University of Milano-Bicocca
Department of Informatics, Systems and Communication
Master's degree program in Data Science

Predicting PM10 concentration using geo-statistical machine learning models

Supervisor: Prof. Matteo Maria Pelagatti

Co-supervisor: Dott. Paolo Maranzano

Master's degree thesis by:
Samir Doghmi
897358

Academic Year 2023-2024

Contents

Abstract

Air pollution is an environmental and public health issue, with particulate matter having a significant impact on human health with long-term exposure that shows to cause several diseases. Therefore, forecasting has also become important to guide individual action and preventive measures, especially in urban contexts, like Milan where pollution levels are increasingly high. This paper aims to forecast daily PM_{10} levels at unobserved locations not covered by official monitoring stations or low-cost sensors using geostatistical machine-learning models. Specifically, we apply General Additive Models and Spatio-temporal Kriging, incorporating meteorological variables and land cover data. By integrating heterogeneous data sources, our approach aims to improve air quality assessment and provide valuable insights into the spatio-temporal variability of PM_{10} levels across different areas of the city.

1 Introduction

Air pollution represents one of the major environmental and health challenges of our time. Particulate matter is one of the most harmful pollutants, causing many serious health effects. In urban environments, where human activities (vehicular traffic, industrial processes, heating) generate high concentrations of particulate matter, it becomes essential to monitor and predict PM_{10} levels to protect public health and promote a more sustainable management of cities. Many studies have shown that northern Italy has a problem with air pollution and recently, IQAir, a Swiss real-time air quality website, ranked Milan as the third most polluted city in the world [13]. Air quality forecasting is crucial for guiding individual actions to limit exposure to PM_{10}, such as choosing between outdoor and indoor activities. However, air quality forecasts are affected by spatially and temporally complex factors [19]. The classical data-driven approach to predicting air quality is the time-series estimation methods, such as AutoRegressive Integrated Moving Average (ARIMA) [70] which captures the temporal dependencies but fails to handle the spatial correlation. To analyze its dynamics across both time and space, we employ a spatiotemporal modeling approach that captures the characteristics of PM_{10} and helps track its diffusion across the city. In particular, we use geo-statistical machine learning models that don't require strict prior assumptions about the data distribution. We employ Generalized additive

models (GAM), which are regression models in which smoothing splines are used instead of linear coefficients for covariates. This approach has been found particularly effective at handling complex non-linear associations within the air pollution field [71] and [72]. Comparatively, Several studies have explored spatiotemporal kriging for urban air pollution estimation in small areas. For instance, Halimi et al., 2016 [73] applied geostatistical interpolation techniques to estimate air pollution in Tehran, comparing various kriging methods. This method relies on spatiotemporal variogram functions to model autocorrelation structures. In our study, Ordinary Kriging (OK) and Universal Kriging (UK) techniques are compared after generating continuous air pollution estimates across space and time. Therefore, we forecast daily particulate matter levels over the entire area of Milan using a framework based on data from ARPA's ground stations and low-cost sensors.

This study proposed a general predictive model for air quality comparing the results of different approaches based on the same covariates. To do so we incorporated various information from monitoring location, meteorological factors, and spatial context around the stations. Model performance is assessed using leave-location-out cross-validation. Then, the forecasts are made on a regular grid of points spaced 500 meters apart within the municipality of Milan, ensuring high-resolution spatial coverage of air quality levels.

The experimental results provide insights into PM_{10} diffusion throughout the city, highlighting the neighborhoods of Milan where the air is more or less polluted.

The paper is organized as follows. In Section 2, we provide an overview of the problem description. Sections 3 and 4 describe the study area and the data used in our analysis, respectively. In Section 5, we propose various methods to handle missing data. In Section 6, we detail the models used, including their formulations. Sections 7 and 8 present the experimental studies and results. In Section 9, we give conclusions and indicate future work.

2 Problem Description

2.1 Urbanisation and Smart cities

Urbanization is one of the most prominent trends of the past century and a key driver of development. Across high-income countries, more than 80% of the population lived in urban areas [1]. According to OECD [2], the spatial concentration of the population in metropolitan regions is expected to continue over the next two decades, reinforcing the past trend.

In this context, it has become crucial for policymakers to control and better understand the cities they govern. The increasingly complex socio-economic processes and rapid changes demand flexible and suitable answers that ensure effective governance. The concept of "Smart city" tries to respond to this need. A "Smart City" leverages various information technologies - including the Internet of Things (IoT), cloud computing, big data, and artificial intelligence (AI) - to facilitate the planning, construction, management, and smart services of cities. The emergence of a network of sensors, cameras, cable, and data centers allows city authorities to deliver essential services more quickly and efficiently [5]. The major goal of Smart Cities is improving the quality of life, reducing social inequality, protecting the environment, and promoting a more sustainable economic growth. These objectives are condensed in the eleventh United Nations Sustainable Development Goal: "Make cities and human settlements inclusive, safe, resilient and sustainable".

In this scenario, sustainability is a critical concept of societal health: Urban sustainability focuses on the persistence of a desirable outcome of urban environments over time [8]. Achieving this requires the efficient use of naturally available environmental resources and smart urban planning. As a result, sustainable cities are cities that manage to maximize energetic efficiency, reduce waste and pollution, support renewable energy, promote sustainable mobility, and preserve urban ecosystems.

Given the shift toward a more people-centric view, the need for smart governance has also become important. Greater citizen participation, access to public information, and the growing acknowledgment of a city as a complex system of system [4] has led to the recognition of the central role of the social, economic, and both institutional and not-institutional forces. With the advancement of ICTs, cities have more tools to improve

certain aspects of the urban environments. On this logic, many institutions improved government services by reducing information asymmetries and promoting open data initiatives, guided by the principle that on the guideline citizen empowerment leads to more virtuous behavior. Information and awareness-raising measures become relevant to encourage more appropriate and positive lifestyles, facilitating urban interventions at the neighborhood scale. These efforts also open up opportunities for collaboration with private agents, researchers, and non-profit organizations, aiming to enhance and make more efficient traditional networks and services through digital solutions.

2.2 Urban air pollution and Particulate matter

City expansion and increasing complexity bring several risks to urban societies, especially in terms of environmental impacts. These constant pressures such as burning fossil fuels, residential heating and in general human footprint without proper accountability on possible effects, have triggered negative impacts on the environment. Air pollution stands out as one of the most pressing issues, with significant health and environmental consequences. The Institute For Health Metrics and Evaluation (IHME) ranked it as one of the leading threats to global health, second only to high blood pressure [25]. The growing threat of climate change and environmental degradation has pushed the adoption of measures and international compromises, such as the Paris Agreement (2015) to strengthen the global response to climate change and reduce anthropogenic emissions. Climate change influences air pollution by altering the frequency, severity, and duration of heat waves, air stagnation events, precipitation, and other meteorological conditions favorable to pollutant accumulation [9].

Exposure to pollutants significantly affects the socio-economic sphere, leading to increased mortality rates, higher healthcare costs, and reduced productivity [46]. According to the World Health Organization (WHO), 4.2 million deaths are attributed to ambient air pollution each year [3]. it also estimated that in 2010, the annual economic cost of premature deaths from air pollution across the countries of the WHO European Region stood at US$ 1.431 trillion [10].

Particulate matter (PM) is one of the most harmful pollutants, causing many serious health effects. It's an air-suspended mixture of solid and liquid particles that vary in

number, size, shape, and origin. It is a pollutant of great concern because of its negative effects on human health [47]. Short-term and long-term exposure contributes to disease through an increase in mortality rate, years of life lost and years lived with disability [11]. The most common indicator of fin PM are $PM_{2.5}$ - particles with an aerodynamic diameter equal to or less than $2.5\mu m$ - and PM_{10} - particles with a diameter of equal to or less than $10\mu m$ are primarily originated by anthropogenic activities in urban areas [26].

In response to the growing number of local monitoring sites and mounting evidence of the harmful effects, the WHO updated its health-based guidelines for outdoor air quality in 2021. The new guidelines significantly lowered the recommended limits for average daily concentrations of particulate matter, setting the target for $PM_{2.5}$ at 15 $\mu g/m^3$ and PM_{10} at 45 $\mu g/m^3$. More stringent annual limits were set with PM_{10} at 15 $\mu g/m^3$ and $PM_{2.5}$ at 5 g/m^3. The spatial and temporal concentration of these pollutants in outdoor air varies according to the spatial distribution of the sources and their pattern of operation (e.g. daily or seasonal), the characteristics of the pollutants, and their dynamics (dispersion, deposition, interaction with other pollutants), and meteorological conditions [7].

The deterioration of air quality questions the urban sustainability of smart cities, requiring a data-driven approach to effectively mitigate emissions at their source and capture pollutants in the air, while protecting public health.

3 Area of study

Delving into the details of our project, Milan is the largest and most populous city in northern Italy, with approximately a population of 1.4 million inhabitants in the municipality and over than 3 million inhabitants in the metropolitan area. Currently, based on a ranking by the European Environment Agency, which orders cities from the cleanest to the most polluted depending on levels of $PM2.5$, the Metropolitan area of Milan is ranked 334th with a concentration of 19.7 $\mu g/m^3$ and it's in last place among cities with at least 1 million inhabitants [69].

Figure 1: Administrative boundaries of Lombardy region And Milan

The air quality in the city results from the interaction of heterogeneous factors. Geographically, it's located in the center of the Padan plain in the Lombardy region, surrounded on three sides by large mountain ranges that severely restrict the circulation of large air masses. The climate of the Padan plain is therefore continental, characterized by rather cold winters, hot summers, infrequent rainfall, and generally high relative humidity.
The annual trend in PM_{10} concentrations, like the other pollutants, shows a pronounced seasonal dependence, with higher values in the winter period, due both to the poorer dispersive capacity of the atmosphere in the colder months and due to the anthropogenic emissions [12].
The latest update of the regional inventory of atmospheric emissions, managed by ARPA

Lombardy, is dated 2021 and shows that road transport is the most relevant source of particulate matter emissions in the metropolitan area of Milan with a combined increase of 12% since 2019, as shown in Table 1. The primary contributors are emissions from diesel as well as vehicle brake, tire, and road surface wear. Followed by the "other sources" macro-sector, which has risen by 49%, including urban fires, wildfires, and the non-industrial combustion macro-sector (-7%), with residential heating, particularly from the consumption of woody biomass. Overall, emissions have increased by 0.52% from 2019.

Macro-sector description	PM10	PM2.5
Combustion in energy and trasformation industries	0.97%	1.21%
Non-industrial combustion plants	21.96%	26.80%
Combustion in manufacturing industry	6.76%	6.49%
Production processes	4.11%	3.63%
Solvent and other product use	6.45%	6.93%
Road Transport	44.50%	38.03%
Other mobile sources and machinery	1.59%	1.98%
Waste treatment and disposal	0.09%	0.11%
Agriculture	0.71%	0.33%
Other sources and sinks	12.86%	14.47%

Table 1: Emission sources of the main pollutants.ARPA, 2021

To address the persistent issue of air pollution and its impact on public health, the European Union has implemented a series of directives to establish a harmonized approach to air quality management and limit concentrations of the most harmful substances. These directives define standardized monitoring methods, assessing, reporting, and exchanging air quality modeling methodologies across member states.

Eu Legislation, through Directive 2008/50/EC set a daily mean value limit for PM_{10} at 50 $\mu g/m^3$, permitting up to 35 days of exceedance per year, and an average annual value of $PM_{2.5}$ at 25 mug/m^3, without a daily binding constraint. Also, the EU has strengthened air quality standards closer to the guidelines of the World Health Organization (WHO) by 2030 with the possibility of derogation until 2035. These updated standards will reduce the PM_{10} daily mean value at 20 $\mu g/m^3$ and $PM_{2.5}$ at 10 $\mu g/m^3$. In addition, The European Commission adopted the EU action plan: Towards Zero Pollution for Air, Water and Soil' to reduce pollution to levels no longer considered harmful to health and natural ecosystem, complementing the 2050 climate-neutrality goal in synergy with the

clean and circular economy and restored biodiversity goals. To achieve these ambitious objectives, the Commission has proposed revising existing legislative frameworks by introducing stricter requirements across several key sectors, including agriculture, industry, transport, and buildings. These revisions aim to address pollution at its source, promoting cleaner technologies, sustainable practices, and enhanced monitoring mechanisms [48].

3.1 PMs impact on communities

Short and long-term exposure to air pollution can lead to a wide adverse health effects depending on the duration, concentration of the exposure and the health status of the affected population. The people most affected by the health consequences of poor air quality are those who spend more time outdoors and are physiologically more vulnerable: children, the elderly, those with chronic illnesses, pregnant women, and unborn children. Focusing on individuals under 20 years of both sexes, the death rate due to air pollution exposure decreased by 86.55% between 1990-2021 but still remains the second cause of death as reported by IHME [25].

However, the quality of life for the entire population is compromised. World Health Organization (WHO) has found evidence of a relationship between particulate matter and lung cancer risk [15]. Based on this acknowledged evidence, they classified it as carcinogenic to humans (Group 1), as it is associated with an increase in genetic damage [17]. Short-term exposure to PM_{10} has been associated with the worsening of respiratory diseases, such as asthma, while the effects of long-term exposure are less clear. $PM_{2.5}$ with a less diameter is much more harmful because it can reach even the deep airways; both short and long-term exposure are associated with premature mortality, chronic bronchitis and heart causes [16].

City - Urban Audit Cities (LAU)	Tot Premature deaths	Premature deaths per 100,000 inhabitants	Years Of Life Lost
Brescia	356	154	3160
Milano (greater city)	4517	124	40100
Bolzano	90	84	798
Como	96	100	853
Gallarate	90	114	796
Bergamo	199	119	1764
Monza	214	125	1908
Milano	1882	127	16710
Varese	102	99	905
Busto Arsizio	115	115	1024
Saronno	56	118	494

Table 2: Premature deaths and years of life lost in 2021 due to $PM_{2.5}$ exposure in Lombardy

Table 2 shows Brescia as the city with the highest premature death roll per 100,000 inhabitants in 2021, based on data retrieved from the European Environment Agency. The years of life lost (YLL) metric highlights the potential loss of years of life, considering factors such as age and sex. While Brescia has the highest rate, the Metropolitan area of Milan records the highest absolute numbers of premature deaths and YLL. In addition to its impacts on human health and social well-being, this severe disease burden results in reduced labor productivity, increased medical costs, and higher hospitalization rates, which can put significant pressure on the national health care system.

3.2 Current perception of air quality information accessibility

A recent survey showed that Milan's citizens are aware of the quality of the air they breathe [40], yet 46% of them report that they rarely get information. The most common method for getting information is to query web search engines, followed by 31% who prefer online magazines, and 22% who directly consult institutional channels (Municipality website/di AMAT/ARPA). However, navigating the AMAT and ARPA websites (visited 21st August 2024), we discovered that the daily air quality reports had not been updated respectively from 30th April 2024 [41] and 21st March 2024 [42].

Moreover, according to Article 18 of Legislative Decree 155/2010 [43], reporting $PM_{2.5}$ concentration levels is not mandatory, unlike PM_{10}. Although the public is informed about PM_{10} level exceedances, both websites lack comprehensive information on air pollution exposure, symptoms to watch for, and recommended actions. This is significant,

as 81% of those surveyed believe it is crucial for Milan's municipality to inform both residents and visitors about air quality [40]. In addition to details on the impact on human health, over than 80% expressed the need for information on how to minimize environmental impact and the sources of emissions [40].

Instead, at the European level, the Environment Agency's European Air Quality Index (EAQI) allows users to understand more about air quality where they live in real time. By accessing the interactive map available on the EAQI website or related app, users can find information about local air quality based on data from the nearest air monitoring station. As shown in the following table, the index is computed on five pollutants ranging from 1 (good) to 6 (extremely poor). For each pollutant, the index is computed separately according on concentration levels, as defined by World Health Organization [7], using 24-hour mean concentrations for PM_{10} and $PM_{2.5}$, and hourly mean concentrations for NO2, O3, and SO2.

Pollutant	Index level (based on pollutant concentrations in µg/m3)					
	Good	Fair	Moderate	Poor	Very poor	Extremely poor
Particles less than 2.5 µm (PM2.5)	0-10	10-20	20-25	25-50	50-75	75-800
Particles less than 10 µm (PM10)	0-20	20-40	40-50	50-100	100-150	150-1200
Nitrogen dioxide (NO2)	0-40	40-90	90-120	120-230	230-340	340-1000
Ozone (O3)	0-50	50-100	100-130	130-240	240-380	380-800
Sulphur dioxide (SO2)	0-100	100-200	200-350	350-500	500-750	750-1250

	Health messages	
EAQI	General population	Sensitive populations
Good	The air quality is good. Enjoy your usual outdoor activities	The air quality is good. Enjoy your usual outdoor activities
Fair	Enjoy your usual outdoor activities	Enjoy your usual outdoor activities
Moderate	Enjoy your usual outdoor activities	Consider reducing intense outdoor activities, if you experience symptoms
Poor	Consider reducing intense activities outdoors, if you experience symptoms such as sore eyes, a cough or sore throat	Consider reducing physical activities, particularly outdoors, especially if you experience symptoms
Very poor	Consider reducing intense activities outdoors, if you experience symptoms such as sore eyes, a cough or sore throat	Reduce physical activities, particularly outdoors, especially if you experience symptoms
Extremely poor	Reduce physical activities outdoors	Avoid physical activities outdoors

Figure 2: The European Air Quality Index and related health messages for each band

The polluted bands are displayed on a color scale ranging from green (indicating good air quality) to purple (indicating extremely poor air quality), reflecting the relative risk associated with short-term exposure to human health. The overall EAQI for each monitoring station is determined by the highest value among the five individual pollutant indices. For better interpretability, the index bands are complemented by health-related messages that provide recommendations for both the general population and sensitive populations.

4 Data Sources and analysis

4.1 Air quality data

Air quality monitoring provides essential information regarding the status of the present air quality. Traditionally, Fixed monitoring stations are the most reliable and highly expensive method for obtaining accurate data. This study utilizes data obtained from ARPA of Lombardy, the Regional Agency for Environmental Protection responsible for local environmental protection and air quality monitoring, collected via the EEAq package in R [4] (accessed on 12 November 2024). The stations are strategically dislocated based on the Evaluation program (PDV), following Directive D.lgs. 155/2010 [27], to be as representative as possible of the air quality status of the area in which they are located. The deployment of fixed stations generally depends on the zoning of the territory and is a function of the population of each zone and the air quality status. Within the city of Milan, there are four urban ground stations located in built-up areas to monitor PM_{10} levels. These stations are classified according to their environmental context and the exposure of the general population. Three of these are traffic stations, mainly influenced by traffic emissions from neighboring roads, while the fourth is a background station situated in a location where pollution levels are not predominantly influenced by specific sources. To ensure the reliability of collected data, ARPA implements a rigorous validation process that includes instrument calibration, automatic validation, blind tests, and performance audits conducted by a dedicated Meteorological Office, which operates independently from the network management personnel. After passing multiple levels of validation, the air quality data are considered definitive and made available in the inventory as daily average concentrations, in conformity with regulatory limits.

Recently, with the development of modern technologies, low-cost sensors (LCS) have gained significant interest. These sensors provide air pollution monitoring at a lower cost than conventional methods and are housed in devices much smaller than ever. By using many sensors, they can offer new opportunities as complimentary monitoring resources for a higher granularity and more accurate spatiotemporal analysis as they offer high-time resolution data in near real-time. Also, these systems allow citizens and communities to monitor the local and personal air and raise environmental awareness in society [13].

However, LCS comes with challenges, including uncertain measurement quality, reliability issues, and inconsistent data availability. While official monitoring stations operate under controlled environments, LCS measurements are sensitive not only to the air pollutants of interest but also to a combination of external factors, including interfering compounds, temperature, humidity, and the limitations of the sensor technology itself. The World Meteorological Organization highlights the critical need for standardized testing protocols to evaluate sensor performance and the necessity of training individuals to ensure effective deployment and operation of these systems [7]. Data is obtained from SensorCommunity [14], a global collaborative project that aims to create a network of Do-It-Yourself (DIY) sensors to monitor air quality. This initiative was started by a community of developers and activists interested in collecting environmental data in an open and shared way. For our purposes, we considered data from seven sensors, each providing measurements at 10-minute intervals (retrieved on 12 November 2024). The website itself sells a complete kit to be assembled, welded, and programmed with all the necessary instructions for a price of around 50 euros [16]. For particulate matter measurements, the system employs the SDS011 sensor, which is known to exhibit increased measurement uncertainty under high humidity conditions [28].

However, the SensorCommunity platform provides only the general location (coordinates) of the sensors, not giving precise information on the type of sensor installed. Furthermore, no details are available regarding the specific conditions of the sensor placement, such as altitude and proximity to traffic. In both sources, the pollutant is measured as $\mu g/m^3$.

Station Name	Zone Type	Station Type	Altitude (m)
Milano v.Marche	URBAN	TRAFFIC	127
Milano v.Senato	URBAN	TRAFFIC	119
Milano Verziere	URBAN	TRAFFIC	119
Milano Pascal Citta Studi	URBAN	BACKGROUND	118
Sensor_24644	URBAN	NA	NA
Sensor_32399	URBAN	NA	NA
Sensor_40256	URBAN	NA	NA
Sensor_44216	URBAN	NA	NA
Sensor_50128	URBAN	NA	NA
Sensor_70169	URBAN	NA	NA
Sensor_22851	URBAN	NA	NA

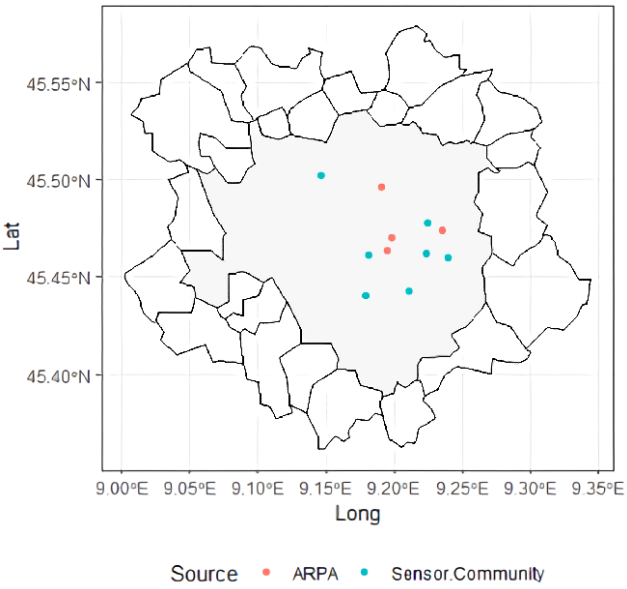

Table 3: Map of air monitoring stations and sensors

As shown in the graph, the spatial distribution of the stations is not uniform: air quality stations cover northern, eastern, central, and southern parts of the city, leaving the western districts uncovered.

To provide a preliminary overview of the analysis, over the two-year interval, both air monitoring stations and low-cost sensors do not show PM_{10} averages exceeding the EU daily limit of 50 $\mu g/m^3$.The STL decomposition reveals a decreasing trend in the concentrations for both types of sources, while the he remainder component exhibits a significant variability around the seasonal shift. The winter season, particularly the January–February window, is identified as the most critical period, with frequent spikes in PM_{10} levels exceeding the daily thresholds. Both air monitoring stations and low-cost sensors display similar seasonal trends, although the smooth line for low-cost sensors is systematically lower than that of the monitoring stations, potentially underestimating pollutant levels. This difference could be due to sensitivity calibration or the combination of other conditions.

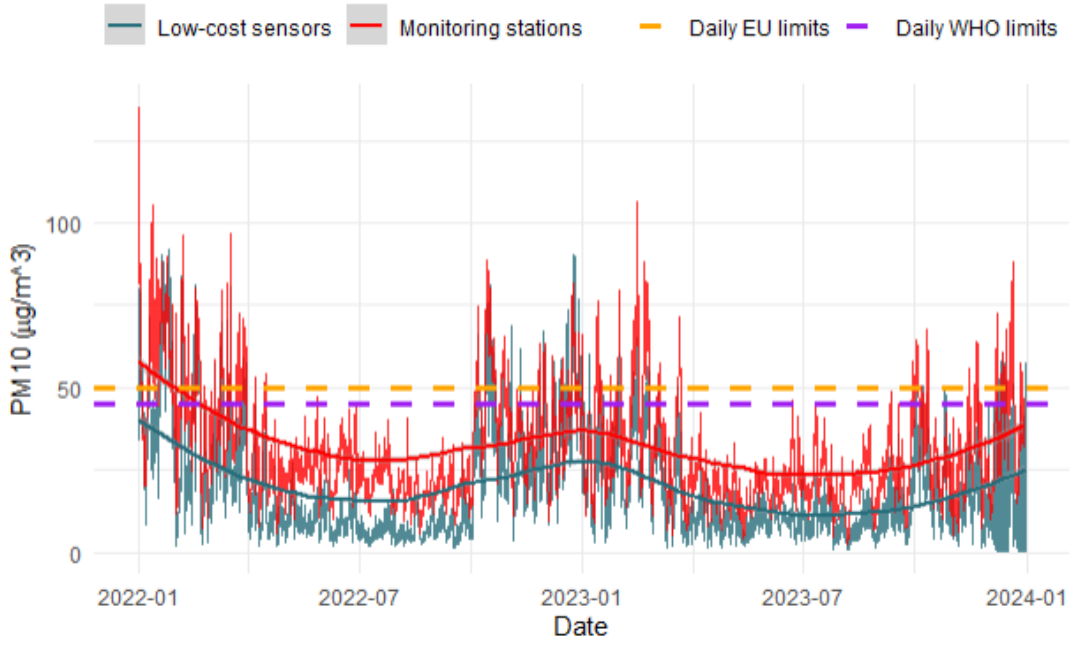

Figure 3: The time series graph shows the PM_{10} concentration levels recorded by both low-cost sensors and official monitoring stations throughout 2022-2023.

The quantitative analysis shows considerable differences between the stations. Government monitoring stations report an average PM_{10} concentration of approximately 31 $\mu g/m^3$, with traffic-type stations (e.g. Milano-Senato and Milano-V.LE Marche) recording the highest concentrations during the reporting period. In comparison, low-cost sensors report an average of 18 $\mu g/m^3$, while the degree of variability around the mean for all the factors stays comparable.

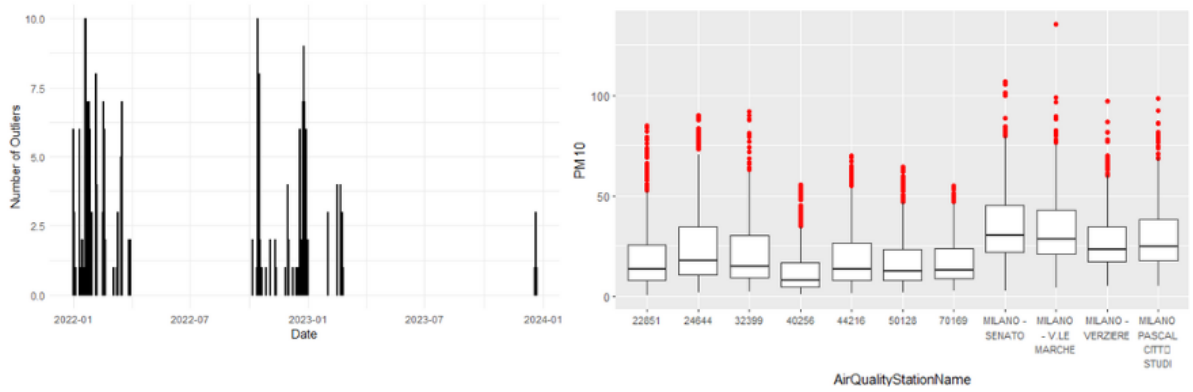

Figure 4: The left plot shows the number of outliers observed per day during 2022-2023 period, while the box plot highlight the variability in air pollution levels across different different stations

Also, Outliers are analyzed using the Interquartile Range (IQR) criterion, which identified 280 data points that exceeded the threshold. Such outliers are consistently observed at all stations, as visualized in the box-plot. In particular, specific days, such as 19 January 2022, show uniform outlier behavior across all stations. This phenomenon suggests an association with the meteorological drivers, as mentioned in the AMAT air quality reports, [18] that revealed unfavorable ventilation conditions during that period, leading to a significant accumulation of pollutants.

4.2 Meteorological factors

Meteorological conditions are the primary factor causing the day-to-day variations in pollutant concentrations and accumulation [19]. The data, sourced from ARPA, consists of hourly measurements. These parameters include hourly mean temperature (measured in degrees Celsius), relative humidity (measured in percentage), wind speed (measured in meters per second), wind direction (in degrees), precipitation (measured in millimeters per day), and global radiation (W/m^2). Similar to the previous arrangement of air monitoring stations, the spatial distribution of the monitoring stations remains uneven: they provide detailed coverage of the city center and the northern neighborhoods, but no stations are installed in the southern areas.

Station Name	Altitude (m)
Milano Lambrate	120
Milano P.zza Zavattari	122
Milano v.Brera	122
Milano v.Juvara	122
Milano v.Marche	129

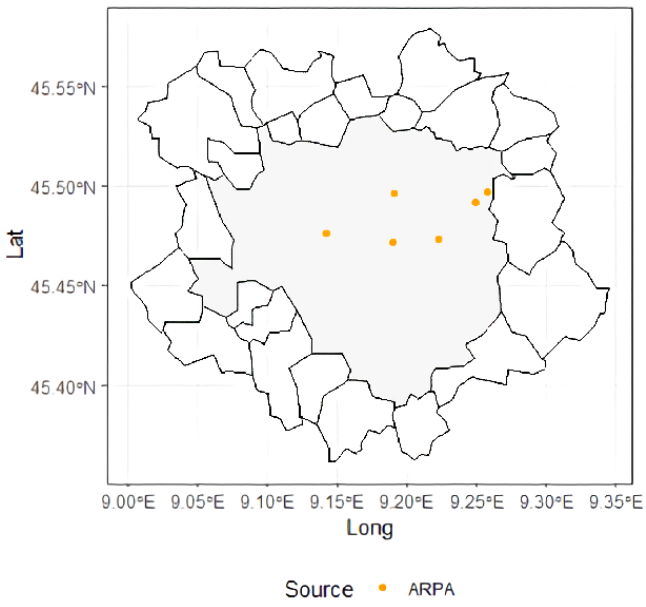

Figure 5: Map of meteorological monitoring stations

From the analysis of the following figure, it's clear a seasonality in temperature patterns, with high temperatures during summer and low temperatures in winter. The STL decomposition reveals several spikes, indicating unusual weather events. According to a report by ARPA Lombardia [20], the year 2022 was the hottest year ever recorded since temperature monitoring began, with data for Milan dating back to 1763. This year was also characterized by a significant rainfall deficit, which favored the accumulation of pollutants in the air.

Global radiation exhibits a trend similar to temperature, showing peaks in summer and declines in winter. In contrast, relative humidity remains consistently high throughout most of the year, without any clearly defined trend. This phenomenon is amplified by the geographic characteristics of the Po Valley, a region prone to air stagnation. During winter, particularly from November to February, dense fog is frequently observed. In summer, high humidity levels often occur when temperatures exceed 30°C, further intensifying the perception of heat. Precipitation, on the other hand, remains generally low and does not display significant trends. Rainfall is primarily concentrated in spring and autumn, whereas thunderstorms are common during hot and humid summers. Therefore, the climate of the Po Valley can be described as humid temperate subcontinental.

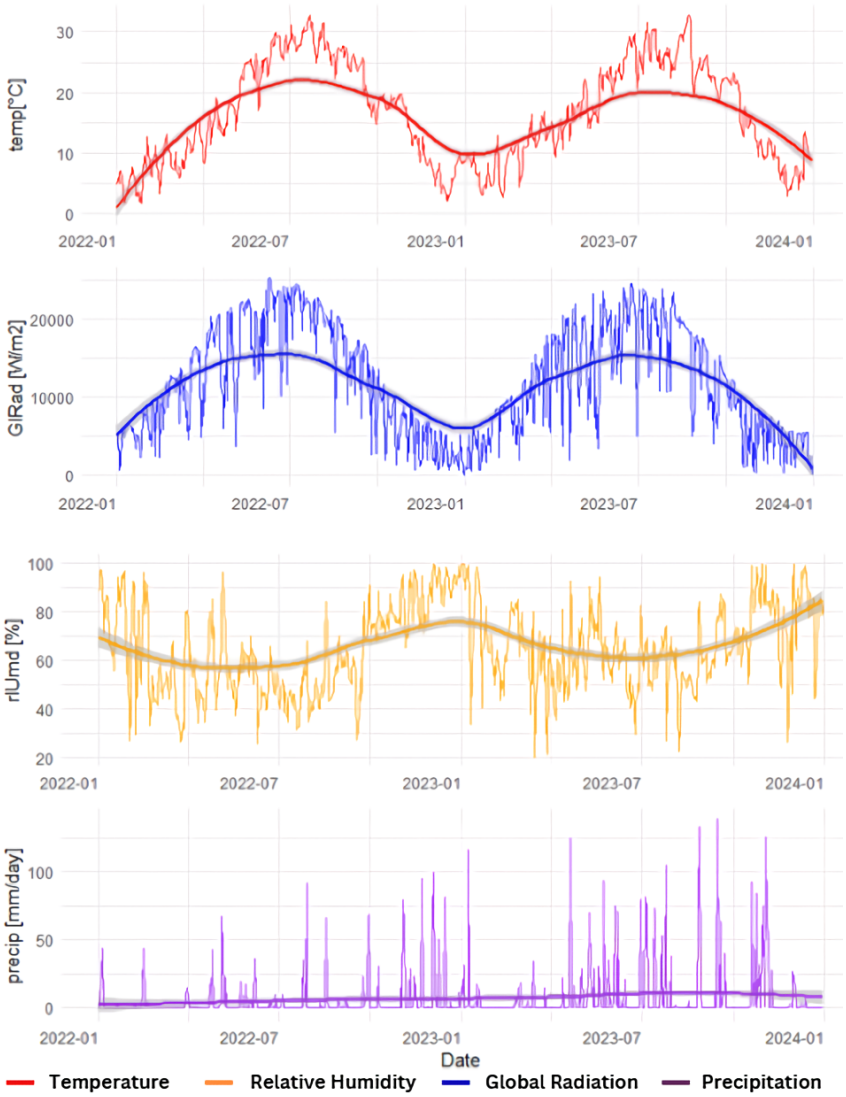

Figure 6: The plot shows the temporal distribution of Temperature, relative humidity, precipitation, and global radiation

Wind direction and wind speed play a crucial role in air quality monitoring. The effects of wind are complex and can take many different forms. For example, weak/calm wind (stagnation) can trap and accumulate pollutants, while moderate to light winds from different directions may introduce cleaner or more polluted air masses. In contrast, strong winds can quickly disperse air pollutants, reducing their concentrations. To analyze these effects in detail, we present a wind rose diagram constructed using hourly wind data, grouped according to astronomical seasonal divisions (defined by solstices and equinoxes). The Wind Rose consists of a circular plot divided into segments that represent the cardinal directions (e.g., north, south, east, west) of the compass. Wind speed is visually represented using a color scale, while the frequency of wind direction is indicated by the length

of the spokes.

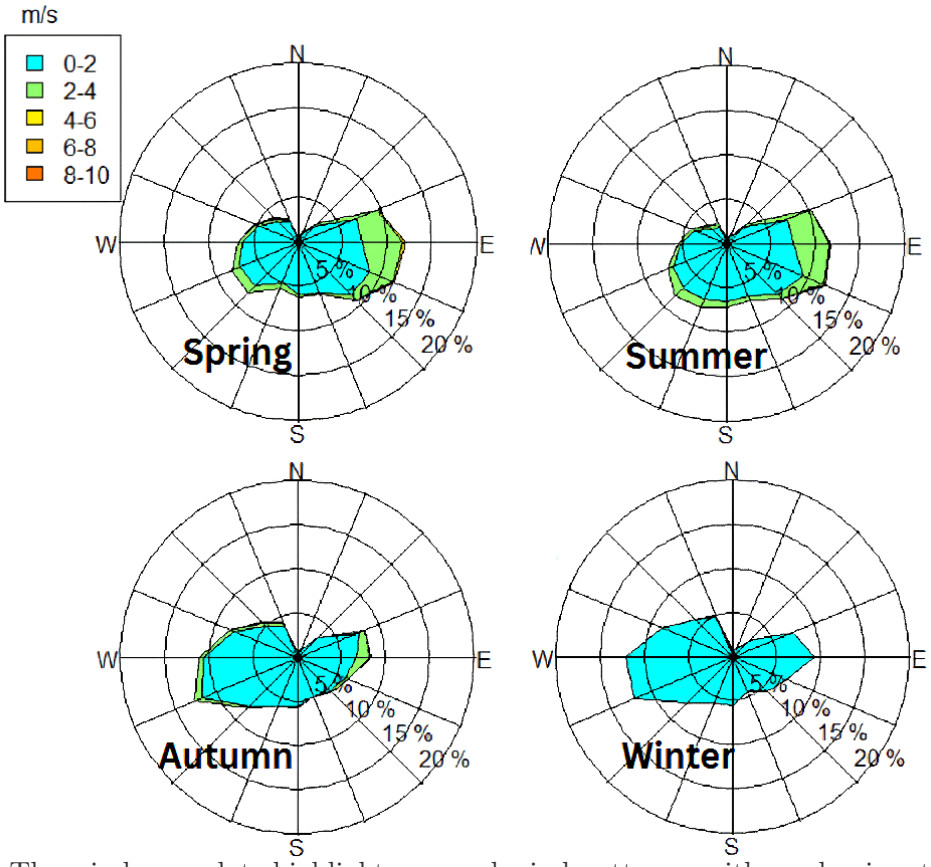

Figure 7: The wind rose plots highlight seasonal wind patterns, with predominant WS winds in autumn and winter and SE winds in spring and summer

During spring, wind predominantly blows from the East (E) and Southeast (SE) directions, with speeds ranging primarily between 0 and 4 m/s. This directional pattern accounts for over 15% of all recorded hourly wind observations. In summer, a similar wind pattern persists, with prevailing winds continuing to originate from the E and SE. However, during autumn and winter, a notable shift in wind direction is observed. The prevailing winds during these seasons predominantly come from the West (W) and Southwest (SW) directions. Wind speeds remain moderate, still ranging mostly between 2 and 4 m/s, but there is a marked increase in directional variability compared to the more consistent patterns observed in spring and summer.

The heat map reveals some expected relationships between meteorological variables. In particular, there is a strong positive correlation between global radiation and temperature. Moreover, the negative correlation of both temperature and global radiation with humidity reflects the well-known inverse relationship, in which hotter conditions are typically

associated with lower relative humidity.

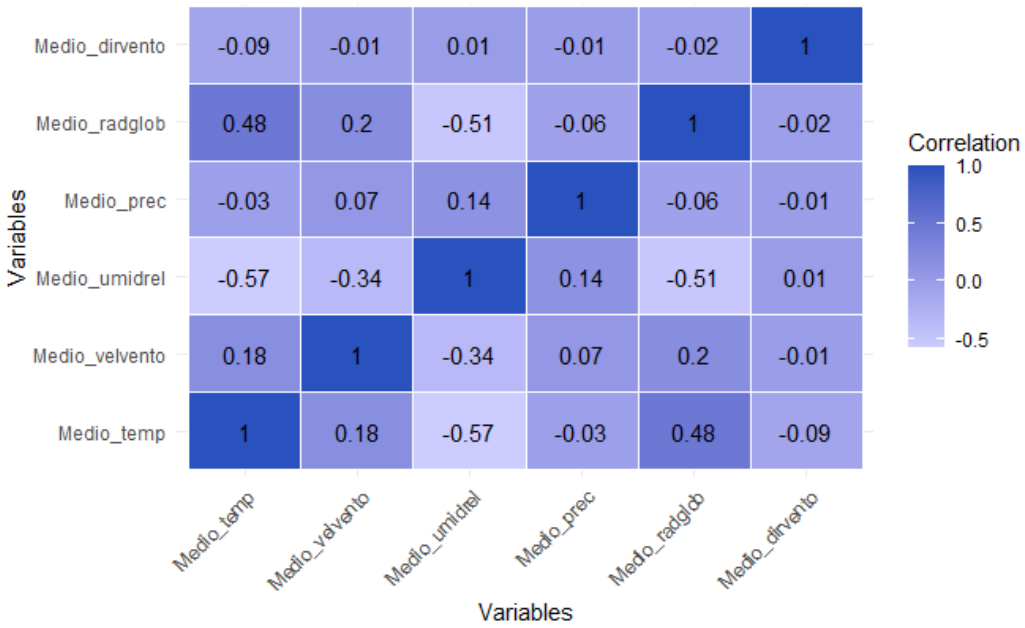

Figure 8: The correlation matrix of meteorological variables shows the bivariate dependency among variables. Interestingly, there is an inverse relationship between temperature and relative humidity.

In terms of differences in measurements of meteorological variables: precipitation, temperature, global radiation, and humidity are highly correlated between stations, indicating a consistent and uniform pattern across the region. However wind-related variables (e.g. wind speed and direction) show moderately high correlations ($r <= 0.75$) between stations, suggesting greater spatial variability. This indicates that these variables are likely influenced by localized environmental factors rather than large-scale atmospheric processes.

4.3 Land Cover

We obtained a fine-scale representation of the study area with CORINE Land COVER 2018 (CLC) provided by the European Union Copernicus program. The dataset is derived from satellite imagery with a spatial resolution of 100x100 meters (accessed on 15 November 2024), using the EPSG:3035 projection system, where each pixel is assigned the most dominant land cover class within that area. Changes are identified through direct mapping by comparing consecutive inventories using an image-to-image comparison methodology. The classification system consists of 3 main categories that are artificial surfaces, forest and semi-natural areas, wetlands and water bodies, Each one is further subdivided into several labels, for a total of 44 distinct classes.

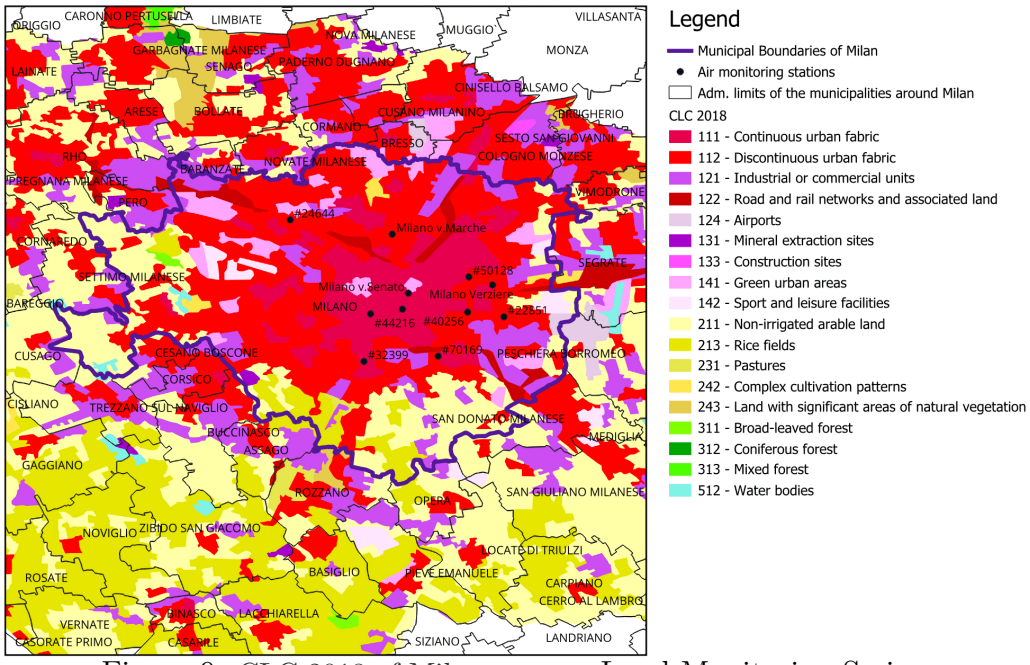

Figure 9: CLC 2018 of Milan source: Land Monitoring Serivce

From the map, we can observe that the center of Milan is primarily composed of continuous urban fabric. According to the definition, this class is assigned when more than 80% of the land surface is covered by artificial structures, such as impervious features like buildings, roads, and when the dominant land use type is residential. Traffic is available as an individual class 'Road and rail networks and associated land' but this is a very selective class as roads and related infrastructure have to cover a minimum area of 100 meters in width. Outside Area C, the territories are marked by the "industrial or commercial units" class and the "discontinuous urban fabric" class. The discontinuous urban fabric covers

areas where artificial surfaces range from 30% to 80% of land cover, and the presence of vegetation (such as gardens, lawns, or trees) is more evident, reflecting a more fragmented urban pattern. Most of the air quality stations are within these area classes, except for Milano V. Senato, which is situated in a green area. Moving southward, the landscape transitions into non-irrigated arable land that includes non-permanent crops and fields with sporadic sprinkler irrigation using non-permanent devices. Additionally, the presence of rice fields can be observed. We chose not to aggregate these classes into broader land use categories, as the observed classes represent the predominant land cover in the area.

To account for the spatial heterogeneity of land cover and its influence on air pollution, a buffer of 500-meter radius was generated around each air quality station. This distance was chosen to represent the local proximity of each station while avoiding excessive overlap between buffers. It was calculated by considering the correlation of particulate measurements among stations and sensors, as well as the fact that traffic air quality stations have at least a 200-meter representativeness, while background stations can cover up to 2 kilometers [22].

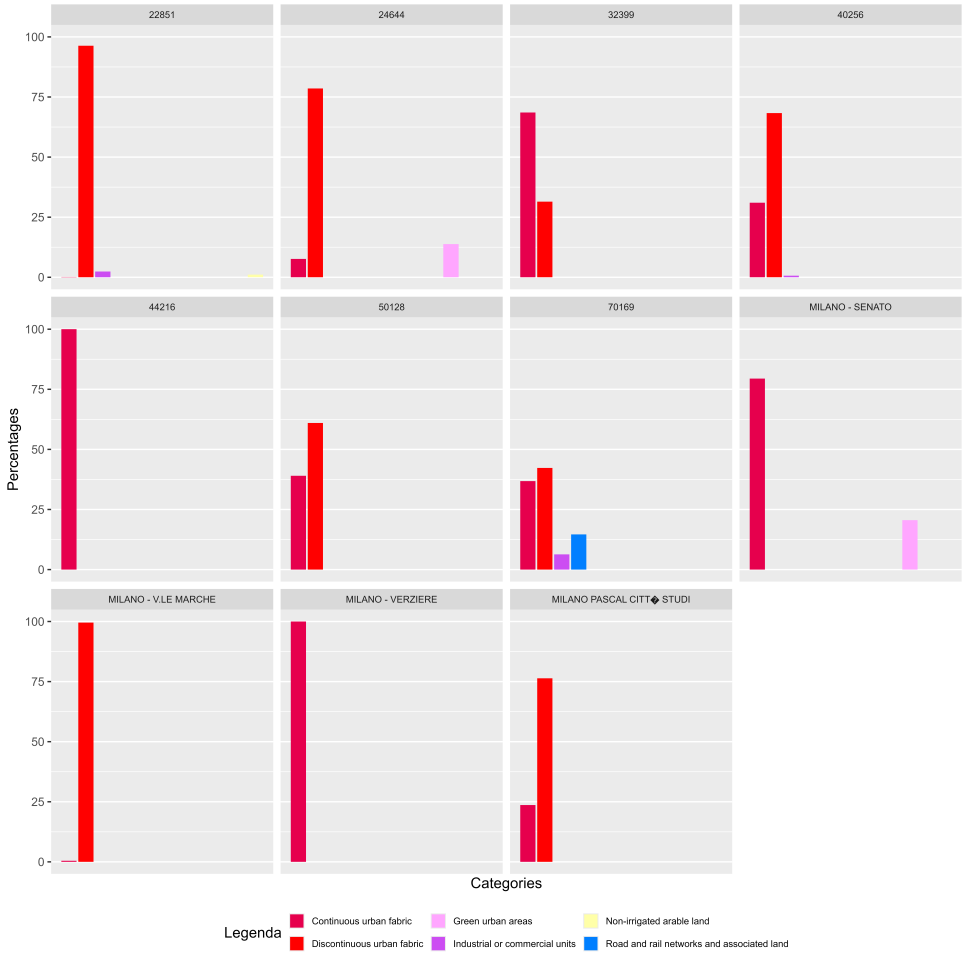

Figure 10: The bar graphs illustrate the composition of land cover within 500-meter buffers around the air quality monitoring stations. The predominance of continuous and discontinuous urban fabric at all stations emphasizes Milan's highly urbanized context

allowed us to summarise the raster values on polygonal areas, enabling the creation of bar graphs in which each represents a specific air quality station. In the figure above, the composition of the land cover within a 500-meter buffer for each station is shown. For stations Milano Senato, Milano Verziere, and Sensor 44261, which are located in the city center, the dominant land cover class is Continuous Urban Fabric, characterized by densely built-up areas, while for all other stations, the dominant class is Discontinuous Urban Fabric, describing the suburban areas of Milan. Stations 24644, located closer to the borders, includes small portions of non-irrigated arable land, pointing to a transition to agricultural land. This highlights the diversity of land cover in Milan, especially in areas further away from the city center.

4.4 Traffic regulations zones

When analyzing emission inventories, road transport, particularly the diesel fleet, has been the main contributor to PM_{10} emissions over the years. To tackle this issue, Milan introduced a Restricted Traffic Zone (ZTL), where vehicle access and circulation are allowed only during specific hours and for particular categories of users or types of vehicles. In 2012, the city launched Area C, which includes the city center and coincides with the ZTL 'Cerchia dei Bastioni', a zone monitored by 40 access points with cameras [23]. As a result of a popular referendum with 79% support, Area C has introduced restrictions on certain Euro emission classes and imposed a pollution charge, whereas free access is granted only to electric and hybrid vehicles, with some exemptions for residents (up to 50 free admissions), healthcare workers and people with disabilities [23]. A report by AMAT states that congestion has decreased by 40.9% since its introduction. Furthermore, Fasso [24] stated that the congestion charge paradigm has had a positive and lasting impact on air quality in the city center.

In February 2019, Milan introduced Area B, which extends to almost the entire urban area, covering 72% of the city's territory with 188 monitored access points. Unlike Area C, Area B focuses on reducing traffic emissions by banning the most polluting vehicle categories. Both areas are active from Monday to Friday, from 7:30 a.m. to 7:30 p.m., while they remain inactive on Weekends and public holidays. That same year, a ZTL was proposed to include the area surrounding the San Siro Stadium, but its activation has yet to take effect.

The overall reduction in atmospheric pollutant emissions following the activation of Area C is largely attributed to the technological renewal of vehicles and the growing adoption of environmentally friendly vehicles [74].

In 2023, the average daily transits in Area C amounted to 74,673, with most registered vehicles being passenger cars, predominantly Euro 6 models, which represent 89.4% of the vehicles entering the zone. Instead in Area B, the average daily transits were 613,380. According to a TomTom report [75], analyzing data from 389 cities, Milan ranks third in the Traffic Index, with an average travel time of 28 minutes and 50 seconds for a 10 km journey, behind only London and Dublin. Additionally, the congestion index—measuring the

percentage increase in travel times compared to free-flow conditions (nighttime)—shows significant peaks during the morning (8:00–9:00) and evening (6:00–7:00) rush hours, coinciding with the opening and closing times of Areas C and B. For traffic, data such as the number of daily transits for a specific road or the average time spent in traffic for our reference period are not freely available. Instead, as covariates, it was decided to consider time-invarying spatial variables, such as the type of area that provides more information on the location of each air monitoring.

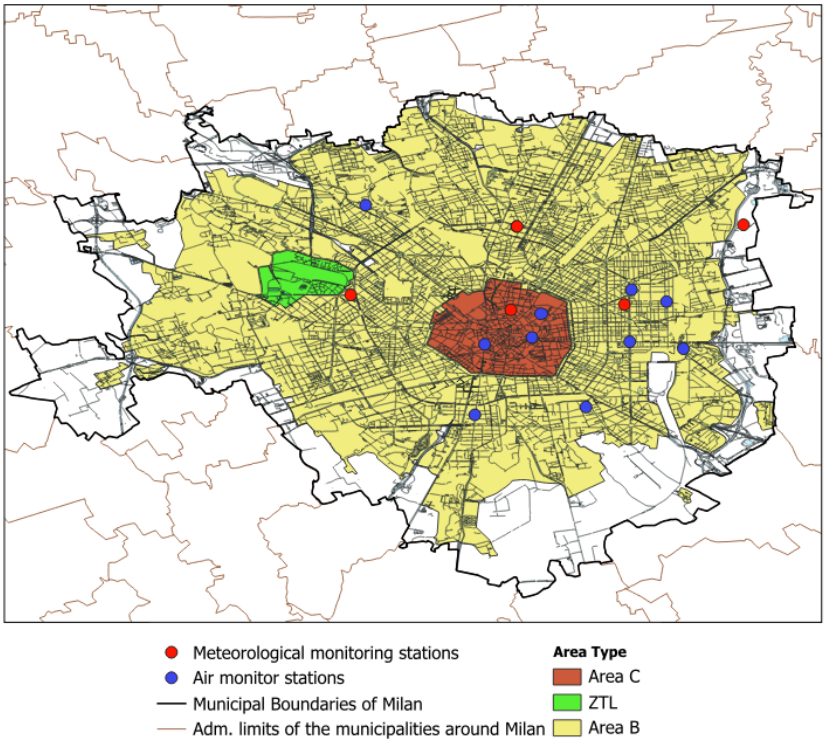

Figure 11: Traffic Regulation Zones in Milan

4.5 ERA5-Land

Meteorological data for areas not covered by air quality stations were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF),on 23 November 2024. Specifically, we utilized the ERA5-Land dataset [29], which is part of the Copernicus Climate Change Service. ERA5-Land is a gridded dataset derived by replaying the land component of ERA5, the fifth-generation global atmospheric reanalysis, with an enhanced spatial resolution of 0.1° (approximately 9 km) with hourly information on surface variables. It describes the evolution of the water, and atmospheric cycles over land providing hourly estimates for more than 70 years of historical atmospheric data,

producing a total of 50 variables. Several studies have validated ERA5-Land as a reliable data source by comparing it with ground-based observations. For instance, Yilmaz et al., 2022 [30] demonstrated consistency between temperature trends derived from ERA5-Land and station observations over Turkey, particularly for long-term trends and seasonal variability. In the region of interest, Maranzano et al., 2023 [31] created Agrimonia, a spatio-temporal dataset on livestock, meteorology, and air quality, further demonstrating the usefulness of ERA5-Land for research purposes. According to our methodology, the grid points that fall within the cell represent that area. If we assume a grid of 0.5 km spatial distance between points, at least 1 point will be representative of that area. To retrieve the data, we focused on the extended region of Milan, adding a buffer zone to define the longitude/latitude grid boundaries as [45.6°N, 9.0°E, 45.4°S and 9.3°W]. We derived the same meteorological variables obtained from the ARPA data and calculated the relative humidity using the temperature at 2 meters and the dew point temperature at 2 meters. The calculation used the August-Roche-Magnusapproximation formula [51]:

$$RH = 100 * exp\left(\frac{17.625 \times T_{dew}}{243.94 + T_{dew}} - \frac{17.625 \times T}{243.94 + T}\right) \tag{5.1}$$

To ensure uniformity between the various datasets, we converted all variables into consistent units before analysis.

4.6 Cams Data

The PM_{10} data used to construct the Lag1 and Lag7 variables at unobserved locations for the prediction step were retrieved from the CAMS European Air Quality Reanalyses dataset, provided by the ECMWF (on 20 January 2025) [32]. This dataset offers annual air quality reanalyses for Europe, based on both validated and unvalidated observations, with a spatial resolution of 0.1° (∼10 km) covering the entire municipality of Milan with 8 cells. The data are produced using 11 state-of-the-art numerical air quality models, ensuring high accuracy. In this project, we rely on the median ensemble observations, and, for each specific location, we simply extract the PM_{10} data by intersecting the point coordinates with the grid of the dataset. For December 2021 and the year 2022, we use validated data, while for 2023 we rely on Interim Reanalysis, as validated data are not yet available.

5 Handling Missing Data

5.1 Missing Data Mechanisms

Missing data in environmental monitoring is a common problem. It may occur for various reasons, including sensor failures and network outages, resulting in data sets missing significant periods of data measurements that can lead to a substantial amount of bias in the prediction phase. Missing data is, at root, a statistical problem and these gaps must be addressed appropriately as they can affect subsequent in-depth analysis work [33]. From the graph below, where the black lines represent the missing data, it can be seen that the stations operated by ARPA have missing data at random with an average of 31 days of missing values and an average gap size of 2.01%. On the other hand, the low-cost sensors exhibit sequential missing observations over some time.

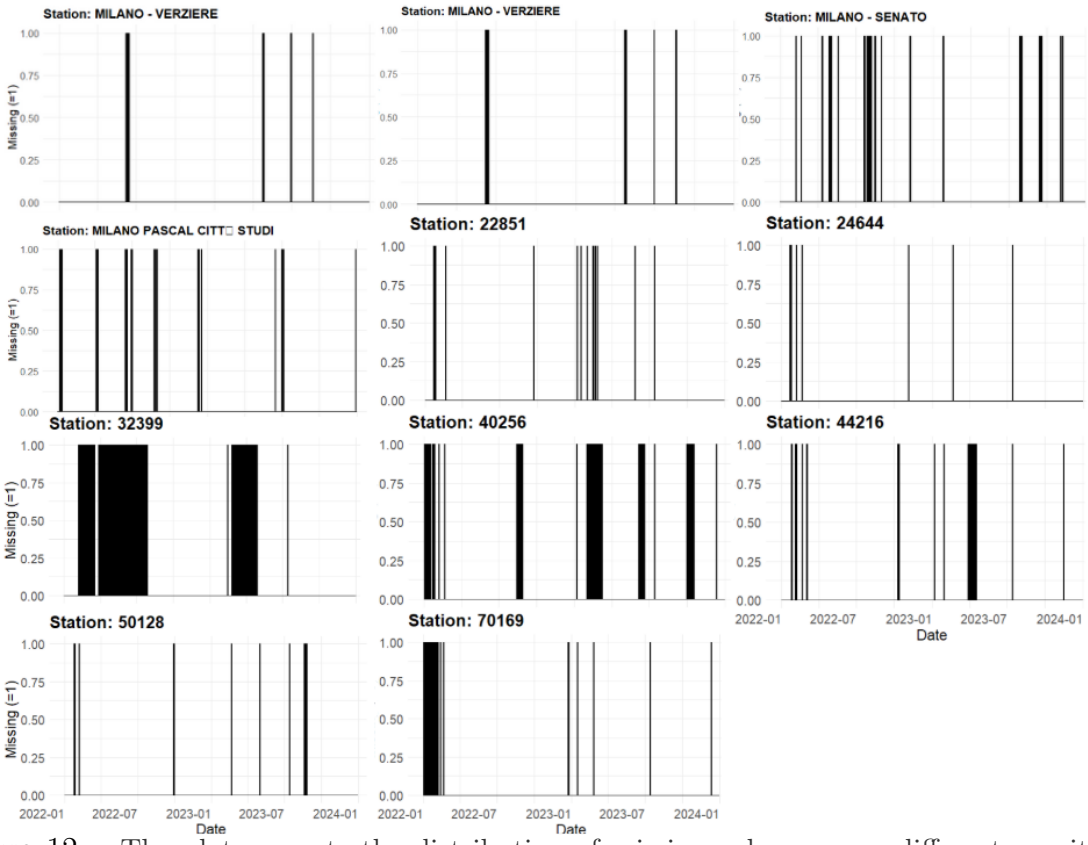

Figure 12: The plot presents the distribution of missing values across different monitoring stations. It is evident that the low-cost sensors exhibit a higher frequency of missing values, particularly for consecutive sequences

Specifically, sensors 32399 and 40256 have consecutive gaps of 122 and 38 missing values,

respectively. While several distributions of the NaN values seem to be randomly distributed, others show a repeated seasonal pattern, particularly during winter months for most low-cost sensors, indicating a possible dependence on the operating conditions of the instruments themselves, particularly during the winter months. However, considering the limited information available on sensor.Community, it is assumed that the mechanism of missing data is Missing Completely at Random (MCAR). This means that the probability of an observation being missing is independent of the other variables and the value of the observation itself.

Also, no significant differences in missing data patterns between ARPA stations and low-cost sensors were observed at monthly and daily aggregation levels.
For dealing with missing data, rather than using the deletion method, we will use the imputation method. Imputation methods for missing data can generally be divided into univariate and multivariate approaches. Univariate methods rely solely on the time series data measured by each station to impute missing values. While these methods can be effective in cases of sparse or isolated missing values, they struggle with the high proportions of missing data [34], as in our case, where missing rates range from 0.9% to 31.4% with the presence of large consecutive gaps observed. In contrast, multivariate methods, which leverage the relationships between different stations and other available sources of data (e.g., meteorological variables), are better suited to this context as highlighted by Liyanage et al., 2020 [34]. For this task, we focus on non-parametric approaches for missing data imputation, specifically exploring k-nearest neighbors (k-NN) model, the random forest-based model, and the multilayer perceptron (MLP) model. Additionally, we utilize the Mice (Multiple Imputation by Chained Equations) package [35], which implements multiple imputations to deal with missing data. This method imputes missing values for each variable sequentially, conditioned on all other variables in the dataset. Unlike traditional approaches that distinguish between training and test datasets, Mice uses the entire dataset for imputation, iterating over a specified number of cycles, generating m imputed datasets (typically 5) where each imputed dataset represents a plausible version of the complete data. In this study, we specifically consider the CART (Classification and Regression Trees) imputation method from the MICE package. Among the four methods employed, the normalization is applied only to K-NN and MLP models.

Then, to evaluate the effectiveness of these imputation techniques, we conducted a simulation study by generating different time series of data, considering the two types of particulate sources, and simulating various missing data ratios, reflecting the real scenarios. For robust validation, we use k-fold cross-validation, where the dataset is split into k subsets (folds) and the model is trained iteratively on $k-1$ folds while being validated on the remaining folds. This approach avoids the need for a strict partitioning between training and testing sets, while ensuring that all data are used in the training phase, especially given the small dataset. Once the model has imputed all missing values we assess the performance of the imputation methods using a standardized framework that involves three evaluation metrics:

Root Mean Square Error ($RMSE$)

$$RMSE = (\frac{1}{m})\sqrt{\sum_{i=1}^{m}(x_i - \tilde{x}_i)^2} \tag{5.1}$$

Mean Absolute Error (MAE)

$$MAE = (\frac{1}{m})\sum_{i=1}^{m}|x_i - \tilde{x}_i| \tag{5.2}$$

and **Coefficient of Determination ($R2$)**

$$R^2 = 1 - \frac{\sum_i (x_i - \tilde{x}_i)^2}{\sum_i (x_i - \bar{x}_i)^2} \tag{5.3}$$

where x_i represents the actual values, $\tilde{x}_i$ are the imputed values, and m is the number of missing observations. $RMSE$ measures the average magnitude of the error expressed as the difference between forecast and corresponding observed values. MAE is similar to $RMSE$ with the exception that the absolute value is taken, thus reducing the bias towards large outliers. The R^2 evaluates the proportion of variance in the observed data explained by the imputation model, with values closer to 1 indicating better performance. After identifying the best-performing model, we retrain it using a combination of original data and synthetic data to ensure optimal imputation performance. Given the strong inter-

station correlations observed in meteorological variables, we incorporated temperature, relative humidity, global radiation, and precipitation as global covariates, assuming they are invariant for each air quality station.

5.2 Simulation of Missing Data

Given the substantial differences in mean PM_{10} concentrations, though with similar variability, between ARPA stations and low-cost sensors, two separate synthetic datasets were constructed. The decision to consider two data sets is related to the analysis of spatial and temporal dependencies of daily PM_{10} concentrations. A pairwise analysis of the linear correlation coefficients reveals that ARPA stations exhibit consistently strong inter-station correlations ($r > 0.9$), and a similar pattern is observed among low-cost sensors. In contrast, the inter-group correlations between the two sources are significantly lower, with a mean of around 0.65. The stations are, on average, 3.9 km apart, with a maximum distance of approximately 8.3 km (between sensor 70169 and station 24644). The graph shows that the correlation is consistent over distance within the same group, apart from sensor 22851, which exhibits a slightly lower association with the other low-cost sensors, although the values are still above 0.85. In contrast, the association between mix sources shows a correlation ranging from 0.6 and 0.75, which is significantly below the intra-group correlations. This may be because the analyzed area around the stations is relatively homogeneous (regional uniformity) in terms of pollution sources, land use, and weather conditions, which may explain why the correlation between stations remains high regardless of distance. The spatial correlation is not as strong as the temporal dependencies that play a more predominant role.

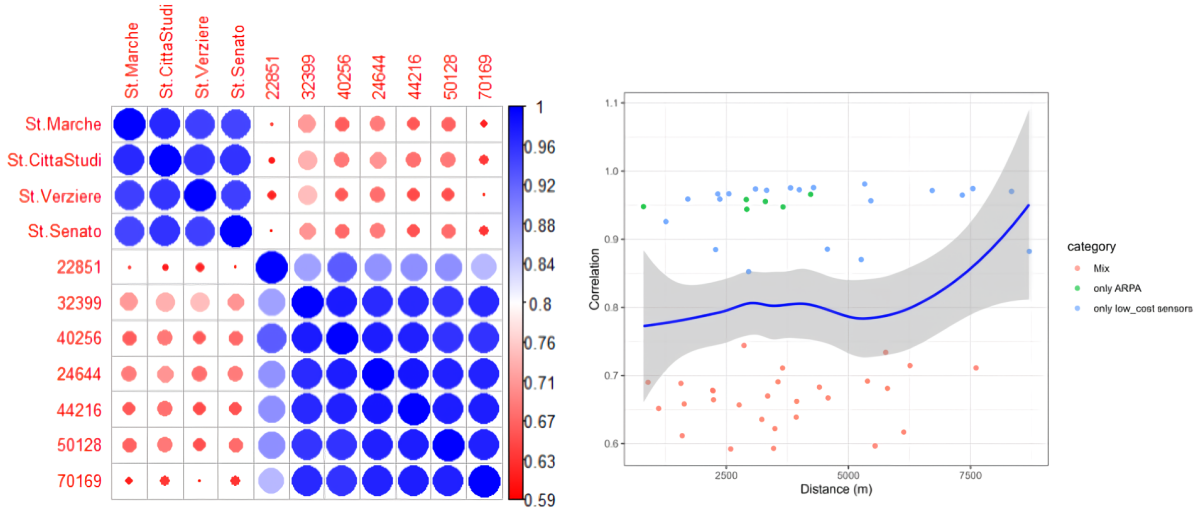

Figure 13: The left plot displays the correlation matrix, clearly showing how data points are grouped based on their source. In the scatter plot, the linear relationship between correlation and distance highlights intra-source differences

The dataset representing the ARPA stations was built from three synthetic time series, while the dataset representing the low-cost sensors was constructed from four synthetic time series collecting the daily observations. To capture the unique characteristics of each source type, we follow the method proposed by Savarimuthu et al., 2021 [36] leveraging the STL decomposition of the available PM_{10} data. We base the construction of the synthetic time series on the seasonal trend, as incorporating the trend component would introduce unrealistic level shifts. A variability factor, calculated as the ratio of the absolute seasonal component to its maximum value, was added to ensure that the variability was higher during periods with a strong seasonal pattern, such as winter, and lower during more stable periods, such as summer. In addition, a random noise was incorporated to simulate the intrinsic randomness of the measurements. The synthetic concentrations were then generated as the sum of the mean PM_{10}, the seasonal component, and the variability-corrected random noise. Reflecting the observed characteristics of the actual data, missing data levels were set at 2%, 6%, and 8% for ARPA stations, while higher missing data levels of 5%, 10%, and 25% were applied for low-cost sensors. Since low-cost sensors often have longer consecutive gaps in data availability, up to 133 days in the real world, a customized function was developed to randomly assign missing values, ensuring that gaps greater than 50 days would begin in January or February of 2022 or 2023. This approach ensures that the synthetic data capture the intrinsic differences in data quality

and measurement accuracy between the two types of monitoring stations.
In addition to the inclusion of meteorological factors as explanatory variables, we also incorporate lag1 and lag2. By analyzing the autocorrelation function (ACF), which measures the degree of similarity between a time series and its lagged versions, we observe a decreasing trend with significant peaks at the first few lags, which is further validated by the partial autocorrelation function (PACF). This aligns with the physical property of PM_{10} dust, which can remain suspended in the air for up to 12 hours [37]. These results emphasize that the current PM_{10} concentration is highly dependent on recent observations, underlining the importance of short-term temporal dependencies. Interestingly, this behavior is consistent with both ARPA stations and low-cost sensors, indicating that despite differences in measurement, both sources capture similar temporal dynamics.

5.3 Models used

k-nearest neighbors (k-NN). It's a supervised learning algorithm used for both classification and regression tasks. In this study, we focus on the application of k-NN for regression problems.The algorithm identifies the k nearest neighbors for a given data point based on a predefined distance metric and leverages their values to estimate the missing value. In order to determine which data points are closest to a given point, k-NN computes a distance measure, typically the Euclidean distance, which helps identify the "decision boundary," partitioning the feature space into regions associated with different sets of neighbors. In R, the caret library [38] provides a convenient interface for implementing k-NN. Before training the data, we standardize the predictor variables, ensuring that all variables have a mean of zero and a standard deviation of 1. Then, we perform a grid search to optimize the hyperparameters k and the weights, which determine the weight assigned to each neighbor, with options like "uniform," where all neighbors contribute equally, or "distance", where closer neighbors are given more importance.

Random forest (RF). Based on the CART concept, it is an ensemble learning for both general-purpose classification and regression methods. It builds a set of uncorrelated and independent decision trees, where each tree serves as a learner. Since it is not distance-based, it does not require feature scaling. Each tree is trained on its own unique training set, created by randomly sampling from the original dataset with replacement, a process known as bootstrapping. Any split within a tree is determined by evaluating a subset of features, selected randomly at each node. To introduce randomness during the training process, each tree considers only a random subset of features at each split, rather than all available features. This further ensures that the trees are uncorrelated and prevents dominance by highly predictive features in the dataset. The trees are grown independently to a stopping point, which could be defined by a maximum depth, minimum number of samples per node, or other criteria. Once all the trees are trained, their predictions are aggregated to make the final prediction. For regression tasks, the predicted values from all trees are averaged. In the function used with the Caret package for this task, we optimise m, the number of variables used to determine the decision at a tree node, ranging from 1 to 10, and the total number of trees, between 50 and 500, for consideration. A tunegrid is constructed to evaluate the combinations of these hyperparameters, with the $RMSE$ used as the performance metric for internal evaluation during the training phase.

Multilayer perceptron (MLP). It is a type of artificial neural network consisting of multiple fully connected layers of neurons where the signal is transmitted in one direction from input to output by a process known as forward propagation. The hidden layers, which are not directly connected to the external environment, process and transform the input data. Each neuron in the hidden layer performs two operations: transforms the values from the previous layer with a weighted linear summation, and then it applies a non-linear activation function to introduce the capacity to model complex, non-linear patterns. In this work, we utilize the Rectified Linear Unit (Relu) activation function, which outputs 0 for any negative input and returns the input value for positive inputs. The learning process involves minimizing the error between the model's predictions and the target values, using MSE as the loss function. The model's weights are iteratively updated through backpropagation, a process in which the error signal is propagated backward from the output layer to the input layer to adjust the weights and biases. This process is repeated

iteratively across multiple training epochs until the loss converges to a minimum using Adam as an optimization algorithm to reduce the loss. The output layer receives the transformed values from the final hidden layer and produces the model's predictions. For this task, we develop a model with an architecture of 4 hidden layers and a total of 155,845 parameters. To improve training efficiency and mitigate overfitting, we incorporate batch normalization, which re-centers and re-scales intermediate layer outputs to speed up convergence. Additionally, dropout and an L2 kernel regularizer are implemented to handle overfitting and make the neural network stochastic, given the relatively small size of our dataset. For optimizing the performance of the model, we trained it several times with different configurations, to minimise the evaluation metrics on the test set.

Classification And Regression Tree analysis (CART). As described by Burgette et al., 2010 [39], CART involves a single binary decision tree, in which each internal node represents a partition based on a single predictor variable. The algorithm recursively splits the data, to minimize the variance of the outcome within each resulting node. While splits in RF are determined using a randomized subset of predictors at each node, CART deterministically selects the best split based on all available predictors.

5.4 Results

Tables 4 and 5 present the experimental results of different imputation methods applied to PM_{10} missing values, with missing rates ranging from 2% to 25%.

Focusing on evaluation metrics, $RMSE$ consistently shows higher values than MAE in all scenarios, reflecting $RMSE$'s greater sensitivity to large errors, as highlighted by Chai et al., 2014 [44], while MAE treats all errors equally, $RMSE$ gives more weight to larger absolute errors.

Random Forest (RF) outperforms the other methods across all three evaluation metrics. For instance, for ARPA sensors with 10% missingness, RF achieves an $RMSE$ of 5.66 and an R^2 of 0.82, while for low-cost sensors with 25% missingness, it records an RMSE of 4.51 and an R^2 of 0.88. In contrast, MLP shows the weakest performance with $RMSE$ reaching up to 9.05 and R^2 as low as 0.63, likely due to the dataset's limited size, which was insufficient for training complex neural networks. K-NN and CART show intermediate performance, with K-NN exhibiting a gradual decline in accuracy as the missingness rate

increases. On the other hand, RF maintains consistent performance due to its ability to combine multiple decision trees, and so reducing the variance of the model [45].

The feature importance analysis, computed using the mean decrease in impurity (MDI), reveals that lag_1 and lag_2 are the most influential features, with normalized importance scores of 100 and 38.25, respectively. This shows that the PM_{10} concentration at a given time is highly dependent on the concentrations observed in previous time steps. Other relevant features include $MaxRad$ (22.59) and day_of_year (21.59), which capture the impact of solar radiation and seasonality on pollutant dispersion, while variables such as month and rainfall have minimal influence.

Random Forest emerges as the most reliable method for imputing missing PM_{10} values (up to 25% of missingness), for both ARPA and low-cost sensors. Moreover, given the relatively small dataset, the computational costs of using Rf are not a concern. RF is adopted as the standard approach for imputing missing values in the real recorded PM_{10}.

Missing rate	Measures	K-NN	RF	MLP	CART
2%	RMSE	6.566	5.0718	7.320	4.695
	MAE	3.313	3.316	5.318	3.331
	R2	0.759	0.870	0.743	0.854
4%	RMSE	5.989	4.347	6.756	4.813
	MAE	3.120	2.834	5.173	3.160
	R2	0.785	0.880	0.729	0.862
6%	RMSE	6.326	5.735	7.696	7.131
	MAE	3.421	3.824	5.988	3.976
	R2	0.774	0.821	0.706	0.730
8%	RMSE	6.858	5.525	9.055	8.089
	MAE	3.659	3.260	7.284	4.476
	R2	0.718	0.810	0.636	0.671
10%	RMSE	7.382	5.665	8.106	7.180
	MAE	3.867	3.620	6.216	4.096
	R2	0.706	0.826	0.633	0.691

Table 4: Experimental results of different missing data imputation methods for ARPA PM10 sensors.RF consistently performs well across all scenarios, showing lower RMSE/MAE and higher R^2 values even at higher levels of missingness.

Missing rate	Measures	K-NN	RF	MLP	CART
5%	RMSE	4.148	3.573	5.885	4.240
	MAE	1.872	1.712	4.110	2.713
	R2	0.836	0.876	0.705	0.863
10%	RMSE	4.408	4.060	7.970	5.775
	MAE	2.197	2.244	5.909	3.540
	R2	0.887	0.905	0.675	0.810
15%	RMSE	6.326	5.735	7.696	7.131
	MAE	3.421	3.824	5.988	3.976
	R2	0.774	0.821	0.706	0.730
20%	RMSE	5.513	4.614	9.032	6.045
	MAE	2.756	2.721	6.611	3.686
	R2	0.829	0.883	0.621	0.805
25%	RMSE	5.207	4.513	7.530	6.975
	MAE	2.537	2.689	5.330	3.971
	R2	0.851	0.887	0.704	0.745

Table 5: Experimental results of different missing data imputation methods for low-cost PM10 sensors.Also here, RF consistently performs well across all scenarios.

6 Geo-statistical machine learning models

The skewed nature of the response variable suggests the need for some transformation. To improve modeling and subsequent predictions, the box-cox transformation [52] has been applied to stabilize the variance of the PM_{10} distribution, especially in the case of universal kriging, where the deterministic component is modeled through regression analysis. This adjustment simplifies the patterns and makes them more consistent across the whole data set by considering both the logarithms and power transformations. In the Box-Cox family, the optimal transformation of the dependent variable is determined by estimating λ over a range from -2 to 2. The results indicate that the best transformation is a Box-Cox transformation with $\lambda = 0.38$, approximately equivalent to applying a cube root transformation. When converting the results back to the original scale, we employ the back-transformation of the forecasts, adjusting the prediction to the mean rather than the median to avoid bias. The formula is:

$$\begin{cases} exp(w_t)\left[\,1+\frac{\sigma_h^2}{2}\right] \\ (\lambda w_t+1)^{1/\lambda}+\left[1+\frac{\sigma_h^2(1-\lambda)}{2(\lambda w_t+1)^2}\right] \qquad otherwise \end{cases} \tag{6.1}$$

where σ_h^2 is the forecast variance that determines the magnitude of the difference between the mean and the median. The larger the variance of the forecast, the larger the discrepancy. With this correction, the PM_{10} concentrations on the original scale become more reliable allowing an accurate comparison between the different districts of Milan. The theoretical framework and formulas presented in the following chapter are entirely based on the methodologies described by Montero et al., 2015 [53] in *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging* and by Wood et al., 2017 [27] in *Generalized Additive Models* book.

6.1 Spatio-temporal Kriging

Spatio-temporal kriging is a geostatistical interpolation technique accounting for spatial and temporal correlation, and it is used to predict the value of an underlying random field $Z(s,t)$ in the continuosn domain $D \times T$ at unsampled locations of interest (s_1, t_1). Each point can be represented by a couple (s,t) where $s \in R^2$ is the two-dimensional Euclidean space, which is a mathematical framework to describe longitude and latitude, and R represents the time dimension. Whether the values are intrinsically ordered for time, the same does not occur with the spatial coordinates used. Instead, the spatial relationship is defined in terms of distance between points. Spatio-temporal continuity is a property that characterizes the relationship between observations at different locations in a spatio-temporal domain. The random field $Z(s,t)$ can be seen as a family of random variables defined at each point in the spatial and temporal domain. Under certain assumptions, it allows the determination of the covariance structure of the stochastic field that is compatible with the variability present in the data of our study. Instead, the covariance is a linear function defined as:

$$C((s_i, t_i)), (s_j, t_j)) = C(Z(s_i, t_i), Z(s_j, t_j)) \tag{6.2}$$

which captures the linear relationship between observations at two locations. To make inferences based on observed realization consistently, it's important to adopt the hypothesis of stationary which guarantees that the observed values at different locations and times in the domains are seen as the same realizations of the same random field $Z(s,t)$. In a strict sense, the joint distribution function of $Z(s_1, t_1), Z(s_2, t_2), \ldots, Z(s_n, t_n)$ with the same joint distribution of $Z(s_1 + \mathbf{h}, t_1 + u), Z(s_2 + \mathbf{h}, t_2 + u), \ldots, Z(s_n + \mathbf{h}, t_n + u)$ is unaffected by any spatial and temporal shift $\mathbf{h}$ and u. However, this condition is quite unrealistic in real-world applications. Instead, second-order stationarity or weak stationarity assumes that: weakly stationary that $\mu(s,t)$ is constant across the domain $\forall(s,t) \in R^2 \times R$, and that the covariance exists for every pair of points and $Z(s,t)$ and depends only on the vector $\mathbf{h}$ and u. The formula is:

$$C\left(Z(s_i, t_i), Z(s_j, t_j)\right) = C(\mathbf{h}, u) \;\; \forall(s,t) \in R^2 \times R \tag{6.3}$$

In other words, a spatio-temporal random field with a stationary covariance structure depends only on the spatial distance $(t_i - t_j)$ and the temporal distance $(s_i - s_j)$, rather than the absolute positions. Further, it implies the symmetric property $C(\mathbf{h}, u) = C(-\mathbf{h}, -u)$. Alternatively, the intrinsic stationarity relaxes the assumption of a well-defined covariance structure. Instead, it ensures that the increments $Z(s+h, t+u)Z(s,t)$ have a constant mean (zero) and a finite variance that depends on any $(\mathbf{h}, t)$. At the same time, it doesn't require the existence of a covariance function. This is sufficient for defining the variogram.

6.1.1 Variogram

The variogram is a key tool for modeling spatial variability, which measures the dissimilarity between two points as a function of their space-time distance:

$$2\gamma((s_i, t_i), (s_j, t_j) = V(Z(s_i, t_i) - Z(s_j, t_j)) \tag{6.4}$$

Under the weak stationary assumption, the semivariogram, defined as half the variogram, whose main properties are the symmetric and negatively defined function, is complementary to the covariance function and satisfies the relation:

$$\gamma((s_i, t_i), (s_j, t_j)) = \frac{1}{2}V\left(Z(s_i, t_i)\right) + \frac{1}{2}V\left(Z(s_j, t_j)\right) - C\left((s_i, t_i), (s_j, t_j)\right) \tag{6.5}$$

Since weak stationarity implies constant variance across the spatio-temporal domain, the semivariogram and variogram depend solely on the relative distances in time and space. The relationship of the covariance function with the semivariogram is only verified if the semivariogram is bounded. This condition ensures the covariance function is well-defined across the domain.

These concepts are essential for constructing the empirical semivariogram, which is a key step in fitting valid spatio-temporal models. Assuming an isotropic and stationary random field, the classical estimator proposed by Matheron [54] is defined as:

$$\hat{y}(\mathbf{h}(l), u(k)) = \frac{1}{2N(\mathbf{h}(l), u(k))} \sum_{(s_i,t_i),(s_j,t_j) \in N(\mathbf{h}(l),u(k))} (Z(s_i, t_i) - Z(s_j, t_j))^2 \tag{6.6}$$

where $\mathbf{h}(l)$ represents the spatial lag distance, grouping them in bins, while $u(k)$ is the

temporal one between pairs of observations (i, j). $N(\mathbf{h}, u)$ is the sets of observation (i, j) that satisfy the spatial and temporal lag conditions. Before applying Kriging, to construct the empirical semivariogram, we use the R packages gstat [54] and spacetime [55], both developed by Edzer Pebesma. In order to adequately represent the observation, we work with a complete spatio-temporal data frame (STDF), since the monitoring stations are fixed and there are no missing daily measurements. For the spatial coordinates, we use the UTM coordinates (EPSG:3035), while the *stplot()*function provides a preliminary visualization of the concentrations.

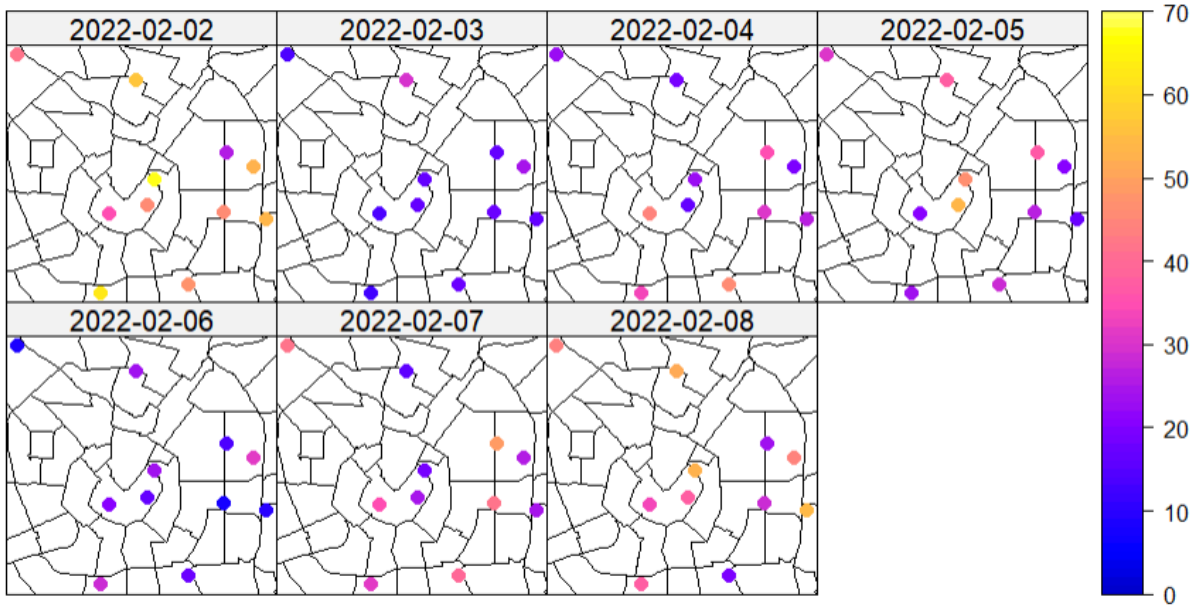

Figure 14: This graph represents a spatio-temporal visualization of PM_{10} concentrations observed at different stations during the period between 1 and 7 January 2022

From the graph, we can see that the variability of PM_{10} is very high in winter compared to summer, as emphasized in the previous chapters.

To determine the empirical semivariogram, we use the function *variogramST()*, which generates a three-dimensional grid of semivariance values. In practice, constructing the empirical semivariogram requires careful consideration of tags and cutoff: the first determines the temporal lags to be included, and the second decides the maximum spatial distance beyond which there is no spatial correlation between the data. The fitting of these parameters is manual based on visual interpretations and knowledge of the underlying phenomenon. A commonly used rule of thumb in geostatistics is that the empirical

semivariogram should be computed for lag distances up to half the diameter of the study domain. This is because the number of observation pairs decreases with increasing distance, reducing the reliability of estimates for large lags. In our case, the minimum and maximum distances between two monitoring stations are 803 meters and 8,702 meters, respectively. Based on this rule, the cutoff is set to half of the maximum distance, approximately 4,353 meters. For temporal lags, we experimented with different numbers of days and selected 6 days. This configuration ensures that the semivariogram is computed using a sufficient number of pairs while excluding negligible temporal and spatial distances.

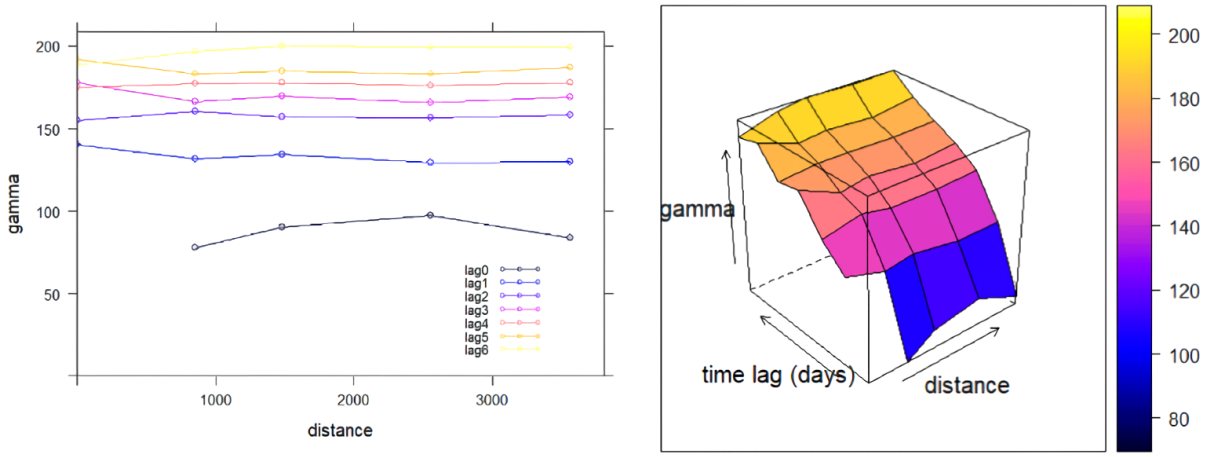

Figure 15: The 2D graph on the right shows that the temporal autocorrelation decreases as the time lag increases. A similar pattern can be seen in the 3D graph, where dissimilarity, measured with the color scale, indicates that the greater the distance and the temporal lag, the higher the semivariance values.

Without incorporating explanatory variables, both graphs show how dissimilarity increases with spatial and temporal distance in different ways. Temporal autocorrelation decreases with increasing temporal lag, with a clear trend visible in all curves. In the 2D graph, the difference in dissimilarity between lag5 and lag6 is less pronounced but still present, suggesting that including additional temporal lags is unnecessary.

6.1.2 Parameters estimation

Before fitting a theoretical covariance model to the empirical semivariogram, it is essential to assess its features and the level of anisotropy. These help describe the structure of spatial and temporal dependence in the data. A theoretical semivariogram model is considered anisotropic if the spatial and temporal dependence vary with the direction. Fitting a linear model to the observed spatial and temporal coordinates transforms it into an isotropic, directionless representation. The predicted spatial-temporal anisotropy represents the optimal scaling that standardizes space and time dimensions, facilitating variogram modeling. Thus, a ratio of 10.44 km/day means that the temporal variance observed in one day corresponds to the variability observed over 10.44 km in space, demonstrating that spatial dependence extends on a larger scale. To accurately fit the spatio-temporal variogram, which is derived from the covariance function, we must estimate three key parameters:

- Nugget: Represents the residual variance not explained by spatial and temporal structure;
- Sill: The total variance of the data, reached as $\mathbf{h}$ increases.
- Range: The distance at which the variogram reaches the sill, beyond which points are considered spatially and temporal independent.

These parameters are critical in defining the shape of the empirical semivariogram and ensuring an optimal fit with the chosen theoretical covariance model.

6.1.3 Covariance Models

In the spatio-temporal framework, our analysis concentrates on four main classes of spatial-temporal covariance models: separable, product-sum, metric, and sum-metric. For each of these models, we test different combinations of isotropic covariance functions, such as Gaussian, exponential, and spherical, to model the theoretical semivariogram and fit it to the empirical variogram. The covariance models presented are non-stationary, isotropic, and non-separable.

Metric model. The assumption behind this model is that the spatial and temporal dependencies are combined into a single metric distance. This means that the structure

of the covariance is uniform across the entire domain and independent of the specific covariance function used. The model is defined as:

$$C(\mathbf{h}, u) = C(\|\mathbf{h}\| + c|u|) \ (\mathbf{h}, u) \in R^d \times R \tag{6.7}$$

where c is a positive constant that ensures isotropy, while $\|\mathbf{h}\|$ is the norm of vector h. The same sill for both components simplifies the interpretation of the model.

Sum-Metric model. It combines the metric model and the sum model. The sum model decomposes the spatio-temporal covariance into two independent functions, one for the spatial component and one for the temporal component, which has the drawback of not ensuring the positive definiteness. This limitation is overcome by incorporating the metric model into the formulation. In semivariogram terms, the combined model is:

$$\gamma_{st}(\mathbf{h}, u) = \gamma_s(\mathbf{h}) + \gamma(\|\mathbf{h}\| + a|u|) \tag{6.8}$$

where a is a positive constant term that ensures isotropy.

Product-sum model. It combines the product model and the sum model. The product model assumes that the structure of temporal covariance remains the same for any given spatial separation. However, this assumption is very restrictive for real space-time phenomena. To address this limitation, De Iaco introduced the sum-of-products model, which considers the spatial and temporal components separately and incorporates an interaction term as the product of two separate covariance functions. The stationary product-sum model is:

$$C(\mathbf{h}, u) = k_1 C_s(\mathbf{h}) C_t(u) + k_2 C_s(\mathbf{h}) + k_3 C_t(u) \tag{6.9}$$

where the k are positive constants that guarantee isotropy.

Seperable model. It is a more general formulation than the product model, which assumes that the spatio-temporal covariance function can be represented as the product of a spatial and temporal term. The formula is:

$$C(\mathbf{h}, u) = C_s(\mathbf{h}) C_t(u) \tag{6.10}$$

This model allows for an additive contribution from the pure spatial variability and the

combined spatio-temporal effect. The spatial and temporal dependence treated independently simplifies the estimation of the covariance parameters used to construct the theoretical semivariogram. Typically, this parameter estimation is carried out using a bound-constrained optimization algorithm such as the BFGS method, implemented by the gstat package. In our study, these models will be compared based on the ability to minimize the MSE between the empirical and theoretical semivariogram. These covariance models will be employed only within the framework of Ordinary Spatio-temporal Kriging.

6.1.4 Kriging

Kriging is geostatistical method named in honor of the South African mining engineer Daniel Gerhardus Krige, used for spatial-temporal prediction. It aims to predict the value of a random field $Z(s,t)$, at non-observed points (s,t) or blocks from a collection of data observed at n points within a domain $D \times T$. By construction, it provides the best linear unbiased predictor (BLUP) of the regionalized variable under study at such non-observed points. The choice of the covariance or semivariogram model has a great impact on prediction and kriging variance. Formally, to predict $Z(s_0,t_0)$, a linear predictor is constructed as follow:

$$Z^*(s_0,t_0) = \sum_{i=1}^{n} \lambda_i Z(s_i,t_i) \tag{6.11}$$

where the coefficients λ_i are chosen so as to minimize the variance of the prediction error. Focusing first on the Ordinary spatio-temporal Kriging (OK), which assumes that $Z(s,t)$ is a second-order stationary process with a constant but unknown mean u and a known covariance function $C(h,u)$. Under the unbiasedness condition which ensures that the sum of the weights is one, the weights λ_i are represented in semivariogram terms as:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(s_i - s_j, t_i - t_j) + \alpha = \gamma(s_i - s_0, t_1 - t_0) \; \forall i = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{6.12}$$

where α is the Lagrange multiplier enforcing the unbiasedness constraint. This model is often referred to as a "pure" model because it exploits exclusively the spatio-temporal structure of the random field without the inclusion of external covariates. Instead, whether $Z(s,t)$ is a spatio-temporal random field with drift so that the rf depends on the spatio-temporal location (s,t), it introduces the concept of stochastic error so that $Z(s,t)$ is viewed as a combination of the deterministic and constant $\mu(s,t)$ and the stochastic error $e(s,t)$.

Instead, the Universal spatio-temporal Kriging (UK) is an extension of OK that handles scenarios where the process mean is not constant but depends on the spatio-temporal location (s,t). The random field can be decomposed into a linear combination of components: a deterministic component $\mu(s,t)$ representing the global trend of the phenomenon (drift) and a stochastic component $e(s,t)$ which captures the residual variation around the trend and is treated as an intrinsically stationary process, with a zero mean and known variogram. Then, in the spatiotemporal context, a local approach is adopted to model the mean:

$$\mu(s,t) = \sum_{H=1}^{p} \alpha_h, f_h(s,t) \tag{6.13}$$

where $f_h(s,t), h = 1, \ldots, p$ are the basis function that defines the drift, which describes how the mean varies over space and time. α_h are constant coefficients indicating the contribution of each basis function and p is the number of bases determining the complexity of the drift model. Once the drift (s,t) has been specified, kriging is applied to the residuals, which are assumed to follow an intrinsically stationary process with a known variogram. The Universal Kriging equations are given by:

$$\begin{cases} \sum_{j=1}^{n(s_0,t_0)} \lambda_j \gamma_e(s_i - s_j, t_i - t_j) + \alpha \sum_{k=1}^{p} \alpha_k f_k(s_i, t_i) = \gamma_e((s_i - s_0, t_1 - t_0)) \\ \qquad\qquad \forall i = 1, \ldots, n(s_0, t_0) \\ \sum_{j=1}^{n(s_0,t_0)} \lambda_i f_k(s_i, t_i) = f_k(s_0, t_0), \forall k = 1, \ldots, p \end{cases} \tag{6.14}$$

where λ_i represent the ringing weights for each observation and γ_e is the semivariogram

of the residual process.

As an initial step, we first apply Ordinary Kriging (OK) to assess the inherent spatio-temporal structure of PM_{10} concentrations. This method serves as a baseline model, as it relies only on spatial and temporal dependencies. By implementing the Box-Cox transformation on PM_{10}, we compute both the theoretical and fitted variograms to analyze the spatial correlation structure. In the prediction step, the function *krigeST* automatically applies the inverse transformation using the estimated λ parameter, ensuring that the predictions are returned on the original PM_{10} scale. Then, we extend our approach by incorporating external covariates p to explain the residuals in the empirical variogram better. In the case of UK, the deterministic component is estimated through a linear regression:

$$\hat{\mu}(s_i, t_i) = \hat{\beta}_0 + \sum_{h=1}^{p} \hat{\beta}_h x_h(s_i, t_i) \tag{6.15}$$

After estimating the regression parameters and computing and predicting $\hat{\mu}(s_i, t_i)$ at each observed space-time location, we analyze the spatio-temporal autocorrelation structure of the residuals, by fitting different covariance models. The best variogram model is selected based on the three metrics used previously. The shape of the semivariogram is modeled in terms of spatial, temporal, and joint spatio-temporal dependencies. Variogram parameter estimation is performed using a bound-constrained BFGS optimization method.

6.2 GAM

An alternative approach to model the complex relationship within spatio-temporal data is to use Generalized additive models (GAMs) which extend Generalized Linear Models (GLMs) by incorporating non-linear functions of explanatory variables in an additive framework. GAMs are particularly effective in capturing spatial and temporal dependencies while maintaining interpretability.Differently from spatio-temporal kriging, which explicitly models spatial autocorrelation through variograms, GAMs approximate spatial and temporal effects through smooth functions and penalty methods. The random component assumes that the response variable $Y(s,t)$ follows a exponential family of distributions $Y(s,t) \sim EF(\mu(s,t), \phi)$ with mean $\mu(s,t)$ and scale parameter ϕ . The expected value of the response variable is linked to a set of covariates through a monotonic link function $g(.)$. In general, the model has the following structure :

$$g(\mu(s,t)) = \beta_0 + f_1(X_1(s,t)) + f_2(X_2(s,t)) + \cdots + f_p(X_p(s,t)) \tag{6.16}$$

where β_0 is the intercept parameter and $f(.)$ are the smoothing functions that shape the non-linear relationships between covariates, estimated directly from the data and are written as basis function such as thin-plate spline or cubic splines. The effect of each predictor variable is assumed to be additive, meaning that the overall prediction is viewed as the sum of the contributions of each predictor to the response variable., while tensor product smooth can be used to model interactions between covariates of different scales, such as meters/day. When incorporating smooth functions into a GAM, identifiability issues can arise. This happens because the smooth functions $f(.)$ are not uniquely determined. By imposing identifiability constraints such as zero-sum constraint, so that the the smoothing terms do not introduce collinearity with the intercept. Another advantage of using GAMs is that the fitting procedure automatically determines the optimal shape of the smooth functions without requiring pre-specified functional form for each predictor. Indeed, the coefficient estimates $\hat{\beta}$ are obtained by minimization of the penalized least squares objective:

$$\|y - X\beta\|^2 + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta + \cdots + \lambda_p \beta^T S_p \beta \tag{6.17}$$

where the first term $\|y - X\beta\|^2$ ensures that the model fits the observed data, while the second term imposes a smoothness penalty to balance overfitting and flexibility. The estimated function's degree of smoothness (or wiggliness) is controlled by the parameters (λ). A widely used method for selecting these parameters is Restricted Maximum Likelihood (REML). Proposed by Patterson et al., 1971 [68], REML improves the estimation of variance components by focusing on the random effects in the model. The key idea is to estimate the smoothing parameter λ by excluding the contribution of fixed effects β that reduce the residual variability. This process ensures that λ is selected by maximizing the marginal likelihood, balancing the model flexibility. It is also more stable with small samples than other methods.

6.2.1 Smoothers

In this section, we introduce different types of smoothers, which are used for penalized regression in GAMs. For the smoothing parameter λ, we assume it remains constant across all directions and variables.

Cubic smoothing spline. Consider a set of points $x_i, y_i : i = 1, \ldots, n$ where $x_i < x_{i+1}$. These points are approximated by the smoother constructing different sections $[x_i, x_{i+1}]$. Thus, a cubic smoothing spline is a piecewise-defined function consisting of cubic polynomials that are joined together to ensure continuity up to the first derivative. The function values at the given points $g(x_i)$ are treated as n free parameters, which determine the estimated function's complexity while balancing the smoothing penalty's influence. The value $g(x_i)$ are estimated by minimizing this function:

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int g''(x)^2 dx \tag{6.18}$$

where the integral term penalizes the wiggliness of the function while the first term represents the sum of squared errors between the observed values and the values predicted by the link function. The number of knots should not be established a priori. However, the main issue is that the number of free parameters depends on the effective degrees of freedom used for smoothing, which can lead to high computational costs.

The cubic regression spline is a more computationally efficient version of the cubic smoothing spline. It is a basic penalty smoother that parametrizes the spline in terms of values at a fixed set of selected nodes rather than using all observed data points. Cubic polynomials are defined piecewise between nodes, with the smoothing penalty applied only to internal nodes. The second derivatives are set to zero at the boundaries.

Thin plate spline (TPS). The advantage of using thin plate splines, like other smoothing splines, is that GAMs do not require any a priori knowledge of the functional form of the data or the relationship of interest. Compared to cubic smoothing splines, TPS allows the estimation of basis functions without requiring the specification of knot positions and is better suited for multivariate problems. A TPS smoother estimates the link function by finding the $\hat{f}$ that minimizes:

$$\|y - f\|^2 + \lambda J_{md}(f) \tag{6.19}$$

where $J_{md}(f)$ is a penalization function that measures the wiggliness of f,) penalizing the squared second derivatives of the smooth function across all dimensions. If m is chosen such that $2m > d$, the function to minimize can be rewritten as:

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(\|x - x_i\|) + \sum_{j}^{M} \alpha_j \phi_j(x) \tag{6.20}$$

where δ and α are vectors of coefficients to be estimated,δ being subject to the linear constraints of identifiability. The term $\eta_{md}(\|x - x_i\|)$ that represents the matrix E depends on the Euclidean norm of the distance between points, ensuring that the penalization term λ maintains isotropy. The M functions $\phi_j(x)$ form a basis of polynomials that span the null space of J_{md}, corresponding to functions considered completely smooth and not penalized. The major challenge with TPS is the computational cost, which increases significantly as n grows. Similarly, the regression version TPRS is a reduced and regularised version of the TPS using a reduced-rank approximation and retaining only the principal components of the wiggly part. Instead of computing the full spectral decomposition of E, TPRS through a truncated eigen-decomposition UDU^T makes it possible to select only the first eigenvectors to construct a reduced matrix U_k of lower-numbered parameters k.

Using the Lanczos algorithm, the computational cost decreases from $\mathcal{O}(n^3)$ to a substantially lower cost of$\mathcal{O}(n^2k)$. TPSs are commonly used for spatial smoothing due to their flexibility and ability to estimate a single smoothing parameter at a time. However, the assumption of isotropic smoothing can be quite restrictive. Thus, this assumption makes TPS sensitive to unit variances, as the scaling of a single covariate can significantly affect the estimated smoothing [27].

Tensor Product bases. It is a much more flexible approach. that allows variables to be smoothed separately before combining their effects, making it particularly well-suited for spatiotemporal modeling and cases where covariates have different scales or units. A tensor product smooth can be visualized as a lattice of flexible strips, allowing for anisotropic smoothing. This approach is built upon univariate smoothers, and supposing that we have three covariates x,y, and t, such as geographical coordinates and time, we construct a smooth function $f_{x,y,t}$ by allowing the parameters of each marginal smooth function to vary smoothly with the other two, leading to a combined function:

$$f_{x,y,t} = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{k=1}^{K} \beta_{ilk} b_k d_l(z) \alpha_i(x) \tag{6.21}$$

where α_i, d_l and b_k are the univariate basis function representing the smooth effects of each covariate while β_{ilk} weight coefficients, determining the contribution of each combination of basis functions to the overall smooth function. This approach makes it possible to combine as many covariates as are required. A fundamental assumption is low-rank basis representation, which ensures the number of parameters to be estimated is smaller than the number of observations of a full-rank model, improving computational efficiency. The willingness penalty is based on the single penalty defined on the marginal functions, while their combination penalty is defined in terms of the conditional penalty of each variable at a time when holding the other two variables stationary. In this way, the complexity of the penalty is controlled not only with respect to the single variable but also concerning how it interacts with the others. In the mgcv package [56], *te()* is used to specify tensor product smooths constructed from any singly penalized marginal smooths, while *t1()* is used to specify only the tensor product interactions with the marginal smooths. We use

both to include the spatio temporal dependence in the model.

Although less critical than the choice of basis K, the basis dimension should be large enough to ensure that the number of effective degrees of freedom is greater than the number of nodes, especially for thin-plate splines and cubic splines. In general,K should be set large enough to allow the smoother varying.

6.2.2 Models Setup and comparisons

The geographical covariates have been transformed into a UTM metric system so that they reflect the real Euclidean distances, facilitating model estimation in Gam. A fundamental aspect is the number of basis functions of each smooth covariate. This number represents the maximum model complexity allowed for each smooth covariate. It should be large enough to ensure that the number of effective degrees of freedom is greater than the number of nodes, especially for thin-plate splines and cubic splines. In general, k should be large enough to allow for a more uniform variation. For temporal variables, instead of considering the datetime in the interval [0,730], we use the temporal covariate day_of_year. Through testing with different values of k, we determined that five basis functions were sufficient to capture the annual trend while avoiding excessive flexibility. The strategy for finding the suited EDF for the terms was to observe the k-index and p-value, which ideally should be close to 1, otherwise, the size of the basis is too small to incorporate all EDFs, restricting the flexibility of the smoothing term. In most cases, no major differences were found that led us to choose one basis function over the other. For the time variable, day_of_the_year (bs=cr), we chose a cubic regression spline, and for all other variables the default thin-plate splines (bs=tp). For source and season, on the other hand, we transformed these into categorical variables and treated them as fixed effects.

As suggested by Wood [27], we implement a diagnostic test based on the correlation between the deviance residuals and the covariates to check whether the model correctly captures the structure of the data. This test is implemented in the *gam.check()* function of the package. Furthermore, although it is possible to incorporate an ARMA structure to model temporal dependencies explicitly, we chose not to include it as the lagged covariates (lag1 and lag7) were already included in the model. The results of the ACF and PACF confirmed that no significant peaks suggested strong temporal autocorrelation in

the residuals. Therefore, the inclusion of additional autoregressive terms was not deemed necessary.

For estimating the model coefficients and smoothing parameters, we used the REML method, which allows for automatic selection of the smoothing parameter λ. Regarding the distributional assumption, we employ both the default setting identity and try the scaled t-distribution, as PM_{10} concentration exhibits a skewed distribution with heavy tails, and *scat()* imposes a stronger penalty on the extreme residuals.

For assessing the relative performance of the different model specifications, we examine the AIC and the summary result as well as whether the EDF values are close to the respective basis dimension k. If the k-index is not approximately 1, this indicates that the model's flexibility is either overfitted or insufficiently penalized. In other words, the trade-off between smoothness and mutability is not balanced.

Models	**Effect**	**DF**	**AIC**	**REML**	**Dev.Res**	**Dev.Exp**	**R2**
Mod1	isotropic	74.36	21735.20	10970	21586.5	70.5%	0.791
Mod2	isotropic	71.32	21799.24	10986	21656.6	70.4%	0.79
Mod3	main+interc	72.59	22111.39	11148	8663.1	79.3%	0.791
Mod4	main+interc	68.43	21814.48	7377.8	21677.6	70.4%	0.789
Mod5	main+interc	84.92	21489.47	10868	21319.6	70.8%	0.796
Mod6	tensor prod	64.54	22398.75	7544.9	9030.8	78.4%	0.783

Table 6: Gam Models comparison

Separating the models into isotropic framework, if we model the covariates using thin splines and cubic regression splines, the results compared on the different link identity between Mod1 (gaussian distribution) and Mod2 (scat distribution) indicates clearly that scat() corrects the skewness of the tails as shown in table 6 with smaller degrees of freedom, but increases the unexplained residual deviance and worsens the AIC value.

Checking the result with the tensor product, it turns out that season is not considered a very relevant predictor with a p-value higher than 0.5. However, the tensor product on the combined geographic coordinates with time is relevant with an F of 4.57, but not as relevant as the Source of PM_{10} recording, which stands at 5.12 10^6 highlighting the descriptive analysis that showed how the stations and low-cost sensors had a difference in measurements. The explained deviance increases to 78.4 % while residual deviance decreases greatly to 9030.8.

Even better performance was found when considering the main effects and the spatio-temporal interaction term modeled using ti() suggesting an improvement in terms of AIC and REML. Between models 3, 4, and 5 the number of k considered and the inclusion of interaction terms vary, elements that may improve the explanatory power. However, model 3, which only considers pure effects, is slightly superior, especially in terms of unexplained residual deviance, due to a less flexible structure than model 5. The spatiotemporal dependency modeled is not considered relevant. Metrics comparisons suggest that Mod3 is better than other models with a good trade-off between explainability and smoothness the use of *ti()* allows for better modeling of spatiotemporal dependencies without assuming isotropy.

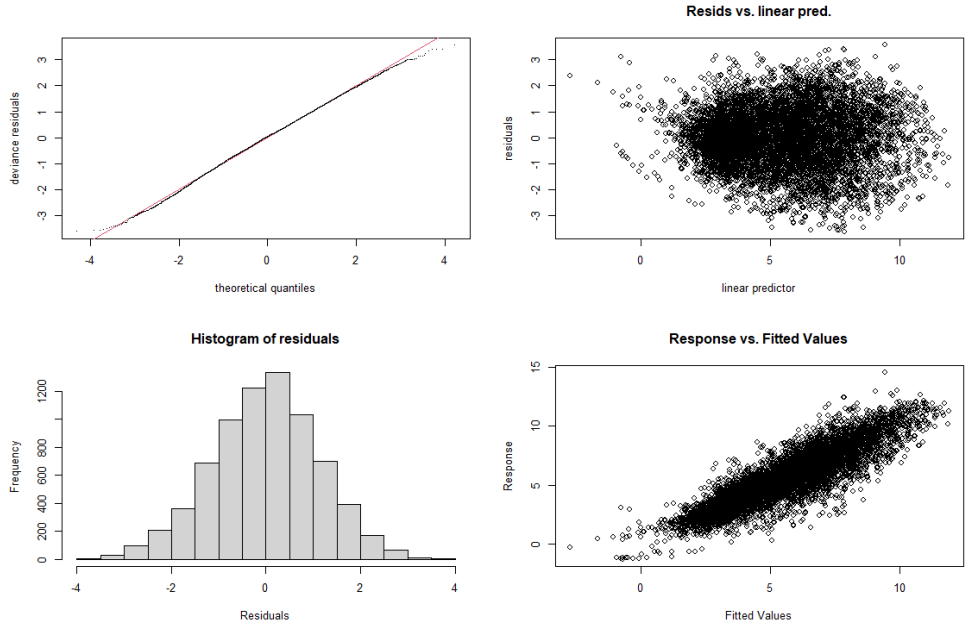

Figure 16: The top left plot shows the normal Q-Q plot for the residuals very close to a straight line. The lower left histogram of residuals and the invariance of the residuals over the linear predictor confirm the pattern in the QQ plot

The plots from *gam.check()* suggest a reasonable model fit. The QQ plot shows a slight deviation from normality in the tails confirmed by the histogram of the residuals which shows a slight asymmetry, while the residuals appear approximately constant as the mean increases. The plot of the response variable against fitted values, shown in the lower right panel, emphasizes the positive linear relationship without a scatter, Regarding the spatial predictions, the contour plot of the linear predictor highlights strong spatial heterogeneity. Higher PM_{10} predictions are concentrated in central and eastern Milan while moving

westward, the increasing density of contour lines suggests a rapid decline in PM_{10} levels as we move away from the city center. This pattern aligns with the distribution of the training data, which is predominantly located in the central and eastern parts of the city. After all these steps, we employ the *predict.gam()* function, which not only provides predicted values but also returns the standard deviation associated with each prediction. This additional information is crucial, as it allows us to apply the Box-Cox back-transformation with bias adjustment.

6.3 Validation

Cross-validation (CV) is the most commonly used validation procedure in the context of spatially correlated functional observations. However, standard CV is based on the fundamental assumption that the training and test data are independent, which may lead to overly optimistic estimates of model generalization when spatial dependencies are present [56]. Therefore, to account for spatio-temporal dependence, we adopt leave-location-out cross-validation (LLO-CV), in which each location is used once as the test set, while all other observations constitute the training set. In addition for Kriging, the evaluation mechanism considers several scenarios, differing in the neighborhood size included in the interpolation process to assess their impact on prediction. To perform local kriging, we tested 10 and 50 neighbors' observations closed in terms of space and time locations. Additionally, to provide a broader perspective, we also evaluated a scenario including all available data, given the relatively small size of our dataset. In contrast, for GAM models, we only considered the scenario using all data.

7 Experimental Results

Fitting a theoretical space-time model to the empirical variogram requires imposing constraints on the search space by setting lower and upper bounds for the spatial, temporal, and joint components. To identify the optimal covariance function, we employed the L-BFGS-B optimization algorithm, which efficiently minimizes the objective function while maintaining parameter constraints within a predefined range.

The initial boundary values were derived from the empirical variogram at lag 0. To en-

hance model performance and convergence, we implemented an iterative feedback loop: parameters obtained from the *optim* function were used to adjust the search boundaries dynamically. This step is crucial, as the choice of initial values significantly influences model performance and optimization speed. Additionally, we tested different weighting schemes and found that method=6, where all values are assigned equal weights, resulted in the lowest $wMSE$. This fitting approach was consistently applied across different covariance functions within each covariance model family using a one-dimensional variogram selection process.

When fitting the variogram without explanatory variables, the sum-metric model emerged as the best-performing covariance model, achieving a mean $wMSE$ of 0.01. The spatial component was modeled using an exponential function, with a nugget effect of 0.1, indicating minimal unexplained variance at zero distance. The small value can be explained by the box-cox transformation which has smoothed out the model reducing the residual variance not explained. The sill was estimated at 9.68 while the spatial range was beyond the area of study, suggesting that PM_{10} levels exhibit strong spatial correlation even at large distances. The temporal component was also modeled with an exponential function, with a nugget effect of 0.1 and a very low sill , with an estimated interval of 58 days, indicating that the temporal correlation remains significant over this period before decreasing. The spatio-temporal joint component was modeled with a spherical function, while the estimated anisotropy of 448 km/day, suggests that the temporal dependence is more relevant.

Table 7 shows the predictive performance of the covariance models, highlighting the influence of neighborhood size in kriging on prediction quality. The numbers of neighbors used in the interpolation, significantly affect the stability of the prediction. Indeed, the differences in performance should be interpreted in the context of the underlying model structure and the sample size limitations.
The sum-metric model consistently outperforms the other covariance models, achieving the lowest $RMSE$ and highest R^2 across all neighborhood sizes. The best estimation is obtained by considering all the data, suggesting that also spatial structure contributes to prediction accuracy than temporal dependencies. Nevertheless, a station-wise cross-validation analysis reveals that the highest $RMSE$ is observed at station 22851, which is

Covariance model	VarComp	wMSE	Neigh.	RMSE	MAE	R2
separable	Gau+Sph	[0.05]	10	15.50	11.41	0.18
product-sum	Exp+Gau	[47.58]	10	12.88	9.26	0.43
metric	Exp	[0.27]	10	15.51	11.41	0.17
sum-metric	Exp+Exp+Sph	[0.01]	10	12.11	8.56	0.50
separable	Gau+Sph	[0.05]	50	16.26	12.54	0.09
product-sum	Exp+Gau	[47.58]	50	15.30	11.63	0.20
metric	Exp	[0.27]	50	16.26	12.54	0.09
sum-metric	Exp+Exp+Sph	[0.01]	50	12.10	8.57	0.50
separable	Gau+Sph	[0.05]	All	17.07	13.35	0.00
product-sum	Exp+Gau	[47.58]	All	11.00	7.97	0.59
metric	Exp	[0.27]	All	17.07	13.35	0.00
sum-metric	Exp+Exp+Sph	[0.01]	All	10.84	7.60	0.60

Table 7: OK Models comparison. Notice that in column VarComp, the first term refers to the spatial decay function, the second term to the temporal decay function, and the third to the spatio-temporal interactions decay function.

located in eastern Milan, near Milano Pacal Città Studi and 40256 station. The results are coherent with the initial evaluation of the variogram fit, showing that the sum metric model has a clear predictive advantage over the others.

When applying Universal UK, the empirical variogram was constructed using only 3 temporal lags. The deterministic component of the model was determined by selecting the most informative covariates, chosen through a stepwise selection strategy based on AIC. Additionally, p-values from standard regression models were examined to further refine variable selection, ensuring that only statistically significant predictors were retained. Within the same modeling framework, the Gaussian function emerged as the best fit for the different covariance components. For the prediction phase, the same explanatory variables could not be used to the risk of a quasi-singular covariance matrix. This problem arises because several covariates especially spatial variables remain constant over time, leading to low variability that can cause numerical instability in the kriging equations. So, we further checked the variability of the covariates and the variance inflation factor (VIF) to minimize collinearity. Additionally, we explicitly exclude the intercept, as indicated by the -1 term in the equation. The final fitting model considered is:

$$PM10 \sim -1 + Temp + U_{media} + V_{media} + Umidit + Season + Lag_1 + Lag_7 + Wind_s + Source \tag{8.1}$$

where $Temp$ represents the temperature, U_media and V_media are the wind speed compo-

nents along the cartesian axes, *Season* is a categorical variable indicating the different seasons, *Source* indicates whether the value is measured by ARPA station or low-sensor, $Wind_s$ represents the average wind speed, and *Umidit* represents the relative humidity. At the same time, Lag_1 and Lag_7 correspond to the PM_{10} concentrations recorded one and seven days before, respectively, capturing the temporal autocorrelation in the data.

Covariance model	**VarComp**	**wMSE**	**Neigh.**	**RMSE**	**MAE**	**R2**
separable	Gau+Gau	[0.001]	10	31.94	16.58	-2.50
product-sum	Gau+Sph	[69.57]	10	27.43	12.98	-1.58
metric	Gau	[2.11]	10	31.94	16.58	-2.50
sum-metric	Gau+Gau+Gau	[0.003]	10	27.43	12.98	-1.58
separable	Gau+Gau	[0.001]	50	10.32	7.19	0.63
product-sum	Gau+Sph	[69.57]	50	9.38	0.70	0.20
metric	Gau	[2.11]	50	10.32	7.19	0.63
sum-metric	Gau+Gau+Gau	[0.003]	50	9.21	6.30	0.71
separable	Gau+Gau	[0.001]	All	9.44	6.63	0.69
product-sum	Gau+Sph	[69.57]	All	12.58	10.13	0.46
metric	Gau	[2.11]	All	9.44	6.63	0.69
sum-metric	Gau+Gau+Gau	[0.003]	All	8.94	6.21	0.73

Table 8: UK Models comparison. Notice that in column VarComp, the first term refers to the spatial decay function, the second term to the temporal decay function, and the third to the spatio-temporal interactions decay function.

Table 8 shows more unstable results when considering a neighbor size of 10, likely due to limited spatial support. However, the predictions become more stable as more information is incorporated when considering a larger number of neighbors (50) or using all available data. The sum-metric model remains the best-performing approach, achieving an MAE of 6.21 and explaining 73% of the variability in the residuals. This result is expected, as the sum-metric model is the only one that explicitly accounts for all three covariance components, leading to better performances. To ensure a robust comparison with the Gams, we consider the same class of covariates used in the UK modeling. In generalization, the isotropic models and those incorporating main effects with interaction produce very similar performance metrics, with $RMSE$ values around 8.8, MAE around 5.85, and a R^2 of around 0.74. In contrast, model 6 using the tensor product formulation performs considerably worse, with a $RMSE$ of 12.76, a MAE of 7.60, and a R^2 of 0.44. We can argue that there is not a significant interaction between coordination and time and not less, does not improve our model, despite Table 6 highlighting the distribution of the residuals was explained more by the joint model. The richness of the joint model does

not lead to meaningful improvements during testing, as assessed through the LOO-CV scheme.

The time series analysis of observed PM_{10} levels and model predictions highlight the performance of the best-performing approaches during the LLO-CV testing phase across different periods. At the ARPA Milano-Verziere station, the GAM model with interaction effect (blue) appears more parsimonious, especially in the presence of outliers, producing a smoother response. The UK captures greater seasonal variability but tends to overestimate the extreme peaks or the immediate changing slope. This pattern is more evident in the second graph, based on data from a low-cost sensor, where the PredUk model struggles to fit the PM_{10} observations.Both models exhibit some degree of limitation depending on the data source, despite achieving overall reasonable performances. Given the result, we employ Mod3 Of Gam with the interaction term for the predictive mapping phase.

Figure 17: Both graphs show the predictions of the GAM and UK Kriging models concerning the observed data. The upper graph corresponds to the Milano Verziere station in the period from 1 January 2022 to 30 April 2022, while the lower graph refers to station 22851 in the period from 1 May 2022 to 30 June 2022.

8 Prediction maps

The analysis of the violin diagrams of the deviation between predicted and observed values shows that the low-cost sensors have a more compact distribution with less variation than the ARPA stations, which exhibit longer tails. This suggests that the model fits the low-cost sensors better, probably because there are more of them in the sample.

To produce the forecast maps, we added other significant variables, that is, *day_of_month* and the *Industrial_commercial_units* to stabilize the forecasts and incorporate more local information. The other land-cover variables did not have a significant impact (with p-values greater than 0.5) since they had low variability. This is mainly due to the 500-meter buffer size considered around each station and the highly urbanized context of the city center, where the training stations are located. Therefore, a grid of low-cost sensor points within the municipality was constructed, and enriched with all necessary information from the ERA5-land, CAMS, and Land Cover programs.

For more informative visual inspection we use the administrative neighborhoods of the city, namely the *Nuclei di Identità Locale* (NILs). Thus, Milan is divided into 88 spatial polygons and each neighborhood contains at least one hypothetical station. For instance, the Chiaravalle area (0.3 km^2) contains only one station, whereas the Parco delle Abbazie area (13.7 km^2) contains 52 stations. Figure 18 shows both the spatial and daily variability in PM_{10} concentrations. The colors used follow the Air Quality Index, ranging from green (indicating good air quality) to purple (indicating extremely poor air quality), reflecting the relative risk associated with short-term exposure for human health. It shows the forecast maps for four different days in the third week of January 2023. Obviously, Monday was a day with PM_{10} levels above 50 $\mu g/m^3$ throughout the city, with northern Milan neighborhoods recording lower levels. The 50–100 range indicates poor air quality, accompanied by a general recommendation to reduce intense outdoor activities. The high concentration is mainly due to lag_1 and $Umidity$. However, conditions improved over the following days, with values in the 20–40 range.

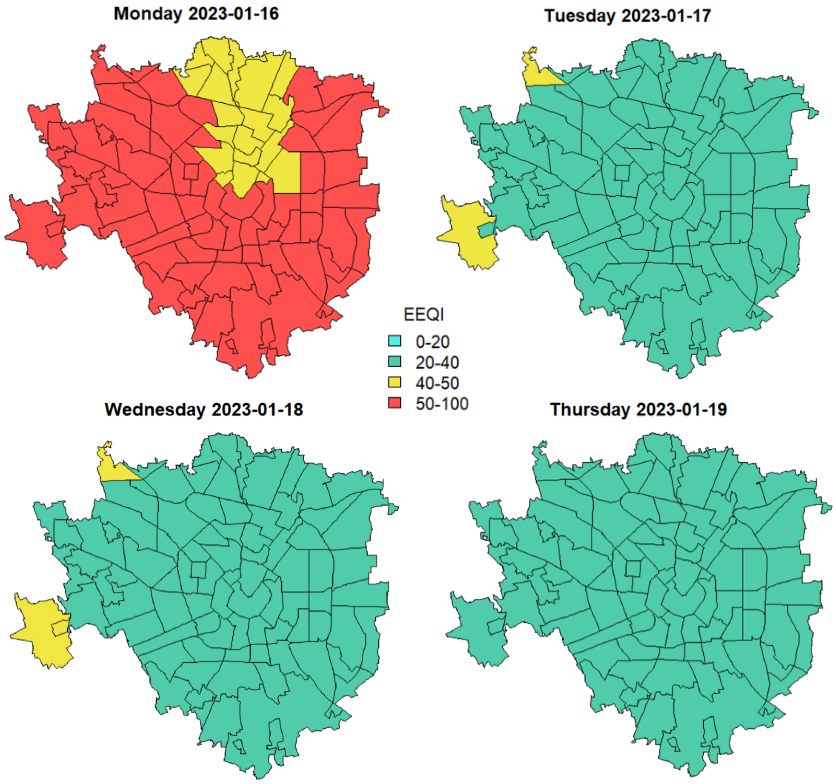

Figure 18: Prediction maps of PM_{10} concentration at four-day stamp in the January month

The following maps show the number of days per year that fall into different bands of the EAQI index. Looking at the map, it can be seen that some western neighborhoods of Milan, such us Trenno and Baggio, record more than 25 days a year with daily PM_{10} level exceeding 50 $\mu g/m^3$. For instance, Cascina Triulza-Expo reaches 35 days, the current maximum EU standards threshold. At the same time, the other neighborhoods, especially those in the center and south-east, show a more attenuated air quality and are within limits, with Parco Monlue - Ponte Lambro and Guastalla stand out among the districts with comparatively cleaner air.

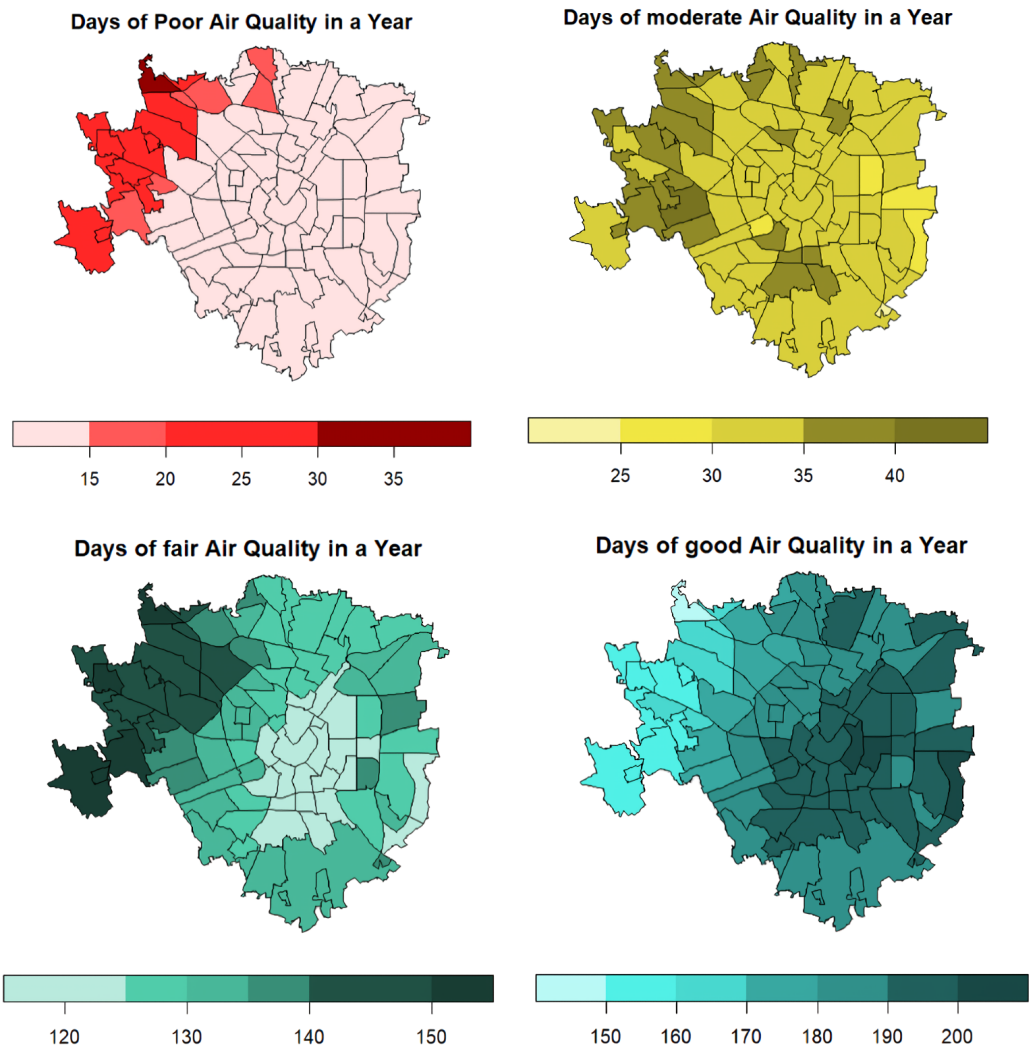

Figure 19: Number of days per year classified into four different air quality categories - poor, moderate, fair and good - across Milan. Each map shows the spatial distribution of the frequency of each category, with darker tones indicating a higher number of days in that category.

9 Discussion and conclusions

9.1 Main mitigation policies Adopted by the city

9.1.1 Emissions Reduction

Milan is part of the international C40 Cities Climate Leadership Group, an organization made up of mayors from greater cities around the world working together to promote urban decarbonization. According to a recent survey by ISTAT [57], only 11.5% of interviewed citizens are satisfied with the air quality, a very low figure compared to other European cities.

As part of its local efforts, Milan has established the Piano Aria e Clima with the ambitious goal of becoming carbon neutral by 2050. The vision refers to a city that is clean and healthy, prosperous and modern, competitive and climate neutral [58]. Willing to improve air quality and enhance the quality of life for its citizens, the public administrators have adopted several policies aimed at reducing major sources of particulate matter. These include promoting the replacement of gas heating systems, renewing the vehicle fleet with fiscal incentives, and reducing traffic and speed limits within residential areas to 30 km/h, a measure introduced as part of the "Città 30" initiative to reduce the resuspension of particulate matter from road surfaces.

One of the main factors that has been developed in Smart Cities is Mobility. In 2012, Milan introduced a traffic restriction zone, Area C, limited to the historical city center and covering 8.2 square kilometers, following a popular referendum with 79% approval, that involved the restriction of many Euro emission classes and a pollution charge that not only has decreased the PM emissions but also traffic accidents and congestion. Furthermore, in 2018, the new Area B was activated. The Area B zone covers almost the entire surface area of the municipality of Milan but with more lax restrictions, contributing to a reduction in traffic flow in the city center and allowing pedestrians to reclaim public spaces and reallocating street space from fast to slow mobility. One instance, the Piazze Aperte program, aims to enhance urban public space as a social space: between 2018 and 2023, 42 tactical urbanism interventions have been completed, including 280 benches, 450 bike racks, 50 picnic tables, 38 ping pong tables, and 380 planters [16].

Parallel to the above measures, policymakers have rethought urban mobility to reduce car ownership by strengthening public transportation and promoting alternative mobility options such as shared vehicles. The pandemic has enforced investment in cycle-pedestrian infrastructure (BiciBus, PediBus), with the city allocating 250 million euros to create 750 kilometers of new cycling corridors between the regional capital and the hinterland [59].

9.1.2 Enhancing Pollutant Uptake

Nature-based solutions (NBS) have emerged in recent years, as a new approach to strengthen urban resilience and sustainability. European Union defines NBS as a strategically planned network of natural and semi-natural areas with other environmental features, designed and managed to deliver a wide range of ecosystem services [61]. These services are the benefits ecosystems provide, which humans rely on for their well-being [62]. Under this broad concept, urban forests, which comprise networks of forests, groups of trees and individual trees [63], can provide important ecosystem services, including air purification (mitigation through carbon sequestration and storage) and adaptation to climate change, restoring of anthropized landscapes, enhancing protected areas and mitigating land consumption. In other words, they contribute to both pollutant removal and exposure reduction. More than just their environmental impact, green spaces serve multiple functions. Beyond their environmental functions, they also have significant social and well-being benefits. The city's policy-makers have created thirteen new parks, each exceeding 10,000 square meters, and have also revitalized existing green spaces to improve urban greenery. Currently, the city's public tree heritage includes 240,000 trees (60% in parks and gardens) [64], with plans to plant 220,000 trees within 2030 in the municipality area. This effort is part of the broader urban forestation program, ForestaMI, which aims to plant three million trees across the entire Metropolitan City [65].

Moreover, the city is promoting the creation of green roofs and facades by offering incentives and deductions to private entities. These initiatives are part of a broader project aimed at the renovation of the building sector, which is central to the ecological transition and improving energy efficiency in Milan's journey toward becoming fully carbon-neutral by 2040 [66].

9.2 Conclusive remarks

Our study demonstrates that urban PM_{10} concentrations in Milan can be effectively predicted using spatio-temporal geo-statistical modes. By integrating heterogeneous data sources and leveraging both the spatial and temporal variability of air pollutant levels, it is possible to construct detailed daily maps that reveal significant intra-urban differences. As it turns out, some neighborhoods, especially those in the north-west have, on average throughout the year, more polluted air than other neighborhoods. Key findings indicate that the implemented geostatistical models provide valuable insights into pollutant dynamics, revealing that concentrations are affected more by previous level concentrations and meteorological rather than spatial factors. Moreover, the comparative analysis of missing data imputation methods shows that Random Forest is a valuable and robust algorithm for reconstructing incomplete time series data when dealing with non linear relationships. In fact, it consistently outperforms other methods across various simulated missingness levels, ensuring reliable predictions even when the consecutive gaps in the data are large. The municipality has implemented numerous initiatives on several fronts to tackle air pollution. However, the daily PM_{10} values revealed the need for targeted and intensive interventions, as well as greater synergy between institutions. For instance, a study conducted in 2010 [16] by Ricerca sul Sistema Energetico highlights that approximately 65% of PM_{10} detected comes from extra-municipal contributions. Therefore, broader plans involving the metropolitan area and adjacent provinces are necessary. At the same time, enhancing public participation in neighborhood-level environmental projects is critical for fostering a sense of belonging and responsibility among residents for their urban environment. For instance, a survey [67] revealed that only 45% of respondents are willing to change their habits to improve the air quality in their city, despite being aware of the harmful effects of air pollution on human health. This underlines the need for environmental initiatives that effectively translate pro-active environmental behavior into the private sphere.

In summary, our work demonstrates that predicting pollution in Milan with high precision is feasible, though it would be better to integrate additional data, such as traffic variables (that we couldn't find available for free) and information on major emission sources near specific locations within the city.

References

[1] Hannah Ritchie and Veronika Samborska and Max Roser.
Urbanization, 2024. *Our World in Data*

[2] OECD. Regions at a Glance, 2013

[3] World Health Organization article

[4] Paolo Maranzano, Riccardo Borgoni, Samir Doghmi, Agostino Tassan Mazzocco. EEAaq R package

[5] Umang Singh, Ajith Abraham, Arturas Kaklauskas, Tzung-Pei Hong. he Role of Smart Sensors in Smart City, 2021 *T.Smart Sensor Networks: Analytics, Sharing and Control: 27-48*

[6] *Decreto Legislativo 13 agosto 2010, n. 155*

[7] World Health Organization.
WHO global air quality guidelines, 2021

[8] World Meteorological Organization (WMO), United Nations Environment Programme (UNEP), International Global Atmospheric Chemistry project (IGAC).
Integrating Low-cost Sensor Systems and Networks to Enhance Air Quality Applications, 2024

[9] Daniel J. Jacob, Darrell A. Winner. Effect of climate change on air quality, 2009. *Atmospheric environment 43(1): 51-63*

[10] World Health Organization. Economic cost of the health impact of air pollution in Europe, 2015

[11] Bart Ostro, Joseph V. Spadaro, Sophie Gumy, Pierpaolo Mudu, Yewande Awe, Francesco Forastiere, Annette Peters. Assessing the recent estimates of the global burden of disease for ambient air pollution: Methodological changes and implications for low- and middle-income countries, 2018. *Environmental research 166: 713-725*

[12] Concessioni autostradali Lombarde. Progetto di monitoraggio ambientale

[13] European Commission. Low-cost sensors offer improved monitoring of air quality, 2024

[14] *sensor.community website*

[15] Hamra GB, Guha N, Cohen A, Laden F, Raaschou-Nielsen O, Samet JM, Vineis P, Forastiere F, Saldiva P, Yorifuji T, Loomis D. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis, 2014.

[16] Municipality of Milan. Piano aria e Clima, 2020

[17] WHO. Outdoor air pollution, 2016

[18] AMAT report.
Rapporto giornaliero sulla qualità dell'aria

[19] Jianjun He, Ye Yu, Yaochen Xie, Hongjun Mao, Lin Wu, Na Liu and Suping Zhao. Numerical Model-Based Artificial Neural Network Model and Its Application for Quantifying Impact Factors of Urban Air Quality, 2016. *Water, Air, & Soil Pollution 227: 1-16*

[20] ARPA report.
Clima, rischi naturali e disponibilità idrica in Lombardia nel 2022, 2023

[21] Ying Li, Alexis Lau, Agnes Wong, Jimmy Fung. Decomposition o he wind and nonwind effects on observed year-to-year air quality variation, 2014. *Journal of Geophysical Research: Atmospheres 119(10): 6207-6220*

[22] Agenzia per la Protezione dell'Ambiente e per i servizi Tecnici (APAT) report. Linee guida per la predisposizione delle reti di monitoraggio della qualità dell'aria in Italia, 2004

[23] Comune di Milano article

[24] AMAT. Report Mobilita Milano 2023, 2024

[25] Institute For Health Metrics and Evaluation.
Report, 2021

[26] Alessandro Fassò. Statistical assessment of air quality interventions, 2013. *Stochastic environmental research and risk assessment 27: 1651-1660*

[27] Simon Wood. Generalized Additive Models, 2007

[28] World Meteorological organization. Integrating Low-cost Sensor Systems and Networks to Enhance Air Quality Applications, 2024. *India Clean Air Summit (ICAS) 2024*

[29] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, Brecht Martens, Diego G. Miralles, María Piles, Nemesio J. Rodríguez-Fernández, Ervin Zsoter, Carlo Buontempo, and Jean-Noël Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, 2021. *Earth system science data 19(3):4349-4383*

[30] Meric Yilmaz. Accuracy assessment of temperature trends from ERA5 and ERA5-Land, 2022. *Science of the Total Environment 856: 159182*

[31] Alessandro Fassò, Jacopo Rodeschini, Alessandro Fusta Moro, Qendrim Shaboviq, Paolo Maranzano, Michela Cameletti, Francesco Finazzi, Natalia Golini, Rosaria Ignaccolo, Philipp Otto. Accuracy assessment of temperature trends from ERA5 and ERA5-Land, 2023. *Scientific Data 10(1): 143*

[32] Copernicus EU

[33] John K. Dixon. Pattern Recognition with Partly Missing Data, 1979. *IEEE Transactions on Systems, Man, and Cybernetics 9(10): 617-621*

[34] W. M. L. K. N. Wijesekara, Liwan Liyanage. *Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index, 2020. Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2: 257-269*

[35] Stef van Buuren and Karin Groothuis-Oudshoorn. Mice: Multivariate Imputation by Chained Equations, 2011. *Journal of statistical software 45: 1-67*

[36] Nickolas Savarimuthu, Shobha Karesiddaiah. An unsupervised neural network approach for imputation of missing values in univariate time series data, 2021. *Concurrency and Computation: Practice and experience 33(9): e6156*

[37] Istituto Superiore di Sanità. PM10 - Particolato atmosferico, 2020

[38] Kuhn and Max. Caret Package, 2008

[39] Lane F. Burgette, Jerome P. Reiter. Multiple Imputation for Missing Data via Sequential Regression Trees, 2010. *American journal of epidemiology 172(9): 1070-1076*

[40] Mosaic report. Risultati del questionario su informazione e percezione della qualità dell'aria a Milano , 2023

[41] Agenzia Mobilità Ambiente e Territorio. Emissions map

[42] Qualità dell'aria - Inventario Emissioni, Arpa Lombardia. *Emissions map*

[43] Legislative Decree 155/2010

[44] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, 2014. *Geoscientific model development 7(3): 1247-1250*

[45] Leo Breiman. Random Forests, 2001. Machine learning 45: 5-32

[46] OECD report. The Economic Consequences of Air Pollution

[47] C. Arden Pope, Douglas W.Dockery. Health Effects of Fine Particulate Air-Pollution: Lines that Connect, 2006. *Journal of the air & waste management association 56 (10): 1368-1380*

[48] European Commission. *Zero Pollution Action Plan, 2021*

[49] Massimo Crespi. Caratteristiche e rappresentatività della metereologia di precisione nel contesto nazionale italiano, 2020

[50] Sensor.Community article

[51] Clausius–Clapeyron relation

[52] G.E.P. Box, D.R. Cox. An Analysis of Transformations, 1964. *Journal of the Royal Statistical Society Series B: Statistical Methodology 26(2): 211-243*

[53] Montero, José-María and Fernández-Avilés, Gema and Mateu, Jorge. Spatial and Spatio-Temporal Geostatistical Modeling and Kriging, 2015. *John Wiley & Sons*

[54] Edzer J. Pebesma. Multivariable geostatistics: the gstat package, 2004

[55] Edzer Pebesma. Spacetime: Spatio-Temporal Data in R, 2012. *Journal of Statistical Software, 51 (7): 1–30*

[56] S.N Wood. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, 2017

[57] Mariana Oliveira, Luís Torgo and Vítor Santos Costa. Evaluation Procedures for Forecasting with Spatiotemporal Data, 2021. *Mathematics, 9(6): 691*

[58] Istat. Qualità della vita nelle città italiane: un confronto europeo, 2023

[59] Municipality of Milan. Piazze Aperte, 2023

[60] Metropolitan City of Milan. Progetto Biciplan, 2021

[61] Article on European Union website

[62] European Union. Ecosystem services

[63] European Cooperation in Science Technology. Memorandum of Understanding for the implementation of the COST Action "European Network for the Integrative Approach of Urban Forestry" (INTUF) CA23148, 2024

[64] Municipality of Milan. Gestione degli alberi

[65] ForestaMI. Report, 2020

[66] Municipality of Milan. Piano aria e Clima, 2020

[67] Mosaic report. Risultati del questionario su informazione e percezione della qualità dell'aria a Milano , 2023

[68] H. D. Patterson, R. Thompson. Recovery of inter-block information when block sizes are unequal, 1971. *Biometrika, 58(3):545-554*

[69] European Air Quality Index

[70] Contreras, Javier and Espinola, Rosario and Nogales, Francisco J and Conejo, Antonio J. ARIMA models to predict next-day electricity prices, 2003. *IEEE transactions on power systems, 18(3): 1014-1020.*

[71] Shtein, Alexandra Kloog, Itai,Schwartz, Joel, Silibello, Camillo Michelozzi, Paola Gariazzo, Claudio Viegi, Giovanni Forastiere, Francesco Karnieli, Arnon Just, Allan C., Stafoggia, Massimo. Estimating Daily PM2.5 and PM10 over Italy Using an Ensemble Model, 2020. *Environmental Science & Technology, 54(1):120-128*

[72] Junyu, Zheng, Jenise, L. Swall, William, M. Cox, Jerry, M. Davis. Interannual variation in meteorologically adjusted ozone levels in the eastern United States: A comparison of two

approaches, 1971. *Atmospheric Environment, 41(4): 705-716*

[73] Junyu, Zheng, Jenise, L. Swall, William, M. Cox, Jerry, M. Davis. Modeling spatial distribution of Tehran air pollutants using geostatistical methods incorporate uncertainty maps, 2016. *Pollution, 2(4): 375-386*

[74] Agenzia Mobilità Ambiente e Territorio (AMAT) Report della mobilità, 2023.

[75] TomTom TomTom traffic index, 2023.