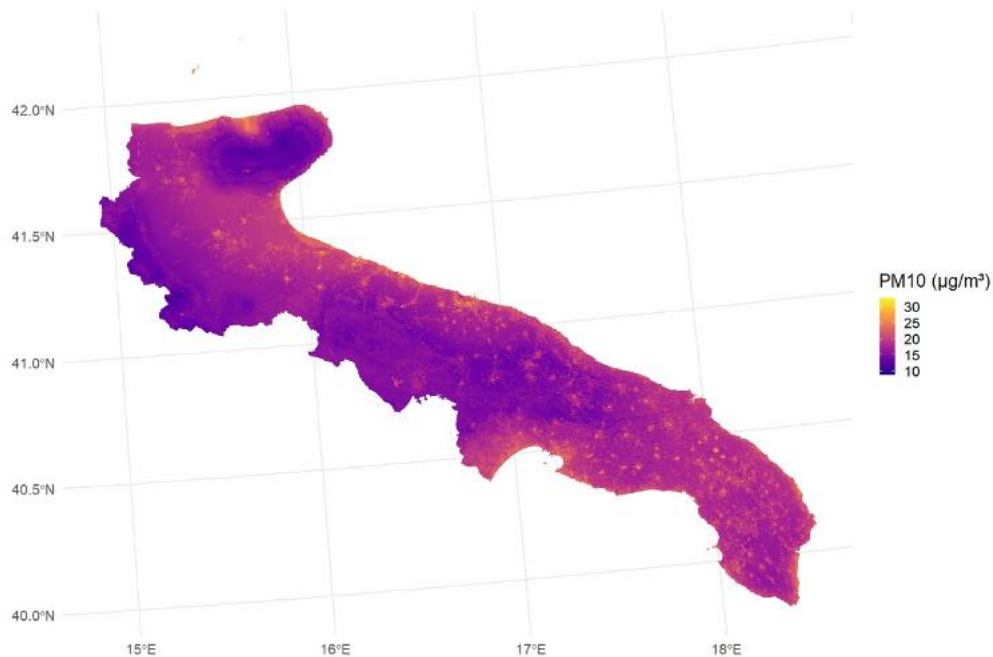


Figure 13-Mean PM10 concentrations predicted by the XGBoost model over the Apulia region at 300 m spatial resolution. Daily predictions were averaged over the 2019–2022 study period. Concentrations are expressed in $\mu\text{g}/\text{m}^3$.

The predicted PM10 field shows a heterogeneous spatial distribution across the region, with higher concentrations mainly associated with coastal, urban and industrial areas. The high-resolution representation allows broad regional patterns and finer local gradients to be visualized within the same modelling framework.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

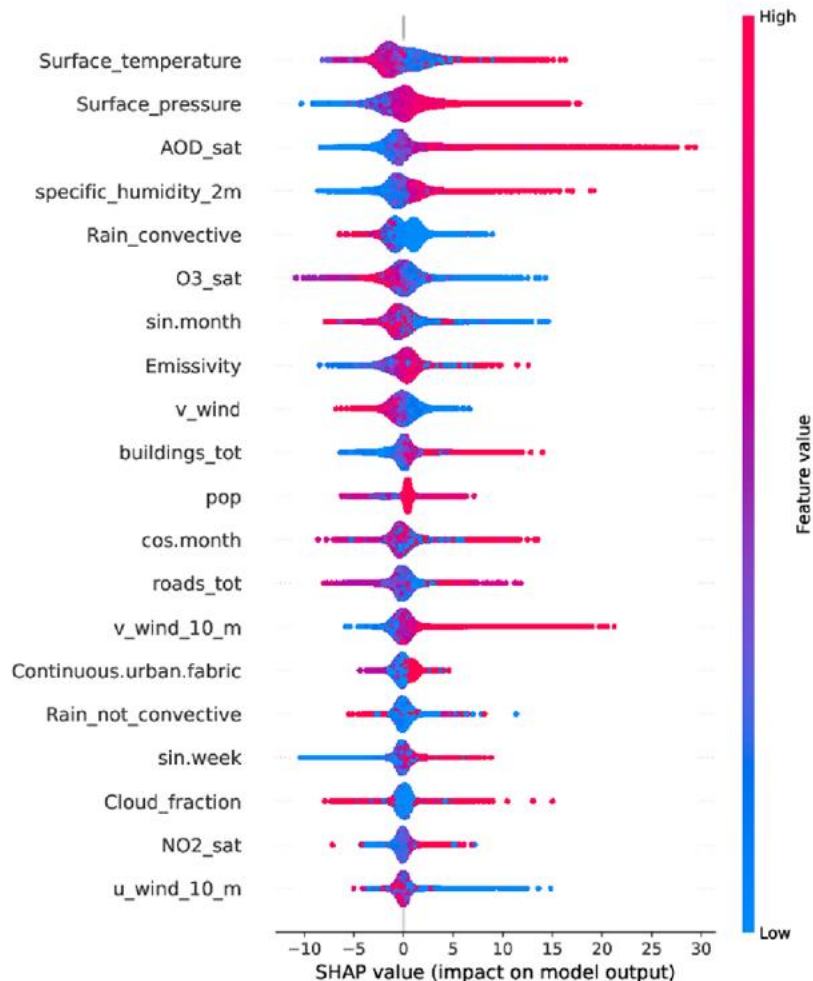


Figure 14-SHAP summary plot for the XGBoost PM10 model. SHAP values quantify the contribution of each predictor to individual PM10 estimates. Positive values increase the predicted concentration, whereas negative values decrease it. Feature colours represent the

The SHAP analysis provides an interpretation of the relationships learned by the model. Aerosol Optical Depth (AOD) has a strong positive contribution to PM10 predictions, supporting its role as an indicator of atmospheric aerosol loading. Meteorological variables also behave consistently with known atmospheric processes: higher wind speed and precipitation generally decrease predicted particulate matter concentrations through enhanced dispersion and wet scavenging. Urban and industrial land-use variables provide additional information on local emission patterns.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

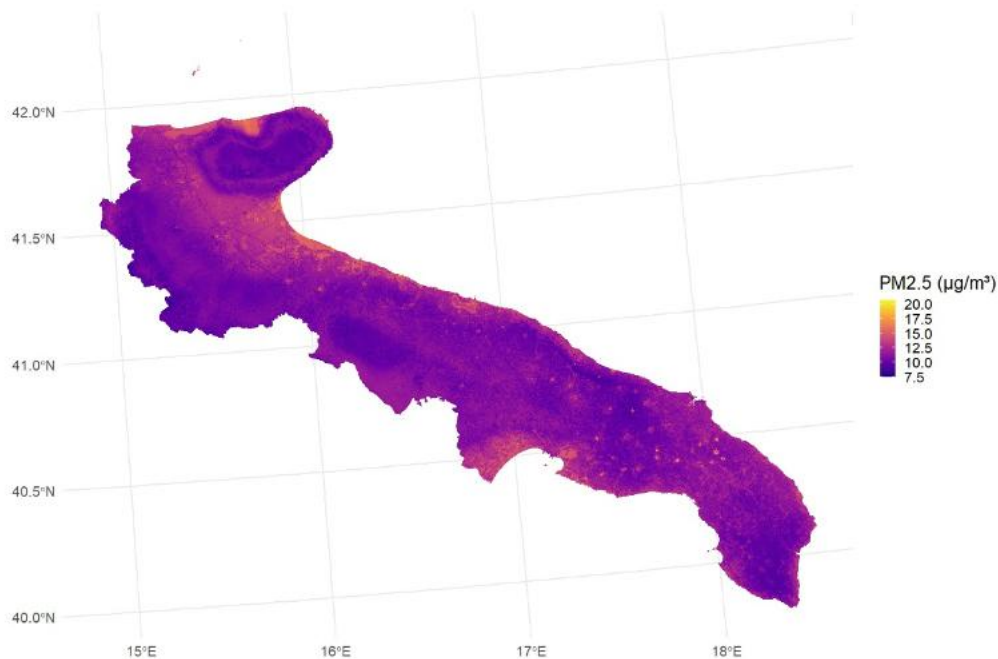


Figure 15-Mean PM_{2.5} concentrations predicted by the XGBoost model over the Apulia region at 300 m spatial resolution. Daily predictions were averaged over the 2019–2022 study period. Concentrations are expressed in $\mu\text{g}/\text{m}^3$.

PM_{2.5} predictions also show spatial variability across the Apulian territory, reflecting the combined effect of regional atmospheric conditions and local anthropogenic characteristics. The resulting concentration fields provide continuous information in areas not directly covered by ground monitoring stations and can therefore support regional-scale exposure assessment.

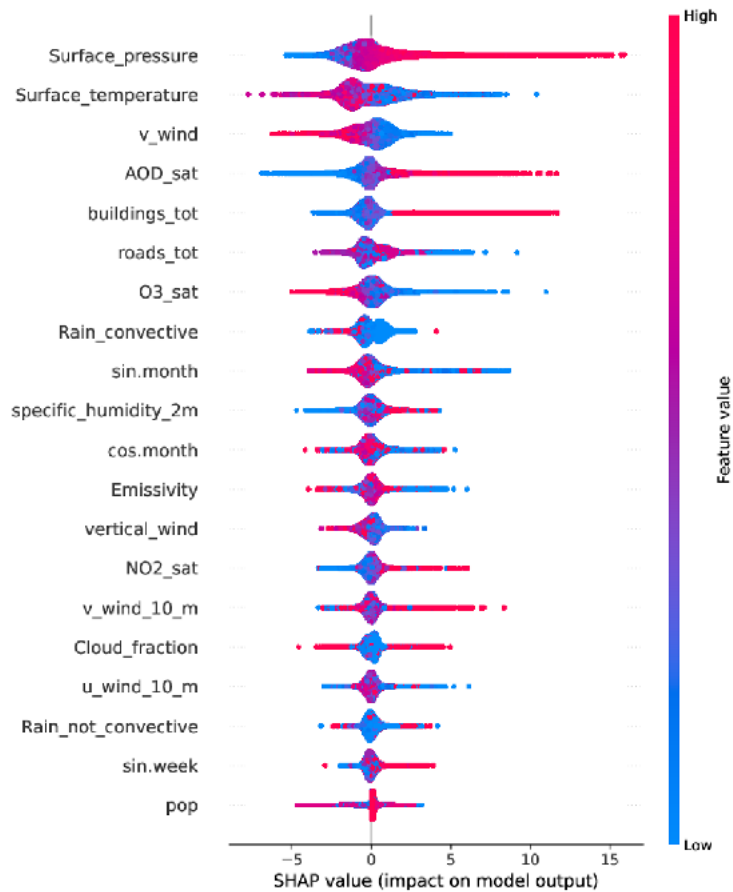


Figure 16-SHAP summary plot for the XGBoost PM2.5 model. The plot shows the magnitude and direction of the contribution of each predictor to the estimated PM2.5 concentration. Positive SHAP values correspond to an increase in the model output, whereas negative values correspond to a decrease.

Similarly to PM10, the PM2.5 SHAP analysis highlights the contribution of satellite-derived aerosol information and meteorological conditions. Higher AOD values generally increase predicted PM2.5 concentrations, while stronger winds and precipitation tend to reduce them. The combination of these dynamic variables with land-use and anthropogenic information allows the model to account for both regional-scale variability and local spatial differences.

Overall, the regional-scale approach demonstrates the feasibility of producing spatially continuous daily PM10 and PM2.5 concentration fields over Apulia at 300 m resolution. The combination of predictive performance, high-resolution mapping and SHAP-based interpretability makes the

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

framework suitable for environmental exposure assessment and for identifying spatial patterns that cannot be resolved by the monitoring network alone.

Nevertheless, the reduction in performance observed under temporal hold-out validation indicates that short-term particulate matter variability and episodic events remain more difficult to reproduce than persistent spatial patterns. Longer observational records and the integration of additional process-related information may further improve temporal generalization and the representation of extreme pollution events.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

4 Intra-urban scale approach

4.1 Database of predictors

4.1.1 Aerosol Optical Depth (AOD) data

Aerosol Optical Depth (AOD) is a satellite parameter measuring the degree to which suspended particles affect the transmission of light by absorption or scattering. Therefore, it is an indirect measure of the particles present in the column of air on a given time. Several AOD products were released by space agencies across word.

PLATFORM	PRODUCT	PROC. LEVEL	SPATIAL RESOLUTION	TEMPORAL RESOLUTION	ACQUISITION TIME	BAND
Sentinel3	S3A_SY_2_SYN, S3B_SY_2_SYN	Level 2	300m	9 images per week	8 - 10 UTC	AOD @ 550 nm
Modis	MCD19A2	Level2	1 km	daily	8-11 UTC	AOD @470 nm & @550 nm
Landsat8	LC08_L2SP	Level2	30 m	16 days	8-10 UTC	SR_QA_AEROSOL
Sentinel2	S2A_MSIL2A	Level2	60 m ricampionat a a 20 m	3 - 5 days	8-10 UTC	Quality_AOT
VIIRS	AERDB_L2_VIIRS_SNPP	Level2	6 km	2 per day	10-13 UTC	AOD @550 nm
Sentinel2 MAJA	SENTINEL2A	Level2	10 m e 20 m	3 - 5 days	8-10 UTC	AOD @550 nm
NASA's TERRA	MISR	Level2	10 km - < 50 km	daily- <weekly	10:30 local time	AOD product from 4 spectral bands: 446, 558, 672, 866 nm

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Sentinel5/5P TROPOMI	L2__AER_AI, L2__AER_LH	Level2	3.5 x 5.5 Km	varying	Varying (multiple times per day)	Aerosol UV absorption index, layer height, optical depth (TBC)
-------------------------	---------------------------	--------	--------------	---------	----------------------------------	---

After a careful evaluation of AOD product benchmark, we implemented 2 case testes, alternatively, using:

- AOD downloaded dataset at lowest spatial resolution: SEN3SYN Aerosol Land Product.
- “AOD improved” dataset modelled according to the method developed in WP2102 and described in [AD 4].

Both datasets are at the same 300 m spatial resolutions.

4.1.2 Meteorological parameters

Meteorological parameters (daily mean air temperature, atmospheric pressure, precipitations, relative humidity, wind speed) and planetary boundary layer height were retrieved by the local running of WRF model. In addition, ECOSTRESS Land Surface Temperature (LST) was considered, it is the skin temperature of the Earth's surface measured from the International Space Station at a high spatial resolution of 70 meters. From ARPA MODEL we retrieved the dust concentrations, a variable strictly connected to PM₁₀.

Data from ERA5 Reanalysis model provided by Copernicus Climate Data Service are indeed used for an exploratory analysis on the feature importance scoring.

Name	Code	Product	Spatial resolution	Temporal Resolution
Wind U component	U_daily_mean	WRF	4 km	Hourly
Wind V component	V_daily_mean	WRF	4 km	Hourly
Wind W component	W_daily_mean	WRF	4 km	Hourly
Wind intensity	MODV	Computed	4 km	Hourly
Accumulated Convective Rain	RAINC_daily	WRF	4 km	Hourly

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Accumulated Not Convective Rain	RAINNC_daily	WRF	4 km	Hourly
Planetary Boundary Height Layer	PBLH_daily_mean	WRF	4 km	Hourly
Relative Humidity	RH_daily_mean	WRF	4 km	Hourly
Temperature at 2m	T2_daily_mean	WRF	4 km	Hourly
Surface Pressure	P_daily_mean	WRF	4 km	Hourly
Total Radiation	TOTRAD_daily_mean	WRF	4 km	Hourly
Net Radiation	NETRAD_daily_mean	WRF	4 km	Hourly
Land Surface Temperature	LST	ECOSTRESS	70m	Daily
Dust Concentrations	DUST_CONC	Model	1 km	Daily

Table 4-List of all meteorological variables collected, with source, spatial and temporal resolution.

4.1.3 Vegetation Index

As first instance, a set of vegetation indices, such as NDVI, MSAVI, NBR, NBR2, NDWI, and BSI, were computed from Sentinel 2 A/B time series extracted from google earth only on the area where a ground monitoring station was available. Observed the partial redundancy across the indices we selected NDVI to be extracted from a full extent stack of Sentinel 2 A/B image. The stack was downloaded from Wekeo api. In both cases the interpolation used was a Gaussian Process Regression shaped as a sum of radial basis function kernel with scale parameter bound between 12 and 365 days and a white noise kernel bound between 0.001 and 10. The approach was implemented in python using a sklearn module spatialized with rioxtarray one.

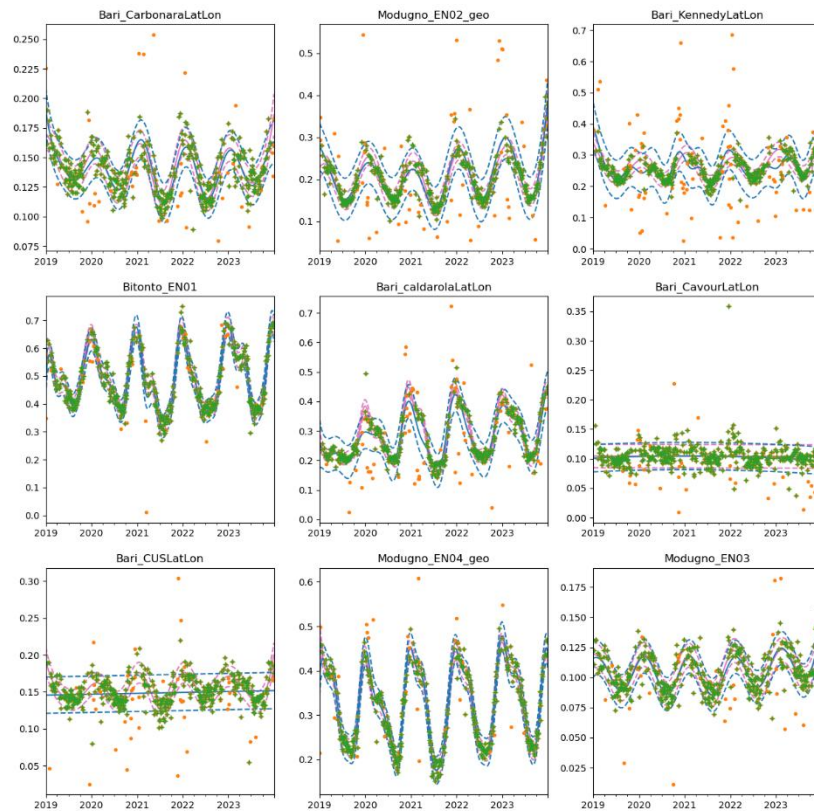


Figure 17-Interpolated NDVI on the Bari stations

4.1.4 Other spatiotemporal data

We computed several spatial predictors at the grid cell level, e.g. variables spatially changing from cell to cell but assumed to be fixed over time which serves to define different microclimatic zones within the urban fabric. To integrate the characteristics with information on the environmental and anthropogenic context, a set of auxiliary urban variables was considered as potential predictors for the PM estimation model.

A Spatial predictor related to demographic density, namely POP grid, has been included. It represents the average population density. This data comes from the permanent census survey of the National Institute of Statistics (ISTAT) ISTAT (2021), 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 plu gin Aquilino et al. (2021) developed within the European Horizon 2020 projects SMURBS and e-SHAPE, for producing the population density map updated to 2021 with a resolution of 100×100 meters. For traffic indicators we used indirect information coming from

the road network downloaded from Open StreetMap source. In particular, the features calculated are the total length of road segments and the number of crossroad points in each element of a regular 300 m grid.

Variable code	Description
HHmean	Mean Building Heigh
POP_grid	Mean Population Density
Length	Mean Streets Length
Count	Number of street nodes
SeaHubDist	Minimum Sea Distance
SVF	Sky View Factor
BSFsum	Building Surface Printing

Table 5- Urban variables used in the intra-urban model

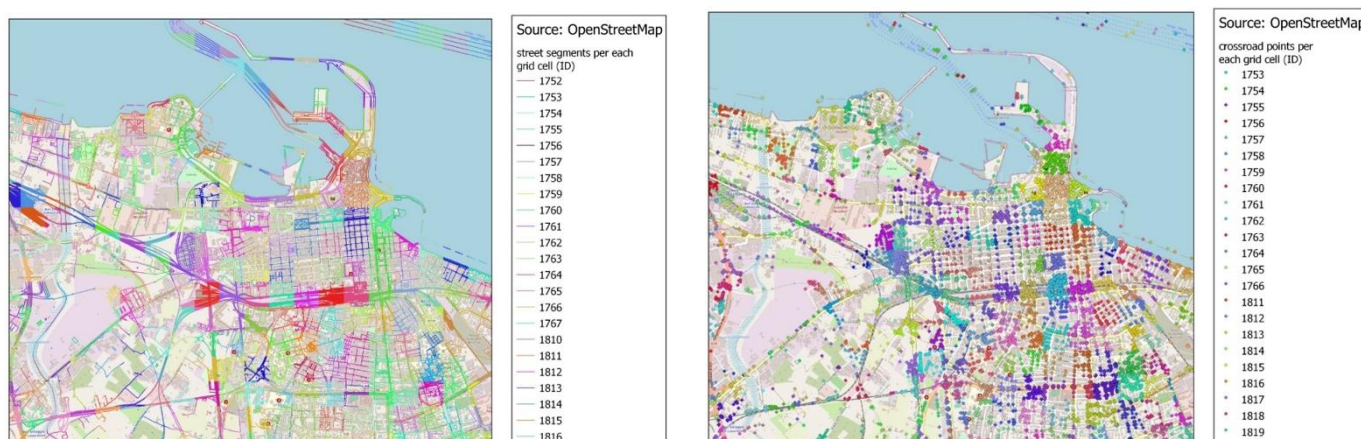


Figure 18-Traffic indicators from OpenStreetmap



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

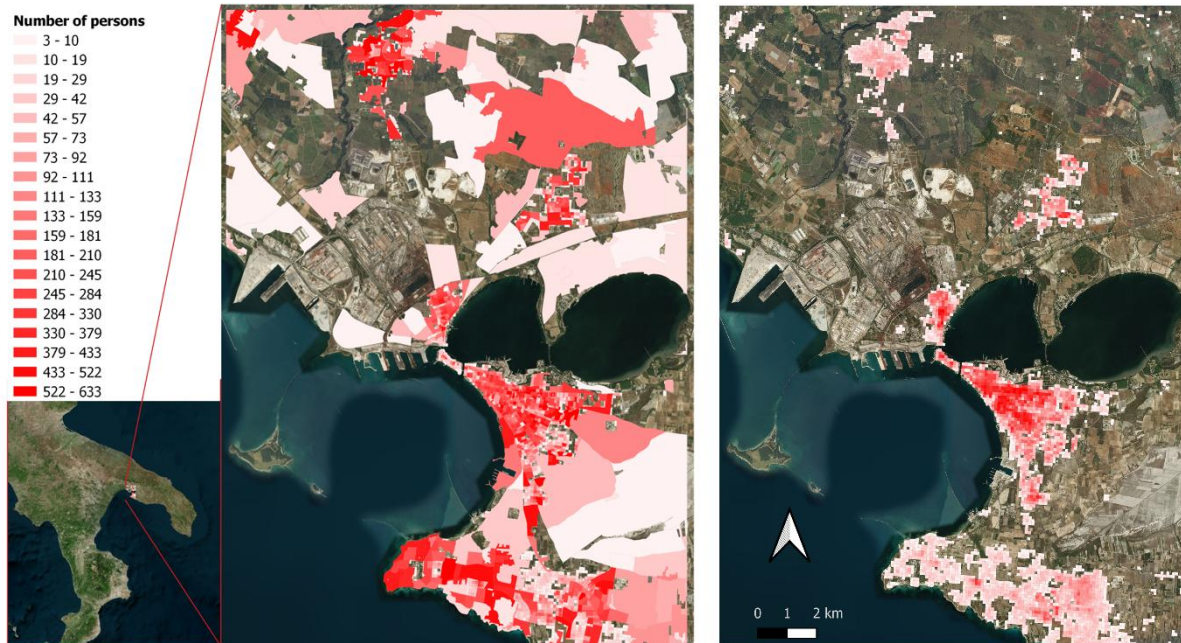


Figure 19-Taranto, a) population per census zone, b) population density map 100x100m

Land cover mapping, at a spatial resolution of 10 m, was obtained using a GEOBIA (Geographic Object-Based Image Analysis) approach applied to multitemporal Sentinel-2 imagery acquired during 2022. The methodology followed a process structured in two distinct phases (C. Tarantino et al., "Combination of GEOBIA and data-driven approach for grassland habitat mapping in the Alta Murgia National Park", *Science of Remote Sensing*, 2025) first, segmentation was performed using the multi-resolution algorithm implemented in the eCognition 9.0 environment. Subsequently, classification was performed using the Random Forest algorithm. The resulting land cover classes, listed in detail in the Table, were then aggregated into three main macro-categories (VEGETATED AREAS, ARTIFICIAL SURFACES, WATER) to simplify their use and because they were considered sufficiently representative of the territorial context.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

Variable code	LC Description
VEGETATED AREAS	Natural Vegetation/Trees or Shrubs Natural Vegetation/Trees or Shrubs Evergreen Natural Vegetation/Herbaceous Graminoid Cultivated Land
ARTIFICIAL SURFACES	Built up urban areas, Roads, Shadows
WATER	Water

Table 6-Land Cover Variables used in the intra-urban model

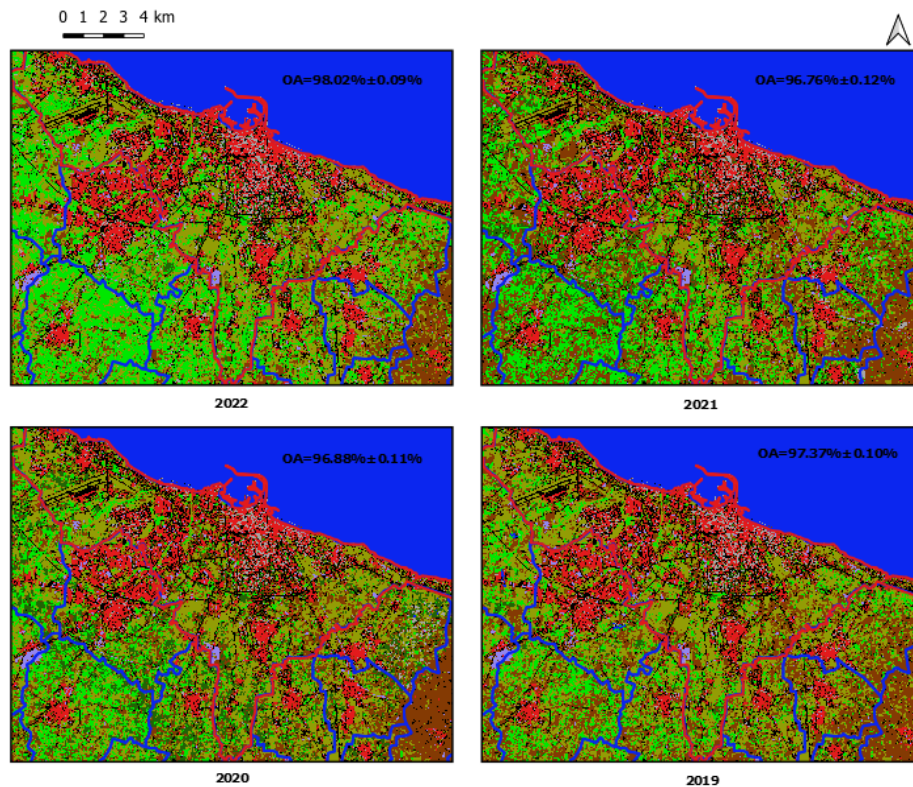


Figure 20-Bari, LC mapping

The urban land use classes were derived from the Copernicus Urban Atlas (UA) Copernicus Land Monitoring Service (2023) service available for 2018. The UA classes were aggregated based on building use typology, following the associations shown in Table, 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.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

Variable Name	Building Use
Buclass_1	Percentage cover of residential area
Buclass_2	Percentage cover of roads
Buclass_3	Percentage cover of other classes
Buclass_4	Percentage cover of rural areas
Buclass_5	Percentage cover of Industrial/Commercial areas

Table 7-Urban Land Use Variables used in the intra-urban model

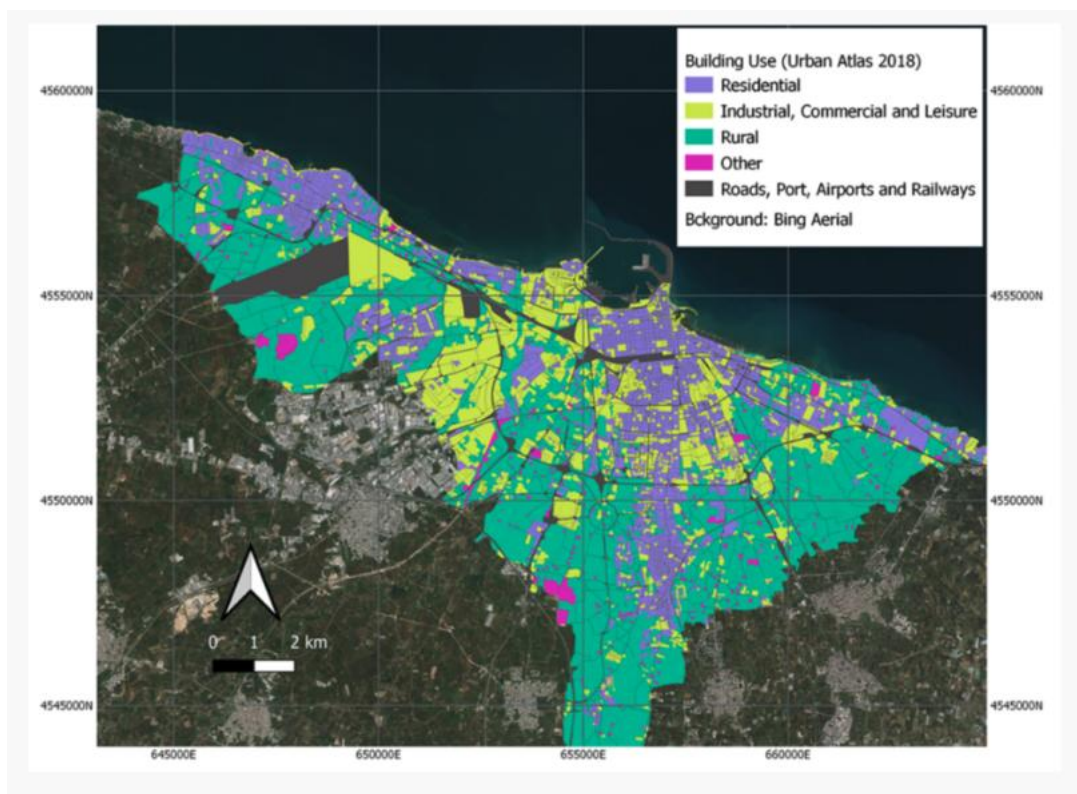


Figure 21-Building use classification over Bari area

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

4.1.5 Spatiotemporal aggregation at grid level

For PM10 estimation, we reported a comprehensive list of type of predictors used to implement ML models, including meteorological variables, vegetation indices, population density, building morphology, traffic indicators and land cover data. This is a heterogeneous dataset and our initial efforts focused on intensively downloading and projecting the various products onto the reference grid at the same resolution as the Sentinel 3 synergistic product: a 300 m x 300 m grid.

To ensure compatibility across all predictor variables, a spatial harmonization procedure was implemented to align all datasets to a common reference grid with 300 m × 300 m resolution. The input datasets were characterized by different native spatial resolutions, ranging from 10 m for some land cover products to several kilometers for meteorological variables.

For variables with spatial resolution coarser than 300 m, bicubic interpolation was employed to generate values at the target resolution. This method uses a weighted average of the 16 nearest pixel values in the original grid, applying a bicubic polynomial function that ensures smooth transitions and preserves local spatial patterns. Conversely, for variables with finer resolution than 300 m different spatial aggregators were performed by calculating the average or the sum within each 300x300 m cell.

TYPE	PROVIDER	SOURCE	SPATIAL RESOLUTION	UPDATE FREQUENCY	AGGREGATOR	SIMILAR DATASET
Meteo	ARPA PUGLIA	WRF Reanalysis model	4 km (Bari); 1km (Taranto)	hourly	Spatial Bicubic interpolation, temporal mean	ERA5 from Copernicus
Vegetation index	Copernicus by ESA	Satellite Sentinel2A/B	10 m	daily (by GPR interpolation)	Mean value for 300m-size grid element	
Traffic	Open Street Map	Open datasets; volunteer	vector	1 up-to-date dataset	Statistics for 300m-size grid element	Local traffic sensors



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

Demographic	ISTAT	Census, local agency	vector	yearly	Statistics for 300m-size grid element	
Building morphology	EUBUCCO, GOVERNMENT	Open database, satellite & aerial surveys	vector	1 up-to-date dataset	Statistics for 300m-size grid element	
LC classified map	Copernicus by ESA	Satellite Sentinel2A/B	10 m	yearly	Percentage cover per each class within 300m-size grid element	Urban Atlas by Copernicus

Table 8-Type,source and aggregation of predictors used in intra-urban model

Meteorological data with temporal updating frequency less than daily were temporally aggregated according by computing the mean value of all the data available. This outcomes a daily data for each variable of interest.

4.2 ML models

In the context of predictive methods, decision tree-based approaches represent a class of algorithms that operate by stratifying or segmenting the predictive space into elementary regions. The principle of these methods lies in their ability to recursively divide the space of input variables into increasingly homogeneous subsets with respect to the target variable. Therefore, they can capture nonlinear relationships between predictive features and the target. To leverage the strengths of tree-based algorithms, ensemble techniques have been implemented to improve predictive accuracy through the combination of multiple learning algorithms. Two distinct ensemble methodologies were selected for this analysis: Random Forest (RF) and Extreme Gradient Boosting (XG-Boost). The RF approach builds an ensemble by training numerous decision trees independently on bootstrap samples of the training data. Feature randomness is introduced at each node split, where only a subset of variables is considered for the optimal partition.

This methodology mitigates model overfitting by exploiting the bias-variance tradeoff, where the reduced variance of the ensemble compensates for the slight increase in bias compared to individual trees. XGBoost implements a sequential boosting strategy in which trees are built iteratively, with each subsequent tree focusing on correcting the residual errors of the previous ensemble. Unlike

	<i>Project:</i> APEMAIA Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

parallel training in Random Forest, XGBoost uses gradient boosting optimization to minimize a loss function, progressively improving the model's predictive accuracy by learning from the errors of previous iterations.

4.2.1 Model implementation and validation

As outlined in Deliverable 11, the workflow for applying and validating machine learning models was structured into several phases aimed at predicting and mapping particulate matter (PM).

Step 1 — Full Model

Random Forest / XGBoost

- All available predictors (full feature set)
- Hyperparameter tuning: Grid Search
- Evaluation: cross-validated performance metrics

Step 2 — Feature Selection with RFECV



Step 3 — Model Comparison

- Full model vs. reduced model (selected features), evaluated under the same cross-validation scheme
- Performance metrics: R^2 , RMSE, MAE and bias; best-performing configuration retained

Step 4 — Interpretability with SHAP Values

- SHAP values computed on the selected model to quantify the contribution of each predictor
- Global feature importance and local explanations by site and by season

Figure 22-Modelling workflow for the intra-urban PMs model

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

The initial phase involved training Random Forest and XGBoost models using the full set of available predictors. This preliminary step established a performance baseline and provided insights into the overall predictive capability of the input variables. Subsequently, a feature selection strategy was tested and evaluated to identify the most informative variables and potentially reduce model complexity. Selection was based on intrinsic feature importance metrics provided by the algorithms themselves, specifically "permutation importance", which assesses a variable's importance by accounting for its interaction with all others. Model performance was evaluated using three cross-validation strategies. In "Leave-One-Station-Out" (LOSO) cross-validation, models were trained on data from n-1 stations and tested on the remaining station, with the process repeated until each station had served as the test set; this approach allowed for the assessment of the model's ability to predict PM concentrations at unmonitored locations. "Leave-One-Year-Out" (LOYO) cross-validation followed a similar principle, training the model on n-1 years and testing it on the excluded year to evaluate temporal transferability. Finally, "k-fold" cross-validation (with k = 5) randomly split the dataset into 10 subsets, using 9 for training and 1 for testing in each iteration. Performance metrics such as R^2 , RMSE,

and MSE were calculated for each validation scheme, enabling a comprehensive comparison of the models. For the purpose of further external validation, SHAP (SHapley Additive ExPlanations) values were used to quantify the contribution of each variable to the model's predictions and to assess whether the rules adopted by the model were physically interpretable, all complemented by a qualitative and visual analysis of the maps.

4.3 Final results

Based on the comparative analysis presented in Section 4.2.1 and in DEL_011, XGBoost was selected as the reference model for PM prediction in both study areas, since it outperformed Random Forest under all the cross-validation schemes considered. This section reports, for Bari and Taranto, the final validation metrics of the selected models, the relevance of the predictors, the physical interpretation of the model behaviour through SHAP analysis, and the resulting annual mean concentration maps of PM₁₀ and PM_{2.5} at 300 m spatial resolution.

The configurations adopted for the four final products are summarised in Table 8. For PM₁₀, in both study areas the final model is the reduced one obtained through the RFECV procedure described in Section 4.2.2: the recursive elimination of the least informative predictors preserved the predictive skill of the full model while reducing its complexity. For PM_{2.5}, in both study areas the full set of predictors was retained, since the elimination procedure did not improve the cross-validated

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

performance. The number of monitoring stations available for each species and study area, reported in Section 2.2 and recalled in Table 8, determines the number of folds of the Leave-One-Station-Out scheme and is therefore an essential element for the interpretation of the results discussed below.

Study area	Species	Model	Predictor set	Stations available (LOSO folds)	Period
Bari	PM10	XGBoost	RFECV-reduced	9	2019–2022
	PM2.5		Full feature set	4	
Taranto	PM10		RFECV-reduced	8	
	PM2.5		Full feature set	5	

Figure 23-Configuration of the four final products: selected model, predictor set and number of monitoring stations available for training and validation.

4.3.1 Interpretation of the validation results

The three cross-validation schemes introduced in Section 4.2.2 each isolate a different source of error and address a different question. Consequently, they are interpreted jointly, following the framework summarized in Table 9, rather than as independent performance metrics.

The k-fold scheme quantifies the model's ability to reproduce observed variability when the sites and years present in the test samples are also represented in the training set. Since daily PM series exhibit strong spatial and temporal autocorrelation, random sample partitioning results in the training set containing days and locations very similar to those used for testing.

The Leave-One-Station-Out (LOSO) scheme is the most relevant regarding the intended use of the products. As the final output is a spatially continuous concentration field, the model must predict PM levels where no measurements are available; LOSO replicates this exact scenario by excluding one station at a time from the training set and testing the model against it. LOSO metrics are therefore adopted here as the benchmark estimate for the spatial accuracy of the maps. Since the number of folds equals the number of available stations, the LOSO statistics for the PM2.5 models, which rely on only four or five stations, are based on a limited number of folds; consequently, they are subject to greater sampling variability than the corresponding PM10 statistics. For this reason, they are always reported alongside station-by-station details.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

The Leave-One-Year-Out (LOYO) scheme tests the temporal transferability of the model, that is, its ability to reproduce a year that was not seen during training. With four years available (2019–2022), each fold removes one quarter of the dataset and, at the same time, an entire portion of the meteorological and emission variability, including the anomalous conditions of 2020.

Cross-validation scheme	Number of folds	Use made of the result
k-fold	5	Upper bound of the achievable performance
Leave-One-Station-Out (LOSO)	One per available station (9 and 8 for PM10; 4 and 5 for PM2.5)	Reference estimate of the spatial accuracy of the maps
Leave-One-Year-Out (LOYO)	4 (2019, 2020, 2021, 2022)	Delimitation of the temporal domain of validity

Table 9 -Cross-validation schemes.

Performance is quantified through the mean squared error (MSE), its square root (RMSE, expressed in $\mu\text{g}/\text{m}^3$ and therefore directly comparable with the concentration levels) and the coefficient of determination (R^2). For each scheme, the mean across folds and the corresponding standard deviation are reported; the latter is an integral part of the result, since a low dispersion indicates a homogeneous behaviour across folds, whereas a dispersion of the same order of magnitude as the mean indicates that the average value is not representative of the individual sites or years. In the absence of regulatory thresholds for this class of products, the following criteria were adopted for the interpretation of the metrics: (i) a LOSO R^2 above 0.60 was regarded as adequate for exposure assessment at the intra-urban scale; (ii) a difference between the k-fold and the LOSO R^2 larger than 0.20 was regarded as evidence of site-specific behaviour that cannot be inferred from the remaining stations; (iii) a fold-to-fold standard deviation of R^2 comparable with its mean was regarded as an indication that the average metric must not be used as a synthesis of the model performance, and that the station-by-station results must be examined instead.

4.3.2 Model performance

4.3.2.1 Bari

Table 10 reports the cross-validation metrics of the final XGBoost models for the Bari study area, for both PM10 and PM2.5.

Species	CV scheme	MSE ($\mu\text{g}^2/\text{m}^6$)	RMSE ($\mu\text{g}/\text{m}^3$)	R ²
PM10	k-fold	25.79 $\pm$ 3.33	5.07 $\pm$ 0.22	0.780 $\pm$ 0.009
	LOSO	33.34 $\pm$ 13.96	5.64 $\pm$ 1.23	0.706 $\pm$ 0.122
	LOYO	77.69 $\pm$ 16.67	8.76 $\pm$ 0.96	0.340 $\pm$ 0.040
PM2.5	k-fold	10.82 $\pm$ 2.34	3.27 $\pm$ 0.33	0.750 $\pm$ 0.040
	LOSO	12.26 $\pm$ 4.01	3.44 $\pm$ 0.58	0.700 $\pm$ 0.120
	LOYO	24.22 $\pm$ 3.66	4.90 $\pm$ 0.37	0.440 $\pm$ 0.020

Table 10 - Cross-validation metrics of the final XGBoost models for Bari

For PM10, the model effectively captures daily variability under the k-fold scheme ($R^2 = 0.78$, $\text{RMSE} = 5.07 \mu\text{g}/\text{m}^3$). Switching to the LOSO procedure results in a slight loss of predictive performance ($R^2 = 0.71$, $\text{RMSE} = 5.64 \mu\text{g}/\text{m}^3$, specifically, a change of -0.07 in R^2 and $+0.57 \mu\text{g}/\text{m}^3$ in RMSE); thus, based on criterion (ii) in Section 4.3.1, the Bari PM10 model is capable of transferring relationships learned at training stations to the excluded ones, a necessary capability for generating maps. However, there is significant dispersion across the various folds (R^2 standard deviation of 0.122), indicating that accuracy is not uniform across the nine stations; station-specific details are provided in Table 11.

A similar pattern is observed for PM2.5, though with lower absolute errors, consistent with the lower concentration levels of the fine fraction: $R^2 = 0.75$ and $\text{RMSE} = 3.27 \mu\text{g}/\text{m}^3$ for k-fold; $R^2 = 0.70$ and $\text{RMSE} = 3.44 \mu\text{g}/\text{m}^3$ for LOSO. The minimal performance drop between the two schemes (-0.05 in R^2) is a positive result, though it must be interpreted considering the limited number of available folds; this makes the average metric less robust than the corresponding figure for PM10.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Under the LOYO scheme, both models lose a substantial portion of their predictive capability ($R^2 = 0.34$ for PM₁₀ and $R^2 = 0.44$ for PM_{2.5}, with RMSE rising to 8.76 and 4.90 $\mu\text{g}/\text{m}^3$, respectively). Excluding a single year entails removing a specific combination of weather conditions, Saharan dust intrusions, and emission levels from the training set; a four-year archive is too short for the remaining years to adequately represent that combination, an effect further amplified by the anomalous emission conditions of 2020.

4.3.2.2 Taranto

Tables 12 and 13 report the cross-validation metrics of the final XGBoost models for the Taranto study area.

Species	CV scheme	MSE ($\mu\text{g}^2/\text{m}^6$)	RMSE ($\mu\text{g}/\text{m}^3$)	R^2
PM ₁₀	k-fold	25.6 $\pm$ 5.26	5.03 $\pm$ 0.52	0.76 $\pm$ 0.03
	LOSO	36.9 $\pm$ 15.0	5.94 $\pm$ 1.26	0.60 $\pm$ 0.13
	LOYO	65.3 $\pm$ 7.59	8.07 $\pm$.46	0.36 $\pm$ 0.08
PM _{2.5}	k-fold	11.41 $\pm$ 2.18	3.36 $\pm$ 0.31	0.670 $\pm$ 0.050
	LOSO	29.19 $\pm$ 23.99	4.87 $\pm$ 2.31	0.084 $\pm$ 0.660
	LOYO	20.17 $\pm$ 4.14	4.46 $\pm$ 0.46	0.400 $\pm$ 0.050

Table 12 - Cross-validation metrics of the final PM₁₀ and PM_{2.5} XGBoost model for Taranto

For PM₁₀, the model reproduces the observed daily variability well under the k-fold scheme ($R^2 = 0.76$, RMSE = 5.03 $\mu\text{g}/\text{m}^3$) similar to those obtained for Bari ($R^2 = 0.78$, RMSE = 5.07 $\mu\text{g}/\text{m}^3$): the predictor database therefore describes the variability of the coarse fraction with comparable effectiveness in the two study areas, despite their different emission regimes.

Under the LOSO scheme the performance decreases to $R^2 = 0.60$ and RMSE = 5.94 $\mu\text{g}/\text{m}^3$. The loss with respect to k-fold (-0.16 in R^2) remains below the threshold set by criterion (ii) of Section 4.3.1, and the fold-to-fold dispersion is moderate (standard deviation of R^2 equal to 0.13): the average value is therefore representative of the behavior of the individual stations, and the model is able to

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

transfer the relationships learned at the training sites to the excluded ones. The result meets the acceptability criterion adopted for exposure assessment, although it lies at its lower limit and remains slightly poorer than the corresponding value for Bari ($R^2 = 0.71$). Under the LOYO scheme the performance drops to $R^2 = 0.36$, with RMSE increasing to $8.07 \mu\text{g}/\text{m}^3$, as for all the other configurations.

For PM_{2.5}, the k-fold scheme yields acceptable performance ($R^2 = 0.67$, RMSE = $3.36 \mu\text{g}/\text{m}^3$), confirming that the model reproduces the overall variability of the observations. The LOYO validation result ($R^2 = 0.40$, RMSE = $4.46 \mu\text{g}/\text{m}^3$) is lower but stable across the various folds and fully consistent with the values obtained for all other models, including those developed for Bari; this consistency indicates that limited temporal transferability is a general characteristic of the approach rather than a site-specific issue. In contrast, the LOSO scheme yields much poorer average performance ($R^2 = 0.08$), associated with high dispersion across the folds (standard deviation of 0.66). Based on criterion (iii) in Section 4.3.1, the average value is not informative in this case; instead, the results for individual stations reported in Table 14 must be examined. Detailed analysis shows that the model behaves in two distinctly different ways: three sites are reproduced from the remaining ones with accuracy ranging from low to good (ARCHIMEDE $R^2 = 0.75$, MACCHIAVELLI $R^2 = 0.69$, CISI $R^2 = 0.13$), whereas two sites are essentially not predicted, exhibiting negative R^2 values and high biases of opposite signs (ALTOADIGE $R^2 = -1.08$, bias = $+5.10 \mu\text{g}/\text{m}^3$; TAMBURI $R^2 = -0.08$, bias = $-5.14 \mu\text{g}/\text{m}^3$).

The two stations where the model fails are those whose local conditions cannot be inferred from the rest of the network. The TAMBURI case is the most significant and informative: the station is directly influenced by the nearby industrial complex and exhibits concentration levels that are consistently higher than those at any other site in the network (Section 2.2). When it is excluded from the training set, none of the remaining stations possess the information needed to reproduce its levels, which the model consequently underestimates by approximately $5 \mu\text{g}/\text{m}^3$; conversely, the Alto Adige site is overestimated to a comparable degree. This behaviour stems not from numerical instability, but rather from a structural limitation of a data-driven approach applied to a network where the industrial source is monitored by a single representative station: the available predictors, with a 300 m resolution, describe the urban and meteorological context but cannot resolve the near-field gradient of an isolated industrial source. The implications for the use of these products are discussed in Section 4.3.6.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Station	MSE	RMSE	R2	BIAS
ALTOADIGE	56.017	7.484	-1.075	5.099
ARCHIMEDE	5.147	2.268	0.745	0.373
CISI	16.813	4.100	0.131	2.718
MACHIAVELLI	7.644	2.764	0.692	-0.433
TAMBURI	59.898	7.739	-0.076	-5.144

Table 109-PM25 LOSO metrics by station in Taranto

4.3.2.3 Comparison across sites and species

The four models outline a coherent and interpretable picture. Under the k-fold scheme the performance is comparable in all cases (R^2 between 0.67 and 0.78), indicating that the predictor database describes the daily variability of both species and both areas with similar effectiveness. The differences appear when the schemes that break the spatial and the temporal dependence of the samples are applied, and they are informative.

In space, Bari behaves substantially better than Taranto: the loss of skill from k-fold to LOSO is limited for both species in Bari (-0.07 and -0.05 in R^2) and dramatic for PM_{2.5} in Taranto (-0.59). The explanation does not lie in the number of stations, which is comparable, but in their representativeness: the Bari network samples traffic, background and industrial contexts whose conditions partially overlap, whereas the Taranto PM_{2.5} network includes a site dominated by a single, spatially confined industrial source. The spatial accuracy of the products is therefore high where the concentration field is governed by urban-scale and meteorological processes, and low in the immediate surroundings of isolated industrial sources.

In time, all four models degrade in a similar way (LOYO R^2 between 0.34 and 0.44, with small fold-to-fold dispersion). Such a homogeneous behavior across two independent areas and two different size fractions points to a common cause, associated with the length of the training archive and with the model configuration, rather than to a site-specific deficiency.

Between the two size fractions the difference in skill is smaller than expected, but the difference in mechanisms is substantial: as discussed in Sections 4.3.3 and 4.3.4, comparable performance is achieved through different predictors, namely mineral dust and boundary layer dynamics for PM₁₀,



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

and seasonality, synoptic conditions and proximity to emission sources for PM_{2.5}. This is the main methodological argument in favor of the size-segregated modeling approach adopted in the project.

4.3.3 Predictor relevance

The relevance of the predictors was quantified through permutation importance computed on the test folds of the cross-validation, so that the resulting ranking reflects the contribution of each variable to the predictive skill on unseen data rather than its role in fitting the training set.

Bari

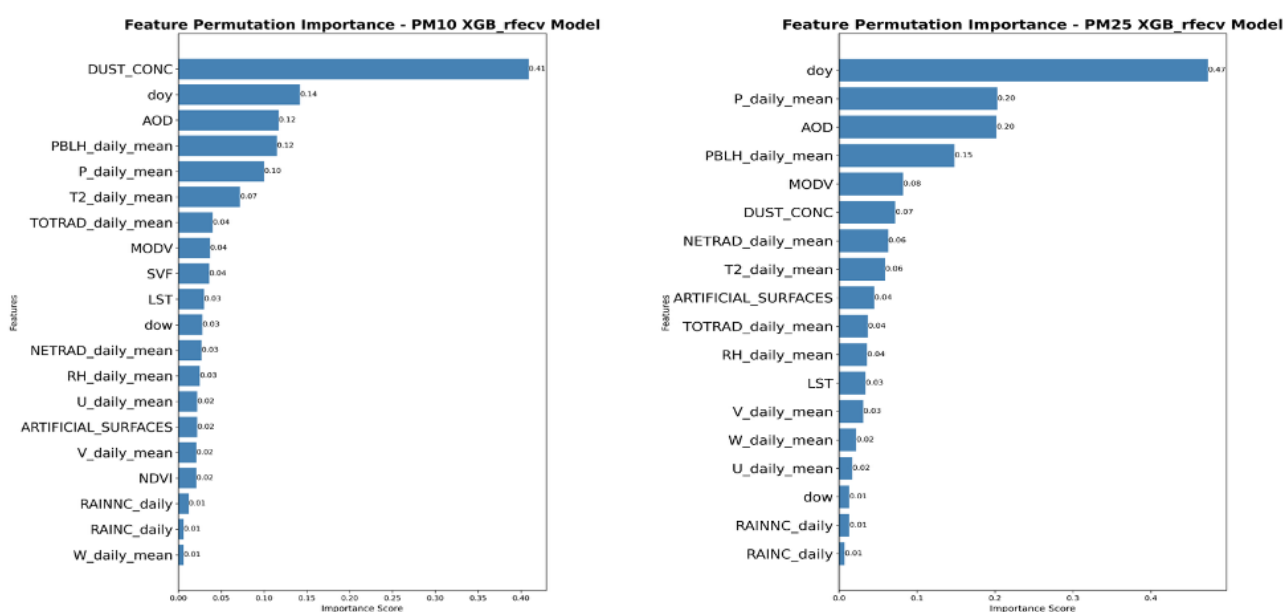


Figure 16 - Permutation importance of the predictors of the final XGBoost models for Bari: PM₁₀ (left) and PM_{2.5} (right).

The permutation importance analysis reveals distinct predictor hierarchies for the coarse and the fine fraction in Bari, consistent with their different sources and atmospheric processes. Mineral dust concentration (DUST_CONC) dominates the PM₁₀ model (importance = 0.41) but contributes marginally to PM_{2.5} (0.07), confirming that the mineral component is mostly retained in the coarse fraction. Conversely, the day of the year (doy) is by far the strongest predictor of PM_{2.5} (0.47), more than three times its importance for PM₁₀ (0.14): the fine fraction is modulated by a pronounced seasonal cycle, driven by winter heating emissions and by the recurrent stagnation conditions of the cold season. Daily mean surface pressure (P_daily_mean) is relevant for both species, with a larger



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

weight for PM_{2.5} (0.10 for PM₁₀, 0.20 for PM_{2.5}), consistent with the role of anticyclonic conditions in the accumulation of secondary fine particles. Aerosol optical depth (AOD) contributes substantially to both models (0.12 and 0.20), confirming that the satellite-retrieved columnar aerosol load is an effective proxy for the surface concentration of both fractions. Planetary boundary layer height (PBLH_daily_mean) has comparable importance for the two species (0.12 and 0.15), indicating that vertical mixing governs the surface levels of coarse and fine particles alike. Temperature, radiation and wind variables play a minor role in both models, as do the precipitation fields (RAINNC_daily and RAINNC_daily, importance ≤ 0.01), whose limited weight reflects the small number of rainy days in the study period rather than the absence of a wet removal process.

Taranto

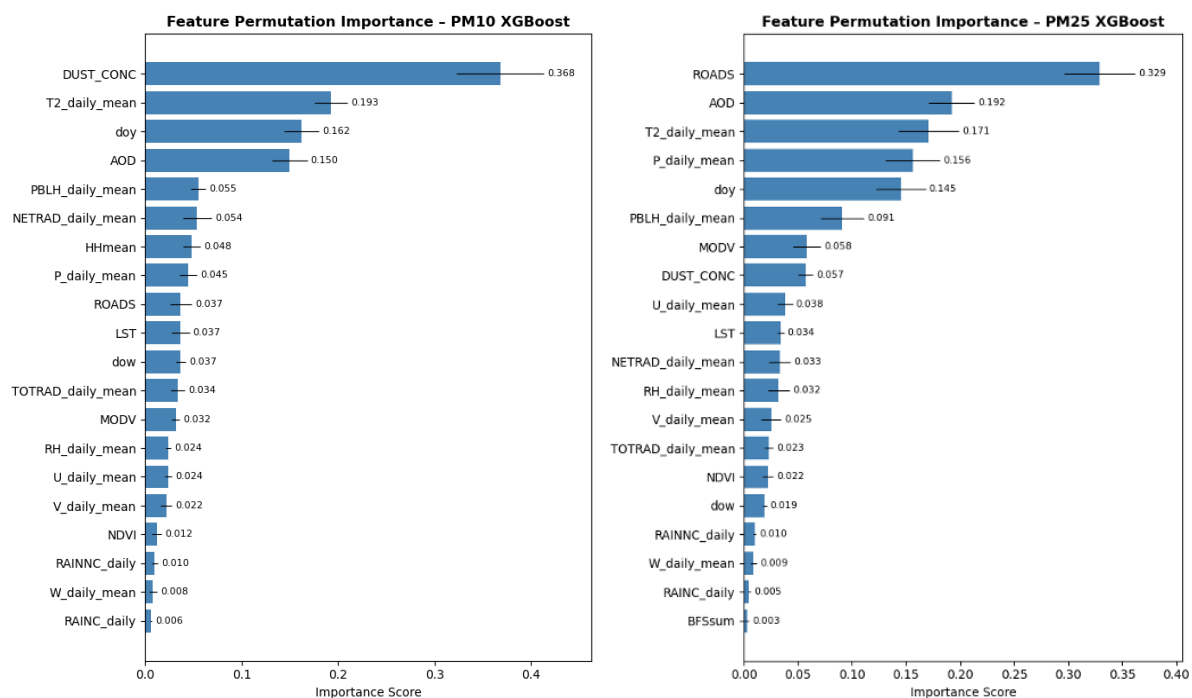


Figure 24-Permutation importance of the predictors of the final XGBoost models for Taranto: PM₁₀ (left) and PM_{2.5} (right).

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

Three consistent patterns emerge from the four rankings, confirming the physical coherence of the models. First, aerosol optical depth is among the four most important predictors in every configuration, with remarkably stable weights (0.12–0.19), and planetary boundary layer height is always retained, although with a lower weight in Taranto (0.06–0.09) than in Bari (0.12–0.15): the two variables describing, respectively, the columnar aerosol load and the volume available for dilution provide significant information regardless of the study area or size fraction. Second, mineral dust concentration is by far the leading predictor of the coarse fraction in both areas, with almost identical weights (0.41 in Bari, 0.37 in Taranto), whereas its importance for PM_{2.5} does not exceed 0.07 in either area: the mineral component is therefore confined to the coarse fraction and is governed by a process, the transport of Saharan dust, acting on the whole region. Third, the divergence between the two areas is confined to the fine fraction and involves the predictors related to the local emission characteristics: seasonality dominates in Bari (0.47), given that the fine fraction largely originates from diffuse urban sources, whereas the road traffic descriptor dominates in Taranto (0.33), where the spatial gradients are more pronounced. None of the models rely on variables lacking physical meaning, a necessary condition for using the predicted fields to assess exposure. Further confirmation is provided in the following paragraph.



4.3.4 Model interpretability and SHAP analysis

The SHAP summary plots confirm the direction of the dependencies inferred from the importance ranking.

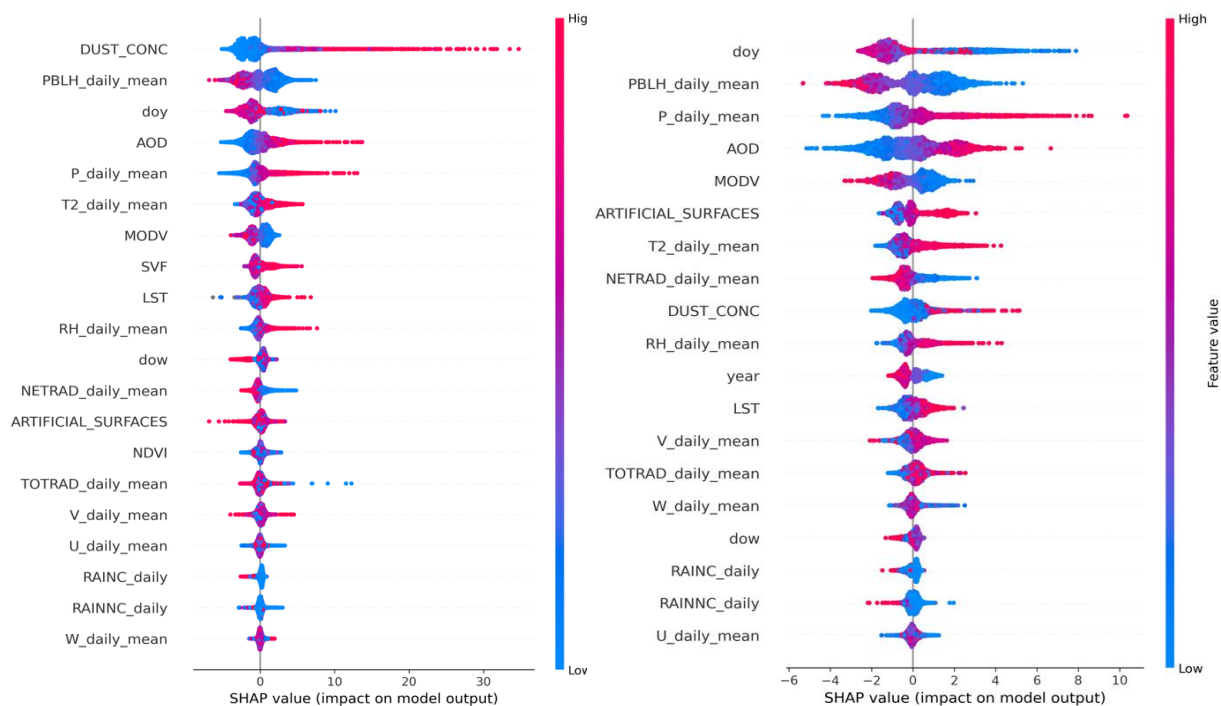


Figure 25-Shap values of predictors in the XGBoost PM10 and PM25 model (Bari)

For PM10, dust concentration shows an almost monotonic positive response, with contributions reaching about $+26 \mu\text{g}/\text{m}^3$ on the most intense days: the Saharan intrusions generate the upper tail of the distribution. Boundary layer height acts in the opposite direction, shallow layers producing positive contributions and deep layers negative ones, consistently with the dilution mechanism, while the day of the year yields positive contributions in the cold season. Mean building height contributes up to $+5 \mu\text{g}/\text{m}^3$ where the urban fabric is highest and most compact, consistently with the reduced ventilation of densely built sectors.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

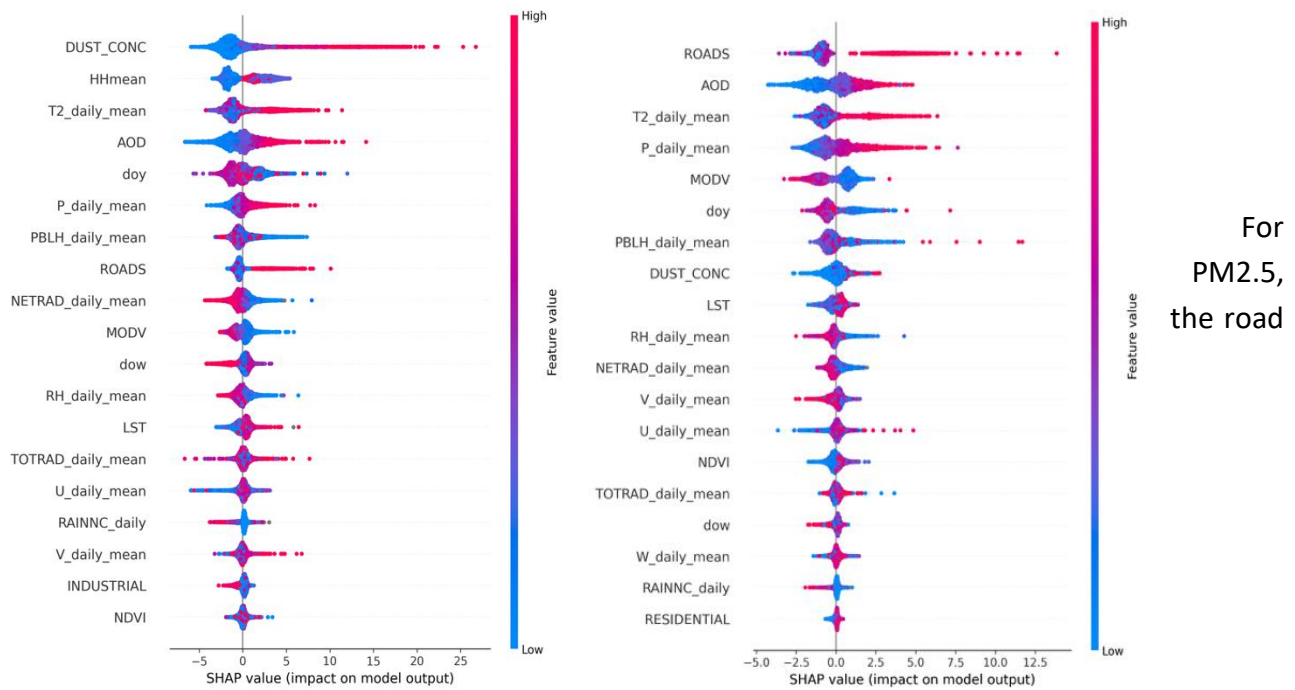


Figure 26-SHAP values of predictors in the XGBoost PM10 and PM25 model (Taranto)

descriptor dominates the amplitude of the contributions (up to about $+13 \mu\text{g}/\text{m}^3$ at the sites closest to the road network) and the vegetation descriptor acts symmetrically in the opposite direction, so that the two variables reconstruct the intra-urban gradient that the meteorological predictors, nearly uniform over the domain, cannot resolve. In both species, aerosol optical depth shows a positive and nearly monotonic response, confirming that the satellite constraint is used consistently with its physical meaning.

4.3.5 Spatial predictions: annual, seasonal and monthly mean maps

The final products consist of annual, seasonal, and monthly mean concentration fields for PM10 and PM2.5 at a spatial resolution of 300 m, obtained by applying the selected models to the full grid of predictors for each day of the year and averaging the resulting daily fields for each of the four years in the dataset.

The complete set of products delivered comprises, for each study area and each species, four annual fields (2019–2022), sixteen seasonal fields and forty-eight monthly fields. Only a subset is reported

57



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

here for illustrative purposes: the annual mean maps of both species for the two study areas, which summarise the spatial structure of the concentration field, and one seasonal sequence, which illustrates its temporal modulation. Within each figure the four years share a common colour scale, so that they can be compared directly.

High-resolution annual mean PM10 over Bari (2019-2022)

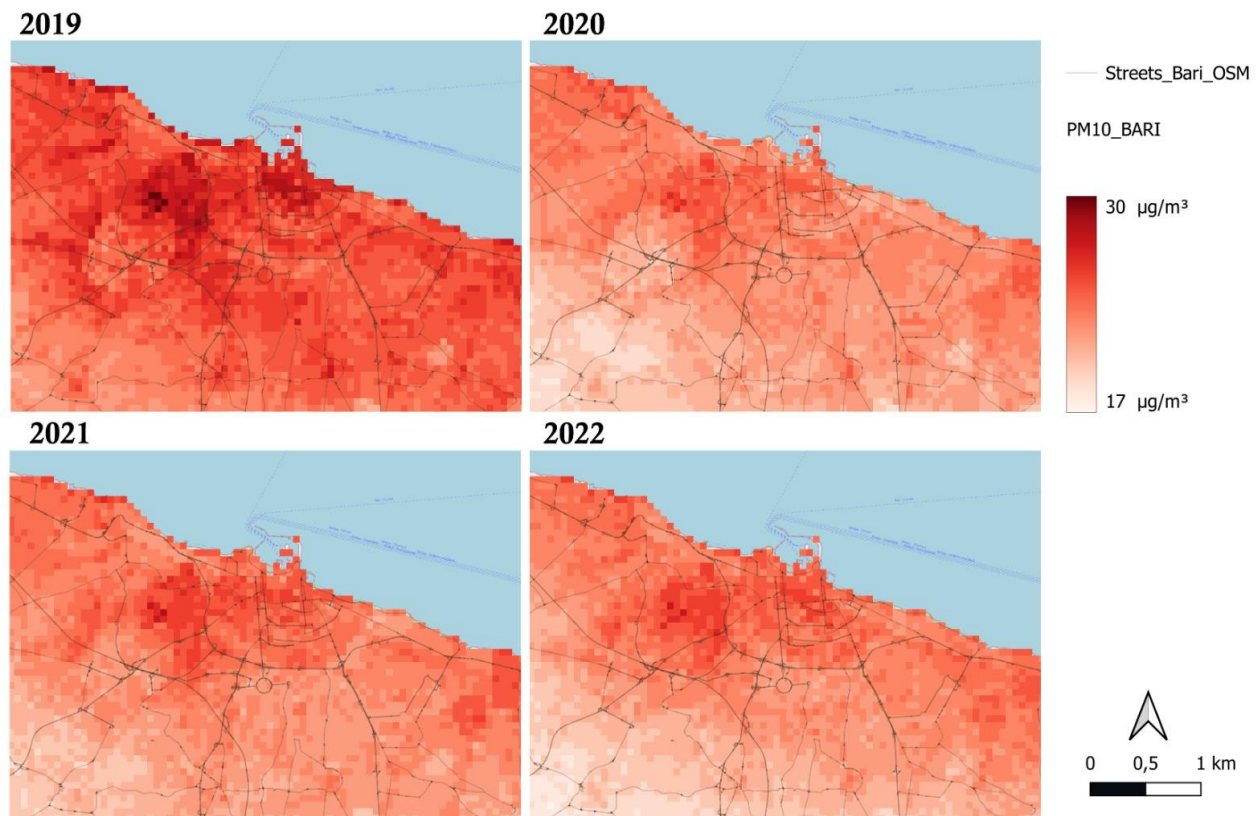


Figure 27-Bari, map of annual average PM10 levels.

In Bari the annual mean PM10 field (Figure 27) reproduces the spatial structure expected from the distribution of the sources: the highest values are found in the western industrial sector, in the area of Modugno, and along the coastal and harbour belt, whereas the lowest values occur in the southern peri-urban and agricultural sectors. The interannual sequence is consistent, with 2019 as the year of the highest annual means and 2021 as the lowest. The PM2.5 field (Figure 18) shows a similar but smoother pattern, with a maximum over the compact urban core and the harbour area

and a monotonic decrease towards the inland periphery; annual means range between 8.5 and 16 $\mu\text{g}/\text{m}^3$, with a gradient of about 5 $\mu\text{g}/\text{m}^3$ between the two extremes, and the interannual variability is weaker than for the coarse fraction.

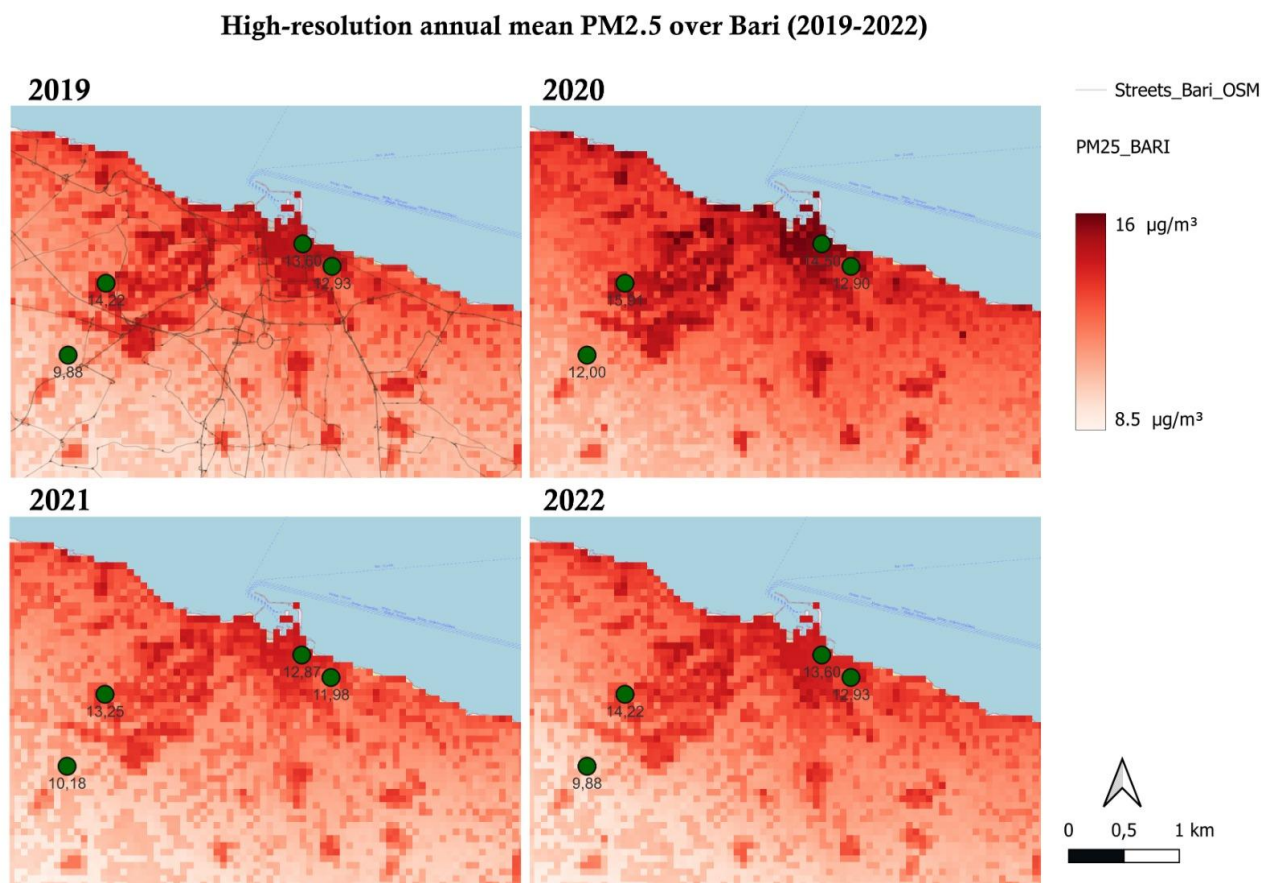


Figure 28-Bari, map of annual average PM_{2.5} levels

In Taranto the annual mean PM₁₀ field ranges between about 14 and 33 $\mu\text{g}/\text{m}^3$ and is organised along two directions: a maximum over the urban and port strip surrounding the Mar Piccolo, extending north-west towards the industrial area, and a minimum over the northern rural sector. The PM_{2.5} field, between about 5 and 21 $\mu\text{g}/\text{m}^3$, has a sharper spatial structure, with the maximum confined to the industrial and densely built sectors and a steeper decrease towards the northern countryside. This difference between the two size fractions is the spatial counterpart of the interpretability analysis of Section 4.3.4: the coarse fraction, driven by mineral dust and by boundary



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

layer dynamics, varies smoothly over the domain, whereas the fine fraction, driven by proximity to local sources, reproduces the pattern of the road network and of the built-up areas. In the PM_{2.5} maps, isolated cells with high values appear along the main roads of the northern rural sector: they follow the road descriptor, which is the leading predictor of the fine fraction, rather than an independently observed gradient, and are therefore to be read as an indication of source proximity rather than as point estimates.

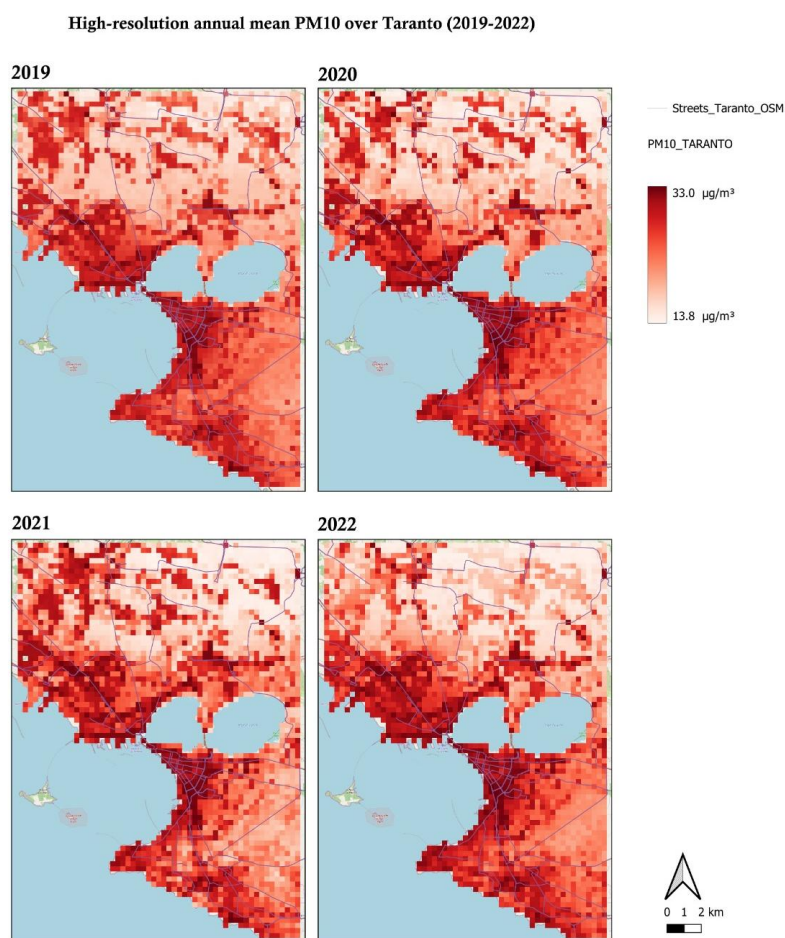


Figure 29-Taranto, map of annual average PM₁₀ levels.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

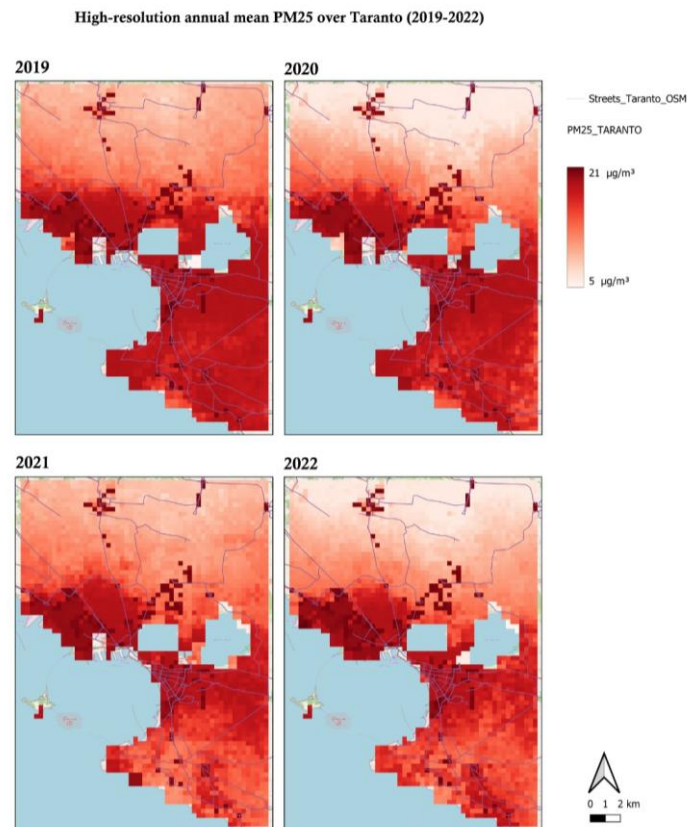


Figure 30-Taranto, map of annual average PM25 levels.

The seasonal fields preserve the spatial structure described above but modulate both its intensity and its degree of spatial contrast. The figure below reports, as an example, the four seasonal means of PM10 over Bari for 2019. The highest values are reached in summer, when the field is also the most spatially uniform, while autumn shows the lowest levels; winter is intermediate in terms of average level but exhibits the strongest intra-urban gradient, with a marked maximum over the compact urban core and the harbour area and much lower values over the southern periphery. Spring is intermediate in both respects, with the industrial sector of Modugno emerging as a local maximum. This behaviour is consistent with the interpretation given in Section 4.3.4. The summer maximum reflects the regional-scale contribution of mineral dust and of dry resuspension, which affects the whole domain in a similar way and therefore raises the background rather than the gradients; the winter pattern reflects instead the local emissions concentrated in the urban core, combined with a shallow boundary layer that confines them close to the sources. The seasonal products therefore convey information that the annual means cannot express in summer the



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

exposure is nearly homogeneous across the city, whereas in winter it depends strongly on the location within the urban fabric.

High-resolution seasonal mean PM10 over Bari (2019)

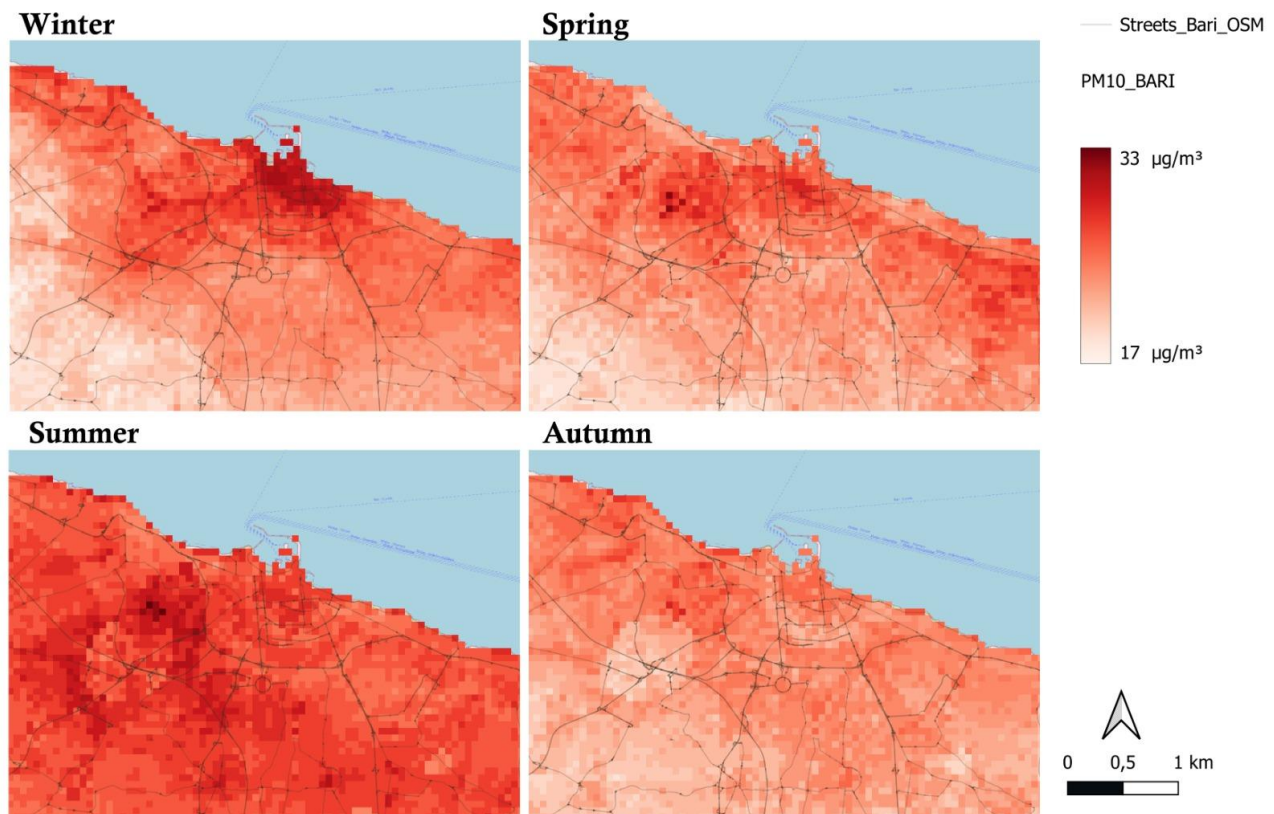
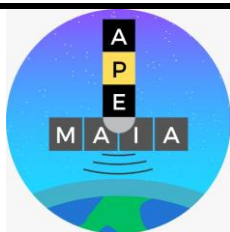


Figure 31-Bari, map of seasonal average PM10 levels (2019)

4.3.6 Preliminary investigation of a multi-city approach

4.3.6.1 Method

The models discussed so far were trained separately for each study area. Since the number of monitoring stations available in each area is limited, and since the operational use of the approach in other urban contexts requires a single transferable configuration, a further experiment was carried out by pooling the training data of Bari and Taranto into one multi-city model per species using for all the model the whole set of features. Preliminary investigation of a multi-city approach



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

The validation protocol is the one defined in Section 4.3.1: the Leave-One-Station-Out scheme was applied to the pooled dataset, so that each station is predicted from a training set containing the stations of both areas. The recursive feature elimination was repeated for each configuration, which therefore do not necessarily share the same predictor set.

4.3.6.2 Results and implications

The station-by-station comparison is reported in *Table 10* for PM₁₀ and in *Table 11* for PM_{2.5}. For both species the two configurations performed comparably in terms of overall accuracy, the mean LOSO RMSE differing by about 0.2 µg/m³, and which of the two ranks first depends on the metric considered. The pooled model gained accuracy at the stations that the local models reproduced worst, and lost it where the local calibration already performed well.

Stations	Multi-city model					Site-specific Model				
	MSE_MC	RMSE_MC	MAE_MC	R ² _MC	Bias_MC	MSE	RMSE	MAE	R ²	Bias
TARANTO										
ALTOADIGE	46.17	6.79	5.73	0.43	4.83	80.50	8.97	7.40	0.05	5.83
ARCHIMEDE	19.20	4.38	3.27	0.75	1.33	16.26	4.03	2.82	0.80	1.31
CISI	29.40	5.42	4.62	0.59	4.18	21.32	4.62	3.64	0.71	2.80
MACHIAVELLI	28.70	5.36	4.11	0.61	-3.02	37.22	6.10	4.76	0.53	-3.63
STATTE	29.62	5.44	4.20	0.65	1.06	45.01	6.71	5.49	0.49	3.10
TALSANO	36.84	6.07	4.06	0.62	1.71	47.11	6.86	4.21	0.54	-1.10
TAMBURI	76.03	8.72	6.13	0.25	-5.59	82.27	9.07	6.36	0.27	-5.99
WIND	25.18	5.02	3.84	0.75	0.25	25.30	5.03	3.69	0.75	-0.97
BARI										
BITONTO_EN01	49.55	7.04	5.13	0.62	-2.17	44.84	6.70	4.93	0.65	-0.06
CALDAROLA	30.03	5.48	4.14	0.72	2.24	16.32	4.04	2.78	0.84	0.61
CARBONARA	50.76	7.13	5.25	0.45	-1.38	51.18	7.15	5.36	0.45	0.18
CAVOUR	17.62	4.20	3.14	0.85	0.44	17.27	4.16	3.02	0.85	-0.90
CUS	26.55	5.15	4.03	0.75	1.63	27.42	5.24	3.99	0.72	2.70
KENNEDY	27.09	5.21	3.94	0.77	0.92	26.02	5.10	3.62	0.78	0.37
MODUGNO_EN02	19.16	4.38	3.17	0.80	-0.67	20.00	4.47	3.38	0.80	-1.36
MODUGNO_EN03	38.86	6.23	4.86	0.66	0.60	44.62	6.68	5.25	0.60	0.03
MODUGNO_EN04	54.59	7.39	4.86	0.63	-3.04	52.38	7.24	4.52	0.66	-2.31

Table 10-Leave-One-Station-Out (LOSO) cross-validation for PM₁₀, XGB model multi-city vs site-specific approach

For PM₁₀ this redistribution is positive in Taranto and neutral to slightly negative in Bari. In Taranto the pooled model improves four of the eight stations, with the largest gains at ALTO ADIGE (R² from 0.05 to 0.43) and STATTE (from 0.49 to 0.65), and it reduces the mean absolute bias from 3.09 to

	<i>Project:</i> APEMAIA Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

2.75 $\mu\text{g}/\text{m}^3$. In Bari the two configurations are substantially equivalent in terms of R^2 , the local model is better at four stations, the pooled one at three, and the two coincide at the remaining two, but the pooled model degrades the bias at the sites that were best calibrated locally, from -0.42 to -2.17 $\mu\text{g}/\text{m}^3$ at BITONTO and from 0.80 to 2.24 $\mu\text{g}/\text{m}^3$ at CALDAROLA, so that the mean absolute bias increases from 1.20 to 1.45 $\mu\text{g}/\text{m}^3$.

For PM_{2.5} the effect is confined to the worst-performing site. In Taranto the only substantial change concerns ALTO ADIGE, whose R^2 rises from -1.08 to 0.04 , while the remaining stations are unchanged or improve marginally and TAMBURI is not reproduced by either configuration. In Bari the local model is better at three of the four stations, with a lower mean RMSE (3.46 against 3.69 $\mu\text{g}/\text{m}^3$) and about half the mean absolute bias (0.69 against 1.24 $\mu\text{g}/\text{m}^3$).

Stazione	Modello unificato					Modello locale				
	MSE_MC	RMSE_MC	MAE_MC	R ² _MC	Bias_MC	MSE	RMSE	MAE	R ²	Bias
TARANTO										
ALTOADIGE	25.47	5.05	4.33	0.04	4.04	56.02	7.48	5.81	-1.08	5.10
ARCHIMEDE	5.52	2.35	1.62	0.75	0.43	5.15	2.27	1.50	0.74	0.37
CISI	17.35	4.17	3.46	0.12	3.17	16.81	4.10	3.42	0.13	2.72
MACHIAVELLI	5.68	2.38	1.65	0.77	-0.21	7.64	2.77	1.90	0.69	-0.43
TAMBURI	53.46	7.31	5.05	-0.05	-4.94	59.90	7.74	5.23	-0.08	-5.14
BARI										
BITONTO_EN01	15.45	3.93	2.98	0.56	1.41	16.97	4.12	2.86	0.53	-0.30
CALDAROLA	8.42	2.90	2.16	0.79	-0.19	7.38	2.72	2.06	0.82	0.70
CAVOUR	10.90	3.30	2.39	0.77	0.70	9.35	3.06	2.21	0.80	0.10
MODUGNO_EN02	21.50	4.64	3.35	0.52	-2.67	15.38	3.92	2.80	0.64	-1.67

Table 11-Leave-One-Station-Out (LOSO) cross-validation for PM₂₅, XGB model multi-city vs site-specific approach

The reduction of the dispersion of R^2 across stations that the pooled model produces — from 0.21 to 0.15 for PM₁₀ and from 0.59 to 0.33 for PM_{2.5} — is almost entirely attributable to ALTO ADIGE: excluding that single site, the dispersion is essentially unchanged, and within the Bari network it does not change at all. Pooling therefore does not make the model uniformly more homogeneous over the network; it limits the error at the sites whose conditions lie outside the range described by the local stations, at the cost of a slight degradation, mostly in the bias, at the sites that the local calibration already reproduced well.



Project:
APEMAIA

Contract n. 2023-
39-HH.0

Doc. N.: DEL_009_CNR_ITIAm_RF_EN

Issue/Rev.:01/01

Date: 25/09/2026

Pages: 66

The residual error follows the same structure in both species and in both configurations: the sites with the lowest observed concentrations are overestimated and the most polluted ones underestimated, by up to about 5 $\mu\text{g}/\text{m}^3$. Pooling accentuates this compression of the spatial contrast, because it widens the range of concentrations that a single model has to reconcile. TAMBURI is not reproduced by any of the four configurations, confirming that the levels of that site cannot be inferred from the remaining network irrespective of the size of the training set.

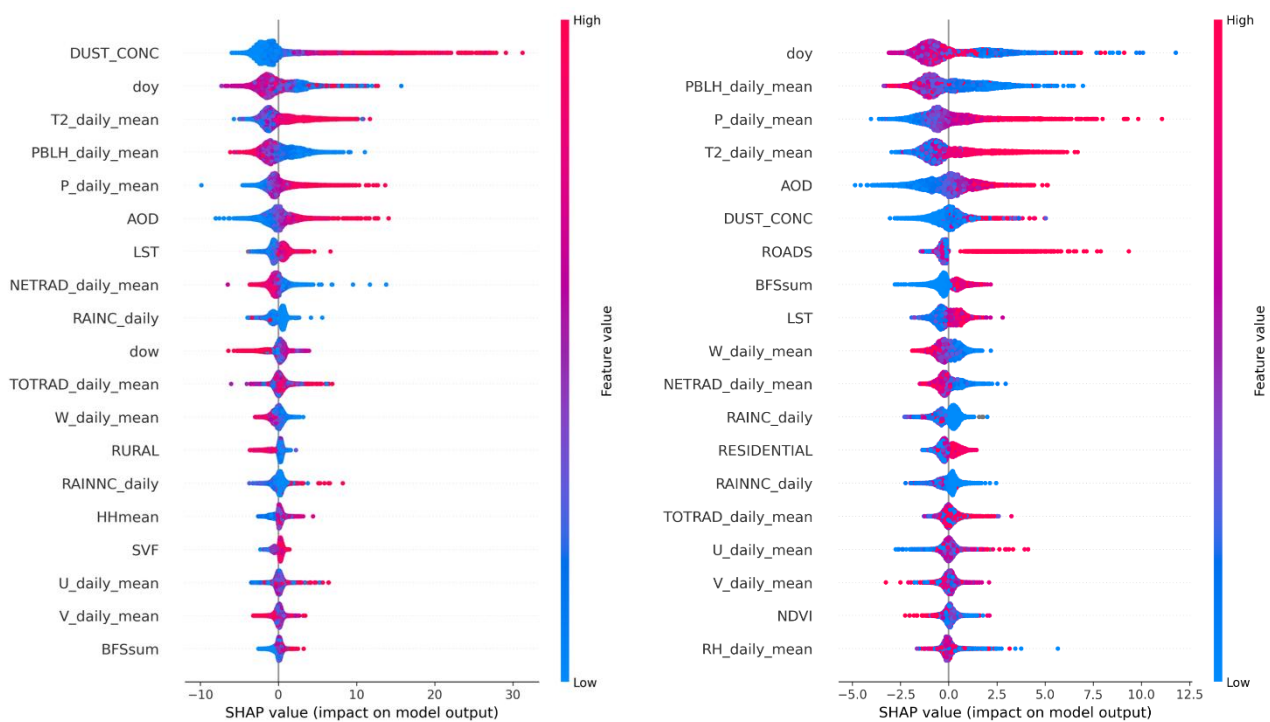


Figure 32-Shap summary plot for the multi-city XGB model for PM10 and PM25

The residual error follows the same structure in both species and in both configurations: the sites with the lowest observed concentrations are overestimated and the most polluted ones underestimated, by up to about 5 $\mu\text{g}/\text{m}^3$. Pooling accentuates this compression of the spatial contrast, because it widens the range of concentrations that a single model has to reconcile. TAMBURI is not reproduced by any of the four configurations, confirming that the levels of that site cannot be inferred from the remaining network irrespective of the size of the training set.

	<i>Project:</i> <i>APEMAIA</i> Contract n. 2023-39-HH.0	Doc. N.: DEL_009_CNR_ITIAm_RF_EN Issue/Rev.:01/01 Date: 25/09/2026 Pages: 66
---	---	--

These outcomes are consistent with the interpretability analysis of Section 4.3.4. The variance of the coarse fraction is dominated by predictors acting over the whole domain — mineral dust, temperature and boundary layer height — so that the relationships learnt at one site remain applicable to the others and a larger, more diverse training set is beneficial where the local network is sparse. The fine fraction is instead governed by descriptors of the local emission regime, whose weight changes between the two areas: pooling then mixes two different spatial structures and smooths the intra-urban gradient that local calibration is able to retain, which is the component of the signal that matters most for exposure assessment.