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

WP2103 – HR AOD 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-IIA)	**DATE SUBMITTED**	N/A
AUTHOR	M. Adamo, M. Aquilino, Marica De Lucia, Cristina Tarantino (CNR-IIA) A. Monaco, R. Cilli, A. Fania (DIF) A. Morabito (ARPA Puglia)	**CONTRACTOR'S REFERENCE**	Issue:01 Rev.:01 Date:25/04/24
PROJECT BEGIN DATE	26/09/2023	**PROJECT END DATE**	25/09/2025

EXECUTIVE SUMMARY			
This document is the preliminary 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. The document represents a preliminary version of the **D009 Deliverable** to be submitted in its final version at Third Release of Advancements (**RA3**) of the project.			
NAME OF ASI PROJECT MANAGER	Name Surname (PM) Name Surname (RUP) Name Surname (Project Scientist)		
STAR CODE		No OF VOLUMES:	

TABLE OF CONTENTS

1 AIM OF THE DOCUMENT 7
1.1 Applicable and reference documents 8
1.2 Applicable documents 8
1.3 Reference documents 8
1.4 Acronyms and abbreviations 8
2 Introduction to the methodological approaches 10
3 Study domain 11
4 Reference Data 11
4.1 Bari sites 12
5 Regional scale approach 17
5.1 Training data 17
5.2 Database of predictors available at regional scale 17
5.3 Method 19
5.3.1 Inverse Distance Weighting 19
5.3.2 Spatial Resample of WRF Data 20
5.4 Results 21
6 Intra-urban scale approach 22
6.1 Overview of the employed predictors 22
6.1.1 Aerosol Optical Depth (AOD) data 22
6.1.2 Meteorological parameters 23
6.1.3 Vegetation Index 23
6.1.4 Other spatiotemporal data 23
6.2 Spatiotemporal aggregation at grid level 24
6.3 Preliminary implementation stages 27
6.4 ML models 28

1 AIM OF THE DOCUMENT

This document represents the preliminary version D009 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 for the **A**ssessment of **P**M **E**xposure 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-IIA), 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 refers 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 Deep Learning techniques with the aim to model daily intra-urban PM_{10} 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.

Our first efforts, reported in this preliminary release of D009, are devoted to implement a procedure to automatically build a database of predictors.

The end of these activities has been planned for the period T0+7M, whereas the final release of this deliverable has been scheduled at T0+20M.

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 (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.

[RD 2] Stafoggia et al. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environmental International (2019).

1.4 Acronyms and abbreviations

Table 1-1 provides the acronym list.

Table 1-1: 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

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 PM_{10} and $PM_{2.5}$.

Traditionally, PM concentration mapping rely 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.

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.

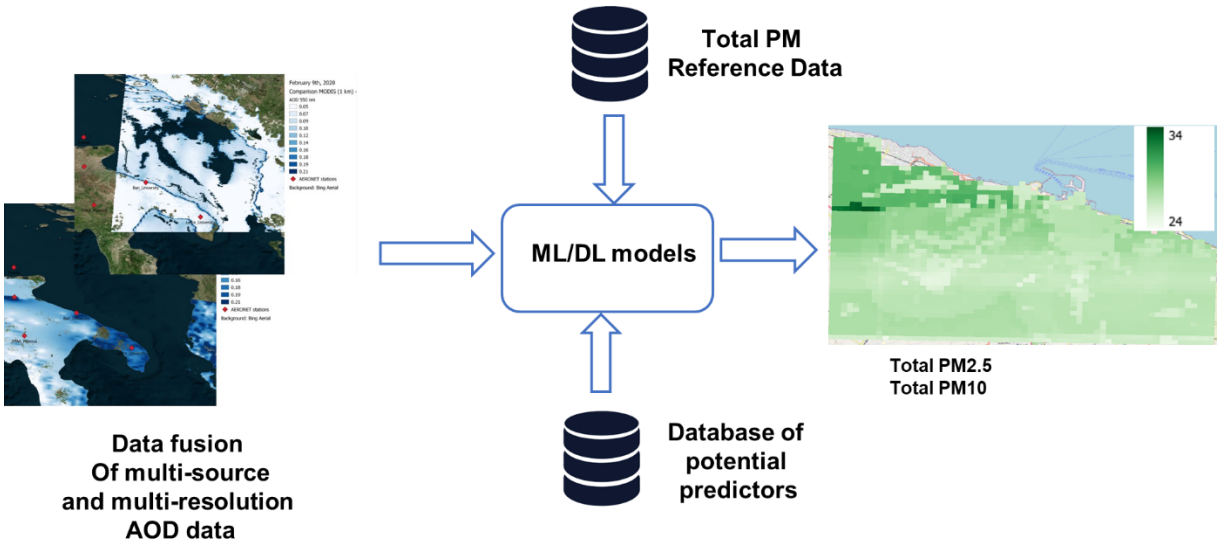


Figure 2-1: Conceptual model outline of APEMAIA project.

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].

3 Study domain

APEMAIA project focuses on primary and secondary target areas identified in support of NASA JPL's MAIA mission. Rome is a Primary Target Area (PTA).

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 3-1- APEMAIA study areas including metropolitan area of Rome PTA and Apulian sites of Bari and Taranto.

4 Reference Data

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 PM_{10} and $PM_{2.5}$ concentrations over the period 2019–2022 from all the available monitoring sites in Bari and Taranto. Additional PM reference data for Rome site will be kindly provided by NASA JPL thanks to the partnership in force with ASI, during the next months of activities.

4.1 Bari sites

Figure 4-1 illustrates the current locations of all PM and weather monitoring stations in Bari and its surrounding areas within the Apulia region. Notably, the Bitonto EN01 station is managed by the private sector. ARPA has responsibility to provide all the validated PM data required for the project.

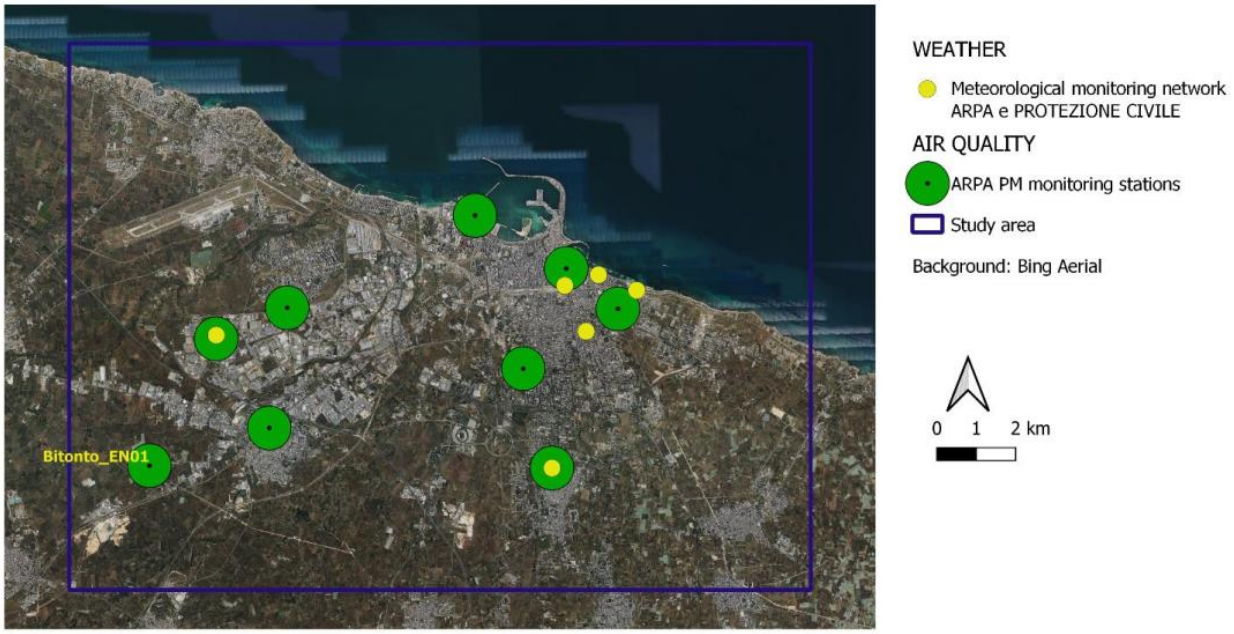


Figure 4-1: Spatial distribution of Bari monitoring stations Bari: 9 stations measure PM 10 while only 4 stations monitor PM 2.5.

Hourly acquisition data for air quality parameters are also available on the webpage https://www.arpa.puglia.it/pagina3139_dati-storici-qualit-dellaria.html. Three stations collect data at an hourly or bi-hourly frequency: Carbonara, Modugno EN02, and Modugno EN04. The other stations provide only daily average data for PM_{10} and $PM_{2.5}$.

For the three stations with higher temporal resolution, an analysis was conducted to evaluate the error introduced when approximating PM_{10} measurements taken at times corresponding to satellite-provided AOD data (the main proxy for PM) with daily average PM estimates.

The morning time window around 10 a.m. is particularly significant for satellite acquisitions. Sentinel-3 primarily acquires data between 8 a.m. and 10 a.m., while MODIS consistently does so around 10 a.m.

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, respectively (Figure 4-2).

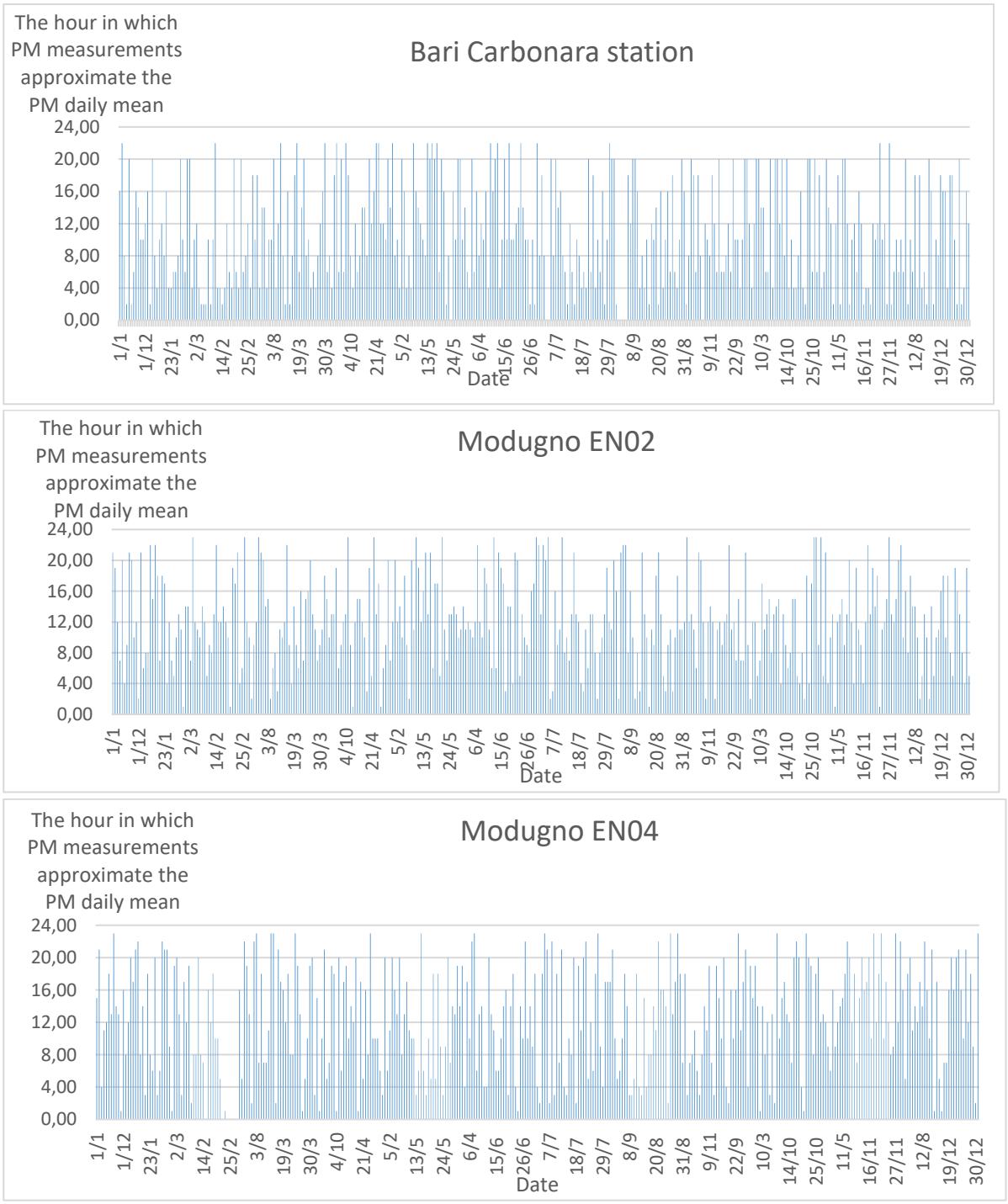


Figure 4-2–Time series of the hours in which PM measurements best approximate the daily averages of PM10 in 2019. Bari Carbonara station at top, ModugnoEN02 in the middle and Modugno EN04 at bottom.

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 4-3).

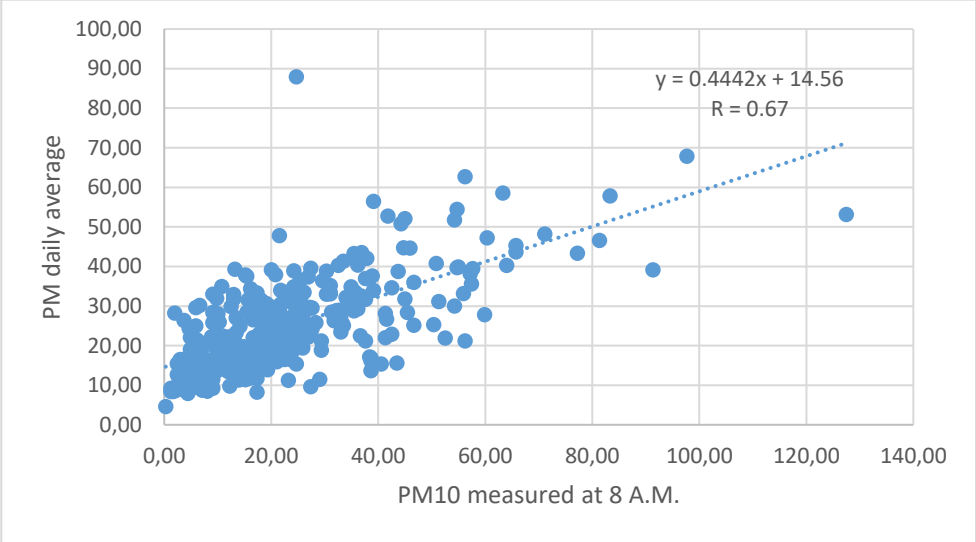


Figure 4-3 Comparison between PM10 measured at 8 a.m. and daily average value in Bari Carbonara station.

For the APEMAIA project, we work exclusively with the daily average measurements provided by ARPA after statistical validation. In the following figures, for both PM_{10} (Figure 4-4, Figure 4-5) and $PM_{2.5}$ (Figure 4-6 e Figure 4-7), we present frequency histograms of these daily averages and box plots indicating the variance of values recorded at each station and across different years.

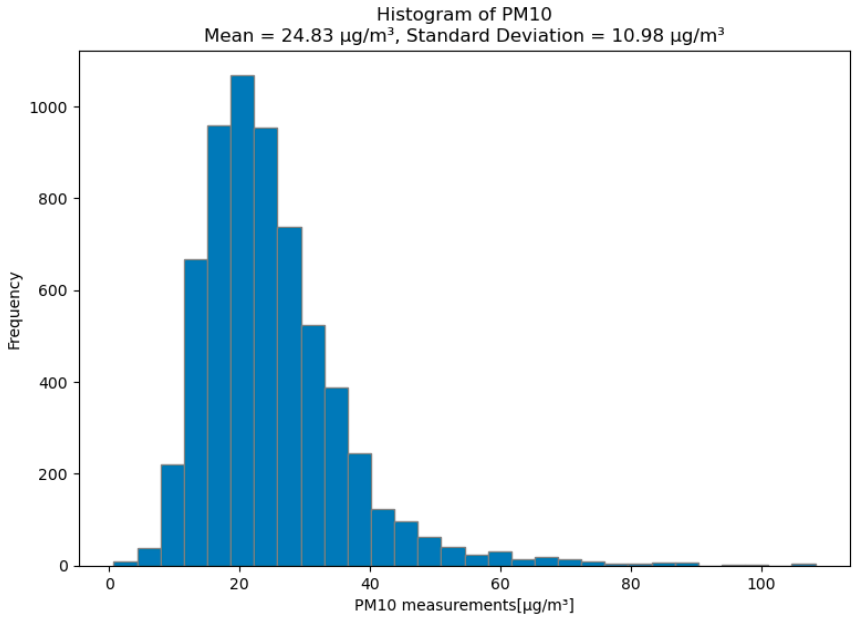


Figure 4-4 Histogram of PM10 value measured at the all the monitoring stations in Bari during the period of study 2019-2022.

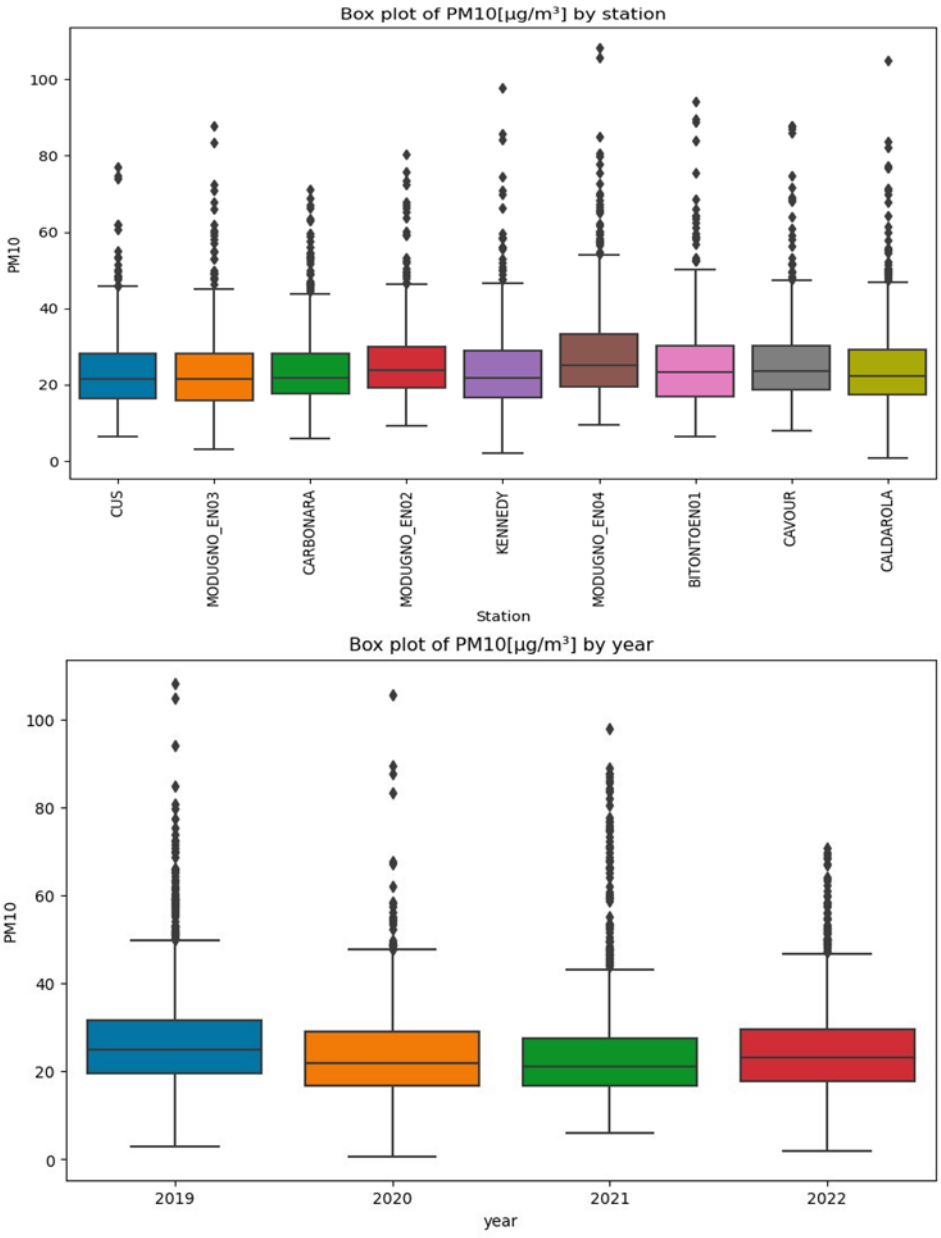


Figure 4-5 Box plot illustrating the variability of PM10 measurements by station (at top) and by year (at bottom).

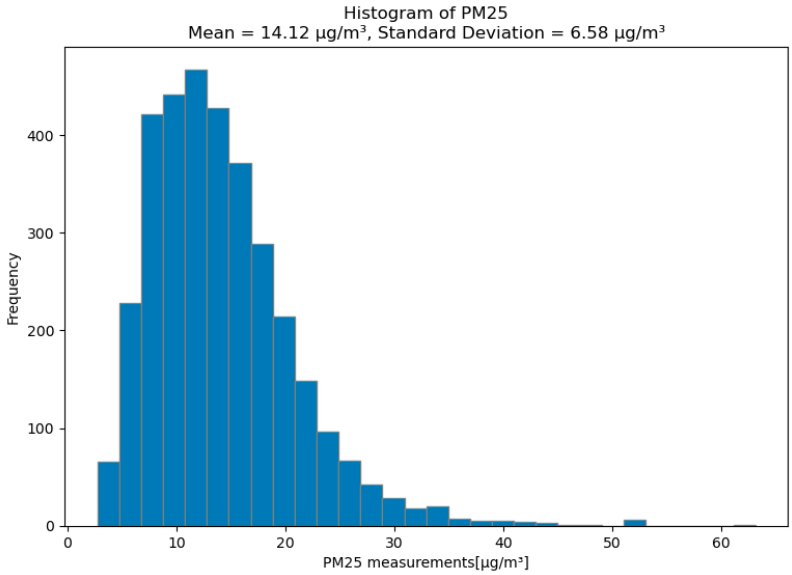


Figure 4-6 Histogram of PM2.5 value measured at the all the monitoring stations in Bari during the period of study 2019-2022.

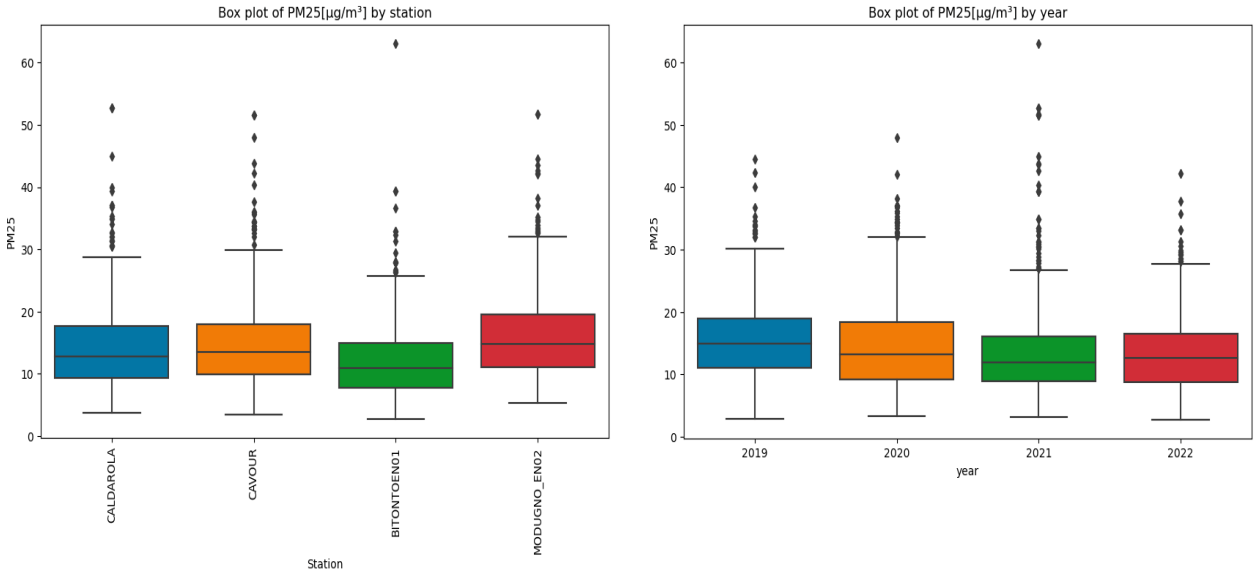


Figure 4-7 Box plot illustrating the variability of PM10 measurements by station (on the left) and by year (on the right).

5 Regional scale approach

5.1 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 4-3 is shown the spatial distribution of these station (53 for PM10, 27 for PM2.5). The measures are provided into µg/m^3. We collected data from 2019 to 2022 and we made a daily average.

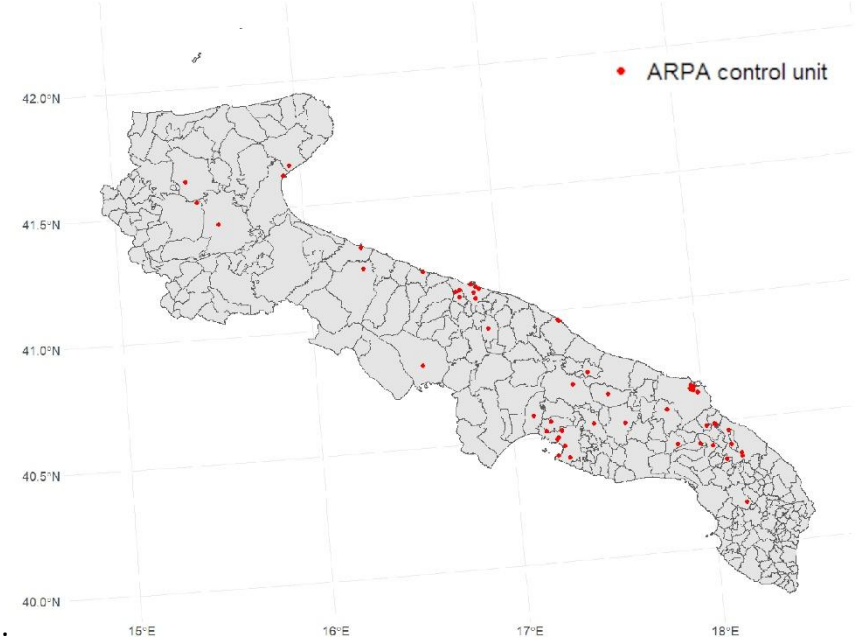


Figure 4-3–Spatial distribution of ARPA Puglia station.

5.2 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 1. 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

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	
Continous_urban_fabric	CORINE	100 m	
Discontinous_urban_fabric	CORINE	100 m	
Cell_id	Computed	300 m	
Date	Acquired		Daily

Tab 1: 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,

as shown in section 5.1.3.2. 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/m^2), 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.

5.3 Method

5.3.1 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 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 3.

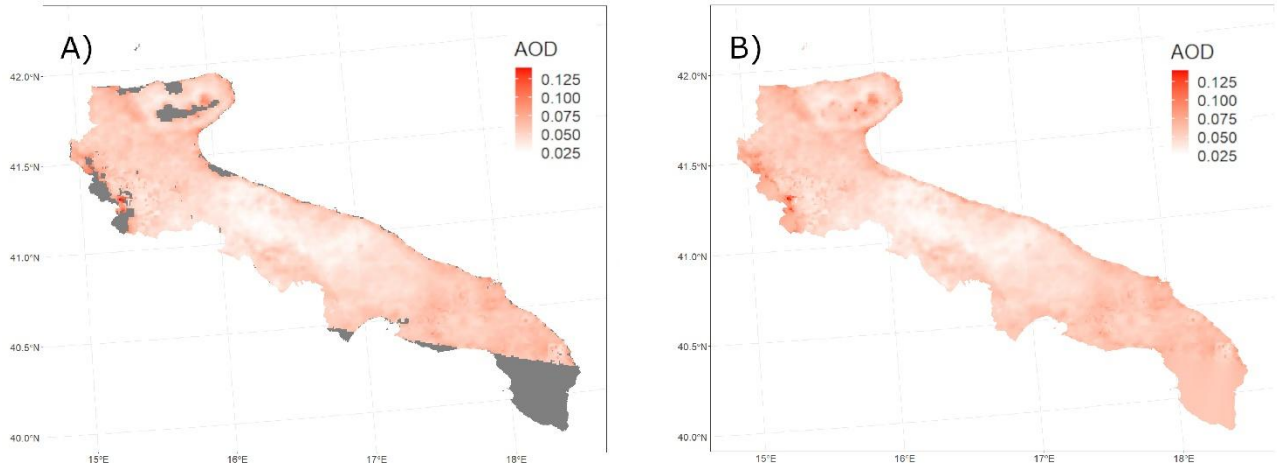


Figure 5-1 Left panel (panel A) AOD values extracted on the grid (the missing values are attributed to cloud covering and orbit of satellite), on the right panel (panel B) the same image in which NA values are replaced using the Inverse Distance Interpolation method

5.3.2 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.

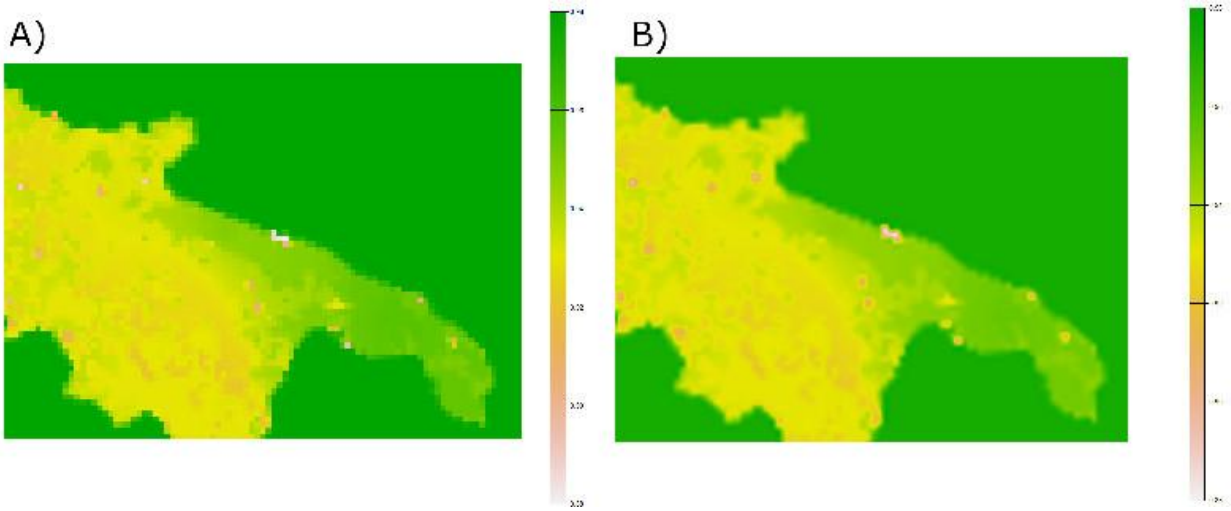


Figure 5-2 Left panel (A) Surface Emissivity predicted by the WRF model (4 km resolution), on the right panel (B) the resampled image to 300 m with bi-cubic interpolation method.

5.4 Results

We obtained a final training dataset of 45.000 rows and 27 columns for PM2.5 and a dataset of 88.000 rows and 27 columns for PM10. The temporal period covered is 2019-2022, with a daily temporal resolution and a spatial resolution of 300 meters. These datasets will be used to fit a Machine Learning model able to predict ground-level PMs concentrations. The reference grid on which make prediction has a cardinality of 250.000.000 total observations. Also, our objective is to highlight the most important variables to the prediction of PMs, using an eXplainable Artificial Intelligence (XAI) method.

6 Intra-urban scale approach

6.1 Overview of the employed predictors

6.1.1 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 ricampionata 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
Sentinel5/5P TROPOMI	L2__AER_AI, L2__AER_LH	Level2	3.5 x 5.5 Km	variying	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.

6.1.2 Meteorological parameters

Meteorological parameters (daily mean air temperature, sea-level barometric pressure, precipitations, relative humidity, wind speed and direction) and planetary boundary layer height were retrieved by the local running of WRF model.

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

6.1.3 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 rioxarray one.

6.1.4 Other spatiotemporal data

We computed a number of 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.

These include demographics census data yearly provided now from the National Institute of Statistics' permanent census survey. An annual update of this dataset was used to generate a 100x100 m population density map for each year in the period of study using a QGIS plugin developed in two previous H2020 projects: SMURBS and e-SHAPE.

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.

6.2 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. These are heterogeneous data, and our initial efforts have been devoted to intensive downloading and interpolation of the various products to a common resolution of the Sentinel 3 synergy product: 300m x 300m grid. Different data required different spatial aggregators to be resampled to the 300 x 300 grid resolution.

Table 6-1: Type and source of data.

TYPE	PROVIDER	SOURCE	SPATIAL RESOLUTION	UPDATE FREQUENCY	AGGREGATOR	SIMILAR DATASET
Meteo	ARPA PUGLIA	WRF Reanalysis model	4 km (Bari); 1km (Taranto)	hourly	Spatio-temporal Nearest Neighbour	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
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 Copernics

Meteorological data with temporal updating frequency less than daily were temporally aggregated according by computing the mean value of all the data available from 8 to 12 a.m. This outcomes a daily data for each variable of interest.

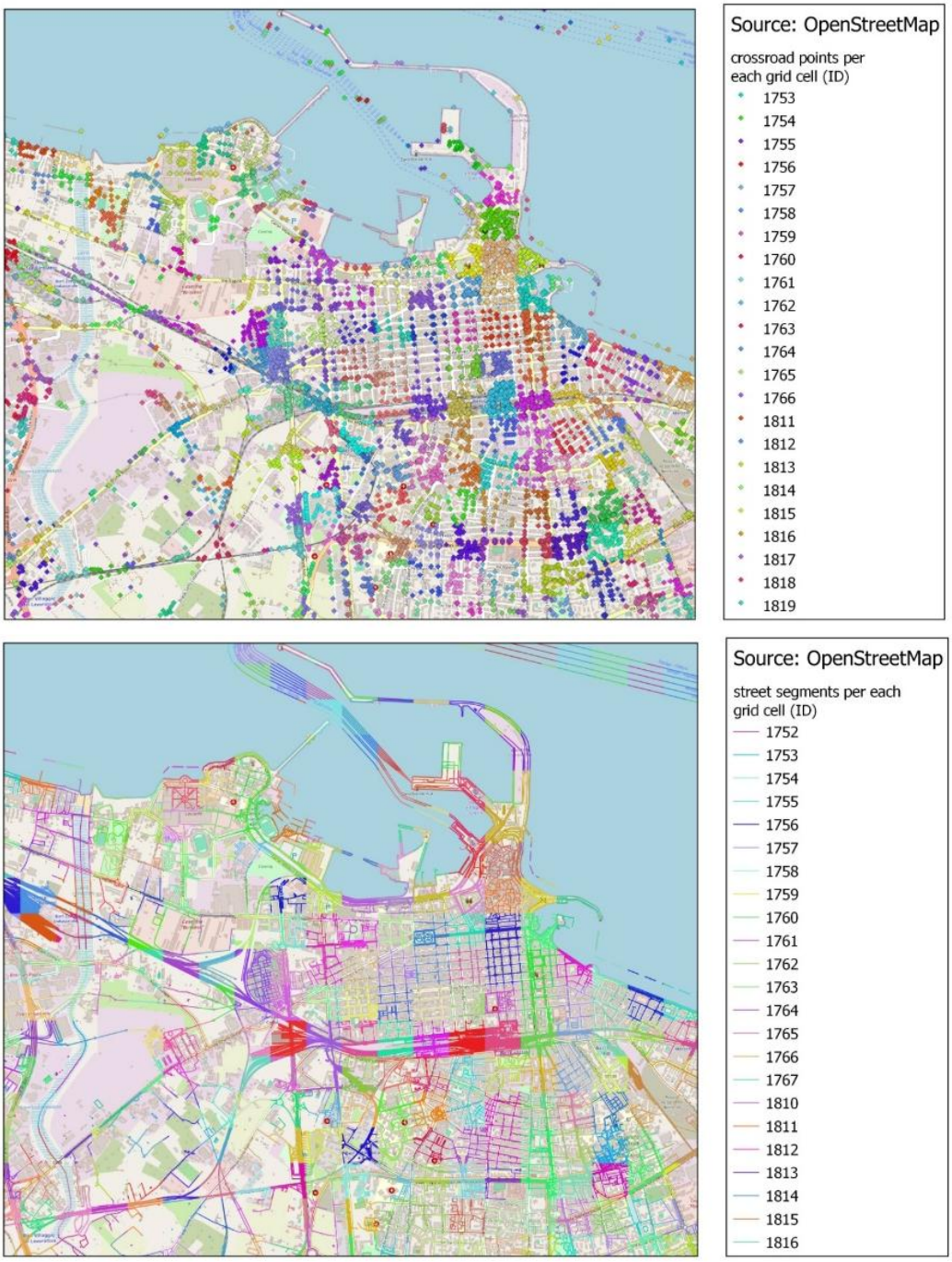


Figure 6-1: Traffic indicators from OpenStreetmap

6.3 Preliminary implementation stages

The implementation steps involve multiple processes that have been translated into open-source code, enabling the automated generation of training datasets for ML algorithms.

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 will be 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 could be shared on a public hub and on the project's official GitHub page: https://github.com/APEMAIAproject.

The architecture of the PM retrieval software (not including AOD improved procedures) is organized 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.
- **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, analyzing trends, and exploring exposure levels.

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 data: It integrates WRF output variables with other datasets (e.g., AOD, geographical data) using a modular and efficient approach.

- Preparing an updated and enriched dataset for use in **3_DATABASE_building**, facilitating the generation of daily data products for subsequent machine learning analyses in **4_ML_models**.

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

6.4 ML models

A comparative analysis of RF and XGBoost was carried out for PM10 in Bari study site. Currently, the ML models training dataset comprises PM reference data and predictors collected for 5 stations in urban/suburban areas, spanning the data acquisition period from 2020 to 2021.

Regarding the city of Bari, the maximum possible intersection of the data available to us has been adopted as an extension of the analysis. The spatial resolution of the Sentinel 3 data at 300 m has been used as a reference for creating a 300 * 300 m grid. The grid has been saved in vector format (generated as a shapefile) to facilitate extraction into interoperable formats (such as CSV, GeoJSON) for different programming environments (e.g., Python, RStudio, etc.).

At this point, it is appropriate to imagine working in a multidimensional space, where in addition to the spatial grid and the feature dimensions for model training (predictors), there is an additional dimension to consider: time. This includes the list of daily PM measurements within the considered interval (2019-2022).

Essentially, our training/validation dataframe initially comprises the list of temporal measurements (validated daily averages) of PM2.5/10 provided by ARPA, corresponding to measurements in the grid only at the positions where monitoring stations are located.

Given the grid, the raster predictor datasets at lower resolution of the reticulate were associated with the grid elements through a Nearest Neighbour interpolation based on the criteria of geographic position (Lat, Lon) and Timestamp (closest value in time to the date of the PM sample used in the training). This is the case for ERA5 meteorological data from Copernicus Climate Data Services, available at a resolution of 9 km. This fundamental dataset includes wind components (v and u), 2m temperature, 2m dewpoint temperature and boundary layer heightand. It has been processed in Python using XArray library.

Higher resolution raster data are aggregated by reporting the mean value (zonal statistics libraries). This applies to vegetation indices calculated from Sentinel 2 provided by ESA Copernicus. Vegetation

indices, such as NDVI, EVI2, MSAVI2, NBR, NBR2, NDWI, NDSI, NDMI, and BSI, were computed from Sentinel 2 A/B time series data at 10 m spatial resolution.

Vector data are also aggregated at the grid cell resolution by calculating the mean, sum, minimum, and maximum values. Depending on the case, the appropriate metric is chosen. For example, for road nodes (point vector), the sum of the points in each grid cell is used. For road density, the lengths of individual segments falling within each grid cell are summed.

The initial set of reference PM data was divided into training (80%) and test (20%) sets for model evaluation. The performance of both models was systematically assessed across different numbers of estimators. In this preliminary investigation, both RF and XGBoost exhibited promising outcomes, with XGBoost showcasing superior performance.

For RF with 100 estimators, the model yielded a coefficient of determination (R2) of 0.66 and a root mean square error (RMSE) of 6.6 µg/m³. Notably, XGBoost demonstrated a significant improvement; utilizing 110 estimators, the model achieved an elevated R2 of 0.72 while concurrently reducing the RMSE to 6.0 µg/m³. These findings suggest that XGBoost outperformed RF in terms of accuracy and precision in PM estimation.