University of Modena and Reggio Emilia

Engineering Department "Enzo Ferrari"
International Doctorate School in Information and Communication Technologies
Cycle XXXVII

Artificial intelligence techniques to tackle urban air pollution

Candidate
Martina Casari

Advisor
Professor Laura Po

Ph.D. Program Coordinator
Professor Luigi Rovati

Tesi di dottorato di ricerca co-finanziata nell'ambito del Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005), risorse FSE REACT-EU, Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione” e Azione IV.5 “Dottorati su tematiche Green”.

Abstract

Nowadays, air quality has become a significant concern due to rising pollution levels, particularly in urban areas, making it a key focus of the 2030 Sustainable Development Goals. According to the European Environment Agency, air pollution causes over 1,200 premature deaths annually among children under 18 in the 32 EEA member countries. Pollution poses serious risks to both human health and the environment. One of the most hazardous pollutants is particulate matter, consisting of tiny solid or liquid particles that can enter the bloodstream when inhaled.

To reduce pollution levels, policymakers implement various countermeasures, such as restricting vehicle circulation in certain urban areas. While these actions are important, it is essential to evaluate their effectiveness and measure their significant impact on pollution levels. To monitor pollution at the urban scale, various types of sensors are employed to detect harmful substances. The concentration levels of particulate detected by sensors could be very accurate if measured by legal stations, developed and installed by environmental agencies; unfortunately, they are also cumbersome and expensive. For this reason, in order to build a finer detection network, cheaper sensors, the so-called low-cost sensors, are used today. This type of technology uses a laser scattering technique to detect the number and concentration of particles, having the advance of requiring a small amount of energy and space. On the other hand, cheaper technology also produces data of lower quality; for instance, because it is not able to reduce the effect of humidity, which, when certain percentages are reached in terms of relative humidity, magnifies the size of the detected particles because they bind to the water vapor present; this problem is called the hygroscopic effect.

This thesis aims to address inaccuracies in data collected by low-cost sensors by leveraging artificial intelligence techniques, in particular by improving anomalous observations resulting from sensor limitations and environmental factors. The literature lacks the ability to cleanse low-cost sensor data of the hygroscopicity problem dynamically with respect to the context and without a legal station as ground truth. In addition, there is a shortage of studies involving different datasets collected in different contexts to explore the capabilities of such models to cleanse data locally but also to generalize air quality properties based on the surrounding environment. To achieve this, machine learning techniques, neural networks, and fuzzy logic are employed to improve the accuracy of the collected data and compensate for these deficiencies. These methods are used to develop models capable of refining the raw data provided by low-cost sensors, ultimately enhancing their reliability for air quality monitoring in various contexts. In addition, a framework capable of reducing the hygroscopic effect, based on past observations has been developed and released as open-source software.

An extensive collection of datasets from multiple cities has been created and made available online for the scientific community. These datasets consist of data from low-cost sensors that have been aligned with legal stations. The data were gathered through collaborations with Italian Environmental Agencies and supplemented by online sources. This comprehensive data collection has facilitated the study of the models' ability to generalize across various contexts. The results are presented in terms of R^2, RMSE, MSE and MAE compared to co-located legal stations. The MitH framework

improves data quality in Turin by 0.3 in R^2. While, general-purpose adjustment models can achieve R^2 improvements of over 0.5, depending on the location and model used.

Contents

Acronyms

AI Artificial Intelligence
AirMLP Air Multilayer Perceptron
ANFIS Adaptive Neuro-Fuzzy Inference System
API Application Programming Interface
AQDR Air Quality Dataset Repository
AQI Air Quality Index

CO Carbon Monoxide
CSV Comma-Separated Values

DNN Deep Neural Network

FNN Feed-Forward Neural Network

GB Gradient Boosting

KNN K-Nearest Neighbors

LCS Low-Cost Sensors
LightGBM Light Gradient Boosting Machine
LR Linear Regression
LSTM Long Short-Term Memory
LUR Land Use Regression

MitH Mitigating Hygroscopicity
MLP Multilayer Perceptron
MLR Multiple Linear Regression

NN Neural Network
NO Nitrogen Monoxide
NO_2 Nitrogen Dioxide

O_3 Ozone
OPC Optical Particle Counters
OPS Optical Particle Sensors

PCA Principal Component Analysis
PM Particulate Matter
PNRR National Recovery and Resilience Plan

R^2 Coefficient of Determination
RF Random Forest
RH Relative Humidity
RMSE Root Mean Square Error

SE Stacked Ensemble
SVR Support Vector Regression

UTC Coordinated Universal Time

XGBoost Extreme Gradient Boosting

Chapter 1

Introduction

Airborne Particulate Matter (PM) is a critical air pollutant with far-reaching implications for the environment, ecosystems, and public health. Composed of a complex mixture of solid and liquid particles suspended in the atmosphere, PM varies in size, composition, and source. These particles are typically categorized by aerodynamic diameter, with classifications such as PM_1, $PM_{2.5}$, and PM_{10} indicating particles smaller than 1 µm, 2.5 µm, and 10 µm, respectively. Smaller particles, particularly $PM_{2.5}$, are of greater concern due to their ability to penetrate deeply into the human respiratory system and even enter the bloodstream, posing severe risks to human health.

The environmental and ecological impacts of PM are equally significant. In the atmosphere, PM contributes to haze, reduces visibility, and disrupts weather patterns by affecting cloud formation and precipitation. When deposited on land and water, PM alters nutrient cycles, pollutes water bodies, and harms ecosystems, particularly in sensitive habitats like wetlands and forests. Toxic components of PM, such as heavy metals, can accumulate in food chains and acidify soils, threatening biodiversity [ZDL+19, RAKB18, VAV+21, XWZ+22, KSHA15].

Addressing the challenges of PM pollution requires effective and reliable air quality monitoring. Traditional monitoring stations, while accurate, are costly and often inaccessible in resource-limited regions. Low-Cost Sensors (LCS) have emerged as a viable alternative, offering affordable and portable solutions that enable widespread air quality monitoring and foster community participation through citizen science initiatives [CFd+23, CBM+22, Ame18, ADMJ22, BGC21, IPdGV+10, KFH+23, ZPBdlC+23, KST+18]. However, LCS measurements are influenced by environmental factors such as temperature and Relative Humidity (RH). High RH levels, in particular, can lead to hygroscopic growth of particles, resulting in overestimated PM concentrations and compromising data accuracy. A more in-depth understanding of particulate matter and low-cost sensors can be found in the Chapter 2.

In summary, this thesis addresses the challenges of improving the accuracy of air quality measurements from low-cost sensors, with a focus on $PM_{2.5}$ data.
Key contributions include:

- The collection and standardization of diverse datasets and the publication of such datasets online to ease the research for open-air quality datasets

- The development of a framework to mitigate hygroscopic effects dynamically without a reference station titled Mitigating Hygroscopicity (MitH)
- The exploration of advanced calibration techniques as a baseline compared to the present literature
- optimized machine learning models for generalization purposes, providing insights to the capability of models to generalize calibration, also providing in the training set data based on some clustering techniques.

- **Collection and Standardization of Air Quality Datasets**: This study gathers and standardizes diverse datasets, making them publicly available to facilitate open research on air quality data. This contribution addresses the challenge of data accessibility and enhances reproducibility in air pollution studies.
- **Development of the MitH Framework**: A novel framework, MitH, is introduced to dynamically mitigate the hygroscopic effects of particulate matter measurements without relying on a reference station. This method improves the reliability of low-cost air quality sensors in varying environmental conditions.
- **Exploration of Advanced Calibration Techniques**: The research investigates and benchmarks state-of-the-art calibration techniques against existing literature, providing a comprehensive comparison and highlighting the most effective approaches to improving sensor accuracy.
- **Application of ANFIS Fuzzy Logic to Air Quality Data**: The study introduces the use of Adaptive Neuro-Fuzzy Inference System (ANFIS) as a novel methodology for air quality data analysis and calibration. This approach leverages the interpretability of fuzzy logic with the learning capabilities of neural networks, providing an innovative solution to enhance the accuracy and reliability of low-cost sensor measurements.
- **Optimized Machine Learning Models for Generalization**: Machine learning models are optimized for calibration generalization, ensuring adaptability across different environmental conditions. Additionally, clustering techniques are applied to enhance training data selection, further improving model performance and applicability in diverse settings.

The research presented in this thesis leverages a diverse dataset collected through collaborations with Italian Environmental Agencies and open-source repositories. Chapter 3 details the Air Quality Dataset Repository (AQDR) [CM24], a standardized repository developed as part of this work. The repository ensures consistency in data processing, including resampling and structuring data for machine learning applications. This comprehensive dataset serves as the foundation for evaluating the impact of hygroscopicity and testing the proposed methods.

A significant contribution of this research is the development of the Mitigate Hygroscopicity (MitH) framework [CP24], a novel approach for reducing measurement inaccuracies caused by high RH. Detailed in Chapter 4, the MitH framework offers several

advantages. It is designed to detect and remove anomalies in data, adapt dynamically to varying environmental conditions, and function effectively in locations lacking reference stations. The framework not only improves the reliability of PM measurements but also serves as a pre-processing step for advanced calibration models. Additionally, the MitH framework is included in the *"Ecosystem for Sustainable Transition in Emilia-Romagna" (ECS_00000033)* part of the ECOSISTER project, funded under the National Recovery and Resilience Plan (PNRR).

To enhance measurement reliability, this thesis explores advanced calibration techniques, incorporating both machine learning models and fuzzy logic systems. A novel approach is presented by integrating fuzzy logic principles with neural network architectures to refine the adjustment of low-cost sensor data [CKP24]. While previous research has examined various machine learning methods for pollution measurement and prediction, the application of ANFIS for calibrating low-cost sensor data constitutes a distinctive contribution to the field. By leveraging the interpretability and flexibility of fuzzy logic alongside the adaptive learning capabilities of neural networks, this study aims to address the limitations of existing methodologies and establish a more robust framework for environmental data analysis. The fusion of fuzzy logic with advanced machine learning techniques highlights an innovative approach to environmental monitoring and underscores the value of interdisciplinary research in tackling critical environmental challenges.

Chapter 5 explores the implementation of Adaptive Neuro-Fuzzy Inference Systems [CKP24] and machine learning models such as LightGBM [CAP24] and multilayer perceptrons (AirMLP) [CPZ23a]. These models undergo rigorous evaluation, focusing on their effectiveness in refining air quality data through anomaly detection, feature selection, and skewness transformation techniques. The findings highlight the potential of these methodologies to significantly improve data accuracy and robustness, reinforcing their suitability for enhancing environmental monitoring systems.

The following chapters of the thesis focus on the generalization of calibration models. Chapter 6 explores methods to enhance the transferability of models across diverse geographical and environmental contexts, ensuring their applicability in new and unseen locations. Complementing this, clustering techniques has been introduced to identify shared characteristics among calibration contexts, supporting the development of generalizable models. Together, these chapters represent a pioneering approach to addressing the challenges of model adaptability and scalability in air quality monitoring.

In Chapter 7, the conclusion highlights the significant advancements achieved in air quality monitoring using low-cost sensors, focusing on improvements in calibration, data standardization, and generalization methodologies. It also emphasizes the practical implications of the developed frameworks, such as MitH, and outlines future research directions, including enhanced sensor integration, advanced clustering techniques, and the development of Digital Twin frameworks for personalized air quality monitoring.

By integrating advanced methodologies and leveraging diverse datasets, this thesis provides a comprehensive study for improving the accuracy and reliability of $PM_{2.5}$ measurements from LCSs. The proposed solutions not only address current limitations

but also lay the groundwork for future advancements in air quality monitoring and environmental research.

Chapter 2

Particulate Matter: Challenges in Monitoring and Regulation

This chapter explores the critical role of low-cost sensors in air quality monitoring and the methods used to address their limitations. While reference stations remain the gold standard for air quality measurements, their high cost and operational requirements restrict their deployment, leaving significant gaps in monitoring networks. Low-cost sensors, despite their inherent inaccuracies, offer a viable solution to enhance spatial and temporal coverage of air quality data.

The chapter is organized as follows:

Air Pollutants: Section 2.1 provides a detailed introduction to particulate matter as a key component of air pollution. It explains its significance, sources, classifications, health impacts, and regulatory guidelines.

Low-Cost Sensors Overview: Section 2.2 introduces the concept of low-cost sensors, focusing on their technological capabilities and the growing importance of their deployment in air quality monitoring. It also highlights the advantages and limitations of these sensors compared to reference stations. A detailed discussion of laser scattering technology, which underpins many low-cost sensors, is accompanied by an overview of commonly used sensors and their specifications.

Impact of Hygroscopicity: Section 2.2.1 delves into the effect of hygroscopicity on particulate matter measurements, particularly at high humidity levels. The challenges associated with humidity-induced inaccuracies in low-cost sensors are illustrated with real-world data, emphasizing the need for correction methods. This section also reviews solutions, such as hardware modifications and advanced calibration techniques, to mitigate the impact of hygroscopicity.

Meteorological Factors: In Section 2.2.2, the chapter examines the influence of meteorological variables on PM concentrations. Factors such as temperature, humidity, wind, precipitation, and atmospheric pressure are analyzed for their dynamic interactions with air quality measurements. Special attention is given to the compounding effects of these variables on sensor accuracy.

Calibration Procedures: Section 2.3 addresses the critical challenge of improving the accuracy of low-cost sensors through calibration. The chapter reviews state-of-the-art calibration methodologies, including machine learning techniques, neural networks, and growth functions. Practical issues such as regional deployment challenges, recalibration frequency, and the trade-offs between number concentration and mass concen-

tration data are also discussed.

The chapter concludes by emphasizing the potential of low-cost sensors to revolutionize air quality monitoring. Through careful calibration, consideration of meteorological impacts, and mitigation of hygroscopicity effects, these sensors can provide reliable data, enabling better-informed decisions for environmental management and public health.

2.1 A Critical Air Pollutant

Air pollution is a major environmental concern, comprising harmful substances in the atmosphere that can affect both human health and the environment. These pollutants can be broadly categorized into gases and particulate matter.

The detection of particulate matter levels is of utmost importance in environmental monitoring and public health. In the atmosphere, particles can originate from both direct emissions and secondary formation processes. Direct emissions refer to particles that are emitted directly from specific sources such as construction sites, unpaved roads, fields, smokestacks, or fires. Industrial facilities, power plants, and vehicles mainly generate these primary pollutants. On the other hand, secondary pollutants are formed through complex chemical reactions that take place within the atmosphere [CL07, LGD+17, SGS+21, DSG+23].

This thesis focuses specifically on the measurement and impact of $PM_{2.5}$, a key pollutant that poses serious health risks, particularly in urban and industrial areas where its concentration is often highest.

Particulate matter is a critical indicator of air pollution due to its ability to adversely impact both human health and the environment. PM is classified based on its aerodynamic diameter, with PM_{10} representing particles with diameters up to 10 micrometers and $PM_{2.5}$ representing finer particles with diameters up to 2.5 micrometers. These classifications are significant because particle size determines how deeply the particles can penetrate into the respiratory system.

PM_{10} particles can reach the upper respiratory tract and are known to cause issues like irritation, coughing, and inflammation. Meanwhile, $PM_{2.5}$ particles are small enough to bypass the body's natural defenses and penetrate deep into the lungs, potentially entering the bloodstream. These fine particles are associated with more severe health effects, including cardiovascular diseases, respiratory illnesses, and even premature death [Uni24].

It is worth noting that the classification of particulate matter into different-size fractions follows a hierarchical structure. In this hierarchy, larger diameter categories, such as PM_{10}, encompass the masses of smaller ones, including PM_4, $PM_{2.5}$ and PM_1, and so forth.

The sources of PM pollution are both anthropogenic and natural [Bel24]. Human activities, such as the burning of fossil fuels, industrial processes, and vehicular emissions, are major contributors to $PM_{2.5}$ and PM_{10}. Natural sources include volcanic eruptions, wildfires, and dust storms. Moreover, particulate matter can be directly emitted from these sources or formed in the atmosphere through chemical reactions involving precursor gases such as sulfur dioxide (SO_2) and nitrogen oxides (NO_x).

2.1.1 Regulations and Reporting Limits

Efforts to reduce particulate matter emissions align with the Sustainable Development Goals (SDGs) [Wor24], specifically those aimed at combating climate change and ensuring sustainable cities and communities.

Air quality regulations and limits play a critical role in monitoring and mitigating air pollution's impact on human health and the environment. These measures are defined by legal and indicative standards, with specific thresholds set for pollutants such as particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and others.

The European Union air quality guidelines have evolved over time, reflecting the growing body of scientific evidence on the impact of air pollution on human health and the environment. Below is an overview of the 2008, 2021, and 2024 guidelines and their key differences.

The 2008 Ambient Air Quality Directive [Eur08] was a significant regulatory update that established legally binding air quality standards for various pollutants, aiming to protect human health and the environment.

- Introduced limit values for key pollutants such as particulate matter (PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), and others.
- Specific limits for PM_{10}:
  - 24-hour limit: 50 $\mu g/m^3$ (with 35 exceedances per year allowed).
  - Annual limit: 40 $\mu g/m^3$.
- Introduced $PM_{2.5}$ exposure reduction targets: Annual limit of 25 $\mu g/m^3$, aiming for reduction by 2015.
- Allowed flexibility for member states, such as extensions for meeting limits.

The 2021 WHO Air Quality Guidelines [Org21] were updated to reflect the latest scientific evidence regarding the health impacts of air pollution. The guidelines introduced more stringent levels for several pollutants, especially particulate matter.

- Reduced the annual limit for $PM_{2.5}$ to 5 $\mu g/m^3$ (down from 10 $\mu g/m^3$ in the 2005 guidelines).
- Set the 24-hour limit for $PM_{2.5}$ to 15 $\mu g/m^3$.
- Reduced the annual and 24-hour limits for PM_{10} to 15 $\mu g/m^3$ and 45 $\mu g/m^3$, respectively.
- These guidelines are not legally binding but serve as the foundation for future policy recommendations and regulatory updates.

The 2024 Ambient Air Quality Directive [PC24] builds on the 2008 Directive and incorporates the latest WHO recommendations. It sets stricter air quality standards for particulate matter and other pollutants and emphasizes cleaner air for all citizens.

- Stricter limits for $PM_{2.5}$:
  - Annual limit: Reduced to 5 $\mu g/m^3$ (aligned with WHO 2021 guidelines).
  - 24-hour limit: Set at 15 $\mu g/m^3$.
- Stricter limits for PM_{10}:
  - Annual limit: Reduced to 15 $\mu g/m^3$.
  - 24-hour limit: Remains at 50 $\mu g/m^3$, but member states are encouraged to adopt stricter measures in high-pollution areas.
- Introduced the concept of Exposure Reduction for $PM_{2.5}$, focusing on reducing population exposure to fine particulate matter.
- Places a stronger emphasis on equity, ensuring that air quality improvements benefit vulnerable and disadvantaged communities.

In Table 2.1 a summary of the key differences between the 2008 Directive, the 2021 WHO Guidelines, and the 2024 Directive.

Aspect	**2008 Directive**	**2021 WHO**	**2024 Directive**
$PM_{2.5}$ Annual Limit	25 $\mu g/m^3$	5 $\mu g/m^3$	5 $\mu g/m^3$
$PM_{2.5}$ 24-hour Limit	Not set	15 $\mu g/m^3$	15 $\mu g/m^3$
PM_{10} Annual Limit	40 $\mu g/m^3$	No update	15 $\mu g/m^3$
PM_{10} 24-hour Limit	50 $\mu g/m^3$ (35 exceed.)	No update	50 $\mu g/m^3$
Exposure Reduction	Introduced for $PM_{2.5}$	Health-based	Stricter reduction for $PM_{2.5}$
Focus Areas	General air quality	Health impacts	Cleaner air for all

Table 2.1: Comparison of the 2008, 2021, and 2024 Air Quality Guidelines

The 2008 Directive established the foundational air quality standards in the EU, the 2021 WHO guidelines reflected the latest health-based evidence and reduced pollutant thresholds, and the 2024 Directive sets even stricter standards for air quality in the EU, with a greater emphasis on equity and population exposure reduction. These evolving guidelines represent the EU's commitment to improving air quality and protecting public health in line with the latest scientific findings.

2.2 Low-Cost Sensors

Although only reference stations (Figure 2.1) provide legally recognized measurements to assess compliance with air quality standards [PC24], the use of low-cost sensors is expanding. These sensors offer a cost-effective means to conduct indicative monitoring in a broader area. In recent years, substantial data has been collected through low-cost sensors and reference stations. Reference stations, developed and installed by environmental agencies, provide high-accuracy measurements. In contrast, low-cost sensors, while less accurate, are increasingly utilized to support established air quality monitoring networks [KANZ22, GSB17, deS22, BCA+21, BRP24].

Figure 2.1: Example of reference station (Rubino station in Turin)

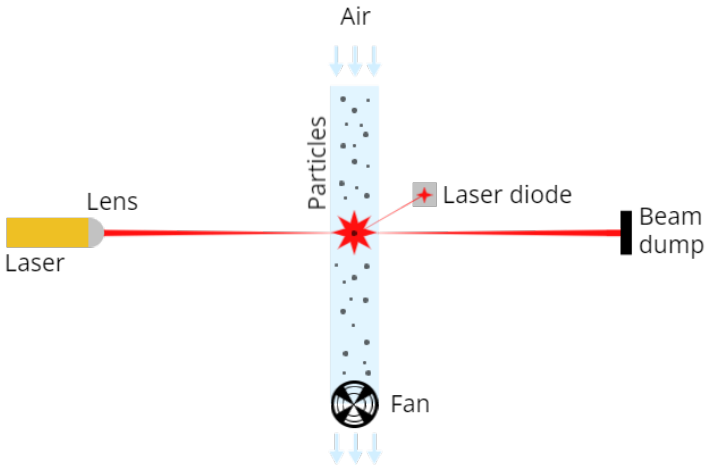


Figure 2.2: Laser scattering technology

In the realm of air quality monitoring, various types of low-cost sensors have been employed to measure pollutants, particularly PM. Instruments used as Optical Particle Sensors (OPS) can be further classified into two primary types: nephelometers and Optical Particle Counters (OPC) [HK20]. Nephelometers operate by measuring particles collectively, capturing light scattered by all particles across a wide range of angles. In contrast, the OPC works by detecting particles individually. A simple laser scattering example is given in Figure 2.2. From the scattered light, with the Mie theory, it is possible to calculate the equivalent particle diameter and the number of particles with different diameters per unit volume. The data collected are then converted into particle mass concentration, expressed in units of micrograms per cubic meter ($\mu g/m^3$).

Table 2.2 compiles information on various low-cost PM sensors commonly studied and discussed in the literature [ABDG+20, GMP+21a, HK20, MTH+20, JLA+20, UOY+20, GTM+16]. For each sensor, the table indicates whether it incorporates a heater designed to generate convective air movement. This convective air movement serves the purpose of directing air within the sensor and, in some cases, heating the

air to reduce PM humidity. It is important to note that in this table, a high cost (on quote) typically refers to sensors equipped with a conditioning mechanism capable of effectively reducing humidity levels in the examined PM, a factor that will be discussed further in Section 2.2.1.

2.2.1 Hygroscopicity

When measuring particulate matter concentrations using low-cost sensors, it is essential to account for the impact of hygroscopicity, which is the ability of particles to attract and retain water molecules from the surrounding air. This phenomenon leads to an increase in the mass of the particles with the absorbed water, which can significantly affect the accuracy of the particle concentration measurements.

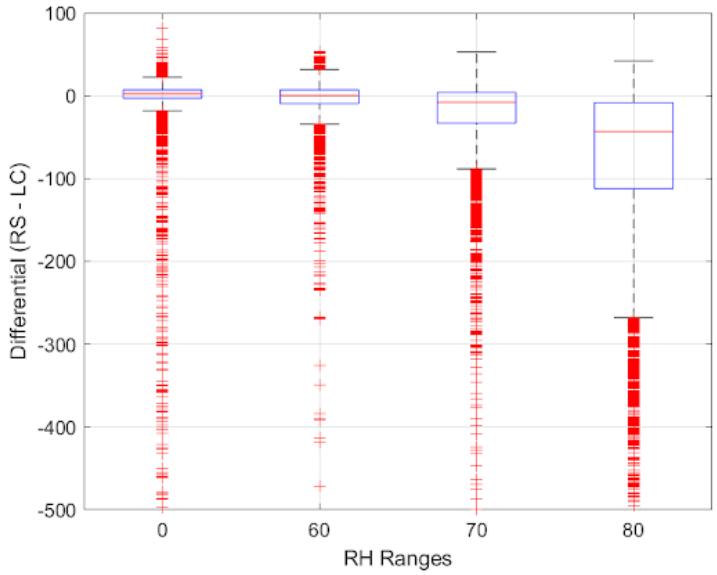


Figure 2.3: Difference between $PM_{2.5}$ mass concentrations detected by the reference station and the low-cost SPS30 sensor, categorized by relative humidity ranges.

A prominent issue with laser scattering technology (Figure 2.2), often employed in low-cost PM sensors, arises from the hygroscopic nature of airborne particles [WOL+21a, Car22]. At elevated humidity levels, these sensors tend to report higher $PM_{2.5}$ concentrations than the actual values. As shown in Figure 2.3, when relative humidity exceeds a specific threshold, the detected $PM_{2.5}$ values exhibit a noticeable increase. This discrepancy occurs because water molecules are absorbed or adsorbed by particles, causing them to expand in size. Consequently, the sensor interprets this increase in mass as a higher concentration of PM, leading to an overestimation.

In Figure 2.3, it is observable that with varying RH ranges, the disparity between low-cost sensor readings and reference station measurements grows with increasing humidity. This demonstrates the critical need for addressing the hygroscopic effect, as it depends on the location and device being used. A prior investigation into the air quality PM concentration dataset is necessary to identify and correct this issue [CP24, PVK+23].

Attempts to resolve this issue by simply eliminating or truncating humidity-induced spikes in $PM_{2.5}$ readings are inadequate, particularly in regions with persistently high

humidity. Such approaches could result in a substantial loss of data. Experiments have shown that sensors equipped with low-cost dryers at their inlets recorded $PM_{2.5}$ concentrations that were, on average, 64% lower during foggy conditions compared to those without dryers [CMLVS22]. However, while effective, adding dryers significantly increases both sensor costs and energy consumption, rendering them impractical for battery- or solar-powered sensors and limiting their feasibility in certain scenarios (see Table 2.2).

More robust solutions involve developing calibration models or employing methods to mitigate the hygroscopicity effect. As demonstrated in Chapters 4 and 5, Adaptive Neuro-Fuzzy Inference Systems, MitH framework or different machine learning models effectively address the hygroscopicity issue. These models not only compensate for inaccuracies induced by humidity but also account for other external factors that contribute to erratic sensor readings, enhancing the reliability of low-cost PM sensors in diverse environments.

2.2.2 Meteorological Impact

Meteorological factors play a crucial role in influencing air quality [OFO+20], particularly in the concentration and dispersion of particulate matter. Temperature, relative humidity, wind speed, precipitation, and atmospheric pressure are among the primary elements that interact dynamically with airborne particles, directly affecting their levels, distribution, and detection accuracy.

Temperature significantly affects PM concentrations through its influence on atmospheric stability and chemical reactions [ZWZK24]. High temperatures promote vertical mixing and dispersion, reducing pollutant concentrations. In contrast, colder conditions, especially during winter nights, often lead to temperature inversions that trap pollutants near the ground, resulting in elevated levels of PM.

As presented in Section 2.2.1, RH is a critical factor in PM behavior due to its effect on particle growth [LZZ+20]. High levels of RH cause hygroscopic particles to absorb moisture, increasing their size and mass, which can lead to overestimated PM measurements, particularly in low-cost sensors. This phenomenon is often observed when RH exceeds a threshold between 60& and 80%, as seen in Chapter 4.

Wind acts as a natural disperser of pollutants. Strong winds dilute PM concentrations by transporting particles away from their sources, while low wind speeds allow pollutants to accumulate locally [OFO+20]. In addition, the direction of the wind determines the spread of pollutants, which could introduce contaminants from external sources into a monitored region.

Rain plays a significant role in the cleansing of the atmosphere by removing airborne particles through wet deposition. This process effectively lowers PM concentrations during and after rainfall events. However, light precipitation may sometimes temporarily increase PM readings as a result of surface particle re-suspension.

Pressure changes, while typically less direct in their influence, can affect air quality by altering weather patterns. High-pressure systems are often associated with stagnant air conditions, reducing dispersion and leading to increased pollutant concentrations. Conversely, low-pressure systems promote airflow and pollutant dilution.

Meteorological factors rarely act independently. Instead, their combined effects and feedback loops complicate air quality dynamics. For example, high RH levels

combined with low temperatures can amplify PM accumulation, while strong winds during a high-pressure system can partially counteract stagnation. Seasonal variations further influence these interactions, requiring region-specific analyzes to obtain accurate predictions.

2.3 Calibration Procedure

Accurate pollution measurement has become a critical objective in environmental science, particularly with the growing use of low-cost sensors [GSB17, deS22, BCA+21, BRP24].

A common strategy to improve accuracy involves co-locating LCS with reference stations for a specified period to collect simultaneous PM data. This process facilitates calibration or adjustment techniques that align the sensor readings with the ground-truth measurements provided by the reference station.

Calibration methodologies have evolved significantly, with machine learning approaches [GMP+21a, CML+21, ABOS22, HZdF+21a, HLT+21, WDWL19, KS21a], including neural networks [PYPL21a, PYPL21b], emerging as leading solutions due to their adaptability and effectiveness. These methods enable robust, data-driven calibration tailored to the specific characteristics of individual sensors and the environmental conditions in which they operate.

Despite advancements, calibration presents several challenges. A key issue arises when sensors are calibrated in one location but deployed in environments with significantly different conditions [DKS+22]. Seasonal variations in environmental conditions and PM concentrations further complicate calibration, often necessitating months of co-located data collection to achieve reliable results.

Additionally, limited access to reference stations in certain regions poses a significant barrier to widespread calibration. To address these issues, alternative approaches, such as growth functions, have been developed [CSP+18, DAPO+18, CP24]. These functions derive correction coefficients based on external factors, such as humidity, which are known to influence PM concentration readings. Importantly, growth functions can be applied without direct calibration against a reference station. Instead, they use parameter optimization to reduce the correlation between external factors (e.g., humidity) and PM measurements, offering a practical solution in resource-limited settings.

Despite their utility, calibration techniques require regular sensor maintenance, with recalibration often needed biannually [BOM+23a]. An additional challenge lies in the debate over whether to prioritize number concentration data directly measured by low-cost sensors or mass concentration values derived through sensor-specific algorithms. This decision significantly impacts data interpretation and application.

The issue of data frequency further complicates calibration efforts. Reference stations typically provide daily averages, which limits the potential for real-time monitoring and fine-grained temporal accuracy. This discrepancy underscores the importance of developing methods that account for both the temporal resolution of low-cost sensors and the reporting standards of reference stations.

In general, calibration remains a critical procedure for maximizing the potential of low-cost sensors. While numerous approaches and techniques exist, addressing contextual challenges and trade-offs is crucial for improving the reliability and applicability

of the collected data.

Table 2.3 showcases notable contributions from the state-of-the-art literature, illustrating various calibration procedures that utilize advanced artificial intelligence techniques. These include machine learning models, neural networks, and other computational methods, which collectively demonstrate the diverse strategies employed to enhance the accuracy and performance of low-cost sensors.

In the table, the results are presented in terms of the Coefficient of Determination (R^2) (Eq. 2.1) and Root Mean Square Error (RMSE) (Eq. 2.2), which are described as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.1}$$

where y_i represents the observed values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ is the mean of the observed values. The R^2 value quantifies the proportion of variance in the dependent variable that is predictable from the independent variables, with a value closer to 1 indicating a better fit.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.2}$$

where y_i and $\hat{y}_i$ are the observed and predicted values, respectively. The RMSE provides a measure of the average magnitude of the prediction error, with lower values indicating a better model performance.

Table 2.3: State of the art of machine learning calibration algorithms for low-cost air quality sensors

Study Title	**Methodology and Approach**	**R^2 Values**	**RMSE Values ($\mu g/m^3$)**
Machine learning techniques to improve the field performance of low-cost air quality sensors [BPL+22]	RF regression during 7 months	0.91	-
Performance Assessment of a Low-Cost $PM_{2.5}$ Sensor for a near Four-Month Period in Oslo, Norway [LSHV19]	Statistical corrections for RH and T using MLR and RF models.	Site 1: 0.80 Site 2: 0.79 Site 3: 0.76	Site 1: 0.80 Site 2: 0.79 Site 3: 0.76

Study Title	**Methodology and Approach**	**R^2 Values**	**RMSE Values ($\mu g/m^3$)**
Evaluation of nine machine learning regression algorithms for calibration of low-cost $PM_{2.5}$ sensor [KS21b]	Best performances using KNN, RF and GB among MLR, Lasso regression, Ridge regression, SVR, MLP and Regression Tree.	Train: 0.99 Test: 0.97 (kNN) 0.96 (RF) 0.95 (GB)	-
Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network [CROD21]	Long-term dataset: RF model considering time-varying co-variates and arterial road length. On-the-fly correction: MLR.	Long-term: 0.75 On-the-fly: 0.78	Long-term: 2.9 On-the-fly: 3.1
Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study [ZSG+20a]	Region-specific multivariate linear regression calibration models for diverse particle sources and meteorological conditions.	Site 1: 0.74 Site 2: 0.95	Site 1:2.46 Site 2: 0.84
Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea [LKV+19]	Land Use Regression (LUR) models using LR, RF and a Stacked Ensemble (SE).	LR: 0.63 RF: 0.73 SE: 0.80	-
Evaluation and calibration of low-cost particulate matter sensors for respirable coal mine dust monitoring [FZZ+23]	A two-layer correction model was introduced, incorporating top-performing models (KNN, RF, ET, XGBoost) and temperature/humidity data.	Sensor 1: 0.97 Sensor 2: 0.98	Sensor 1: 80 Sensor 2: 91

Study Title	**Methodology and Approach**	**R^2 Values**	**RMSE Values ($\mu g/m^3$)**
AirMLP: A Multilayer Perceptron Neural Network for Temporal Correction of $PM_{2.5}$ Values in Turin (Italy) [CPZ23a]	MLP	0.932	
Calibration of Low-Cost Particle Sensors by Using Machine-Learning Method [CKC+18]	LR, SVR and Feed-forward NN.	- Uncalibrated: 0.618 LR: 0.728 SVR: 0.85 FNN: 0.905	
Assessment and Calibration of a Low-Cost $PM_{2.5}$ Sensor Using Machine Learning (HybridLSTM Neural Network) [PYPL21c]	HybridLSTM model combining a deep neural network and an LSTM	- Raw data 0.59 MLR: 0.80 DNN: 0.90 HybridLSTM: 0.93	

Model	Make	Built-in Heater	Technology	PM Detected	Output	Approximate Cost (USD)
DSM501 [DSM]	Samyoung	Yes	LED nephelometer	> 1 micron	Particle counter density	6
GP2Y1010AU0F [GP2]	Sharp	No	IRED and phototransistor nephelometer	> 0.5 micron	Lo Pulse Occupancy	15
PMS1003/3003/ 5003/7003 [PMSb, PMSc] [PMSd, PMSe]	Plantower	No (fan)	Laser scattering nephelometer	PM 1, PM 2.5, PM 10	Particle mass concentration	20
SDS011 [SDS]	Nova	No (fan)	Laser scattering OPC	PM 2.5, PM 10	Particle mass concentration	30
SPS30 [SPS]	Sensirion	No (fan)	Laser scattering OPC	PM 1, PM_4, PM 2.5, PM 10	Particle count and mass concentration	50
HPMA115C0	Honeywell	No (fan)	Laser-based light scattering	PM 1, PM 2.5, PM 4, PM 10	Particle mass concentration	80
HPMA115S0 [HPM]				PM 2.5, PM 10		
PMS1 [PMSa]	Shinyei	Yes (fan)	Light scattering	PM 2.5	Particle mass concentration	?
OPC-N2 [OPCa]	Alphasense	No (fan)	Laser scattering OPC	PM 1, PM 2.5, PM 10	Particle mass concentration	500
OPC-N3 [OPCb]		No (fan)	Laser scattering OPC	PM 1, PM 2.5, PM 4, PM 10		500
NPM2 [NPM]	Met One	Yes (fan)	Light scatter laser nephelometer	PM 2.5, PM 10	Particle mass concentration	2000
10000/12000 [100, 120]	Particle Plus	No (humidity and condensation control)	Optical light scattering	PM 0.3, PM 0.5, PM 1, PM 2.5, PM 5, PM 10	Fine particle counts and mass concentration	On quote
AM520 [AM5]	SidePak	Inlet conditioner	Light-scattering laser photometers	PM 0.8, PM 1, PM 2.5, PM 4, PM 10	Particle mass concentration	On quote
POPS [POP]	NOAA/Handix	No	Light scatter OPC	> 130 nanometre and <3 micron	Particle count	2500
AQMesh [AQM]	Environmental Instruments	Yes	Light-scattering OPC	PM 1, PM 2.5, PM 10	Particle mass concentration	For rent / 8000

Table 2.2: Table showing most explored PM air monitoring low-cost sensors.

Chapter 3

Data Collection

The datasets used in this study were gathered from multiple sources. As part of my PhD collaboration with the industrial partner Wiseair SRL[1], a Milan-based start-up dedicated to improving air quality policies, several SPS30 low-cost sensors were deployed across different Italian cities: Aosta, Trento, Reggio Emilia, and Turin. In each city, agreements were established and formalized with ARPA (the Italian Regional Environmental Agency) to access data from the reference stations where the low-cost sensors were co-located. Notably, Turin was the only location where sensors were already in place prior to the project's inception, whereas the installations in the other cities were developed and implemented during my doctorate.

In addition to the datasets collected in Italy, additional data were sourced from publicly available online platforms. These datasets feature various types of low-cost sensors co-located with reference stations in diverse locations worldwide.

This chapter is organized to present a comprehensive overview of the methodologies, analyses, and augmentations applied to the air quality datasets. It is divided into several key sections, each focusing on a specific aspect of the research:

Wiseair Partnerships: Section 3.1 introduces the datasets used in the study, their sources, and the partnership with Wiseair SRL, which played a crucial role in collecting air quality data from Italian cities using low-cost sensors. A detailed account of the technical aspects of the Arianna devices, data acquisition processes, and characteristics of the Sensirion SPS30 sensors, which form the core of the measurement systems, are also described.

Data Sources: Section 3.2 introduces the datasets used in the study, sourced from online platforms.

Datasets Specifications: Section 3.3 summarizes all datasets collected for the study, including those made publicly available through the Air Quality Datasets Repository. It highlights key features, requirements for dataset inclusion, and their significance for advancing research in air quality modeling and analysis.

Standardization Pipeline: Section 3.4 details the steps taken to align, standardize, and enhance the datasets. It explains the rationale behind retaining only the most relevant features, reducing the data to an hourly frequency, and enriching the datasets

[1] https://wiseair.vision/

with meteorological variables obtained through the Visual Crossing API.

Technical Validation: Section 3.5 validates the data quality and insights derived from it. It includes distribution analyses of $PM_{2.5}$ concentrations across various locations, comparisons of Air Quality Index values, and correlation studies between features. The impact of relative humidity on particulate matter measurements, particularly the hygroscopicity of $PM_{2.5}$, is also explored through comparative examples and summarized in tabular format.

Code Availability: Section 3.6 outlines the tools and resources available for replicating and extending the analyses. It provides information on the programming environment, repository details, and specific Python functions included in the AQDR repository.

Feature Augmentation: Section 3.7 discusses the inclusion of additional contextual variables such as NDVI, industrialization indices, population density, and other environmental and demographic factors. These variables are introduced to enrich the datasets further, providing a more comprehensive understanding of the interplay between air quality and its influencing factors.

By following this structure, the chapter ensures a coherent flow from data preprocessing to advanced augmentation and facilitates replication and further exploration of the datasets.

3.1 Wiseair SRL Partnership

Figure 3.1: Arianna device developed by Wiseair, containing Sensirion SPS30, temperature and relative humidity sensors.

My PhD is part of the PON "Ricerca e Innovazione 14-20: D.M. n. 1061 del 10-8-2021)" program in the ICT field. This project involves a partnership with Wiseair, a company dedicated to innovative air quality monitoring solutions.

In the context of my PhD partnership with Wiseair, air quality data was collected using Arianna devices (Figure 3.1) developed by Wiseair. These devices are equipped

with Sensirion SPS30 sensors and feature an integrated solar panel, allowing them to operate autonomously by harnessing sunlight for power. The Sensirion SPS30 sensor utilizes a laser scattering measurement principle (explained in Section 2.2) and includes contamination-resistant technology. It complies with the MCERTS Performance Standards for Indicative Ambient Particulate Monitors (Version 4, August 2017) [2] and is certified under the MCERTS Performance Standards for Ambient Particulate Monitors [3].

The data from the Arianna devices are collected via a private platform developed by Wiseair or through their API, which encompasses the following variables:

- PM_1, $PM_{2.5}$, PM_4, and PM_{10} mass concentrations ($\mu g/m^3$): these represent different size fractions of particulate matter suspended in the air.
- Relative humidity (%): RH measures the amount of moisture in the air, for instance, a relative humidity of 80% indicates that the air holds 80% of the maximum water vapor it could hold at that particular temperature.
- Temperature (°C): temperature reflects the thermal conditions of the environment.
- Wind speed (m/s): Wind speed denotes the rate at which air is moving.
- Atmospheric pressure (hPA): Atmospheric pressure represents the weight of the air above the sensor.
- Rain (mm): Rainfall indicates the amount of precipitation over a specific time period, measured in millimeters.

These devices record data at 15-minute intervals, although the frequency of data collection can be influenced by factors like battery levels, which depend on the sunlight received by the solar panel. The final evaluation is conducted on an hourly basis to match the granularity of the Arpa reference stations. This is done by calculating the hourly mean of the data collected by each sensor.

In contrast, $PM_{2.5}$ data from Arpa reference stations, located in Aosta, Reggio Emilia, Turin, and Trento, are acquired through direct requests to the agency or through their publicly accessible open portals. These stations provide validated hourly data processed using proprietary methodologies. For example, their techniques often include a standardized approach to managing concentrations below the detection limit, estimating values as high as half of the detection limit for the specific substance under measurement.

Wiseair Arianna devices, on the other hand, use a corrective algorithm designed to enhance measurement accuracy. This includes applying thresholds, a fixed-parameter correction function, and a regression model optimized for autumn conditions. However, for the purpose of this study, the raw data collected by these devices are utilized, bypassing the applied corrections to focus on an unprocessed dataset comparison.

[2] https://sensirion.com/products/catalog/SPS30/

[3] https://sensirion.com/media/documents/3A3BF572/616540E1/Sensirion_PM_Sensors_Datasheet_SPS30_MCERTS-Certificate_2020.pdf

3.1.1 SPS30 Sensor

The sensors used by Wiseair are the Sensirion SPS30 sensors, which are integrated into a device called Arianna, developed and deployed by Wiseair. In addition to the SPS30 sensor, the Arianna device is equipped with relative humidity and temperature sensors, also developed by Sensirion, to provide supplementary environmental data along with particulate matter measurements.

Parameter	Conditions	Value	Units
Mass concentration range	-	0 to 1000	$\mu g/m^3$
Mass concentration size range	$PM_1.0$	0.3 to 1.0	μm
	$PM_{2.5}$	0.3 to 2.5	μm
	PM_4	0.3 to 4.0	μm
	PM_{10}	0.3 to 10.0	μm
Mass concentration precision for PM_1 and $PM_{2.5}$	0 to 100 $\mu g/m^3$	± 10	$\mu g/m^3$
	100 to 1000 $\mu g/m^3$	± 10	$\%m.v.$
Number concentration precision for PM0.5, PM_1 and $PM_{2.5}$	0 to 1000 $\#/cm^3$	± 100	$\#/cm^3$
	1000 to 3000 $\#/cm^3$	± 10	$\%m.v.$
Lifetime	24 h/day operation	>10	years
Maximum long-term mass concentration precision limit drift	0 to 100 $\mu g/m^3$	± 1.25	$\mu g/m^3/year$
	100 to 1000 $\mu g/m^3$	± 1.25	$\%m.v.$ / year

Table 3.1: SPS30 specifications.

Parameter	Recommended Operating Conditions
Temperature	10 to 40 $°C$
Relative humidity	20 to 80 %

Table 3.2: SPS30 recommended operating conditions.

This sensor is frequently used in the literature [KPDWP24a, JXF+24, KPDWP24b] because it is MCERTS certified, has an affordable price, has good electrical parameters, reduced physical dimensions, and uses contamination resistance technology using Sensirion technology. Furthermore, the SPS30 sensor demonstrated high linearity for $PM_{2.5}$ ($R^2 = 0.95$) [NNLV21]. The accuracy of the SPS30 sensor was greater than 95% for $PM_1.0$ mass concentrations below 100 $\mu g/m^3$, but this accuracy decreased to approximately 77% for $PM_1.0$ mass concentrations above 100 $\mu g/m^3$. For $PM_{2.5}$ mass concentrations, the accuracy remained relatively stable, ranging from 81% to 96%. In laboratory experiments conducted at 20$°C$ and 40% relative humidity, the SPS30 sensors generally overestimated the $PM_1.0$ and $PM_{2.5}$ measurements compared to the GRIMM reference instruments [AS].

The specifications given are reported in Table 3.1 and the recommended operating condition in Table 3.2. Furthermore, the SPS30 has a lifetime of ten years in continuous operation, with a start-up time of 30 s. It has a built-in fan to facilitate air transportation. SPS30 sensors operate based on optical particle counting principles utilizing laser scattering (Figure 2.2). Ambient particles are directed to a measurement cell containing a light source and a photodetector. When particles interact with light, some of it scatters to the photodetector. The collected signal is processed to obtain real-time particle count and mass concentration values, which are expressed in units of particles per cubic centimeter ($\#/cm^3$) and micrograms per cubic meter ($\mu g/m^3$), respectively. It is worth noting that the PM_4 and PM_{10} outputs of Sensirion's PM sensors are estimated from measurements of PM0.5, PM_1.0 and $PM_{2.5}$. These estimates take into account typical aerosol profiles rather than being based solely on real raw data events from larger particles.

Even if the sensor itself has a sampling frequency of 1 ± 0.04 seconds, the data collected are transmitted to the Wiseair server at intervals of every 15 minutes. Nevertheless, the frequency of data transmission may vary depending on the battery life of the device, which is charged by a small solar panel located on the device's surface. The air enters the device without the aid of a pump, passing through a grid designed to prevent the entry of insects and larger particles.

The Arianna device, which incorporates the SPS30 sensor, is continuously being refined to enhance its ability to collect more accurate data, such as by adding components to prevent direct contamination from wind or rain.

3.2 Open Source Data

Several, though not many, datasets have been published in the literature, or on open repositories, in which low-cost sensors have been co-located with a legal station to collect PM levels at a frequency of at least 1 hour. Given the difficulty of finding such co-located PM datasets, this thesis also aims to collect and present the available data, which have been retained if they meet certain minimum requirements.

The mandatory rules to maintain the datasets are the following:

- The LCS data must have been collected during a co-location period with the RS.
- The LCS and RS data must both be present and aligned.
- The detection frequency must be an hourly or lower.
- The dataset must have the $PM_{2.5}$ column in the LCS and RS data.
- The collection period must be longer than 3 weeks.
- Datasets must originally be shared under a CC BY 4.0 license.

Access to diverse PM datasets is crucial for advancing future research, as it allows for refining models by examining correlations and patterns in PM behavior across different locations and contexts, including environmental and territorial factors. This comparative approach enables deeper insight into how particulate matter behaves under various conditions, enhancing the robustness and accuracy of the model.

These datasets have been published openly in a repository called Air Quality Datasets Repository [CM24], so other researchers around the world can contribute by uploading new datasets following the listed rules. In the repository, the name of the dataset consists of the surname of the first author of the original work and the city where the data were collected, as presented in Table 3.3, along with the types of LCS used, the sampling period and frequency, and the features released.

The datasets presented in AQDR were collected by searching in the literature on Google or Scopus using keywords such as *air quality*, *PM*, *low-cost*, *correction* and *calibration*.

3.3 Datasets Specification

In this section, a detailed overview of the collected datasets is provided, and a summary is presented in Table 3.3.

These datasets have been standardized by merging LCS and RS readings when they were not originally in the same file, resampling frequencies and adding features via an API service [Cro] to facilitate future research.
The levels of PM concentration detected by low-cost sensors could be influenced by abnormal sensor behavior or physical effects (see Section 2.2.1), such as the hygroscopic effect caused by high levels of relative humidity, typically when RH is about 70% [PVK+24, CP23b].

Although some of the proposed datasets are more affected by anomalous sensor behavior than others, whether the hygroscopic effect is present can be easily seen by plotting the reference station against low-cost sensor PM data while also maintaining RH levels. Where RH clearly affects the data, the LCS series increases, creating a gap from the lower RS series (for further details, refer to Section 3.5.1). In this work, the hygroscopic effect is maintained but can be reduced using several techniques [CP24, HZdF+21b, WOL+21b].

It is worth mentioning that gaps in the data have been preserved, which means that researchers looking for a continuous time series will have to apply interpolation techniques to fill these missing values.

For each city, the dataset variables are outlined and explained below.

3.3.1 Aosta

The dataset used for the analysis comes from various sources. Data from low-cost sensors were obtained via the Wiseair API, while data from reference stations were provided through a collaboration with Arpa Aosta. The monitoring location was Piazzale Plouves, located in Aosta (11100 Aosta AO).

Data collection occurred between February 4, 2024, and June 30, 2024, covering a period of approximately four months. The dataset was stored in CSV format.

- **Low-cost sensors**: Model SPS30.
  The dataset includes data from two sensors, identified as ari-2432 and ari-2433. From May 22, 2024, onward, data from sensor ari-2432 were excluded from the analysis due to systematically underestimated values compared to the reference station.

Dataset Ref	City	LCS Type (quantity)	Sampling Period	Sampling Frequency	All Features Involved
Arroyo - Badajoz [AGSSL21]	Badajoz (Spain)	OPC-N3 (2)	2021-03-12 2021-05-17	LCS aligned with RS: 10min and 1h	NO, NO_2, O3, CO, $PM_{2.5}$, PM_{10}, RH, temperature
Bulot - Southampton [BOM+23b, Bul22a]	Southampton (UK)	SPS30 (8) PMS5003 (8)	2020-07-01 2021-07-03	LCS: 2min RS: 2min	PM_1, $PM_{2.5}$, PM_4, PM_{10}, RH, dew point, pressure, temperature
Campmier - Delhi Hamirpur Bengalore [CGS+23a, CGS+23b]	Delhi, Hamirpur, Bengalore (India)	PurpleAir PA-II (PMS5003 inside)	Delhi: 2018-07-24 2020-01-03, Hamirpur: 2020-03-22 2021-01-11, Bengalore: 2019-06-21 2020-07-31	LCS: 1h	$PM_{2.5}$, RH, dew point, temperature
Casari - Torino [CP23a]	Turin (Italy)	SPS30 (5)	2022-03-01 2022-12-30	LCS: 15min RS:1h	PM_1, $PM_{2.5}$, PM_4, PM_{10}, RH, temperature, pressure, wind speed, rain mm, cloud coverage
Casari - Aosta [CP23a]	Aosta (Italy)	SPS30 (2)		LCS: 15min RS:1h	PM_1, $PM_{2.5}$, PM_4, PM_{10}, RH, temperature, pressure, wind speed, rain mm
Minxing - Calgary [SXDD20, Si19]	Calgary (Canada)	PMS5003 (1)	2018-12-07 209-04-26	LCS aligned with RS: 1h	$PM_{2.5}$, RH, temperature, wind direction, wind speed
Villanueva - Lima [VEC+23]	Lima (Peru)	IQAir (4), AirBeam (3) (PMS7003 inside)	IQAir: 2021-11-24 2021-12-30 AirBeam: 2021-11-15 2022-01-08	IQAir: 15min AirBeam: 1min RS: 1h	$PM_{2.5}$, PM_{10}, RH, temperature

Table 3.3: Table showing the datasets collected with the raw properties as found online. Each of these datasets is subject to CC BY 4.0. All datasets have LCS and RS co-localized.

- **Reference station**: Classified as an "urban background" station, this station is designed to assess the average exposure of the population in Aosta. It is located in an area that is not significantly influenced by specific emission sources (e.g., industries, traffic, or residential heating), but rather by the general contribution of all environmental sources.

The final dataset contains 6068 records and 18 columns after the removal of a few

missing values to ensure completeness. The original columns were:

- *created_at_utc*: Measurement timestamp (in UTC).
- *device_id*: Sensor identifier.
- *pm2p5_ug_m3*: $PM_{2.5}$ concentration, measured in micrograms per cubic meter ($\mu g/m^3$).
- *pressure_hpa*: Atmospheric pressure, measured in hectopascals (hPa).
- *relative_humidity_pc*: Relative humidity, expressed as a percentage (%).
- *temperature_celsius*: Temperature, measured in degrees Celsius (°C).
- *wind_speed_m_s*: Wind speed, expressed in meters per second (m/s).
- *rain_mm*: Rainfall amount, measured in millimeters (mm).

These columns were renamed to standardize the dataset structure as follows: *valid_at, sensor_id, pm2p5_x, pressure, relative_humidity, temperature, wind_speed,* and *rain*. A subsequent join operation merged the low-cost sensor data with reference station data, adding a column named *pm2p5_y*, which represents the reference $PM_{2.5}$ value (ground truth), essential for comparative analysis. In the final version of the dataset, the *valid_at* feature was decomposed into *hour, day, month,* and *year*, while retaining the original version. To provide additional context, new features were added, including *city, altitude, latitude, longitude, density_of_population,* and *type_sensor* to specify the sensor type. This naming convention was adopted to standardize features across the various analyzed datasets. The data collected by low-cost sensors has a sampling frequency of 15 minutes, while reference station data are recorded hourly in GMT+1 timezone. To ensure temporal consistency, the reference station data timestamps were shifted back by one hour. Subsequently, timestamps were rounded to the nearest hour, retaining only the row with the smallest temporal difference from the rounded hour for each observation.

3.3.2 Badajoz

The dataset was obtained from the publication [AGSSL21]. The monitoring location was Av. Antonio Masa Campos, 30, in Badajoz (Spain). The measurements were taken from March 12, 2021, to May 17, 2021, covering a period of approximately two months. The data was initially obtained in XLSX format and then converted to CSV format.

The sensor types considered were:

- **Low-cost sensors**: Model stam OPC-N3.
  The dataset includes data from two sensors, identified as FEC01 and FEC02.
- **Reference station**: The data was sourced from the Air Quality and Protection Network of Extremadura (REPICA), managed by the Department for Ecological Transition and Sustainability of the regional government of Extremadura.

The final dataset contains 3152 records and 18 columns, after the removal of a few missing values to ensure completeness.

The original data used for the analysis comes from a single file that includes both low-cost sensor data and reference station data. The original columns are as follows:

- *Date*: Measurement date.
- *Muestra*: Measurement time.
- *UM_NO, UM_NO2, UM_O3, UM_CO, UM_PM25, UM_ PM_{10}, UM_TMP, UM_RH*: These represent the values of Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Ozone (O_3), Carbon Monoxide (CO), $PM_{2.5}$, PM_{10}, Temperature, and Relative Humidity from the reference station.
- *FEC01_NO2, FEC01_O3, FEC01_NO, FEC01_CO, FEC01_ PM_1, FEC01_ $PM_{2.5}$, FEC01_ PM_{10}, FEC01_TMP, FEC01_RH, FEC01_Lat, FEC01_Long*: These represent the values of Nitrogen Dioxide (NO_2), Ozone (O_3), Nitrogen Monoxide (NO), Carbon Monoxide (CO), PM_1, $PM_{2.5}$, PM_{10}, Temperature, Relative Humidity, Latitude, and Longitude for sensor FEC01.
- *FEC01_T_OPC, FEC01_H_OPC, FEC01_SFR_OPC*: These represent the temperature, humidity, and flow rate measured by the optical sensor for FEC01.
- *FEC01_WE_NO2, FEC01_WE_O3, FEC01_WE_NO, FEC01_WE_CO*: These represent the raw signals from the electrochemical sensors (Working Electrode) for detecting NO_2, O_3, NO, and CO for FEC01.
- *FEC01_AE_O3, FEC01_AE_NO2, FEC01_AE_NO, FEC01_AE_CO*: These represent the compensation signals (Auxiliary Electrode) for the respective electrochemical sensors for FEC01.

The features related to sensor FEC02 are the same as those for FEC01.

All pollutant values are measured in micrograms per cubic meter ($\mu g/m^3$), while temperature is measured in degrees Celsius (°C) and humidity is expressed as a percentage (%).

For the analysis, only the following columns were considered: Date, Muestra, UM_PM25, UM_TMP, UM_RH, FEC01_ $PM_{2.5}$, FEC02_ $PM_{2.5}$. These were renamed to: *valid_at* (combining Date and Muestra), *sensor_id* (to specify the reference low-cost sensor, i.e., FEC01 or FEC02), *pm2p5_x* (for the particulate matter value from the sensor), *pm2p5_y* (for the particulate matter value from the reference station, i.e., the ground truth), *relative_humidity* and *temperature*.

Remaining meteorological features such as *pressure, wind_speed,* and *rain* were downloaded from the Visual Crossing API [Cro] with hourly frequency. To align with the feature structure, the *valid_at* column was also decomposed into *hour, day, month,* and *year*, while keeping the original version. Additionally, new features such as *city, altitude, latitude, longitude, density_of_population,* and *type_sensor* were added.

Low-cost sensor and reference station data have a sampling frequency of 10 minutes. To ensure temporal consistency across all datasets, timestamps were rounded to the nearest hour, and for each hour, only the row with the smallest temporal difference from the rounded hour was kept.

3.3.3 Bangalore, Delhi, and Hamirpur

These three datasets were obtained from the same study [CGS+23a]. The monitoring location for the Bangalore dataset was at 10th, No.18, Mayura Street, Cross, Papanna Layout, Nagashetty Halli, R.M.V. 2nd Stage, Bengaluru, Karnataka 560094 (India), and the measurements were taken from June 21, 2019, to July 31, 2020, covering approximately 1 year and 1 month.

The monitoring location for the Delhi dataset was H5WP+MQQ, Panchsheel Marg, Shantipath, Chanakyapuri, New Delhi, Delhi 110021, (India), with measurements taken from July 24, 2018, to January 3, 2020, covering approximately 1 year and 5 months.

The monitoring location for the Hamirpur dataset was Bharat Uday Gurukul Campus, -RURI 210301, Para, Uttar Pradesh (India), with measurements taken from March 22, 2020, to January 10, 2021, covering approximately 9 months.

The data was obtained in XLSX format and converted to CSV.

The sensor types used were the same across all three datasets:

- **Low-cost sensors**: Model PMS5003.
  The dataset includes one sensor for each location.
- **Reference station**:
  - BAM1022: for Bangalore and Hamirpur.
  - BAM1020: for Delhi.

The final dataset for all three cities consists of 18 columns, with varying numbers of rows: 7441 for Bangalore, 7504 for Delhi, and 3638 for Hamirpur. The original data used for the analysis comes from separate files, with each dataset corresponding to a single file containing both low-cost sensor and reference station data.

The original columns are as follows:

- *Index*: Timestamp of the measurement.
- $PM_{2.5}$: The concentration of particulate matter ($PM_{2.5}$) in micrograms per cubic meter ($\mu g/m^3$).
- *Temperature*: Temperature in degrees Celsius (°C).
- *Humidity*: Humidity in percentage (
- *Pressure*: Atmospheric pressure in hectopascals (hPa).
- *WindSpeed*: Wind speed in meters per second (m/s).
- *Rainfall*: Rainfall amount in millimeters (mm).

The dataset was standardized by renaming the columns: *valid_at, sensor_id, pm2p5_x, pressure, relative_humidity, temperature, wind_speed, rain.* After merging the low-cost sensor and reference station data, a new column *pm2p5_y* was added, representing the reference $PM_{2.5}$ values.

The *valid_at* column was decomposed into *hour, day, month, year*, and new features such as *city, altitude, latitude, longitude, density_of_population,* and *type_sensor* were added. As in other datasets, meteorological data were sourced from the Visual Crossing API [Cro].

3.3.4 Calgary

The dataset was obtained from the publication [SXDD20]. The monitoring location was 6120 2 St SE, Calgary, AB T2H 0P3 (Canada). The data collection took place from December 7, 2018, to April 26, 2019, covering a period of approximately five months. The data was provided in CSV format.
The sensor types considered are:

- **Low-cost sensors**: Model PMS5003.
  The dataset includes data from a single sensor.
- **Reference station**: SHARP Thermal Fisher Scientific 5030.

The final dataset contains 3026 records and 18 columns after removing a few missing values to ensure completeness.
The original data used for analysis came from a single file containing both the low-cost sensor data and the reference station data. The original columns are listed below:

- *TimeStamp*: Timestamp of the measurement.
- *PM_AirShed*: Concentration of $PM_{2.5}$ from the reference station, measured in micrograms per cubic meter ($\mu g/m^3$).
- *PM_Sensor*: Concentration of $PM_{2.5}$ from the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *RH_Sensor*: Relative humidity, expressed as a percentage (%).
- *Temp_Sensor*: Temperature, measured in degrees Celsius (°C).

All original columns were retained and, for consistency with other datasets, were renamed to: *valid_at*, *sensor_id*, *pm2p5_x* (sensor particulate value), *pm2p5_y* (reference station particulate value, i.e., the ground truth), *relative_humidity*, and *temperature*. The remaining meteorological features, namely *pressure, wind_speed*, and *rain*, were retrieved from the Visual Crossing API at an hourly frequency. To maintain consistency with the feature structure, the *valid_at* column was decomposed into *hour, day, month, year*, while also keeping the original format. Additionally, new features such as *city, altitude, latitude, longitude, density_of_population*, and *type_sensor* were added.

The data collected from both the low-cost sensors and the reference station have an hourly frequency.

3.3.5 Lima

For Lima, two datasets were obtained, which differ only in the type of sensor used, as both datasets were derived from the research [VEC+23], where two different particulate sensors were tested at the same site and under the same conditions.
The monitoring location was the rooftop of the Palacio Municipal in Lima, located in the main square of Lima (Peru). The data collection took place from November 15, 2021, to January 14, 2022, covering a period of approximately two months. The data was originally provided in XLSX format and was later converted to CSV.
The sensor types considered are:

- **Low-cost sensors**:

    1. **IQAir AirVisual model**.
       The dataset includes data from four sensors, identified as 4V49RUR, TRW49SR, TUXVAJK, and VSK74UJ. The period before 17:00 on November 23, 2021, was removed for all sensors due to anomalous readings, while for sensor 4V49RUR, data after 09:00 on December 29, 2021, was removed because it underestimated the reference station's values.
    2. **PMS7003 model**.
       The dataset includes data from three sensors, identified as 40a, 4E5, and 5e8.

- **Reference station**: FED - Teledyne API T640 Mass Monitor.

The final dataset for both types contains 18 columns, but for the IQAir sensors, it includes 4279 records, while for the PMS7003 sensors, it includes 3619 records.
The original data used for the analysis came from separate files, with structure varying by sensor type. For the low-cost sensors, each file corresponds to a single device, and the format is uniform within each sensor type. In contrast, the reference station data is organized into two separate files, one for pollutants and another for meteorological variables. The original columns in the files are listed below.

IQAir:

- *Timezone*: Timezone.
- *Datetime*: Timestamp of the measurement.
- *AQI US*: Air quality index in the United States.
- *AQI CN*: Air quality index provided by the Air Quality Index China Network.
- $PM_{2.5}$: Concentration of $PM_{2.5}$ for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- PM_{10}: Concentration of PM_{10} for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- PM_1: Concentration of PM_1 for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *CO2*: Carbon dioxide level, expressed in parts per million.
- *Temperature*: Temperature, measured in degrees Celsius (°C).
- *Temperature*: Temperature, measured in degrees Fahrenheit (°F).
- *Humidity*: Relative humidity, expressed as a percentage (%).

PMS7003:

- *Session_Name*: Sensor name.

- *Timestamp*: Timestamp in ISO 8601 format.
- *Latitude*: Latitude of the sensor.
- *Longitude*: Longitude of the sensor.
- *Temperature*: Temperature, measured in degrees Fahrenheit (°F).
- *PM1*: Concentration of PM_1 for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *PM10*: Concentration of PM_{10} for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- $PM_{2.5}$: Concentration of $PM_{2.5}$ for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *Humidity*: Relative humidity, expressed as a percentage (%).

Reference station:

- *Date*: Timestamp of the measurement.
- *Temperatura*: Temperature, measured in degrees Celsius (°C).
- *Humedad*: Relative humidity, expressed as a percentage (%).
- *Velocidad*: Wind speed, measured in meters per second (m/s).
- *Direccion*: Wind direction, expressed in letters.
- *Presion Atmosferica*: Atmospheric pressure, measured in hectopascals (hPa).
- *Precipitacion*: Rainfall, measured in millimeters (mm).
- *Radiacion solar*: Solar radiation, measured in W/m^2.
- *Indice UV*: UV index.
- *PM10 Conc*: PM_{10} concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- $PM_{2.5}$ *Conc*: $PM_{2.5}$ concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).

To align with the features considered for all other datasets, only the $PM_{2.5}$ pollutant-related columns and the sensor identification were retained for the low-cost sensors. These were renamed as *pm2p5_x* and *sensor_id*, while the remaining features were extracted from the reference station files and renamed as: *valid_at*, *pm2p5_y* (the reference station particulate value, i.e., the ground truth), *relative_humidity, temperature, pressure, wind_speed, and rain*. To maintain consistency with the feature structure, the *valid_at* column was also decomposed into *hour, day, month, year*, while keeping the original format. Additionally, new features such as *city, altitude, latitude, longitude, density_of_population*, and *type_sensor* were added.

The data collected from both the low-cost sensors and the reference station have an hourly frequency.

3.3.6 Southampton

For the city of Southampton, as with Lima, two datasets were created, which differ only in the type of sensor, since both were obtained from the research [BOM+23c], where two different types of particulate sensors were tested in the same manner.
The monitoring took place at the National Oceanography Centre, European Way, Southampton SO14 3ZH (United Kingdom), from July 1, 2020, to July 3, 2021, covering a total period of approximately one year. The data was obtained in CSV format.
The sensor types considered are as follows:

- **Low-cost Sensors**:

  1. **PMS5003 Model**.
     The dataset includes data from eight sensors, identified as PMS5003-2018062702491, PMS5003-2019031808075, PMS5003-2019031807860, PMS5003-2019031807862, PMS5003-2019031807863, PMS5003-2019031807865, PMS5003-2019031807856, and PMS5003-2019031806686. The last two sensors ceased functioning at 01:00 on December 23, 2020, while sensor PMS5003-2019031807860 was completely excluded from the analysis due to erratic values compared to the others.
  2. **SPS30 Model**.
     The dataset includes data from eight sensors, identified as SPS30-60767820-21, SPS30-60767820-38, SPS30-60767820-19, SPS30-60767820-31, SPS30-60767820-36, SPS30-60767820-41, SPS30-60767820-37, and SPS30-60767820-40. The sensor SPS30-60767820-38 ceased functioning at 01:00 on December 23, 2020, while the last sensor in the list was completely excluded as it was not operational.

- **Reference Station**: Fidas 200, an optical instrument that meets the equivalence criteria for particulate monitoring and is certified by the Environment Agency's MCERTS for particulate monitoring in the UK, as well as MCERTS for Continuous Ambient Air Monitoring Systems.

The final dataset has 18 columns for both sensor types, with the SPS30 dataset containing 47,343 records, and the PMS5003 dataset containing 43,523 records.
The original data used for the analysis comes from separate files, which vary depending on the sensor type. For the low-cost sensors, there is a single file, while the data from the reference station is organized into two separate files—one for pollutants and one for meteorological variables. The original columns in the files are listed below.

SPS30 and PMS5003 Sensors:

- *sensor*: Sensor identifier.
- *site*: Indicates the monitor on which the sensor was placed.
- *median_ PM_1*: PM_1 concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *median_ PM_{10}*: PM_{10} concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).

- *median_$PM_{2.5}$*: $PM_{2.5}$ concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *median_ PM_4*: PM_4 concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$), available only for the SPS30 sensor.
- *median_n05*: Number concentration of particles between 0.3 and 0.5 µm, available only for the SPS30 sensor.
- *median_n1*: Number concentration of particles between 0.3 and 1 µm, available only for the SPS30 sensor.
- *median_n10*: Number concentration of particles between 0.3 and 10 µm, available only for the SPS30 sensor.
- *median_n25*: Number concentration of particles between 0.3 and 2.5 µm, available only for the SPS30 sensor.
- *median_n4*: Number concentration of particles between 0.3 and 4 µm, available only for the SPS30 sensor.
- *median_gr03um*: Number concentration of particles >0.3 µm, available only for the PMS5003 sensor.
- *median_gr05um*: Number concentration of particles >0.5 µm, available only for the PMS5003 sensor.
- *median_gr100um*: Number concentration of particles >10 µm, available only for the PMS5003 sensor.
- *median_gr10um*: Number concentration of particles >1 µm, available only for the PMS5003 sensor.
- *median_gr25um*: Number concentration of particles >2.5 µm, available only for the PMS5003 sensor.
- *median_gr50um*: Number concentration of particles >5 µm, available only for the PMS5003 sensor.
- *median_ $PM_{10}0$_cf1*: Mass concentration of PM_{10} with cf1 calibration for PMS5003.
- *median_ PM_{10}_cf1*: Mass concentration of PM_1 with cf1 calibration for PMS5003.
- *median_pm25_cf1*: Mass concentration of $PM_{2.5}$ with cf1 calibration for PMS5003.
- *date*: Timestamp of the measurement.

Reference Station:

- *datetime_cut*: Timestamp of the measurement.
- *$PM_{2.5}$*: $PM_{2.5}$ concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).

- *PM_{10}*: PM_{10} concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *PMtot*: Total concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *PM_1*: PM_1 concentration for the sensor, measured in micrograms per cubic meter ($\mu g/m^3$).
- *rh*: Relative humidity, expressed as a percentage (
- *ws*: Wind speed, expressed in meters per second (m/s).
- *wd*: Wind direction, expressed in degrees (°).
- *precipitation_intensity*: Precipitation intensity, expressed in millimeters (mm).
- *precipitation_type*: Type of precipitation.
- *dew_point_temperature*: Dew point temperature, measured in degrees Celsius (°C).
- *air_pressure*: Atmospheric pressure, measured in hectopascals (hPa).
- *temperature*: Temperature, measured in degrees Celsius (°C).

To align with the features considered for all other datasets, for the low-cost sensors, only the features related to $PM_{2.5}$ were retained (in particular, the values defined as median_$PM_{2.5}$ for the SPS30 sensor and median_pm25_cf1 for the PMS5003 sensor), and the sensor identification was retained. These columns were renamed to *pm2p5_x* and *sensor_id*, while the remaining features were considered in the reference station files and renamed to: *valid_at*, *pm2p5_y* (for the reference station $PM_{2.5}$ value, i.e., the ground truth), *relative_humidity*, *temperature*, *pressure*, *wind_speed*, and *rain*. To maintain consistency with the structure of the features, the *valid_at* column was reworked into *hour*, *day*, *month*, and *year*, while the original *valid_at* column was also kept. Additionally, new features such as *city*, *altitude*, *latitude*, *longitude*, *density_of_population*, and *type_sensor* were added.

The data collected by both low-cost sensors and the reference station had a recording frequency of every 2 minutes. The timestamps were rounded to the nearest hour, and only the record with the smallest time difference from the rounded hour was kept.
During preliminary analysis, it was found that the reference station measurements followed the UTC+1 timezone, adapting to both daylight and standard time, while the data from the low-cost sensors was recorded in UTC without seasonal variations. To standardize the temporal data, all timestamps from the reference station were converted from UTC+1 to UTC, while the low-cost sensor data, already in UTC, was maintained in this format to ensure consistency with the standardized time zone.

3.3.7 Other Datasets

The datasets collected from Reggio Emilia and Trento are not detailed in this thesis as they were gathered during the final phase of the PhD and have not yet been tested. Data collection is still ongoing, and these datasets will be made available in the AQDR once a sufficient volume has been accumulated.

Additionally, the AQDR repository includes other datasets that were not utilized in this analysis due to organizational challenges. However, these datasets could prove valuable for future research. For reference, they are as follows:

- **Albarracín** - Bogotá [Alt23]
- **Yatkin** - Antwerp, Oslo, Zagreb [VPSP+23]
- **Karaoghlanian** - Beirut [KNS+21]

The purpose of the repository is to be enriched not only by my contributions, but also by those of other researchers who aim to share new knowledge about air quality data.

3.4 Standardization Pipeline

The analyzed datasets underwent several standardization steps to produce homogeneous datasets, to facilitate analyses and comparisons without considering the difference in reading frequency. After the standardization, those datasets can be easily used as input in machine learning or deep learning models. The pre-standardization procedure is the following:

Location	PM_1	$PM_{2.5}$	PM_4	PM_{10}	Pressure	RH	Temperature	Wind Speed	Rain	Dew Point
Aosta	x	x	x	x	x	x	x	x	x	
Badajoz	x	x		x	x	x	x	x	x	
Bangalore		x			x	x	x	x	x	x
Calgary		x			x	x	x	x	x	
Delhi		x			x	x	x	x	x	x
Hamirpur		x			x	x	x	x	x	x
Lima Airbeam	x	x		x	x	x	x	x	x	
Lima IQAir	x	x		x	x	x	x	x	x	
Torino	x	x	x		x	x	x	x		
Southampton	x	x		x	x	x	x	x	x	x

Table 3.4: Features Present in Sensor Datasets for Various Locations

- Data obtained from low-cost sensors and reference stations were aligned.
- Data were reduced to hourly frequency.
- The data were enriched with meteorological features downloaded from the Visual Crossing API [Cro] to create homogeneous datasets.
- Only useful features for this analysis have been retained, as described in Table 3.4.

These standardized datasets were released online in *CSV* file format on the Git repository [CM24]. Concerning the last point, attention has to be paid to two different aspects: the first is that where some datasets have several bins of particulate matter (e.g. PM_1, $PM_{2.5}$, PM_4 and PM_{10}) others have less, only $PM_{2.5}$ and PM_{10} have been retained. Researchers can apply interpolation to obtain a complete time series if necessary. However, I chose not to apply interpolation in this study to preserve the original raw values, ensuring the dataset remains as flexible as possible for diverse applications.

3.5 Technical Validation

First, of all for each dataset the $PM_{2.5}$ distributions are presented in Figure 3.2. As it is possible to grasp from this, Delhi suffers from higher pollution levels, reaching hazardous concentrations, followed by Hamirpur, Turin and Bangalore.

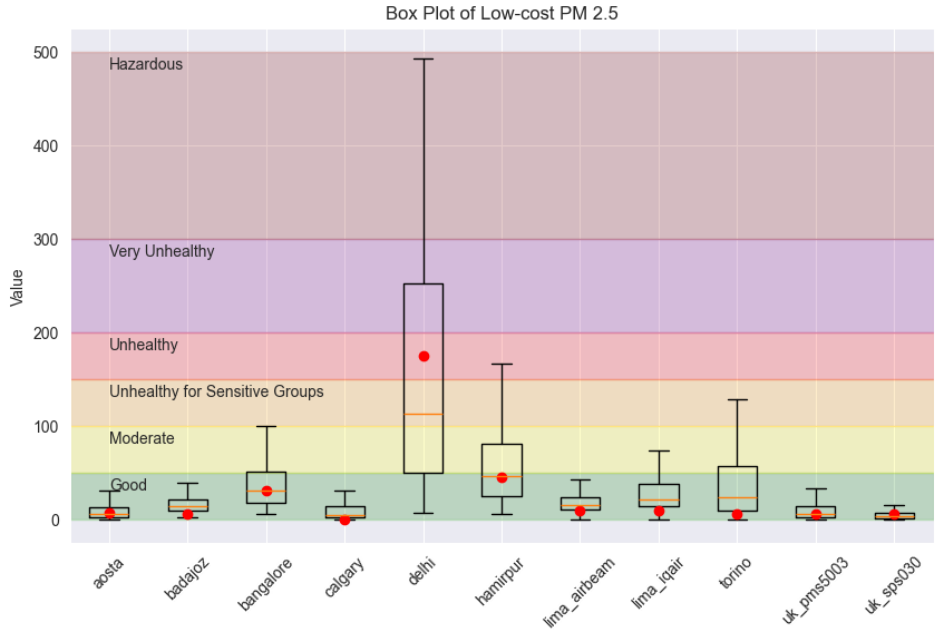


Figure 3.2: PM 2.5 distribution in each location. The background is colored based on WHO guidelines. The red dots refer to PM 2.5 concentration on 2024-10-21.

To reinforce the data insights, on 21 October 2024 the level of the Air Quality Index has been checked. The AQI is an equation expressing an aggregate value of the pollutants in the locality. According to IQAir [IQA24], the AQI are as follows: 290 in Delhi ($175\mu g/m^3$ of $PM_{2.5}$), 240 in Hamirpur ($46\mu g/m^3$ of $PM_{2.5}$), 127 in Bangalore ($31.2\mu g/m^3$ of $PM_{2.5}$), 92 in Lima ($10\mu g/m^3$ of $PM_{2.5}$), 52 in Badajoz ($6.7\mu g/m^3$ of $PM_{2.5}$), 37 in Southampton ($6.7\mu g/m^3$ of $PM_{2.5}$), 34 in Aosta ($8\mu g/m^3$ of $PM_{2.5}$), 33 in Turin ($6\mu g/m^3$ of $PM_{2.5}$) and 8 in Calgary ($0\mu g/m^3$ of $PM_{2.5}$). These specific readings reflect the broader trend noted earlier.

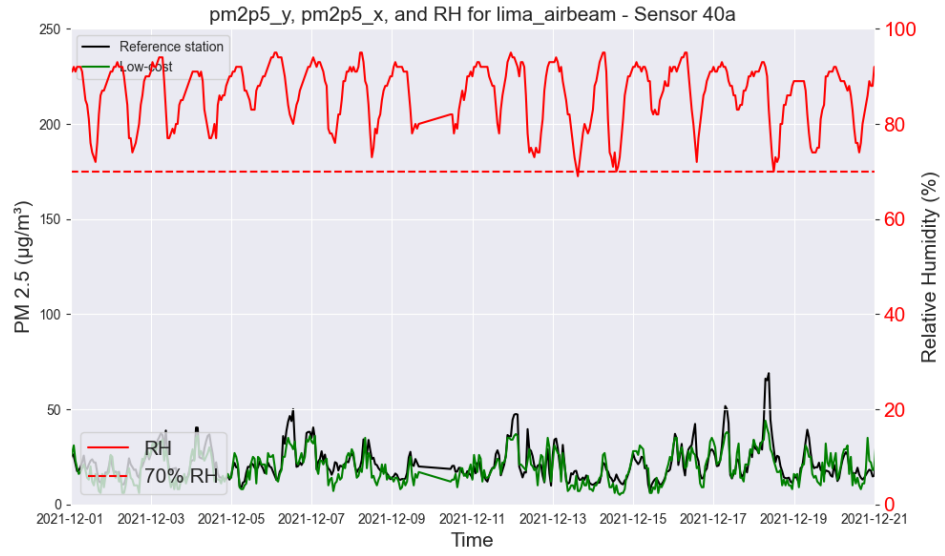
To give clearer context to the observed concentrations, the WHO guidelines [W+21] state that annual average concentrations of $PM_{2.5}$ should not exceed 5 $\mu g/m^3$, while 24-hour average exposures should not exceed 15 $\mu g/m^3$ more than 3 - 4 days per year.

Figure 3.3: Distribution of correlations for each feature pair across locations, with feature pairs displayed on the x-axis.

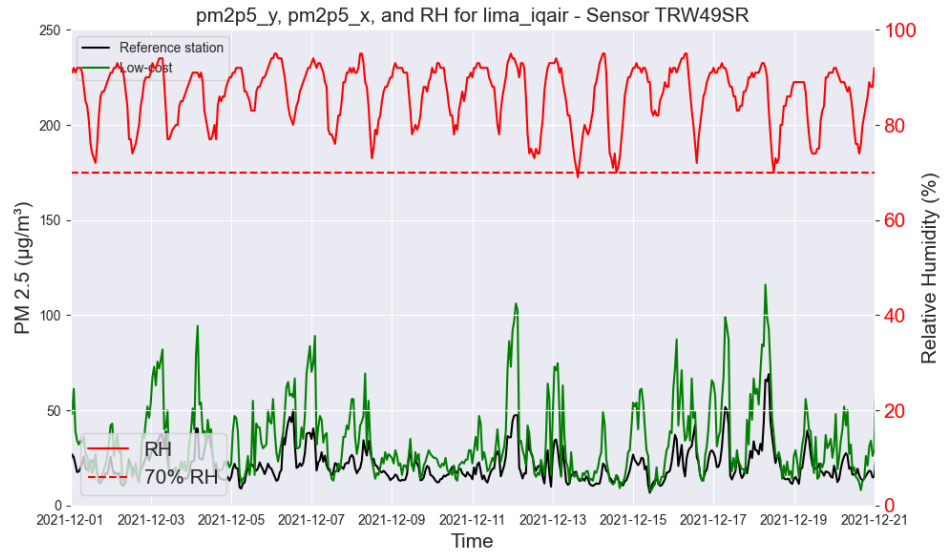
The correlation between features is studied by examining how pairs of features relate to each other across different datasets and locations. For each dataset, correlation coefficients are computed for specified feature pairs, as shown in Figure 3.3. These correlations are visualized using a box plot, which illustrates the distribution of correlation values. The correlation between $PM_{2.5}$ measurements from the LCS and the co-located RS is strong and positive. In contrast, temperature and relative humidity show a strong negative correlation, likely because warmer air can hold a greater amount of water vapor, which inversely affects RH. Relative humidity and $PM_{2.5}$ correlate more positively in LCS data due to particle hygroscopicity, unlike reference sensors, which correct for humidity effects (Section 2.2.1 for further details).

3.5.1 AQDR's Hygroscopicity

In order to study hygroscopicity, which leads to an increase in particle size due to high RH, it is necessary to assess the growth of $PM_{2.5}$ read from the LCS according to the relative humidity value. If the comparison is made with the reference station, the gap between the time series of the two types of sensors, LCS and RS, should be evident based on RH levels. The ability of the RS to reduce such an effect is often due to the station's internal mechanism to detect dried particles. Suppose hygroscopicity is not present in a dataset, in that case, the explanations are that the LCS can reduce the effect of humidity on the particles, or that there is not the right combination at that location to generate aggregation between water vapor and PM, as there may be insufficient hygroscopic substances to retain moisture.



(a) Lima's AirBeam sensor shows no hygroscopicity.



(b) Lima's IQAir sensor shows hygroscopicity

Figure 3.4: A comparative example of two different sensors at the same location, showing and not showing hygroscopy. The black line is for both the same reference station. The relative humidity is in red with the 70% line dotted.

In Figure 3.4 it is possible to observe the hygroscopic effect in two different cases. For the same location and the same period, two different low-cost sensors, compared to the co-located legal station, do not show hygroscopicity (Figure 3.4a) and hygroscopicity (Figure 3.4b), respectively. In Lima, at that time, the RH is above 70% almost all the time, and for the case 3.4b the line representing the LCS is constantly above the RS, narrowing the gap when the RH is closer to the 70% limit. The effect of humidity is not recorded in the data collected from the AirBeam LCS, probably due to some differences during the installation procedure that might involve some drying, although not explicitly provided for in the documentation linked to the original work.

To visualize hygroscopicity and the impact of relative humidity across datasets,

RH Range	Aosta SPS30	Badajoz OPC-N3	Bangalore PMS5003	Calgary PMS5003	Delhi PMS5003	Hamirpur PMS5003	Lima AirBeam	Lima IQAIR	UK PMS5003	UK SPS30	Torino SPS30
0-50	-0.73	2.24	9.08	-1.06	53.79	22.52	-	-	1.33	-3.02	12.23
50-60	0.79	9.49	10.48	4.02	86.08	29.37	-	-	1.34	-3.36	18.56
60-70	1.96	13.42	13.97	7.42	29.79	29.79	-4.93	-2.60	2.65	-3.17	13.45
70-80	4.98	14.33	16.96	11.15	-	31.62	-6.17	-0.93	2.72	-3.00	41.54
80-100	11.53	11.53	7.19	22.57	-5.17	56.74	-2.11	10.82	4.05	-2.26	111.70

Table 3.5: Mean Gaps Between $PM_{2.5}$ LCS and RS Across Locations and PM Ranges

Table 3.5 displays the mean gap between $PM_{2.5}$ concentrations from LCS and RS, categorized by RH ranges. Where values are absent, a '-' is shown. Notably, most locations exhibit an increasing concentration gap, except for Delhi, Lima AirBeam, and UK SPS30. In Turin, the gap is significant, indicating serious issues with LCS data reliability for future analysis.

The available datasets allow for analysis and comparisons, enabling the calibration of models based on contextual factors. My work focuses on exploring ways to reduce hygroscopicity and exploiting these datasets for further studies.

3.6 Code availability

The code for analyzing air quality data is available in the *Code* directory of the AQDR repository [CM24]. This directory contains all the Python functions needed to evaluate the data, calculate features' importance, and generate visualizations. Users can access the code and modify it as needed for their analyses or integrate it into their projects.

Data analysis was conducted using Python in Jupyter Notebooks. The code was executed on a Windows computer, and Python 3.10 scripts to reproduce the examples.

3.7 Feature Augmentation

To further enhance the Air Quality Datasets Repository, additional contextual indices can be integrated. This involves leveraging satellite imagery and raster data, widely accessible online, to extract and include variables that provide a deeper understanding of the environmental and urban context.

One such index is the **Normalized Difference Vegetation Index (NDVI)**, a widely used indicator in remote sensing. The NDVI is calculated using the following formula:

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$$

where NIR refers to the reflectance in the near-infrared band, and RED refers to the reflectance in the red band of the electromagnetic spectrum. The NDVI values range from -1 to $+1$, with higher positive values indicating dense vegetation and negative values typically representing water bodies or barren landscapes. By incorporating NDVI, it is possible to examine the influence of vegetation on air quality, particularly its role in mitigating particulate matter concentrations.

In addition to NDVI, other variables capturing urban and demographic contexts can be included:

- **Industrialization Index:** Measures the density and activity level of industrial facilities within a given region, often correlated with higher emissions of pollutants such as $PM_{2.5}$ and PM_{10}.
- **Population Density:** Provides insights into the human activity intensity, which influences vehicular emissions, heating emissions, and other sources of air pollutants.
- **Traffic Flow:** Quantifies vehicular movement patterns to analyze their contribution to pollution levels in urban areas.
- **Land Use Categories:** Examines the proportion of land dedicated to residential, commercial, industrial, agricultural, and green spaces to assess their respective impacts on air quality.
- **Topographical Variables:** Includes altitude, slope, and proximity to natural barriers such as mountains or bodies of water, which can affect pollutant dispersion and accumulation.
- **Weather Variability Indices:** Expands meteorological analysis by incorporating long-term climate variability, urban heat island effects, and extreme weather event frequencies.

These indices, when integrated with existing air quality data, can provide a more comprehensive understanding of the multifaceted factors influencing air quality.

Chapter 4

MitH Framework

One of the most significant limitations of LCS in measuring PM is their sensitivity to humidity [JLT+18] (Section 2.2.1). This is because particles present in the atmosphere may absorb water and increase in size, leading to an overestimation of the PM concentration detected by the sensor. This is known as the hygroscopic property of the particle. High humidity can also cause false readings due to the formation of water droplets, leading to inaccuracies in measurements [Sku13, MIF+20, HK20, CSP+18, MZHF16]. This is especially problematic for laser-based sensors, which measure particle size and concentration using laser scattering. It is important to highlight that water vapour is not harmful to human health. Therefore, it is essential to model and remove this artefact to obtain accurate pollution level measurements comparing the LCS measurements with regulatory station measurements. As a matter of fact, the EU air quality standards [EEA21], the WHO air quality guidelines [W+21], and other governmental organizations, measure pollution impact based on the dry PM concentration.

To ensure accurate measurements with LCS, researchers and manufacturers have developed procedures and algorithms to mitigate the effects of relative humidity and improve sensor design. Typical corrective methods utilized measurement data and a reference station to estimate growth function or regression model parameters for relative humidity correction, or integrated a dryer into the sensor to counteract the effects of humidity [ORC20, LK, ZSG+20b, GMP+21b, SCC+17, DAPO+18, HNS+22].

For low-cost PM sensors where integrating a dryer is not feasible, existing approaches face significant limitations, primarily due to their heavy reliance on time and location [Str17, RB20]. Methods trained in one location may not accurately capture the environmental specifics when the sensor is moved to a different location, requiring parameter adjustment and re-training [CSK+20, JLW+22]. Additionally, methods that require LCS to be stationed near reference stations for prolonged periods limit their practical use for other applications. Furthermore, approaches based on the k-Köhler theory rely on knowing the elemental analysis of the air, which may not always be readily available or subject to frequent changes, limiting their applicability in certain scenarios.

In this thesis, I present the **MitH framework** [CP24], a novel approach designed to model and mitigate the effects of hygroscopicity on air quality measurements.

The main features of MitH are:

- **Anomalies-tolerant.** The framework can effectively handle anomalies in the data it processes. It identifies, removes and replaces outliers and spikes from

sensor data.

- **Dynamic and adaptable.** This approach allows for continuous updates and refinements of the growth function, thanks to rolling window data, making it context-aware and capable of accommodating variations across different environments, also in isolated and humid locations, where traditional calibration techniques may be limited.

- **Reference station-agnostic.** Unlike existing approaches that require LCS to be stationed near reference stations for extended periods, the MitH framework can be applied in any context, boosting its usability when calibration is challenging or when a model is not exploitable due to different environmental conditions.

- **Real-time.** By providing timely and context-aware corrections, MitH offers a practical and efficient solution to address the hygroscopicity challenge in real-world applications and research endeavors.

MitH excels with its dynamic adaptability, managing anomalies effectively. It operates independently of reference stations, making it versatile in diverse environments. MitH's real-time capabilities offer practical solutions for hygroscopicity challenges in both real-world applications and research.

Thanks to its distinctive features, the MitH framework offers valuable advantages. It enables the evaluation of modelled air quality data across a larger number of locations, overcoming the limitations of conventional simulation techniques that heavily rely on data from reference stations or high-precision sensors [BNV+23, HDQ+22]. Moreover, integrating MitH into interpolation techniques allows for the incorporation of more accurate air quality data, resulting in more comprehensive and representative modelling outcomes. Furthermore, MitH's effectiveness in visualizing and managing environmental sensor data provides valuable insights into air quality patterns and trends. This supports informed decision-making and environmental management [HRJM15].

It is important to note that while MitH addresses the specific challenge of hygroscopicity, it does not negate the need for calibration techniques in LCS. It is acknowledged that low-cost sensors can encounter various limitations, such as sensor drift, ageing, temperature sensitivity, cross-sensitivity to other pollutants, and inherent measurement biases. These factors can introduce errors and inaccuracies in the sensor readings. Therefore, it is crucial to incorporate appropriate calibration techniques to account for these factors and ensure accurate and reliable measurements [GMP+21b, Zhi21].

4.1 Methodology

MitH is organized into four modules, as illustrated in Figure 4.1, along with a final evaluation step:

1. *Data collection*: In the first module, data are collected from LCS. Each subsequent module is performed separately for each sensor's data.

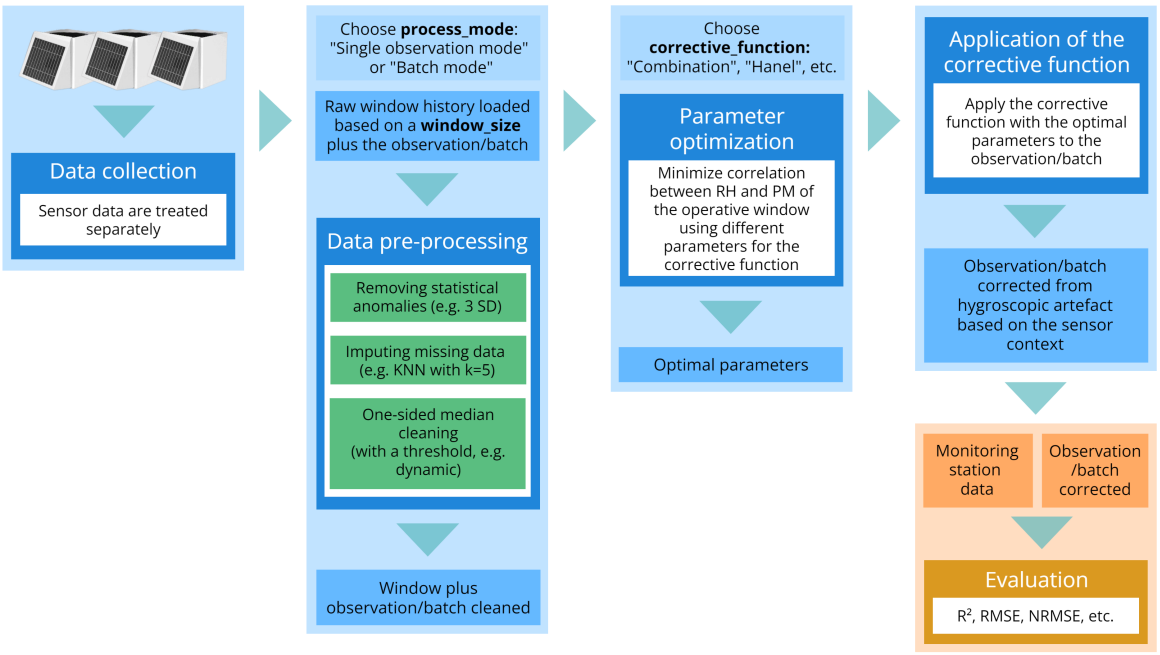


Figure 4.1: Flowchart displaying the presented MitH (Mitigating Hygroscopicity) framework, which consists of four main modules: data collection, data pre-processing, parameter optimization, and application of the growth function.

2. *Data pre-processing*: In this module, the chosen *process_mode* and *window_size* values are utilized to pre-process the operational window, composed by the new observation(s) and the relative window history. This pre-processing aims to eliminate any significant statistical anomalies present in the data.

3. *Parameter optimization*: The third module involves optimizing the parameters of the growth function. This optimization is carried out to ensure that the function accurately captures the relationship between relative humidity and particulate matter based on the operative window data obtained from the previous module. As a result, the function is context-aware. The optimal parameters obtained during this optimization are retained for the next module use.

4. *Application of the correction function*: In the final module, the optimized growth function is applied to the newly measured data, in which RH is above a certain threshold. This application serves to correct any potential hygroscopic biases caused by humidity in the particulate matter measurements.

Upon completion of the four MitH modules for each sensor, a conclusive assessment is conducted by comparing the corrected observations with those obtained from the reference station. It is important to note that the reference observations are exclusively used at this stage; they are not employed in either the optimization or correction phases. Indeed, the framework refines sensor observations without relying on any reference data.

4.1.1 Data Collection

The study was conducted in Turin, Italy, which is the capital city of the Piedmont region in northern Italy. Turin has a population of approximately 847,000 inhabitants and covers an area of 130 square kilometres. In 2019, Wiseair initiated a Citizen Science project in Turin to increase awareness among the city's population regarding pollution levels. As part of this project, over 20 Arianna devices were strategically deployed across the city. Some of these devices were co-located close to a reference station operated by Arpa (the regional environmental protection agency), situated in a public garden, away from heavy traffic (see [AT] for topology specification). The reference station is a Tecora Sequential Unit, that utilizes gravimetric technology, enabling accurate measurements of PM_{10} and $PM_{2.5}$ concentrations. The low-cost devices were positioned on the perimeter fence surrounding the reference station, elevated at a height of 4 meters.

The Arpa reference station provides hourly data of validated PM 2.5 levels. The agency conducts the validation using the medium-bound technique, estimating concentrations below the detection limit as half of the detection limit for the target substance. In contrast, the Wiseair Arianna devices collect data at 15-minute intervals. Wiseair has its own corrective algorithm, which includes a threshold, a correction function with fixed parameters, and a regression applied for the autumn season. In this thesis, the data used are the raw ones. The final evaluation is conducted hourly to match the Arpa reference station granularity, taking an hourly mean of the data collected from each sensor.

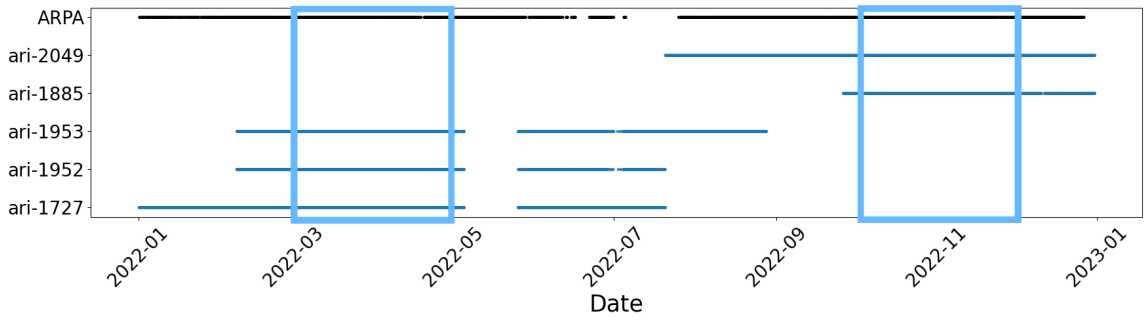
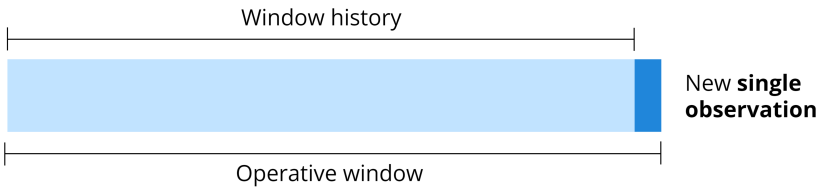


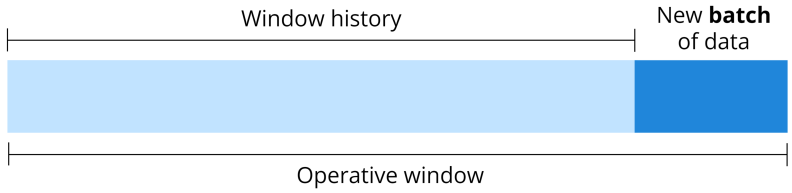
Figure 4.2: Operating periods of LCS and ARPA reference station in 2022

The study was conducted over two distinct periods. The first period took place from March to April 2022, during spring. The second period occurred from October to November 2022, representing autumn. By analyzing data from both seasons, the study aimed to capture variations in RH levels and evaluate the performance of LCS under different environmental conditions. During these specific periods, multiple LCS were available for comparison with the Arpa reference station. The activity of the sensors is illustrated in Figure 4.2. Despite having a more extended duration for data collection in the autumn period, the decision was made to use an equal amount of time for both periods, ensuring consistency and comparability.

A total of 13,000 observations were collected from five Arianna devices. The devices provided a comprehensive set of data, including various environmental parameters and pollutant concentrations. The recorded data from the Arianna devices consisted of the



(a) Single observation mode



(b) Batch mode

Figure 4.3: Two different approaches for data pre-processing and optimization modules: considering each observation individually during pre-processing and optimization (a), or processing multiple observations simultaneously as a batch (b).

following variables for each observation: date and time in UTC, temperature in degrees Celsius, relative humidity in percentage, pressure in hPa, cloud coverage in percentage, wind speed in meters per second, wind direction in degrees, and concentrations of PM_1, $PM_{2.5}$, PM_4, and PM_{10} in micrograms per cubic meter ($\mu g/m^2$). On the other hand, the reference station operated by ARPA provided a dataset of 3,000 hourly measurements. The data included date and time in GMT+1 and measurements of NO_2, NO, NOx, PM_{10}, and $PM_{2.5}$ concentrations in micrograms per cubic meter ($\mu g/m^2$).

4.1.2 Data Preprocessing

The pre-processing is applied to the new observation(s) along with its historical window. If there is only one observation, in the study is referred to as *single observation mode*, and if there are multiple observations, as *batch mode*. In both cases, the new observation(s) undergo pre-processing together with the preceding historical window of length *window_size*. This combined window, consisting of the historical window and new data, is now referred to as the operative window (see Figure 4.3).

Excluding the single observation/batch into the context used in the pre-processing and optimization is possible even if not recommended. By excluding the new data, the operative window is equivalent to the window history. In the *single observation mode*, the downside of not keeping them is less evident because the window history is more similar to the operative window. In the *batch mode*, the window history itself could not be enough to explain the context of the new observations. Section 4.2.3 shows what happens when the batch data context is excluded from the operative window.

The pre-processing considers various factors, including the identification and removal of statistical anomalies, imputation of missing data using the k-Nearest Neighbors (KNN) method, and a one-sided median cleaning process for correcting right-skewed data. These steps are tailored to enhance the data quality and prepare an operative

window for further analysis.

After having chosen the *processing_mode*, as a single observation or batch, and the *window_size*, the pre-processing is performed over the data loaded. The pre-processing, as illustrated in Figure 4.1 second box, is composed of three phases:

1. *Removing statistical anomalies*: The phase involves identifying anomalies in the data by calculating the standard deviation and replacing any events that fall outside 3 standard deviations with NaN values. This helps to eliminate rare and unlikely data points, which are likely to be caused by a misreading of the sensor. It's worth noting that this step might raise concerns, especially considering that high values are crucial when studying hygroscopicity. However, a prior analysis of the raw data revealed rare and very high peaks, exceeding 1000 $\mu g/m^3$. These peaks are not indicative of concentrations affected by hygroscopicity, but rather anomalous data. Applying a 3-standard deviation threshold helps retain peaks induced by hygroscopicity while removing anomalous peaks caused by device errors. Alternatively, a threshold could be used if prior knowledge is available. When handling raw data unaffected by humidity, setting a threshold becomes more straightforward, especially when there's a reference limit for PM data available. However, in the hygroscopic environment, determining a suitable threshold is challenging due to the lack of prior knowledge regarding the maximum concentration levels detectable. Therefore, the 3-standard deviation approach is considered appropriate for filtering out only anomalous peaks, where "anomalous" does not refer to hygroscopicity-induced values but rather to exceptional values.

2. *Imputing missing data*: The second phase is filling in any missing data points. This is done by replacing the NaN values using a k-Nearest Neighbours imputation method with a chosen k-value. This results in a minor loss of data points and the missing data are refilled using similar contexts.

   The missing value $\hat{x}_{i,j}$ for feature j in instance i is estimated as:

   $$\hat{x}_{i,j} = \frac{\sum_{p \in \mathcal{N}_k(i)} w_p \cdot x_{p,j}}{\sum_{p \in \mathcal{N}_k(i)} w_p}$$

   where:

   - $\mathcal{N}_k(i)$ represents the set of the k nearest neighbors of instance i, determined based on a chosen distance metric (e.g., Euclidean, Manhattan).
   - $x_{p,j}$ is the observed value of feature j in the neighboring instance p.
   - w_p is the weight assigned to the neighbor p, often calculated as the inverse of the distance between the instance i and its neighbor p:

     $$w_p = \frac{1}{d(i,p)}$$

     where $d(i,p)$ represents the distance between instance i and the neighbor p. This weighting ensures that the closer neighbors have a greater influence on the imputed value.

KNN imputation assumes that similar instances (as determined by the distance metric) have similar feature values. By utilizing the values from the most similar instances, KNN imputation helps to estimate missing data in a way that maintains the underlying structure and distribution of the dataset. The choice of k (number of neighbors) and the distance metric significantly impacts the accuracy of the imputation.

Of course, the imputation algorithm could be chosen from the classic set of filling algorithms. The choice reflects the capacity of the KNN to use any other available variables to fill empty values.

3. *One-sided median cleaning*: The final phase is a one-sided median cleaning process, which is used to correct right-skewed data. A sliding window is used to determine the median value of the data. Any data points that fall outside the median plus a calculated threshold are replaced with the median value of the window. In the study, the threshold is calculated using statistical properties of the available data, but it should also be determined based on other environmental-specific knowledge. This process is performed only on the right side of the data, hence the term "one-sided" median cleaning, to ensure a real-time approach. It is important to note that the window used should be a temporal arch and not a number of preceding observations unless the frequency of reading of the sensor is ensured.

During each phase, adjustments can be made to various parameters. This includes the standard deviation in the first phase, the k-value used in the KNN algorithm in the second phase, and the number of hours of the rolling window and the threshold in the last phase.

It is important to note that after the removal of anomalous data during pre-processing, the data may not include certain events, for instance, fires, that initially appear as anomalies and are subsequently modified during pre-processing. As time passes and the event persists, the data that were initially considered anomalous may no longer be considered as such and become useful for building a more accurate context. However, this possible delay must be taken into consideration.

As depicted in the flowchart, at the conclusion of the pre-processing, the process yields an operative window cleared of significant statistical anomalies.

4.1.3 Parameter Optimization

After the completion of the three pre-processing steps, the historical data window is ready for the growth function optimization. Various growth functions, often referred to as growth functions, have been proposed in the literature to address the issue of hygroscopicity [DM01, GRS03, CMTS22]. These functions are designed to calculate a corrective coefficient that can be applied to reduce the concentration level of particulate matter based on the relative humidity level, as in Eq. 4.1.

$$PM_{dry} = \frac{PM_{wet}}{gf(RH)} \tag{4.1}$$

The PM results are said corrected with respect to the problem of hygroscopicity. In the equation, PM_{wet} represents the PM concentration detected by LCS and PM_{dry} represents the PM concentration after the application of the growth function gf.

In [Str17] a list of possible growth functions is presented, among with the *Hänel* (Eq. 4.2), and a new proposal called Combination (Eq. 4.3).

$$gfhanel = \frac{1}{(1 - RH)^{\beta}} \tag{4.2}$$

$$gfcombo = 1 + \alpha \cdot \frac{RH^2}{(1 - RH)^{\beta}} \tag{4.3}$$

[SCT+14] provide two other growth functions, the first proposed in [CFDS04], call *Chakrabarti* equation (Eq. 4.4) and the second, *Richards*'s humidity adjustment (Eq. 4.5), originally proposed in [RAM+99].

$$gfchakrabarti = \alpha + \beta \cdot \frac{RH^2}{1 - RH} \tag{4.4}$$

$$gfrichards = exp(\alpha + \beta \cdot Ln(1 - RH)) \tag{4.5}$$

Streibl's work provides insights into the different behaviors of growth functions optimized over months. He also emphasizes the importance of pre-processing, which is similar to the one applied in this study. That is, in the pre-processing module phases, a range was used to reduce anomalies instead of SD, and a non-context-aware algorithm was employed in the filling algorithm. The final phase involved a median window smoother, although details were not provided.

Regarding the optimization of the growth function, Streibl argues that since RH and PM are not correlated in reality, two possible approaches can be considered:

1. Minimizing the Fourier coefficient: This approach takes into account the periodic component with a 24-hour period present in temperature, humidity, and particulate matter growth. By minimizing the absolute value of the corresponding normalized Fourier coefficient, the influence of humidity can be compensated to the best possible extent.

2. Minimizing the correlation factor: This approach aims to minimize the correlation between the corrected PM and RH. By achieving a minimal correlation, the influence of humidity can be effectively compensated.

In the case of *Chakrabarti* and *Richard*'s equations, Soneja et al. explored the original parameters proposed in *Chakrabarti*'s original work for the first, as well as new parameters fitted using simulated cooking test data for both. Additionally, they compared the application of these three growth functions to the entire dataset versus applying them only to data above a certain threshold. The findings suggested that using a threshold for applying the growth functions was a better approach. Furthermore, it was observed that a threshold of 60% relative humidity appeared to be generally too low.

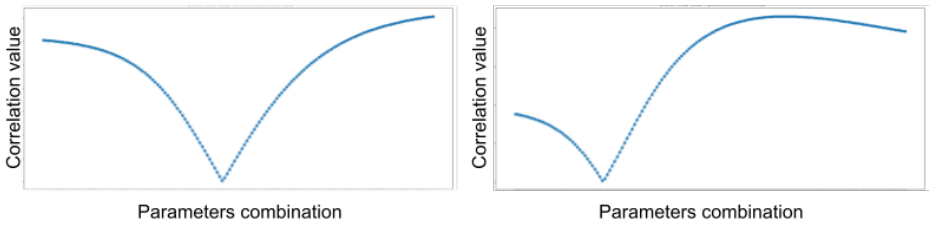


Figure 4.4: Two real examples of parameter optimization for reducing Relative Humidity correlation to Particulate Matter concentration levels.

Typically, as for Soneja, the optimization process involves finding the values of the growth function that provide the best match between the observed data and the reference station values. Once the growth function has been optimized, it is used to correct the raw sensor data. This fitting procedure relies on the fact that the optimal parameters for a specific context are explanatory in general.

In contrast, MitH does not rely on the use of a reference station to optimize the parameters of the growth function. Instead, two distinct approaches were explored. The first approach involved minimizing the correlation between relative humidity and particulate matter measurements, as suggested in Streibl. The second approach focused on minimizing the difference between the distribution of the original data below a chosen threshold and the data corrected above the same threshold.

Two real examples of parameter optimization modules, based on the correlation minimization approach, are shown in Figure 4.4. The x-axis, which remains the same among the two subplots, represents the combined parameters used for optimization, and the y-axis represents the correlation obtained. The figure highlights two different periods, to demonstrate the difference in parameter selection even for closely spaced time periods.

Given a curve, it is possible to see how different growth function parameters yield different correlation values between RH and PM; the same happens when using the distribution difference minimization approach. In the figure, it is evident that there exists an optimal point where the correlation is minimized, which is the one obtained with the parameters used as optimal in the next module. Consequently, by applying the correlation (or distribution difference) minimization over the operative window, the growth function can be customized to the specific characteristics of the sensor and account for the impact of relative humidity on the sensor readings within the specific environmental context. As a result, this dynamic approach enables MitH to effectively adapt to environmental changes over time, without the need for a reference station.

To effectively implement this approach, there is a prerequisite for a significant volume of historical data to ensure the precise optimization of the growth function. This entails periodic reevaluation of the parameters involved in the correction process. Despite these demands, the approach proves especially beneficial for LCS that may lack access to reference station data for calibration.

4.1.4 Application of the Growth Function

In the fourth module of the process, the optimal growth function found is applied as a corrective coefficient to the current observation(s), as in Eq. 4.1, if the associated relative humidity is above a certain threshold. As a result, the corrected data provides a more accurate representation of the particulate matter levels, compensating for the influence of humidity on the sensor readings.

After obtaining the corrected data, the evaluation was conducted by comparing them to the data received from the Arpa reference station. The evaluation was performed using metrics such as R^2, RMSE, and NRMSE. It's noteworthy that the granularity of the data remained consistent with that collected by LCS, approximately every 15 minutes, during the evaluation process. However, for the purpose of comparison with the measurements from the Arpa reference station, the data were re-sampled to reflect the hourly granularity imposed by the reference station. This adjustment facilitated a meaningful assessment of the accuracy of the MitH framework.

4.2 MitH Application Results

In this section, several key insights are presented, focusing on evaluating the performance of MitH. The results section is divided into the following subsections to provide a comprehensive analysis:

1. *Sensors Behavior Analysis*: An examination of the behavior of sensors during the corrective process, aiming to identify specific patterns or challenges encountered (see 4.2.1).

2. *Window History Size*: Exploration of the impact of different window history sizes on the performance of the growth function. This involves assessing how the choice of $window_size$ influences the accuracy of the corrected data (see 4.2.2).

3. *Context Inclusion*: Discussion of the advantages of including the context of new observations alongside the window history in the parameter optimization process. Incorporating current environmental conditions aims to enhance the accuracy of the growth function (see 4.2.3).

4. *Step Performance*: Evaluation of the performance at different stages of the approach, including raw data, pre-processed data, and corrected data (see 4.2.4).

5. *Comparison with Existing Approaches*: Presentation of the results of the presented approach and a comparison with outcomes achieved by applying approaches described in the literature. This comparison provides insights into the effectiveness of MitH in addressing the hygroscopicity issue and improving the accuracy of sensor data (see 4.2.5).

6. *Wiseair approach comparison*: In the last subsection, a comparison is provided between the Wiseair corrective method approach and MitH.

The presentation of these subsections aims to offer a comprehensive understanding of the performance and effectiveness of the proposed approach in addressing hygroscopicity in low-cost sensor data.

In the following, the application results are presented in terms of RMSE (Root Mean Squared Error), NRMSE (Normalized Root Mean Squared Error), and R^2. Additionally, some of the results are visualized as time-series plots to demonstrate practical outcomes and capture some peculiarities of the process. The evaluation metrics, such as RMSE and NRMSE, provide quantitative measures of the model's performance in terms of accuracy, estimating how well the model can predict the target value. A lower RMSE and NRMSE indicate better agreement between the corrected values and the ground truth, in addition, NRMSE may be useful to make the evaluation scale-free. Furthermore, the R^2 value assesses the goodness of fit of the corrected data compared to the observed data. A higher R^2 value indicates a stronger correlation and better predictive capability of the correction model.

4.2.1 Sensors Behavior Analysis

An important aspect to consider when working with LCS is that they can exhibit a wide range of anomalies. Figure 4.5 clearly demonstrates this phenomenon, where the sensor labeled ari-1727 (the blue line) consistently shows higher concentration levels compared to the other sensors. This discrepancy is visually evident in the plot, as the data from ari-1727 consistently deviates from the overall trend observed in the other sensors (orange and green lines), which approximates the reference station (black line) better.

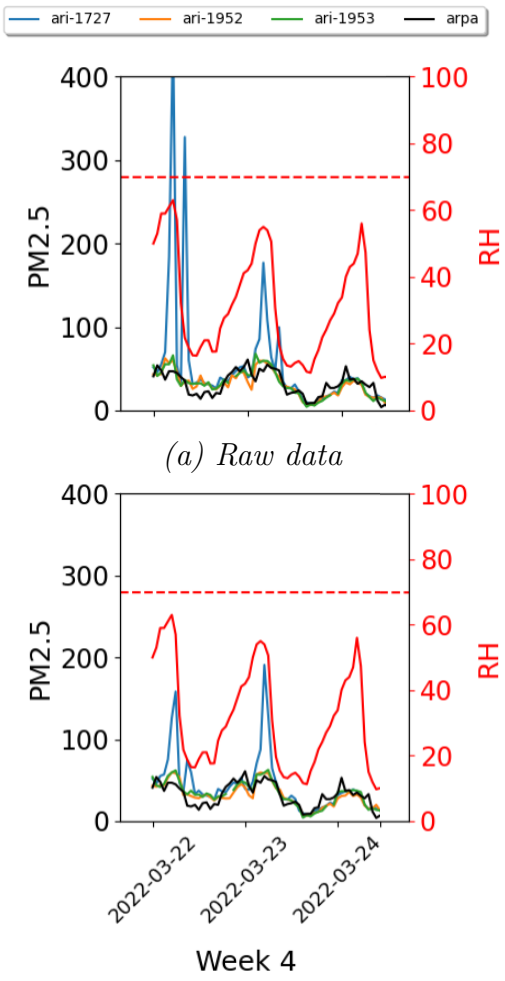
Although some peaks are reduced after pre-processing (see sub-figure b), the overall signal from the ari-1727 sensor remains compromised. Moreover, such anomalies are not addressed by the growth function used in this study, as the RH threshold chosen for correction is higher than the RH levels detected during these measurements. This suggests that the observed anomaly is likely unrelated to hygroscopic effects. Consequently, addressing this anomaly would require additional pre-processing steps specifically designed to handle such anomalies. It's important to mention that despite these anomalies, the sensor remains useful as it generally aligns with PM concentrations from other sensors for the majority of the time, justifying its inclusion in the study.

However, this emphasizes the complexity and challenges associated with correcting and processing data from LCS, especially when dealing with anomalous sensor behaviour. Further investigation and refinement of pre-processing techniques may be necessary.

4.2.2 Window History Size

The choice of the *window_size* parameter is crucial in contextualizing the growth function and obtaining optimal corrective parameters for the specific period under consideration. Initially, it was believed that a larger window size would lead to better performance. However, the findings contradict this assumption.

It has been observed that a shorter window history is sufficient to optimize the growth function parameters, and using longer window periods diminishes performance.



(a) Raw data

(b) Pre-processed data

Figure 4.5: Anomalous sensor behavior with elevated concentration levels: Raw data with visually prominent readings (a) and cleaned data after pre-processing (b).

	Spring period			Autumn period		
Window size	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE
12h	0.778	**11.132**	**0.505**	0.626	12.219	0.573
1d	**0.781**	11.157	0.506	**0.636**	**12.010**	**0.563**
2d	0.779	11.231	0.509	0.577	12.766	0.598
1w	0.776	11.577	0.525	0.543	14.378	0.674
2w	0.760	11.749	0.533	0.577	13.751	0.644
3w	0.702	12.866	0.583	0.549	14.104	0.661

Table 4.1: Comparison of LCS performance with reference station across varying window sizes - Parameters: batch dimension of 1 day, optimization method correlation minimization, growth function *Combination*, RH threshold 70%.

This is illustrated in Table 4.1, where the performance metrics are plotted against

different window sizes.

The performance of LCS is evaluated by comparing the sensors' concentration with the reference station for both the spring measurement period and the autumn measurement period.

The results demonstrate that a briefer window history (12 hours or 1 day) yields better results in terms of corrective parameter optimization. However, it was observed that if the window history is too short, it may not contain enough representative context data, including the classical cyclic relative humidity pattern. This observation is particularly relevant when the process mode is set to *single observation mode*. On the other hand, if the *batch mode* is preferred and a sufficiently large batch is used, these issues can be mitigated to some extent. Therefore, the choice of process mode and the size of the batch can have an impact on the effectiveness of the corrective parameter optimization if a small *window_size* is used.

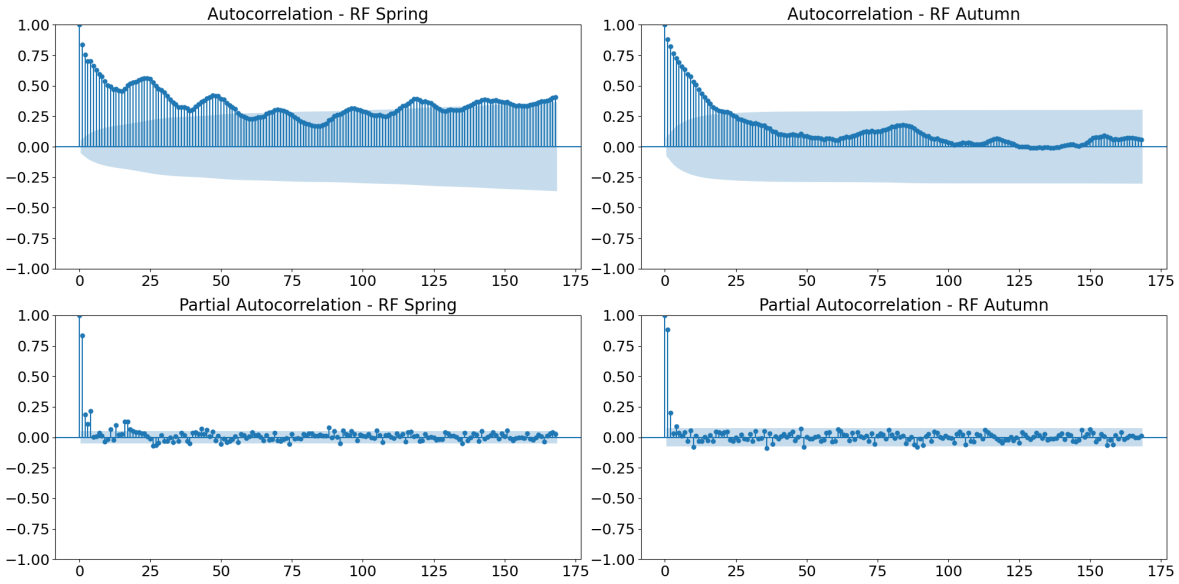


Figure 4.6: Autocorrelation and partial autocorrelation plots for hourly $PM_{2.5}$ concentrations at the reference station.

An additional interesting insight can be extracted from the examination of the autocorrelation (ACF) and partial autocorrelation (PACF) plots above hourly PM 2.5, during both the spring and autumn seasons. In Figure 4.6, the disparities in ACP between spring and autumn regarding $PM_{2.5}$ concentrations likely arise from variations in environmental conditions, human activities, and seasonal influences. During spring, a noticeable seasonal pattern after one day may signify cyclic environmental events. Conversely, autumn lacks a discernible seasonal pattern, and there is a swift decline in autocorrelation after one day, indicating a more dynamic and responsive system.

It is noteworthy that, despite the observed seasonal ACF in spring, the PACF plot suggests that employing a window size of 1 day might be adequate for optimizing meaningful contextual insights. However, considering the observed dynamics, exploring a dynamic window based on the season could be a valuable option for further investigation.

4.2.3 Context Inclusion

The decision to include or exclude the observation(s) to be corrected, along with their corresponding relative humidity levels, during the optimization step of the growth function can significantly affect the obtained results. It is important to note that when the process mode is set to batch and the observations are not included in the operative window (as shown in the *batch mode* Figure 4.3), there is a contextual information gap. This means that the overall context for the entire batch is missing, which can potentially result in a less accurate outcome. However, when the correction process is performed for each new observation, in the *single observation mode*, the impact of non-inclusiveness is reduced. This is because the historical data used for each observation is more closely related and relevant to the specific context of that individual observation.

Importantly, it should be clarified that including new data does not equate to using redundant, inappropriate information, or introducing bias. On the contrary, it is vital to ensure the improved performance of the corrective process.

In this study, a comparison between the two approaches was conducted. Specifically, the *Combination* function was applied to the data both with and without incorporating the observation(s) to be corrected during the pre-processing and parameter optimization steps. Figure 4.7 illustrates the differences in the results.

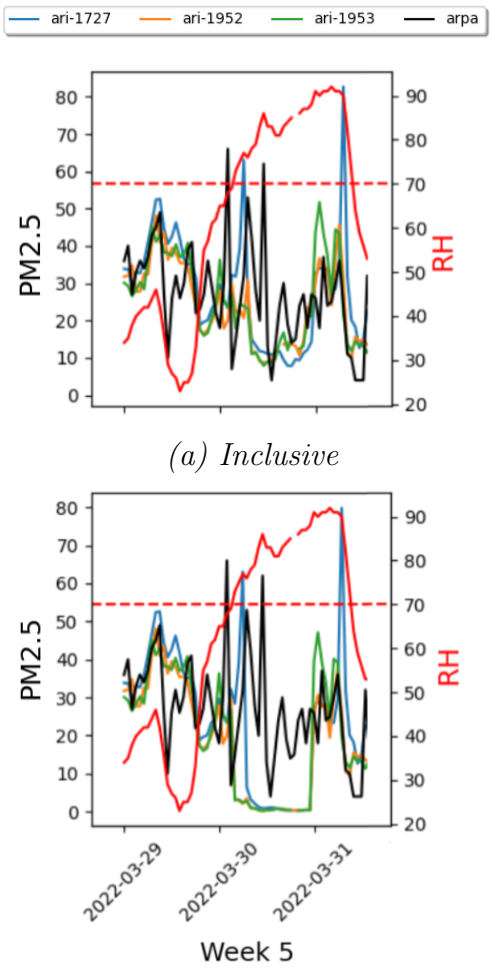
In subplot (a), the presence of the plateau observed in subplot (b) is not evident. This discrepancy can be attributed to the difference in the context considered during the parameter optimization. When the optimization technique was applied to the whole batch at the same time, on day 2022-03-30 (batches of 24 hours were utilized), the previous period (window history) did not experience an RH greater than 70%. Consequently, the growth function lacked optimization for the upper range of RH thresholds. This led to the corrective coefficient potentially exhibiting an overly aggressive impact upon application. However, by including the observations to be corrected during the pre-processing and parameter optimization steps, this negative effect was reduced. Therefore, the more comprehensive context provided by including the relevant observations resulted in a more accurate correction.

4.2.4 Modules' Steps Performance

During the evaluation process, the results were analyzed at different stages of the correction process, namely raw data, pre-processed data, and corrected data (see Table 4.2). R^2, RMSE and NRMSE values were calculated for each module's step, and the results were presented separately for each sensor, comparing the concentration data from these sensors with those from a reference station, enabling a thorough assessment of their performance across different periods.

One notable finding from the evaluation is the beneficial impact of the pre-processing on anomalous sensors, such as ari-1727. These sensors exhibited improved performance after undergoing this step. However, for other sensors, the pre-processing did not yield significant improvements in the results. Its impact on their performance was relatively low.

In Figure 4.8 a comparison is presented between the raw data and the pre-processed and corrected data, displaying four weeks of data collected from LCS and the reference instrument (Arpa) at the monitoring site in Turin. The data for each week is presented



(a) Inclusive

(b) Non-inclusive

Figure 4.7: Comparison of correction approaches within the same time frame: by incorporating new observations in the operative window (a), and without the incorporation of new observations (b).

as a time series, with the black line representing the reference instrument and LCS shown in different colours. The red line on the graph represents the level of relative humidity, with a dotted threshold of 70% RH shown.

During the analysis of the data, it was noted that certain weeks did not meet the predefined threshold for changes in the measured parameters. As a result, these weeks were included in the overall results but were not specifically highlighted in the figures. This implies that the growth function was not applied during those weeks since the threshold for correction was not surpassed. While these weeks may not demonstrate significant changes or corrections, they are still considered in the evaluation process to provide a comprehensive assessment of the sensor performance.

As depicted in the figure, a notable observation is an evident increase in PM concentration when the RH level surpasses the threshold of 70%. This finding aligns with previous studies in the literature, reinforcing the clear correlation between high RH levels and elevated PM concentrations detected by LCS. Despite discovering that a

(a) Spring period

	Raw data			Pre-processed data			Corrected data		
Sensor ID	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE
ari-1727	0.562	29.440	1.335	0.631	25.261	1.145	0.724	14.138	0.641
ari-1952	0.599	22.877	1.037	0.601	22.927	1.039	0.810	9.552	0.433
ari-1953	0.554	26.908	1.220	0.558	27.269	1.236	0.809	9.782	0.443

(b) Autumn period

	Raw data			Pre-processed data			Corrected data		
Sensor ID	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE
ari-2049	0.177	116.174	5.444	0.168	115.019	5.390	0.647	12.600	0.590
ari-1885	0.232	75.202	3.524	0.218	74.798	3.505	0.625	11.419	0.535

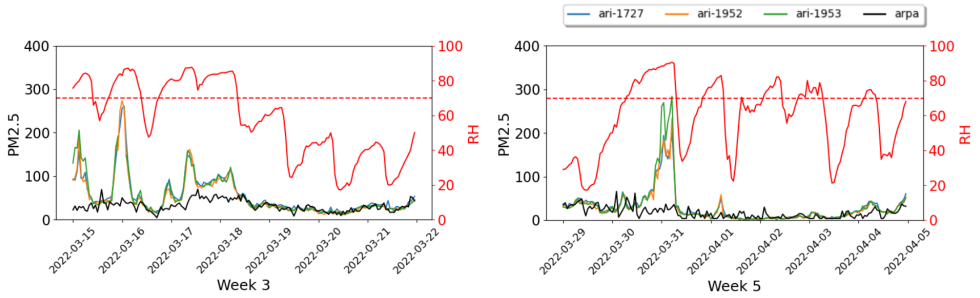
Table 4.2: Evaluation of LCS data during spring (a) and autumn (b) periods.

threshold slightly higher than 70% resulted in improved performance with data, the decision was made to maintain a general threshold commonly used in other studies. This choice was motivated by the goal of developing a procedure that can be applied even in situations where no reference station is available. In any case, the selected threshold of 70% for the SPS30 sensors appears to effectively capture the occurrence of hygroscopicity events. Indeed, it is important to acknowledge that the sensitivity to humidity may vary among different sensors. This is particularly important when dealing with sensors from different manufacturers, as their designs and manufacturing processes can introduce additional variations in their performance. This variation in humidity sensitivity could explain the discrepancies observed between different sensors in their response to RH levels. Therefore, it becomes crucial to carefully evaluate and interpret the data from each sensor in the context of its specific characteristics and limitations.

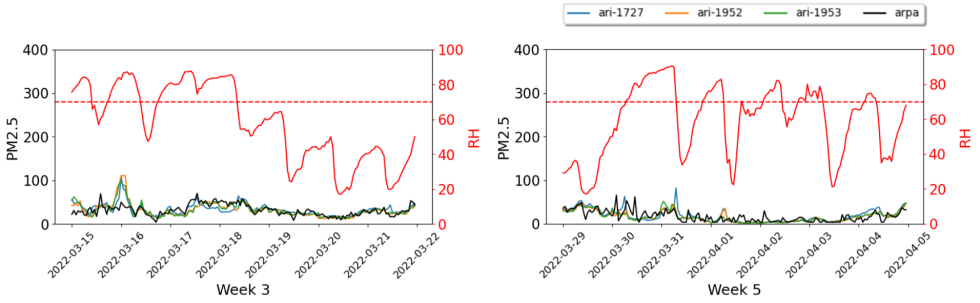
Another significant observation from the figure is the clear cyclic pattern of humidity throughout the day. This cyclic pattern represents the regular fluctuations in humidity levels that occur over a day. These fluctuations can be influenced by various factors, including temperature variations, diurnal weather patterns, and human activities. It is important to note that taking a window size that is too small, such as just a few hours, may lead to the omission of important information regarding the classic behaviour of RH. By using a larger window size, encompassing a longer period, it becomes possible to capture and analyze the complete cyclic pattern of humidity (this aspect is presented also in sub-section 4.2.3).

For a more comprehensive understanding of the growth function's impact, an illustrative example is provided in Figure 4.9. This example pertains to the spring period, specifically weeks 3 and 5, where LCS data is depicted before and after correction. Examining the scatter plots reveals that raw data collected under high relative humidity conditions have undergone reduction through the growth function, resulting in improved alignment with the reference station data.

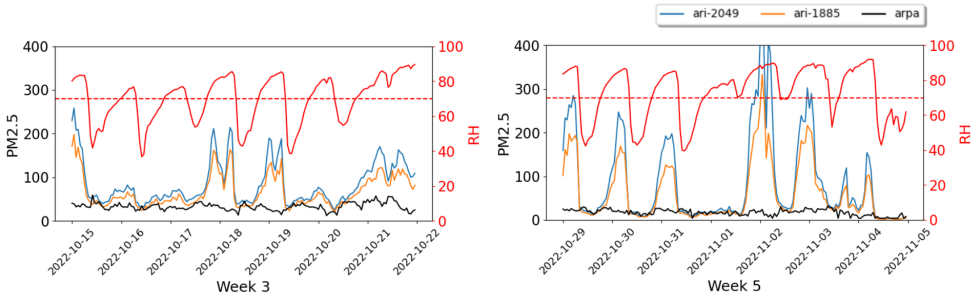
Another type of evaluation involves assessing how many times the LCS observations exceeds the maximum value recorded by the reference station in the period under examination (see Table 4.3). This evaluation can be performed even if the reference station data is not available on an hourly or daily basis. Instead, it is sufficient to have



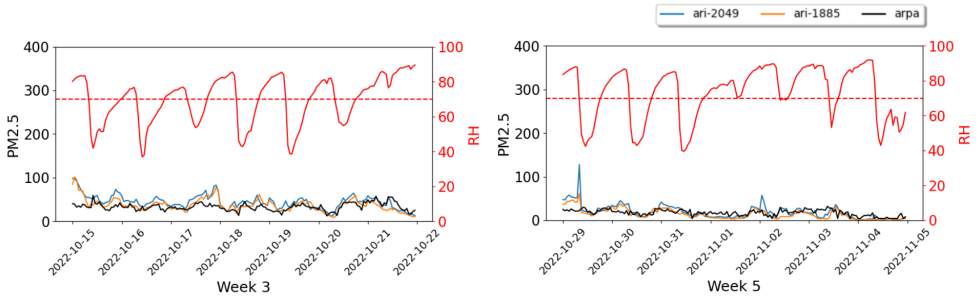
(a) Spring period raw data instance



(b) Spring period corrected data instance



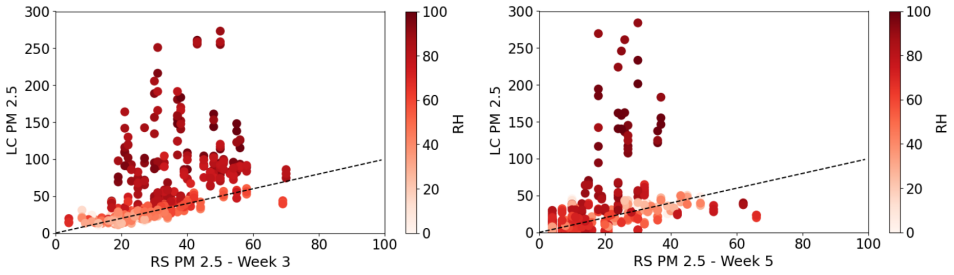
(c) Autumn period raw data instance



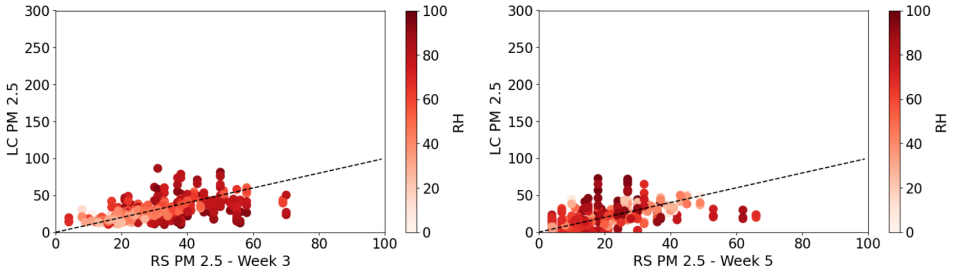
(d) Autumn period corrected data instance

Figure 4.8: Performance overview of the MitH framework during spring and autumn periods, comparing original raw data with corrected data

information regarding the maximum value detected and reported by the local agency.



(a) Spring period raw data



(b) Spring period corrected data

Figure 4.9: Low-cost sensor data (y-axis) during the spring period, depicting the comparison with the reference station (x-axis) before and after correction. The colour gradient corresponds to relative humidity levels, and the black dotted line signifies perfect alignment.

Sensor ID	Before correction	After correction
ari-1727	7.6%	2.6%
ari-1952	5.3%	0.8%
ari-1953	6.0%	1.1%
ari-1885	35.4%	7.3%
ari-2049	27.6%	6.3%

Table 4.3: Percentage of LCS observations exceeding the maximum values detected by the reference station.

4.2.5 Comparison with Existing Approaches

In this section, a comparative analysis between the approach presented in this study and existing methods proposed in the literature is conducted. Specifically, the growth functions used in [Str17] (Equations 4.2 and 4.3) and [SCT+14] (Equations 4.4 and 4.5) were considered. For the equations of *Chakrabarti* and *Richard*, fixed parameters were provided, and they were applied as such. However, in the analysis, each equation was re-fitted using the rolling window history procedure, enabling a more tailored and adaptable approach. This comparison allows for an assessment of the performance and effectiveness of MitH framework in comparison to existing techniques.

Regardless of whether fixed parameters were used or the equations were re-fitted, it is important to note that the pre-processing module's step was always applied. The

data underwent the three phases of module 2 (see Figure 4.1), which involved statistical anomaly removal, data imputation, and smoothing. This ensured that the data used in the analysis were properly pre-processed and free from gross statistical anomalies, providing a reliable basis for comparison and evaluation of the different growth functions.

	Spring period			Autumn period		
Growth function	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE
Combination	0.781	**11.157**	**0.506**	**0.636**	**12.010**	**0.563**
Hänel	0.716	12.676	0.575	0.494	14.846	0.696
Richard	0.723	13.496	0.612	0.530	14.980	0.702
Chakrabarti	0.741	12.496	0.567	0.536	15.081	0.707
Richard *	**0.793**	11.530	0.523	0.383	35.554	1.666
Chakrabarti *	0.787	11.987	0.543	0.395	31.661	1.484

Table 4.4: Performance comparison of different growth functions - Parameters: *window_size* of 1 day, optimization method correlation minimization, batch dimension of 1 day - the asterisk (*) indicates the use of fixed parameters as described in the original work.

The results in Table 4.4 provide an assessment of the performance of various growth functions when evaluated against the reference station data. To ensure consistency in the evaluation, the LCS data were re-sampled at an hourly frequency to match the granularity of the reference station data. The *Combination* growth function shows the best overall performance, with high R^2 values and low RMSE and NRMSE values. This indicates that the *Combination* function corrects for the influence of relative humidity on the sensor readings, leading to more accurate results.

In the spring period, *Chakrabarti* and *Richard*'s equations, with fixed parameters, also show relatively good performance, with R^2 values higher than the *Combination* function. However, in the autumn period, their performance is noticeably worse. This suggests that these equations may not be as effective in capturing the influence of relative humidity during different humidity conditions. This highlights the importance of considering the specific environmental context and optimizing the growth function parameters accordingly. In this way, parameters are optimized for the location and context, but remain dynamic, and the approach could be exploited in each location, also without having a reference station co-located.

Hänel's growth function shows the lowest performance among the evaluated functions, indicating that it may not adequately address the hygroscopicity issue caused by relative humidity.

4.2.6 Comparison with the Wiseair Approach

The current study originally aimed to propose an alternative algorithm to the existing one used by Wiseair. In the existing approach, the pre-processing module involves masking PM values above a certain concentration threshold and filling them with the last available value. The algorithm then differentiates between summer and autumn months and applies a *Combination* growth function with fixed parameters, chosen to

use reference station data. Finally, regression is applied specifically for autumn observations. It is important to note that the existing algorithm relies on the use of a reference station for data correction, in fact, it focuses on removing the concentration of above-limit RH rather than cleaning it. This approach introduces complexity, as the regression model used during autumn is trained using data from a different period and location.

In comparing the performance of MitH to that of Wiseair's, the *single observation mode* was employed, tailored for real-time applications, allowing the correction of each observation upon arrival. The outcomes demonstrated notable performance enhancements with the MitH framework. Specifically, during the spring period, there was an increase in R^2 values by approximately 15%. In the autumn period, the improvement was even more remarkable, with an increase in R^2 values of around 30%. Correspondingly, RMSE and NRMSE values diminished.

4.3 MitH Framework Implications

Based on the preceding discoveries, several noteworthy observations and implications can be discussed.

As anticipated, the comparison between the raw data and the corrected data reveals a significant impact of relative humidity on the PM concentration levels detected by LCS. When the RH exceeds a certain threshold, there is a noticeable increase in PM concentration. This observation emphasizes the importance of considering RH as a significant factor in data correction and analysis.

Additionally, the presence of an anomalous sensor, ari-1727, stands out in the analysis. This sensor consistently displays higher concentration levels compared to the other sensors. It is important to address such anomalies during the pre-processing, as they may not be directly related to hygroscopicity. The growth function, in this case, does not adequately address this anomaly, and further investigation and pre-processing phases may be required to address the signal compromise.

When it comes to optimizing the parameters of the growth function, the choice between including or excluding the observations to be corrected during the optimization module is critical. By including the observations, a more comprehensive context is considered, resulting in improved correction results. Conversely, excluding the observations may lead to a less effective correction, especially when RH levels exceed the threshold.

Furthermore, when comparing different measurement periods, such as spring and autumn, it becomes apparent that there are seasonal variations in PM concentration and RH levels. It is evident that these variations can impact the performance of the growth function. The results show that the autumn period tends to yield worse performance in terms of R^2, RMSE, and NRMSE compared to the spring period. This can be attributed to long periods where RH exceed the threshold levels.

Additionally, the cyclic pattern of humidity throughout the day is an interesting observation. This cyclic pattern, which is likely influenced by temperature variations, diurnal weather patterns, and human activities, can contribute to fluctuations in the PM concentration levels. It is important to consider this cyclic pattern when determining the appropriate window size for the rolling window correction. Choosing a window size that aligns with the duration of the humidity cycle can help capture the relevant

variations and improve the accuracy of the growth function.

When comparing MitH with existing approaches, MitH demonstrates competitiveness. It is important to note that the goal is to develop a growth function that does not rely on a ground truth, such as a reference station. In the literature, parameters for existing approaches are often found using regression techniques. In this study, growth functions were optimized by minimizing the correlation between RH and PM, following the suggestion in [Str17]. Additionally, a technique involving the minimization of the distance between the original PM distribution under 70% RH and the distribution of data above 70% RH has been considered. Although this approach served as a starting point, it required further refinement. As a result, the corrected data using this approach performed better for the spring period but worse for the autumn period. The metric used for minimization was the standardized mean difference (SMD), and alternative approaches may yield better results. The decision was made not to pursue further exploration in this direction; however, it is important to recognize that alternative approaches could potentially provide better results.

Furthermore, an exploration was conducted to combine the *Combination* equation with the *Chakrabarti* equation. However, the results exhibited fluctuations, indicating that this combination might not be effective in addressing hygroscopicity.

Overall, these findings highlight the complexity and nuances involved in correcting LCS data for PM concentration. It underscores the importance of considering multiple factors, including RH, anomalies, seasonal variations, and optimization strategies, to improve the accuracy and reliability of the growth function. Further research and refinement of the corrective algorithms are necessary to address these challenges and enhance the performance of LCS in monitoring and assessing air quality.

4.4 MitH as Pre-Processing

I aim to explore the integration of this procedure as an initial step, followed by the application of other techniques, including well-known machine learning and deep learning models, to further refine and adjust the data.

The models used are the following:

- **Linear Regression Models with Regularization**: Linear Regression, Ridge Regression, Lasso Regression, ElasticNet
- **SVR**
- **ARIMA**
- **Decision Tree and Boosting Models**: Decision Tree, Random Forest, XGBoost, LightGBM
- **AirMLP**: This is a family of Multi-Layer Perceptron neural network architectures [CPZ23a] described in Section 5.3.3.

Table 4.5 presents the performance metrics of the listed models when applied to raw data, while Table 4.6 illustrates their performance after the data have been processed using the MitH framework.

In these tables, results are presented in terms of R^2 (see Eq. 2.1), RMSE (see Eq. 2.2), MAE, MSE, and MAPE:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.6}$$

where y_i are the observed values and $\hat{y}_i$ are the predicted values. The Mean Absolute Error (MAE) measures the average of the absolute errors, providing a clear idea of the magnitude of errors in the predictions.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.7}$$

where y_i are the observed values and $\hat{y}_i$ are the predicted values. The Mean Squared Error (MSE) calculates the average of the squared errors, emphasizing larger errors due to the squaring operation.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{4.8}$$

where y_i are the observed values and $\hat{y}_i$ are the predicted values. The Mean Absolute Percentage Error (MAPE) expresses the error as a percentage of the observed value, providing an interpretable measure of error relative to the magnitude of the actual value.

Model	**R-squared**	**MAE**	**MSE**	**RMSE**	**MAPE**
Linear Regression	0.3068	9.3604	150.8997	12.2841	87.3521
ElasticNet	0.2855	9.5801	155.5424	12.4717	90.3214
Ridge	0.3068	9.3610	150.9153	12.2848	87.3611
Lasso	0.2870	9.5760	155.2270	12.4590	90.3296
DecisionTreeRegressor	0.0201	9.1985	213.3142	14.6053	65.6560
RandomForestRegressor	0.6101	6.5613	84.8784	9.2129	51.7430
SVR	0.4266	8.2613	124.8309	11.1728	63.6250
XGBRegressor	0.5746	6.7409	92.6067	9.6232	51.7413
ARIMA	-0.8675	15.2304	406.5509	20.1631	132.9522
LightGBM	0.6143	6.4859	83.9601	9.1630	49.9900
AirMLP_7	0.740	4.824	63.663	7.979	38.5220

Table 4.5: Performance metrics of various models applied to the raw dataset.

Across all models, the R^2 values are consistently higher in Table 4.6, indicating a better model fit to the data. For example, the AirMLP model improves from 0.740 to 0.804, showcasing enhanced predictive accuracy.

The MitH framework positively influences all evaluated metrics, enhancing accuracy, reducing errors, and improving overall model performance. The AirMLP model, in particular, demonstrates the most substantial improvements, reinforcing the utility of the MitH pre-processing step for optimizing predictions.

Model	R-squared	MAE	MSE	RMSE	MAPE
Linear Regression	0.4292	8.1876	125.1118	11.1853	74.5252
ElasticNet	0.4256	8.2331	125.9014	11.2206	75.8111
Ridge	0.4292	8.1876	125.1117	11.1853	74.5253
Lasso	0.4257	8.2408	125.8674	11.2191	75.9676
DecisionTreeRegressor	0.2862	8.7211	156.4534	12.5081	62.1568
RandomForestRegressor	0.6295	6.5298	81.2141	9.0119	52.9281
SVR	0.4424	8.1958	122.2100	11.0549	66.6813
XGBRegressor	0.6131	6.7246	84.8112	9.2093	52.4557
ARIMA	-0.8035	14.9074	395.2983	19.8821	128.4221
LightGBM	0.6417	6.4427	78.5272	8.8616	50.7500
AirMLP_7	0.804	4.473	48.854	6.990	32.9650

Table 4.6: Performance metrics of various models applied to the dataset underwent the MitH framework.

4.5 Future Steps

The Mitigating Hygroscopicity (MitH) framework is a central component of the AIQS project, where I serve as the work package leader for WP1: "Sensor Data Analysis and TRL Advancement of the MitH Framework." This work package focuses on improving the accuracy of air quality data collected by low-cost sensors.

The broader objective of the AIQS project (Dec 2024 - September 2025) is to integrate highly accurate air quality information into routing algorithms, enabling pedestrians to navigate urban areas using the least polluted and greenest paths. The initiative is funded through an internal call of the larger ECOSISTER project [1] , which aims to support the ecological transition of the Emilia Romagna region. Its objective is to integrate digital transformation and sustainability while promoting employment, improving people's well-being, and protecting the environment.

[1] https://ecosister.it/ **Funding acknowledgment:** "Ecosystem for Sustainable Transition in Emilia-Romagna", Code: ECS_00000033 – CUP E93C22001100001. Funded under the National Recovery and Resilience Plan (PNRR) – Mission 4 "Education and Research," Component 2 "From Research to Enterprise," Investment 1.5 "Creation and Strengthening of Innovation Ecosystems – Building Territorial *R&D* Leaders," financed by the European Union – NextGenerationEU. Ref: MUR Notice 3277/2021.

Chapter 5

PM Data Adjustment and Calibration

As explained in Section 2.3, a wide range of techniques have been explored to improve the accuracy of particulate matter data from low-cost sensors. The methodologies span a spectrum of approaches:

- **Traditional Statistical Methods:** Techniques such as linear regression and multivariate statistical analysis that have been foundational in sensor calibration.
- **Artificial Intelligence Approaches:** Advanced methods like neural networks, fuzzy logic systems, and other machine learning techniques that leverage data-driven insights for calibration and anomaly detection.
- **Hybrid Models:** Combinations of traditional and AI-based methods, offering a balance between interpretability and predictive power.

In this context, several models have been applied to the air quality data gathered during my PhD research. These models include linear regression, ridge regression, lasso regression, elastic net, support vector regression, decision tree, random forest, XGBoost, LightGBM, recurrent neural network, multilayer perceptrons and fuzzy logic.

This chapter is divided into three main sections.

- **Data Preparation**: Section 5.1 outlines the preprocessing steps applied to the dataset, including partitioning strategies and the use of statistical methods for anomaly detection.
- **Adaptive Neuro-Fuzzy Inference System (ANFIS)**: Section 5.2.2 presents a detailed discussion of the Fuzzy Inference System and the principles of ANFIS, followed by its specific applications in air quality data adjustments.
- **Machine Learning Techniques**: Section 5.3 introduces the methodologies employed for advanced data analysis.
  - *Splitting Strategies*: A review of splitting strategies is provided for dataset partitioning. 5.1.1

- *Anomaly Detection Methods*: Various methods for detecting anomalies in the dataset are discussed in this part. 5.1.2
- *Model Performance Analysis*: The evaluation of model outcomes is covered in this section, assessing the performance of different machine learning models. 5.3.1
- *LightGBM*: This section focuses on the use of LightGBM, including feature selection and skewness transformation. 5.3.2
- *Multilayer Perceptron*: Detailed results from the application of Multilayer Perceptron are presented. 5.3.3

Overall, this chapter integrates both traditional adjustment frameworks and cutting-edge machine learning methods to enhance the quality and usability of air quality data.

5.1 Data Preparation

5.1.1 Splitting Strategies

In this section, the study conducted to determine the optimal strategy for splitting the dataset into training and testing sets is described. The initial consideration was the level of granularity for the split: whether to treat each individual instance (in the current dataset configuration, an instance corresponds to an hourly record) as an independent unit or to group data at the level of entire days, thereby preserving a broader temporal structure.

Both approaches have their advantages and disadvantages:

- A split at the instance level provides a larger amount of data for training and testing, enabling the model to learn more comprehensively. However, it risks breaking temporal dependencies.
- Conversely, splitting at the day level preserves daily patterns, which is important as sensors often exhibit non-linear responses to daily particulate matter fluctuations.

Before settling on a final configuration, various splitting strategies were explored to identify the optimal one for this case study. The strategies ranged from simple random splits across the entire time series to more complex splits that accounted for the temporal characteristics of the data.

The splitting strategies considered are as follows:

1. **Instance-Level Split**
   - **Random Day Split (RDS):** The dataset is grouped by days, and for each day, 75% of the instances are randomly assigned to the training set and 25% to the test set.
   - **Random Month Split (RMS):** The dataset is grouped by months, and for each month, 75% of the instances are randomly assigned to the training set and 25% to the test set.

- **Random Total Split (RTS):** The entire time series is considered, and 75% of the instances are randomly assigned to the training set and 25% to the test set.

2. **Day-Level Split**

   - **Random Total Split (RTS):** The entire time series is considered, and 75% of the days are randomly assigned to the training set and 25% to the test set.
   - **Random Month Split (RMS):** The dataset is grouped by months, and for each month, 75% of the days are randomly assigned to the training set and 25% to the test set.

	Instance			Day	
Dataset	**RDS**	**RMS**	**RTS**	**RMS**	**RTS**
Aosta	**0.89**	0.88	**0.89**	0.79	0.75
Badajoz	0.80	0.83	**0.87**	0.59	0.60
Bangalore	**0.91**	0.90	0.89	0.88	0.87
Calgary	**0.98**	**0.98**	0.97	0.83	0.86
Delhi	0.88	**0.91**	**0.91**	0.86	0.87
Hamirpur	**0.96**	**0.96**	**0.96**	0.95	0.96
Lima IQAIR	0.77	**0.80**	0.72	0.64	0.76
Lima AIR-BEAM	0.77	**0.81**	0.78	0.70	0.76
UK PMS5003	0.23	**0.84**	0.49	0.24	0.20
UK SPS30	0.26	**0.82**	0.47	0.25	0.22

Table 5.1: Comparison of R^2 for different splitting techniques using LightGBM.

Table 5.1 presents the results of the various combinations of granularity and temporal splits tested, using a LightGBM regression model for training.

The results show that, for all datasets, the best performance is obtained with instance-level granularity combined with the *Random Month Split*. This choice is further supported by the fact that instance granularity captures critical patterns and takes full advantage of data variability, ultimately improving the model's generalization.

5.1.2 Anomaly Detection Methods

This section presents the results obtained by applying the three different anomaly detection methods, to the datasets. The results are summarized in Table 5.2. For each dataset, the initial raw data R^2 value is reported alongside the best-performing case for each method, including the time window that led to this performance.

Interquartile Range (IQR) The Interquartile Range, or IQR (Figure 5.1), is a measure of variability calculated by subtracting the first quartile (25th percentile or Q1) from the third quartile (75th percentile or Q3). Tukey's method is a common

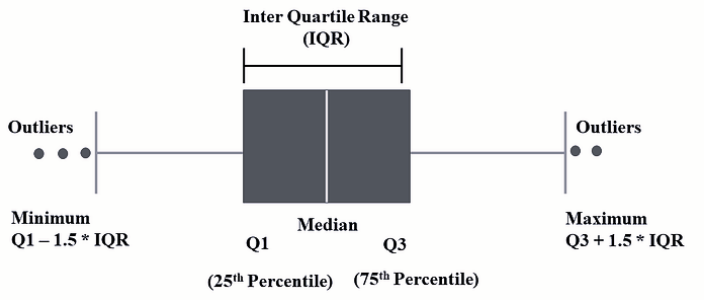


Figure 5.1: Illustration of the Interquartile Range (IQR).

approach that uses the IQR to identify outliers, defining any value outside the range of 1.5 times the IQR as an outlier.

3 Sigma Rule The three-sigma rule is a statistical technique and one of the most commonly used methods for detecting outliers. It assumes the data follows a normal (Gaussian) distribution. The three-sigma limits are calculated using the mean (μ) and standard deviation (σ) of the data. The formula for the three-sigma limits is:

$$L_{\text{lower}} = \mu - 3\sigma, \quad L_{\text{upper}} = \mu + 3\sigma$$

A value greater than the upper limit is considered a positive outlier, while a value lower than the lower limit is considered a negative outlier.

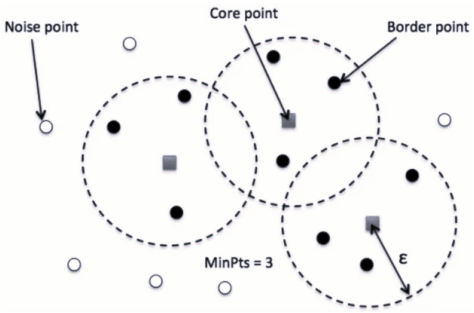


Figure 5.2: Representation of DBSCAN point types.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) DBSCAN (Figure 5.2) is a clustering algorithm based on two hyperparameters:

- **minPts**: Represents the minimum number of points required to form a cluster. As a general rule, $minPts = 2 \cdot D$, where D is the number of dimensions in the dataset.
- **epsilon (ε)**: Defines the maximum distance between two points for them to be considered part of the same cluster. The value of ε can be determined automatically using the *K-distance* method, where $K = minPts$. The computed K-distances are sorted in ascending order and plotted. The optimal value of ε corresponds to the point of maximum curvature on the plot (also known as the "elbow"), where the slope changes most rapidly.

After clustering, the data points are classified into three categories:

- **Core points**: Points with at least minPts neighbors within their ε-neighborhood.
- **Border points**: Points that are not core points but are within the ε-neighborhood of a core point.
- **Noise points (outliers)**: Points that are neither core points nor border points. These are identified by having fewer than minPts neighbors within a distance of ε.

Anomaly Detection Results

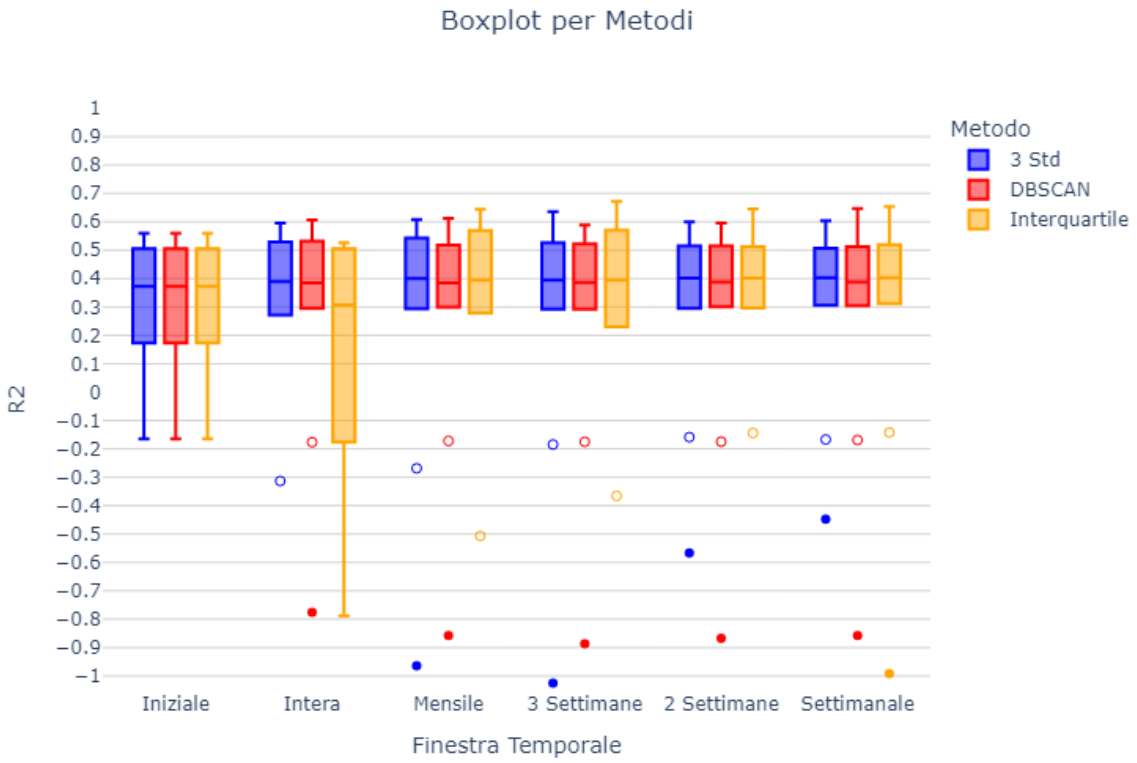


Figure 5.3: Boxplot for the three methods: 3 Sigma, Interquartile, and DBSCAN.

Figure 5.3 illustrates the R^2 metric boxplots for each time window across the three methods.

The results demonstrate that for datasets with performance improvements (excluding Delhi, where the values remained constant, and Lima AIRBEAM, UK SPS30, and UK PMS5003, where performance declined), the best results were achieved using the Interquartile method. For time windows, the weekly interval was chosen for consistency, despite half the datasets improving with a three-week window. Notably, one dataset showed performance degradation with the three-week window. As a result, the Interquartile method with a weekly window was applied to all datasets showing improved performance and used for calibration.

	R^2			
	Time Window			
Dataset	**Initial**	**Interquartile**	**3 Sigma**	**DBSCAN**
Aosta	0.17	0.67	0.64	0.65
	-	3 Weeks	3 Weeks	Weekly
Badajoz	-0.16	-0.14	-0.15	-0.16
	-	Weekly	2 Weeks	Weekly
Bangalore	0.45	0.46	0.45	0.45
	-	Weekly	Weekly	3 Weeks
Calgary	0.50	0.58	0.54	0.52
	-	3 Weeks	Monthly	3 Weeks
Delhi	0.51	0.51	0.51	0.51
	-	Weekly	Weekly	Weekly
Hamirpur	0.56	0.57	0.56	0.56
	-	3 Weeks	Weekly	Weekly
Lima IQAir	0.28	0.31	0.30	0.30
	-	Weekly	Weekly	Weekly
Lima AIRBEAM	0.35	0.33	0.34	0.31
	-	2 Weeks	Weekly	Weekly
UK PMS5003	0.38	0.34	0.32	0.32
	-	Weekly	Weekly	2 Weeks
UK SPS30	0.15	-0.99	-0.44	-0.85
	-	Weekly	Weekly	Weekly

Table 5.2: Anomaly detection results for each dataset.

5.2 Adaptive Neuro-Fuzzy Inference System

Fuzzy logic is a computational paradigm that mimics human decision-making under uncertainty. It has found diverse and impactful applications across various facets of environmental sciences [PK13, SG16, DTA22, PDTD24]. Its adaptability to handle imprecise and vague information makes it a valuable tool in addressing the inherent complexity and uncertainty prevalent in environmental systems [BCC98, KJK20].

One of the applications of fuzzy logic lies in environmental modelling, where it serves as a bridge between traditional deterministic models and the unpredictable nature of ecological processes. By incorporating fuzzy logic, researchers can capture the nuances of environmental variables that resist precise quantification, providing a more realistic representation of the intricate relationships within ecosystems. This approach enhances the accuracy of predictive models, contributing to more effective decision-making in areas such as climate change projections, land-use planning and biodiversity conservation [DA23, Rah20, BSPP21, CLL+16].

Furthermore, fuzzy logic plays a crucial role in the field of air and water quality monitoring [BGBD23, TTK+22, MBC+22, GA18]. Environmental data, often characterized by inherent uncertainties and variations, can be challenging to interpret accurately. Fuzzy logic-based systems excel in processing and analyzing this data, offering a robust framework to account for imprecision in pollutant measurements. This methodology

proves particularly beneficial in discerning pollution levels near regulatory thresholds, aiding in timely interventions and ensuring compliance with environmental standards [W+21].

In environmental risk assessment, fuzzy logic provides a nuanced approach to evaluating the potential impacts of contaminants [SSC19]. Traditional risk assessment methods often rely on deterministic assumptions, neglecting the variability in exposure scenarios and ecological response values. In the field of air monitoring, fuzzy logic has been successfully applied to predict Air Quality Index levels, focusing on gases and particulate matter pollutants, particularly PM_{10}, or encompassing gases alone.

The motivation behind this study stems from the growing importance of ensuring accurate pollution measurement in environmental science, particularly with regard to data derived from low-cost monitoring stations. As human activities continue to exert increasing pressure on air quality, there is a pressing need for precise and reliable data to inform effective mitigation strategies and protect public health [NK24]. The inadequacies of traditional monitoring methods and the proliferation of low-cost sensor technology underscore the significance of addressing this challenge. Using advanced methodologies such as the Adaptive Neuro-Fuzzy Inference System to adjust sensor data at low-cost, this study aims to deepen understanding of environmental dynamics and improve monitoring capabilities. This, in turn, will facilitate more informed decision-making in environmental management.

The novelty of this research is demonstrated through several key aspects: the collection and analysis of low-cost $PM_{2.5}$ data from six SPS30 Sensirion sensors co-located with a reference station, ensuring a comprehensive dataset; the development of a hybrid methodology that integrates fuzzy logic and neural networks for air quality adjustment, offering a unique perspective on sensor calibration; a comparative evaluation of machine learning techniques and ANFIS, shedding light on the strengths and advantages of the proposed approach; and an in-depth investigation into data and model optimization strategies to enhance the accuracy and reliability of ANFIS in air quality analysis. These contributions advance the current state-of-the-art in environmental data processing and monitoring.

To provide insights about the Adaptive Neuro-Fuzzy Inference System applied to the problem of air quality, the objective is to juxtapose established models such as linear regression [Wei05], decision trees [Loh11], random forests [JBLL14], SVR [CM04], and MLP neural networks [IL67] employed as baseline models. The neural network includes an input layer, a batch normalization layer, seven subsequent dense layers of 1500 neurons each (all with ReLU activations), and a final output layer [CPZ23a]. The ANFIS model represents a novel approach for adjusting low-cost PM data; in this way, these machine-learning techniques were selected as benchmarks due to their proven efficacy in calibration procedures.

Understanding the operation of the ANFIS model requires exploring the principles of fuzzy logic and Fuzzy Inference Systems (FIS). These concepts are explored in more detail in the next subsections.

5.2.1 Fuzzy Inference System

Fuzzy logic introduced by Lotfi A Zadeh in 1965 [ZFS96, Zad78] differs from conventional logic, which operates in binary terms (`TRUE` or `FALSE`), through offering a paradigm shift by representing truth not as a binary state but rather as a continuum of truthfulness, ranging in a limited space. This approach allows for a more nuanced representation of real-world phenomena that may not have clear-cut boundaries and do not strictly adhere to Boolean truth. Instead of absolute certainty, fuzzy logic implements degrees of certainty, often represented as linguistic terms (e.g., `low, fair, high` or `certainly yes, possibly yes, possibly no, certainly no`).

A Fuzzy Inference System [MA75] consists of several essential components. An illustration is included in Figure 5.4 to enhance the explanation of the subsequent stages:

1. Fuzzification module: This initial stage involves the transformation of crisp (numeric) input values into fuzzy variables, often represented linguistically, through assigned membership functions (MFs) (Figure 5.4a). The resulting output is represented as a vector from the fuzzification process, which is subsequently used during the inference step (Figure 5.4b). Generally, given a universe X in which the variable x is defined, the fuzzy set A in X comprises ordered pairs, as expressed by:

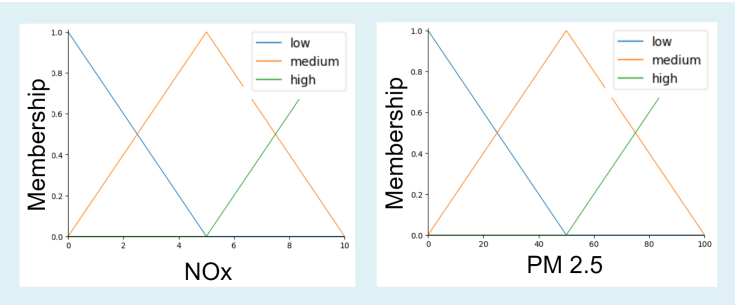
   $$A = \{(x, \mathrm{MF}) \mid x \in X\}$$

   Here, the MF represents the membership function that maps each element of X to a membership value between 0 and 1. The MF can assume various shapes, depending on which best describes the universe under consideration, including linear, Gaussian, sigmoid, quadratic and cubic polynomials, or simpler forms composed of straight lines like triangular, trapezoidal, linear ascending or linear descending.

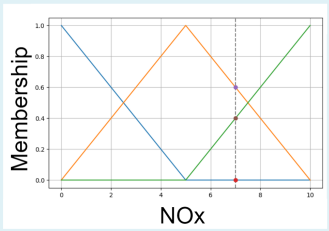
2. Knowledge Base: The knowledge base of a FIS comprises a set of expert-provided rules in the form of `IF-THEN` statements (Figure 5.4c). Each rule specifies conditions (antecedents) based on input variables and corresponding actions (consequents) based on output variables. Each rule can comprise logical operators (*AND*, *OR*, and *NOT*) when combining multiple states regarding different variables. The Boolean logic operators *AND*, *OR*, and *NOT* are typically defined within the scope of fuzzy logic, as operators of minimum, maximum and complement; in this case, they are also called *Zadeh operators* [Zad65] and are defined as follows:

   $$\mathrm{NOTx} = 1 - \mathrm{MF(x)xANDy} = \min(\mathrm{MF(x)}, \mathrm{MF(y)})\mathrm{xORy} = \max(\mathrm{MF(x)}, \mathrm{MF(y)})$$

   It is worth noting that as the number of input fuzzy variables increases, the number of rules typically grows, often showing exponential expansion. While the sheer number of rules might suggest the system's complexity, it is crucial to recognize that a system with fewer membership functions per variable could be more complex, especially when incorporating more variables [GSV17].



(a) Input variables membership functions



NOx Level: 7
Membership Degree - Low: 0.0
Membership Degree - Medium: 0.6
Membership Degree - High: 0.4

PM2.5 Level: 75
Membership Degree - Low: 0.0
Membership Degree - Medium: 0.5
Membership Degree - High: 0.5

(b) Fuzzification

Rule 1: IF NOx[low] AND PM2.5[low] THEN AQI Level[good]

Rule 2: IF NOx[medium] OR PM2.5[medium] THEN AQI Level[moderate]

Rule 3: IF NOx[high] OR PM2.5[high] THEN AQI Level[unhealthy]

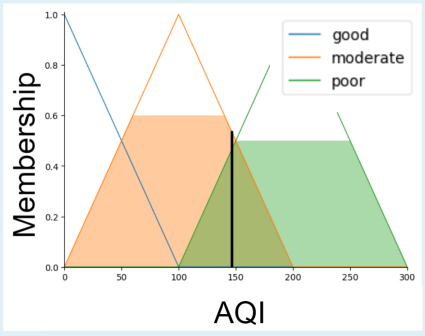
(c) Rules

Rule 1 inference: min(NOx[0.0], PM2.5[0.0]) AQI Level[0.0]

Rule 2 inference: max(NOx[0.6], PM2.5[0.5]) AQI Level[0.6]

Rule 3 inference: max(NOx[0.4], PM2.5[0.5]) AQI Level[0.5]

(d) Inference



AQI Level: 146.64

(e) Aggregation and defuzzification

Figure 5.4: The subfigures illustrate the sequential steps and components involved in the fuzzy inference process.

3. Inference engine: The inference engine simulates the human reasoning process by performing fuzzy inferences based on the inputs and `IF-THEN` rules. Each rule

may carry a weight, typically ranging from 0 to 1, to increase or decrease its effect, or all rules can be assigned a weight of 1 to have equal importance. Each involved variable is assigned a degree derived from its membership function (Figure 5.4d), and the resulting rule output is inferred.

4. Aggregation of inference outputs: In an FIS, decisions are made by testing all the rules, and the outputs of these rules need to be combined. This aggregation process merges the fuzzy sets representing the output of each rule into a single fuzzy set (Figure 5.4e).

5. Defuzzification module: The fuzzy set obtained from the inference engine is converted into a crisp value through defuzzification. Defuzzification is necessary to derive a single output value from the set, one common method is centroid calculation, which determines the centre of the area under the aggregate fuzzy set (Figure 5.4e).

Fuzzy Logic has been used in different fields and it has been proved to deal with the uncertainty and subjectivity of environmental problems adequately, as in [ODFHDS06]. Breaking down the Fuzzy Inference System into systematic steps enhances user understanding, providing clarity on how the output corresponds to input values and rules. This improves the system's interpretability, building trust in its decision-making process. Additionally, the utilization of linguistic variables and rules facilitates domain expert involvement, empowering them to refine the system's performance as necessary.

5.2.2 Adaptive Neuro-Fuzzy Inference System

The Adaptive Neuro-Fuzzy Inference System integrates fuzzy reasoning principles with the structural characteristics of neural networks, enabling it to learn and adapt from data dynamically [Say21, CYP+22].

Initially, ANFIS constructs a FIS with a basic framework, lacking a comprehensive understanding of membership functions or rules. However, it iteratively refines and optimizes these rules and functions to minimize output error or to enhance the explanation of complex system behaviors. This optimization is achieved through the adjustment or tuning of membership function parameters using hybrid learning algorithms or backpropagation techniques applied to specific input-output data patterns [KK19]. Through this integration, ANFIS effectively constructs fuzzy *IF-THEN* rules and membership functions, enabling accurate modelling of input-output relationships.

The resulting model remains highly interpretable, with easily understandable rules. This characteristic is particularly beneficial for systems where verification and certification play a crucial role.

5.2.3 Application of ANFIS

The dataset employed in this study is the Turin one, which underwent meticulous pre-processing to enhance data quality and consistency. Key pre-processing steps include the following:

1. Standardizing data frequency to 1 hour: The granularity of the data obtained from the reference station is at one-hour intervals. To ensure consistency, the data derived from the SPS30 sensors has been resampled hourly using the nearest approximation method. In the context of Python resampling, the nearest approximation involves assigning each new timestamp the value of the existing data point closest to it in time, ensuring a synchronized temporal alignment between the low-cost sensor and the reference station data.

2. Outlier reduction beyond 3 standard deviations: Data points exceeding 3 standard deviations from the mean were deemed outliers and subsequently set as *null*. This step aids in eliminating data stemming from potentially malfunctioning instrumentation.

3. Interpolation utilizing kNN technique: Missing data points were imputed using the k-nearest neighbours (kNN) interpolation technique. This method leverages the entire feature vector to estimate *null* values, with the parameter k set to 5, ensuring a robust estimation of missing data points.

4. Left-side median cleaning smoothing technique: To further refine the dataset, a left-side median cleaning technique was applied for smoothing purposes. This involves using a window of preceding hours to the current data point. If the data point deviates above or below the median by a specified threshold, it is adjusted to the median value of the window, promoting data consistency and reducing noise.

5. Normalization: The final step involved normalizing the dataset across all features. This normalization process was carried out after splitting the dataset into training and test sets as required.

In order to ensure the integrity of the data separation and avoid overfitting issues, a specific approach was adopted. Instead of randomly splitting the data into training and test sets, each day's data was treated as a separate batch. This approach was chosen due to the temporal nature of the data, which exhibited daily variations in PM values. By grouping the data into daily batches, the risk of including overly similar data points in both sets was mitigated.

Two distinct strategies were employed to handle the data batches:

- Sequential approach: the first 75% of days from each month were allocated to the training set, while the remaining 25% were assigned to the test set.

- Random approach: 75% of random days from each month were allocated to the training set, the remaining to the test set.

Both strategies were evaluated to determine their effectiveness in training the model and their ability to accurately generalize to unseen data. This comparison provided insights into the optimal data separation method for the fuzzy system, informing its subsequent implementation and evaluation.

5.2.4 ANFIS Results

The overall performance of the different models tested is detailed first, highlighting their performance using metrics such as R^2 (Eq. 2.1) and RMSE (Eq. 2.2).

This is followed by an exploration of the pre-processing and dataset-splitting outcomes. Subsequently, the exploration shifts to the ANFIS fuzzy system configuration, encompassing aspects such as the type of membership functions utilized, the number of membership functions allocated to each feature, and the optimization algorithm employed for ANFIS training.

Overall methods results

The study results showcase the effectiveness of various methods in adjusting $PM_{2.5}$ data acquired from low-cost sensors, compared to the reference station data. These methods include linear regression, decision trees, random forests, support vector machines regression, the Fuzzy Adaptive Neuro-Fuzzy Inference System and an MLP neural network.

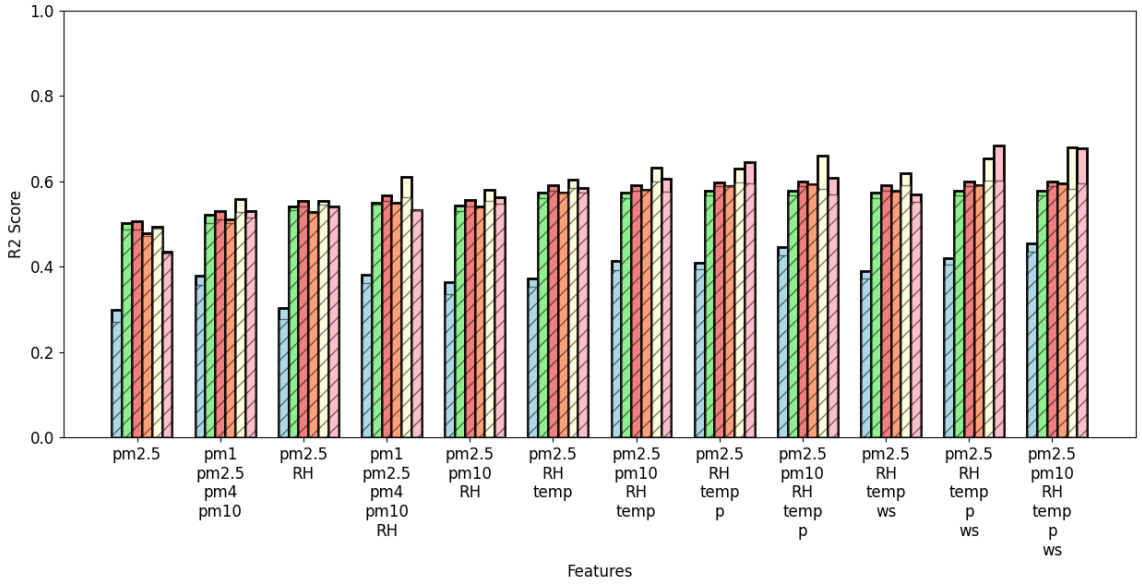
Figure 5.5 displays the R^2 and RMSE scores of the models trained on the preprocessed data, with both train and test values shown. The ANFIS system demonstrated promising results, particularly concerning the inclusion of $PM_{2.5}$, PM_{10}, RH, and temperature, with no notable performance improvements observed when additional features were included in the model. Additionally, a tendency for overfitting was observed in the fuzzy system with increased features.

In contrast, linear regression consistently lagged, never surpassing an R^2 of 0.5. Among the models, random forest performed relatively better apart from the fuzzy ANFIS system.

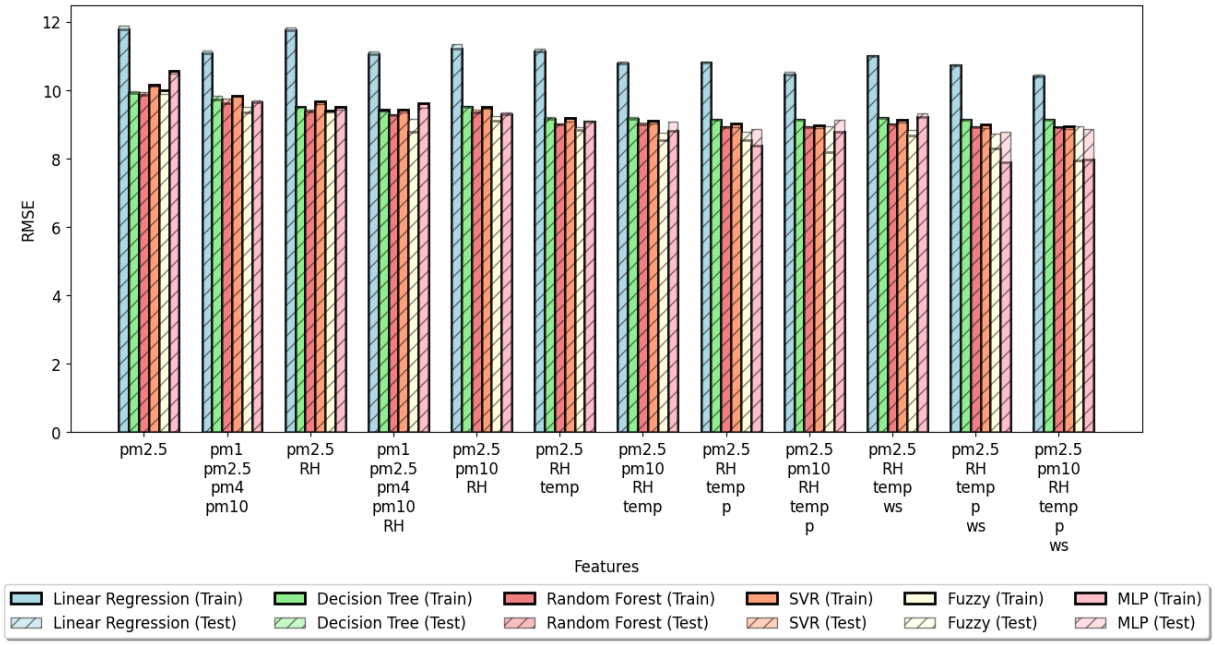
The MLP neural network exhibits comparable performance in terms of test results, surpassing the R^2 of the ANFIS method only when utilizing the full set of features. Furthermore, the RMSE tends to be slightly higher compared to ANFIS. Nevertheless, ANFIS was chosen for its interpretability and explainability, which can be advantageous in certain scenarios. Even if the NN had higher performance, there would still be cases where ANFIS is preferable due to its transparency and ease of understanding.

In general, compared to the studies proposed in Table 2.3, classical machine learning methods performed worse in this study, possibly due to the greater complexity of the data. The SPS30 sensor's significant hygroscopicity effect necessitates consideration when working with this data. Models trained on data without this effect may have an advantage.

The results obtained in our study when compared to the findings of [PGG16], exhibit consistency, with a decrease in performance observed when using fewer features. It is worth noting that the comparison is drawn between hourly data in our analysis and daily data in theirs. Additionally, the variation in performance may be influenced by factors such as the types of sensors utilized and the specific context of sensor deployment. Furthermore, it is important to highlight that while their work focuses on forecasting, ours is centered on data adjustment. These factors collectively contribute to a nuanced understanding of the comparative results and underscore the importance of contextual considerations in interpreting research findings. Nevertheless, the fact that the results are comparable despite these variations is a promising outcome, suggesting



(a) R^2



(b) RMSE

Figure 5.5: Comparative R^2 and RMSE scores for various models (including the Fuzzy Inference System) across different features ($PM1$, $\mathrm{PM}_{2.5}$, $PM4$, PM10, RH for relative humidity, $temp$ for temperature, p for pressure, and ws for wind speed).

the robustness and generalization of the ANFIS method across different contexts and methodologies.

Time Interval	Training Set				
	R^2	MAE	MSE	RMSE	R^2
2h	0.5590	6.7183	90.1888	9.4317	-0.0988
3h	0.5581	6.7391	90.4346	9.4403	0.4080
4h	0.5556	6.7721	90.9021	9.4618	0.5074
5h	0.5519	6.8135	91.7828	9.5046	0.5002
12h	0.5234	7.0669	97.2473	9.7971	0.4661

Table 5.3: Performance metrics on training and test sets changing one-sided median cleaning window

Preprocessing and dataset splitting

As discussed in Section 5.2.3, this study went under a meticulous pre-processing phase, which included the removal of 3 standard deviations to eliminate gross anomalies and the application of interpolation using k-nearest neighbors to fill in the missing data. Subsequently, a one-sided median cleaning procedure has been employed with a window size of 4 hours to smooth the data.

The choice of the window size was crucial, as it influenced the smoothing process. After experimenting with different window sizes (see Table 5.3), a window size of 4 hours was identified as yielding the highest R^2 score on the test set, indicating a superior fit to the data. This optimized window size was then used for all subsequent analyses.

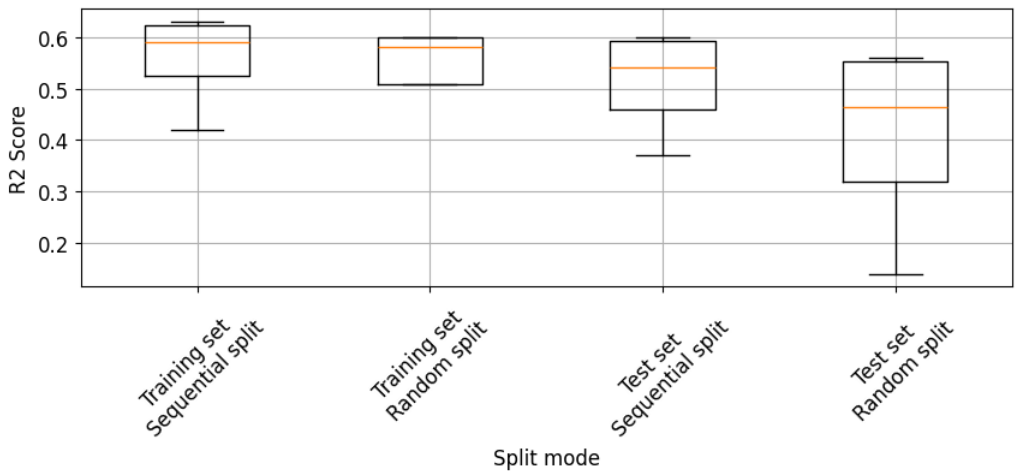


Figure 5.6: R^2 scores for training and test sets obtained by the fuzzy method using sequential and random splits.

As elaborated in Section 5.2.3, a critical consideration was the dataset-splitting methodology. Figure 5.6 illustrates that the sequential method appeared to yield better results compared to the random method, particularly concerning the test set. Therefore. the sequential method was selected as the preferred splitting approach for all subsequent analyses.

Exploration of the ANFIS configuration

Following the dataset pre-processing and splitting phases, the ANFIS was trained and tested using different configurations of the membership functions shape, and number. In addition, the optimization algorithm was tested between GridSearch and SubtractiveClustering.

Membership functions One of the key aspects of the ANFIS system is the type and number of membership functions used for each feature. Different types of membership functions, such as Gaussian, triangular, and trapezoidal, were experimented with to determine their impact on the adjustment process. Additionally, the number of membership functions for each feature was systematically varied to assess its impact on the performance of the ANFIS fuzzy system. The features considered were PM_1, $PM_{2.5}$, PM_4, PM_{10}, RH, temperature, pressure, and wind speed, which were kept in the same order throughout the trials. The number of membership functions tried were:

- Run 1: *[3, 3, 3, 3, 2, 3, 3, 3]*, where only RH was set to 2.
- Run 2: *[2, 6, 2, 2, 2, 3, 3, 3]*, with the number of PM features reduced apart from $PM_{2.5}$, which was set to 6.
- Run 3: *[2, 6, 2, 2, 2, 2, 3, 3]*, with the number of temperature-related features reduced.
- Run 4: *[2, 6, 2, 2, 2, 2, 2, 2]*, with the number of membership functions reduced to 2 for all features except $PM_{2.5}$.

Each configuration was tested, and the performance of the ANFIS was evaluated to determine the optimal number of membership functions for each feature, as shown in Figure 5.7.

In general, reducing the number of membership functions helped to avoid overfitting. Consequently, in the final ANFIS configuration, the Run 4 setup is retained.

The results, depicted in Figure 5.8a, reveal that the triangular membership function consistently yielded the most stable performance, with an R^2 score never dropping below 0.3. This robust performance is further illustrated in the zoomed-in view provided in Figure 5.8b. It is worth noting that [PGG16] also found triangular functions to be optimal for air quality data, where they restricted the number of membership functions to 3, thus reducing computational costs, suggesting a consistent pattern across studies.

Optimization algorithms During the training of ANFIS, optimization algorithms play a crucial role in efficiently handling the numerous combinations required for optimization. Rather than exhaustively attempting every combination, these algorithms aim to identify the optimal solution by intelligently sampling only a small subset of the entire solution space. In this study, two distinct optimization algorithms have been explored: GridSearch [PAB+16] and SubtractiveClustering [Che13].

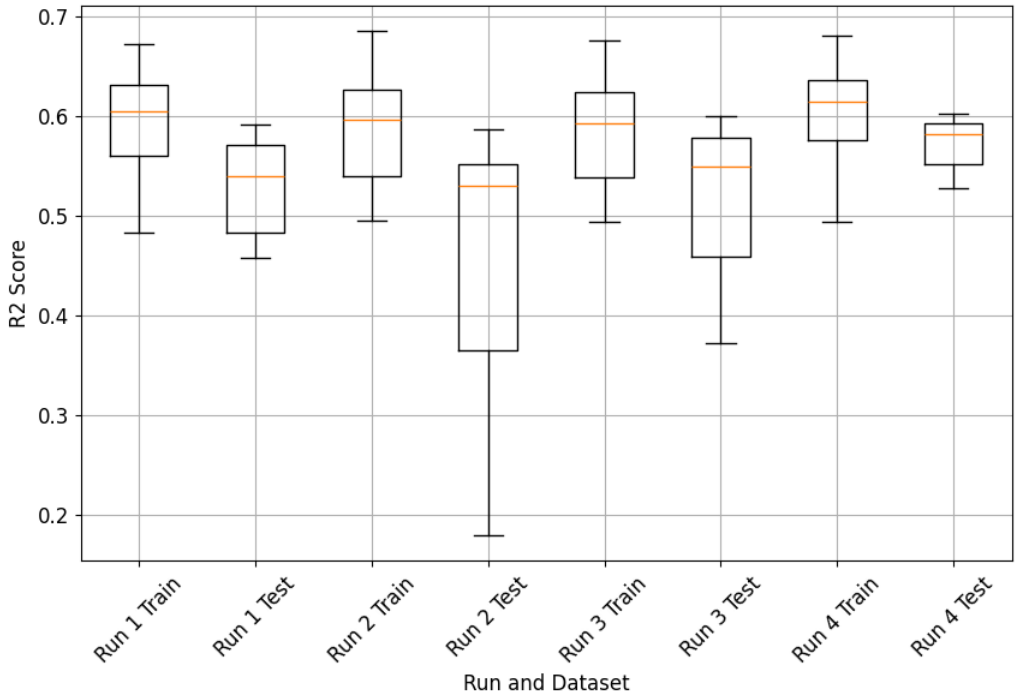


Figure 5.7: The obtained R^2 scores for both the training and test sets across varying numbers of membership functions.

While GridSearch rigorously explores the entire parameter space to find the best solution, SubtractiveClustering dynamically adjusts to the data distribution, providing a more flexible and potentially stable optimization approach.

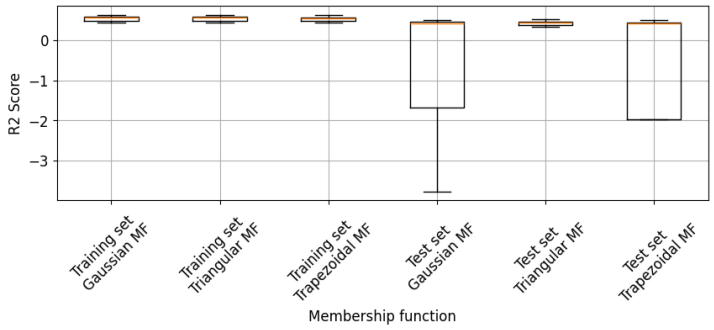
It was found that the GridSearch algorithm generally exhibited better performance in terms of optimizing the fuzzy system. However, when dealing with a larger number of selected features, the SubtractiveClustering algorithm demonstrated greater stability, see Figure 5.9.

Illustrative result

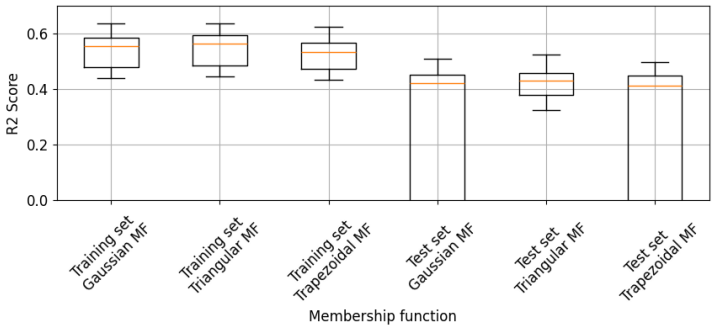
An illustrative example is presented in Figure 5.10 using three variables: $PM_{2.5}$, relative humidity and temperature, with membership functions. Triangular membership functions are chosen for each variable. It is noted that the number of membership functions retained is crucial to avoid overfitting. In this case, 6 membership functions are selected for $PM_{2.5}$ to accurately capture various behaviors across different PM ranges, while relative humidity and temperature each employ 2 membership functions.

Figure 5.11 illustrates time series data obtained from a low-cost sensor, showcasing $PM_{2.5}$ readings from the low-cost sensor itself, the reference station, and predictions generated by the ANFIS model for both the training and test sets. This comparison provides significant insights into the performance and accuracy of the ANFIS model in predicting $PM_{2.5}$ levels. Notably, both figures demonstrate that ANFIS can mitigate the hygroscopic effect and replicate the behavior of the reference station in both training and test sets. These results are satisfactory and provide insights into the effectiveness of ANFIS when applied to $PM_{2.5}$ hourly data in a high RH context.

Advantages and disadvantages of fuzzy logic Fuzzy logic modeling, while



(a) No y-axis limitation



(b) y-axis limited between 0 and 0.7

Figure 5.8: In (a), a boxplot displays the training and test set R^2 scores for different membership function types. No y-axis limit is imposed in this sub-figure. In (b), a similar boxplot is shown, but with the y-axis lower limit set to 0 to provide a clearer representation of R^2 scores above 0.

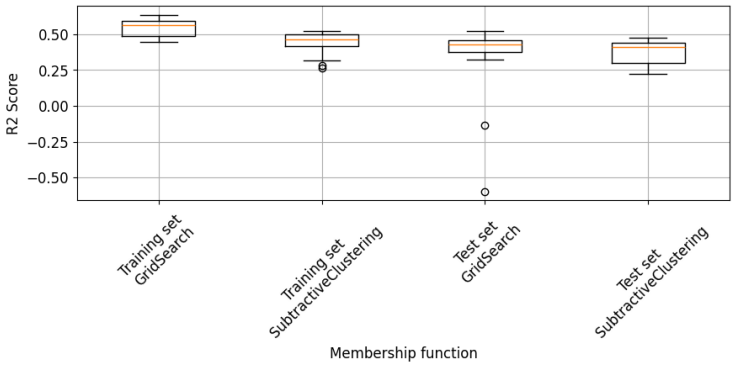


Figure 5.9: Training and test set R^2 scores obtained using the GridSearch and SubtractiveClustering optimization algorithms.

highly versatile and adaptable, presents limitations that are crucial to consider. One of the primary drawbacks is its inherent subjectivity; defining fuzzy sets and rules relies on human judgment, which can introduce bias and inconsistency. This subjectivity can lead to ambiguity and a lack of precision, particularly in complex systems where clear, objective data might be preferable. Additionally, as the number of rules and variables increases, fuzzy logic systems can become quite complex, making them difficult to

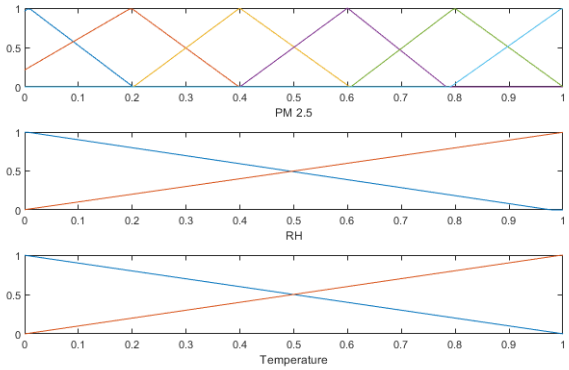
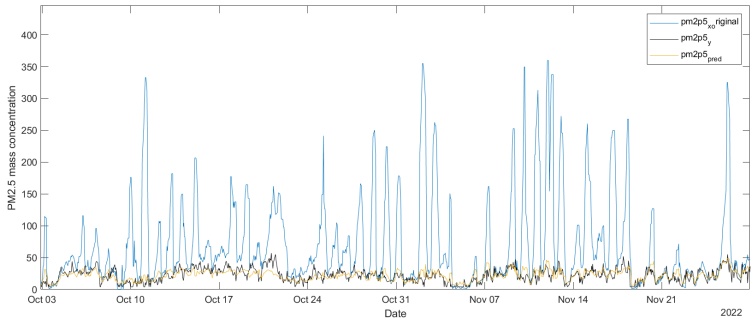


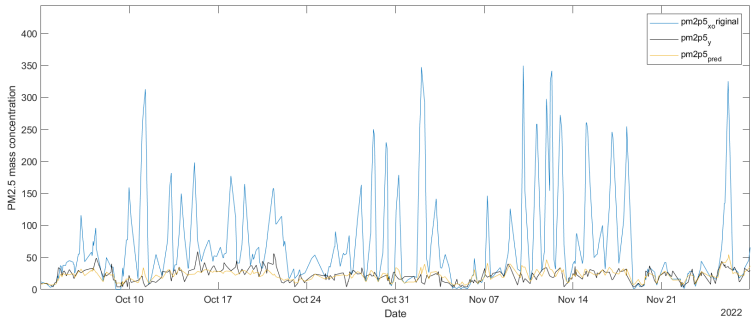
Figure 5.10: Membership functions for the variables $PM_{2.5}$, relative Humidity and temperature within the Fuzzy Inference System (FIS).

manage and optimize effectively.

On the other hand, fuzzy logic offers significant advantages that make it a powerful tool in many applications. Its ability to handle uncertainty and imprecise information allows it to mimic human reasoning more closely than traditional binary logic systems. This makes fuzzy logic particularly useful in situations requiring human-like decision-making, such as in control systems, robotics, and consumer electronics. The flexibility of fuzzy logic enables it to adapt to new data and changing conditions without requiring extensive reprogramming, saving time and resources in dynamic environments.



(a) Training set



(b) Test set

Figure 5.11: Comparison of time series data depicting $PM_{2.5}$ levels obtained from a low-cost sensor (blue line), a reference station (black line), and predictions generated by the Fuzzy Inference System (yellow line).

Furthermore, its interpretability and ease of integration with other AI techniques, such as neural networks, enhance its potential for creating robust, adaptive systems.

Applying the Adaptive Neuro-Fuzzy Inference System to adjust low-cost sensor PM concentrations highlights both the strengths and limitations of AI methods in environmental monitoring. Unlike deterministic approaches that offer rigorous proof of correctness, AI methods, including ANFIS, rely on test procedures involving random selection and repeated validation using various datasets. This introduces an inherent uncertainty, as the lack of formal proof means that the reliability of the results is heavily dependent on the quality and representativeness of the test data. However, fuzzy logic, central to ANFIS, provides a bridge between AI's complex computations and human interpretability by mimicking the way humans perceive and process information. This human-like reasoning capability allows for greater transparency and understanding of how decisions are made within the system.

Despite these advantages, relying on fuzzy logic can also be seen as a drawback, as its interpretability can lead to subjective conclusions that may not always align with objective accuracy. Moreover, the scalability of the ANFIS algorithm is a notable advantage, allowing for two primary approaches: repeating the entire optimization process or expanding the rule base with new, interesting cases specific to different locations or devices. This flexibility is beneficial for adapting to diverse environmental conditions and sensor characteristics, but it also necessitates careful management to avoid overfitting and maintain generalization. While ANFIS and fuzzy logic introduce complexity and potential for subjective bias, their adaptability and interpretability make them valuable tools for fine-tuning sensor data for improved monitoring of air quality.

5.2.5 Data and code availability

The data and MATLAB software utilized in this study can be accessed through [CPZ23b] and [CK24], respectively.

5.3 Machine Learning Techniques

This section presents the methodologies and results obtained during the research, focusing on key aspects critical to the study.

The approach was applied to a set of multiple datasets, making it a standardized method that enabled the processing of datasets created by different authors and with varying measurements, all following a predefined data flow. For the different analyses, a subgroup of these datasets was selected based on their availability at the time of the research and the relevance of the datasets for the study.

This Section includes the following analysis:

- **Splitting Strategies:** A detailed exploration of methods for partitioning the dataset into training and testing sets, with careful consideration of temporal dependencies and data structure.

- **Anomaly Detection Methods:** The evaluation of techniques for identifying and handling anomalies in the dataset, ensuring cleaner and more reliable data for model training.
- **Model Performance Analysis:** An assessment of techniques for adjusting the air quality data, which have been cleaned of anomalies and split into training and test sets.
- **LightGBM:** A thorough evaluation of LightGBM, identified as the most effective model for this data. This section also explores feature selection and skewness transformation applied to the variables from the Turin and UK datasets.
- **Multilayer Perceptron:** The initial phase of testing a customized neural network, an MLP, on the air quality data, specifically applied to the Turin dataset.

The subsequent subsections delve into these topics, presenting the strategies, methodologies, and outcomes in detail, supported by quantitative results and visual aids where appropriate.

5.3.1 Model Performance Analysis

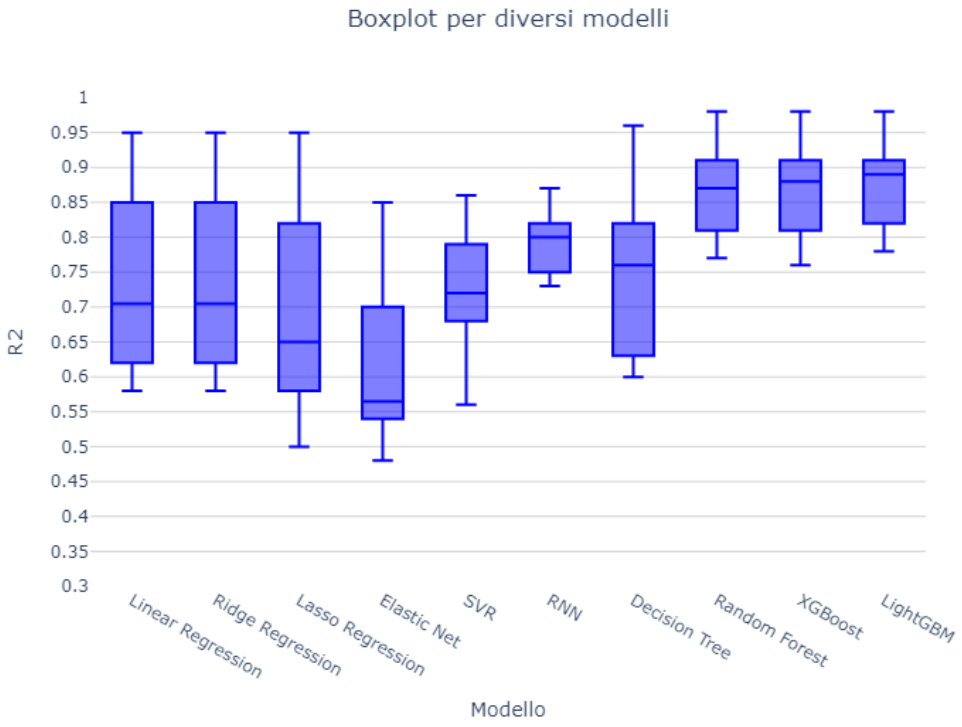


Figure 5.12: Boxplot summarizing model performances across datasets.

This section presents the results obtained by applying various machine learning models, to each dataset. The reported values represent the R^2 scores achieved on the test set, providing a clear measure of model performance, and are summarized in Figure 5.12,

Model	Aosta	Badajoz	Bangalore	Calgary	Delhi	Hamirpur	*Lima IQAir*	*Lima AIRBEAM*	*UK PMS003*	*UK SPS30*
LinR	0.71	0.59	0.86	0.81	0.85	0.95	0.62	0.58	0.66	0.70
RR	0.71	0.59	0.86	0.81	0.85	0.95	0.62	0.58	0.66	0.70
LR	0.64	0.50	0.82	0.69	0.84	0.95	0.58	0.58	0.62	0.66
EN	0.56	0.48	0.70	0.61	0.71	0.85	0.54	0.55	0.54	0.57
SVR	0.79	0.56	0.82	0.86	0.72	0.68	0.65	0.73	0.69	0.72
DT	0.80	0.72	0.78	0.96	0.82	0.93	0.63	0.60	0.74	0.61
RF	0.89	0.85	0.90	0.98	0.91	0.96	0.77	0.78	0.81	0.81
XGBoost	0.90	0.86	0.91	0.98	0.91	0.96	0.81	0.76	0.80	0.82
LightGBM	0.90	0.88	0.91	0.98	0.91	0.96	0.81	0.78	0.83	0.82
RNN	0.75	0.73	0.79	0.73	0.82	0.87	0.80	0.65	0.82	0.81

Table 5.4: R^2 scores for each model across datasets.

From Table 5.4, the following insights can be drawn:

- **Tree-based models** (RF, XGBoost, LightGBM) consistently demonstrate the highest performance across most datasets. Their ability to model non-linear relationships and capture complex patterns in the data makes them well-suited for this application. For instance, in datasets like Calgary and Delhi, LightGBM achieves an R^2 of 0.98 and 0.91, respectively.
- **Linear regression models** (LinR, RR, LR, EN) tend to under-perform compared to non-linear methods. In datasets such as Badajoz and Lima IQAir, the best R^2 scores from linear models are 0.59 and 0.62, significantly lower than LightGBM's scores of 0.88 and 0.81. This suggests that the datasets contain complex relationships that linear models fail to capture effectively.
- **Geographical and temporal factors** heavily influence the results. For instance, some datasets (e.g., Calgary) achieve R^2 scores as high as 0.98, while others (e.g., Lima IQAir) only reach a maximum of 0.81. These disparities are attributed to differences in data collection periods (ranging from several months to over a year) and the geographic location of sensors, where meteorological factors play a significant role in sensor performance.

5.3.2 LightGBM

LightGBM (Light Gradient Boosting Machine) is a powerful and efficient gradient-boosting framework developed by Microsoft researchers in 2017 [KMF+17]. It is designed to be efficient and scalable, making it particularly well-suited for large datasets and high-dimensional feature spaces. It utilizes the boosting framework, building an ensemble of weak learners (decision trees) sequentially to minimize the overall prediction error, thus ultimately combining multiple weak models to create a strong predictive model. Unlike depth-first tree growth in traditional gradient boosting frameworks like XGBoost [CG16], LightGBM adopts a leaf-wise tree growth strategy which chooses the leaf with the maximum delta loss to grow, which can lead to faster convergence and reduced computational cost. The trees are then used as usual, choosing the path that maximizes the information gain which is evaluated via the variance score of each node. Other characteristics are that it includes a feature selection process by itself and the loss used usually is the Mean Squared Error (MSE) Loss, Eq. 5.1.

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

The objective of this section is to demonstrate the effectiveness of the LightGBM algorithm in accurately forecasting $PM_{2.5}$ levels using cost-effective sensors and various environmental parameters. Additionally, the study explores the applicability of the method across different locations, examining both homogeneous and heterogeneous approaches. The training process relies on $PM_{2.5}$ measurements from reference stations, enabling the resultant model to predict and adjust measurement readings effectively.

The datasets considered in this study are created by a collection of measurements captured in two different geographical areas, both by using SPS30 low-cost (LC) sensors as input and the co-located legal stations as reference:

- Turin (Italy): LC sensors capturing records with 15-minute frequency, reference station (RS) with hourly frequency based on Arpa weather stations [CPZ23b];
- Southampton (UK): LC sensors capturing records with 2 minutes frequency, RS sensors with hourly frequency based on Fidas200s weather stations [Bul22b].

The data was obtained through individual sensor measurements, which were then used to construct the raw datasets for both Turin and Southampton. Subsequently, a thorough analysis of the LC and RS data was conducted to create a dataset linking each reference record with a low-cost measurement. To achieve this, the input datasets were resampled to match the hourly frequency of the reference datasets.
Initially, the resampling technique employed was averaging all the LC data over the RS hourly record. However, due to significant variations in the data within an hour, it was decided to assign the closest available LC record to each RS record instead. After this process, the raw datasets for both Turin and Southampton were created, and pre-processing techniques [CP24] were applied to uniformly adjust the data, preparing them for the training step. In the performance evaluation, just the preprocessed dataset was considered for comparison.
Incorporating contextual features based on time into the feature extraction process has allowed for a more thorough understanding of the data. This approach not only captures the original features but also encodes information about the time axis, enabling a fine and accurate representation of patterns that unfold over time. Ultimately, this results in more insightful and precise outcomes.

The final set of features included in the datasets comprises " PM_1," "pm2p5," "pm2p5 RF target," " PM_4," " PM_{10}," "wind speed," "pressure," "temperature," "relative humidity," "month," "day of the week," and "hour." The correlation matrix is depicted in Figure 5.13.

Feature Selection

In this phase, the most representative features of the problem were extracted. Since there is a relatively low number of features, to begin with, the selection was done using a simple correlation method where the resulting features are the ones which correlate with the target variable higher than a chosen threshold.

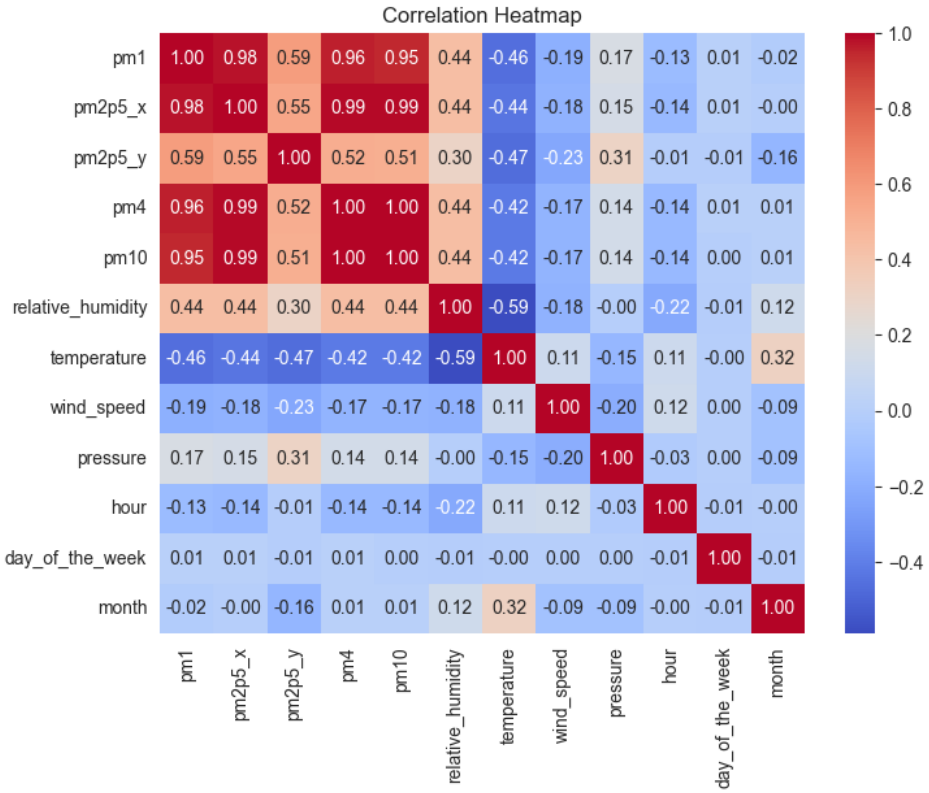


Figure 5.13: Correlation matrix of all the features.

Feature	**Absolute Correlation**
PM_1	0.588402
pm2p5	0.546849
PM_4	0.521004
PM_{10}	0.511054
temperature	-0.473645
pressure	0.306547
relative humidity	0.297370
wind speed	-0.227094
month	-0.159161
day of the week	-0.014929
hour	-0.006393

Table 5.5: Correlation of features with target variable

As evident from Table 5.5, both "day of the week" and "month" exhibit weak correlations with the target variable.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{5.2}$$

The Pearson Correlation Coefficient (r) was utilized to assess these correlations, as indicated by Equation 5.2. Consequently, even if a negative correlation with the target variable is obtained using this formula, it remains valuable as it signifies an inverse correlation, akin to inverse proportionality. Ultimately, the features selected by this method are those for which $|r| > 0.1$. where a coefficient below 0.1 indicates a negligible correlation, while higher values correspond to weak, moderate, or strong correlation. It is worth noting that this method only captures linear relationships, which can be limiting when evaluating feature relevance in complex, non-linear datasets. However, in this specific case, only two features were excluded based on this criterion: *day of the week* and *hour*, both of which exhibited negligible linear correlation with the target variable.

Skewness Transformation

Skewness is a statistical measure that describes the asymmetry of the probability distribution of a real-valued random variable. In simpler terms, it measures the degree and direction of skew (departure from horizontal symmetry) in a dataset. A skewness value of 0 indicates a perfectly symmetrical distribution, see Eq. 5.3. Positive skewness indicates a longer right tail, while negative skewness indicates a longer left tail.

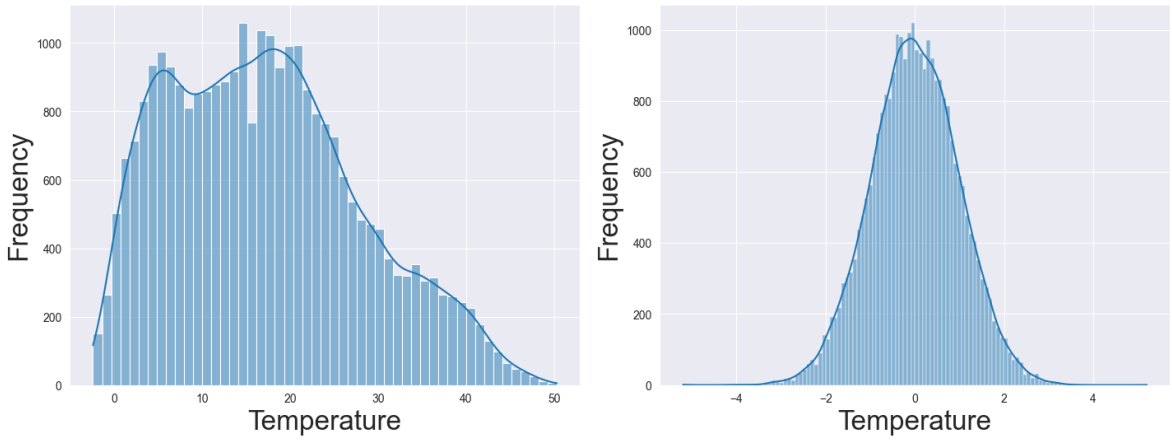
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{5.3}$$

When dealing with regression problems, addressing highly skewed variables is crucial as they can impact the model's fit. This is primarily due to the assumption of linearity made by most regression algorithms, which presupposes linear relationships between features. By applying transformations such as power or logarithmic functions, this effect can be mitigated, especially considering that the chosen model inherently possesses nonlinear properties.
Additionally, highly skewed predictor variables can make the model overly sensitive to extremely high values, potentially resulting in a poor fit for the majority of the data.
To tackle this issue, a skewness transformation was incorporated into the pipeline. This transformation applies a predefined set of transformations to each feature in order to reduce its skewness. The set of transformations includes:

- Logarithm: $f_t = \log(f)$;
- Exponential: $f_t = e^f$;
- Square Root: $f_t = \sqrt{f}$;
- Quantile: $f_t = F^{-1}(f)$;

Feature	Best Transformation
PM_1	Log Transformation
pm2p5_x	Log Transformation
pm2p5_y	QuantileTransformer
PM_4	Log Transformation
PM_{10}	Log Transformation
relative_humidity	QuantileTransformer
temperature	QuantileTransformer
wind_speed	QuantileTransformer
pressure	QuantileTransformer
month	QuantileTransformer

Table 5.6: Best transformation for each feature



(a) Before (b) After

Figure 5.14: Distribution comparison with skewness transformer

For each feature in the dataset, all transformations are tested, and the one selected is the transformation that minimizes the feature's skewness to 0.

An example of feature skewness transformation is depicted in Figure 5.14, illustrating the distribution of temperature data. The second figure demonstrates the attainment of a Gaussian distribution after applying the Quantile Transformer. Table 5.6 shows the best transformation found for each feature.

LightGBM Results

By applying all the aforementioned techniques, the final pipeline is created and then trained on the preprocessed Turin dataset with the RDS split method.

Results here are presented in terms of R^2 (see Eq. 2.1), RMSE (see Eq. 2.2), MAE (see Eq. 4.6), and MDAE:

$$\text{MDAE} = \text{median}(|y_i - \hat{y}_i|) \tag{5.4}$$

where y_i are the observed values and $\hat{y}_i$ are the predicted values. The Median Absolute Error (MDAE) represents the median of the absolute errors, offering a robust measure of error that is less sensitive to outliers compared to MAE.

Metric	**Turin Train**	**Turin Test**
MAE	0.3023	0.3369
RMSE	0.1467	0.1846
MDAE	0.2508	0.2775
R^2 Score	0.7435	0.6735

Table 5.7: Performance metrics obtained from training the LightGBM model on the Turin preprocessed dataset.

As evident from Table 5.12, the selected pipeline demonstrates strong performance on both the Turin training and test sets.

In Figure 5.15, the feature importance ranking for the constructed model is depicted. Observing the significance of meteorological features for the model's predictions is notable.

Metric	**UK Dataset**
MAE	6.4039
RMSE	82.9644
MDAE	4.5478
R^2 Score	-0.9130

Table 5.8: Southampton (UK) performance metrics obtained from training the LightGBM model on the Turin preprocessed dataset.

The results presented in Table 5.8 highlight the performance achieved when applying the model to a distinct dataset, the Southampton dataset. Here, it is evident that the model's prediction of outcomes is unsatisfactory. This suggests that while the

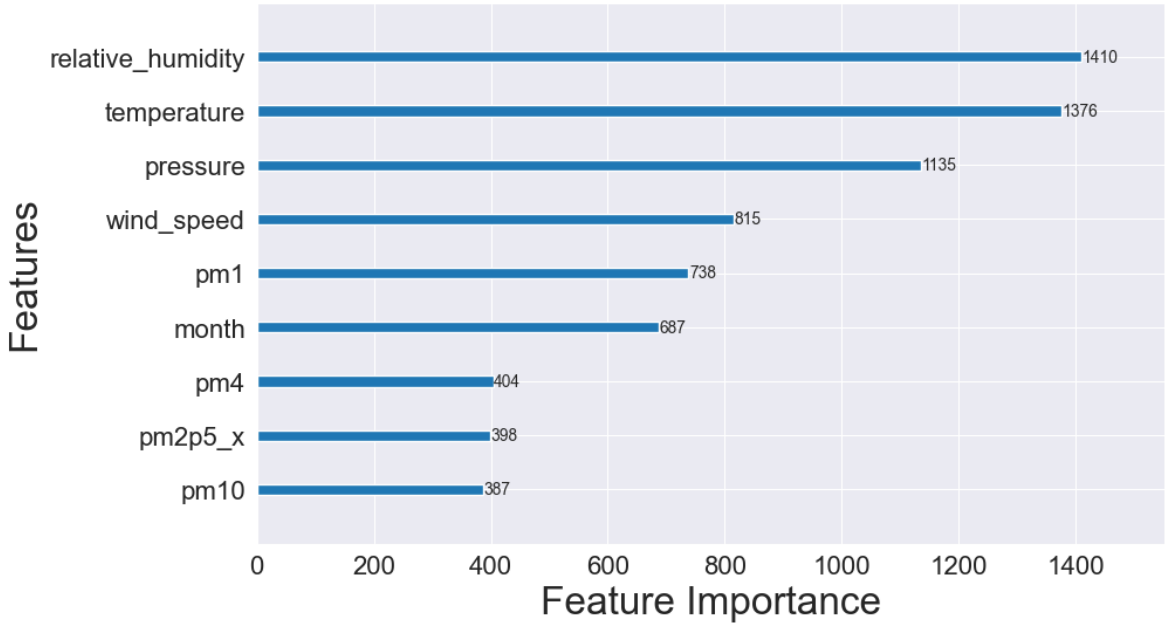


Figure 5.15: Feature importance ranking

model reliably predicts where results should fall within their value range, it struggles to accurately forecast how they are distributed over time. Consequently, it can be inferred that the geographic location under study exerts a significant influence on PM forecasting.
To tailor forecasting models to specific geographic zones, it is essential to incorporate the studied area as a feature or consider creating independent models for each area under consideration. The challenge faced by the model in this scenario may stem from several factors, including the distinct nature of the datasets, their unique contextual considerations, and the temporal misalignment despite both datasets covering an entire year. Furthermore, the placement of the SPS30 sensors within different devices for Southampton and Turin introduces significant variability in the collected data due to positional and rotational differences.

To delve deeper into this issue, an additional test was performed by merging records from both the Southampton and Turin datasets. This merged dataset served as the comprehensive training and testing dataset with the RDS split and was subsequently processed through the aforementioned pipeline. The objective of this test was to develop a model capable of addressing both challenges simultaneously, by incorporating data from both geographical areas concurrently.

Metric	**MAE**	**RMSE**	**MdAE**	**R^2**
Merged Dataset	3.52	5.78	2.08	0.78

Table 5.9: Performance Metrics

As we can see from the results in Table 5.9, this test provided surprisingly good results all across the board, with great values both in the distance metrics and in R^2.

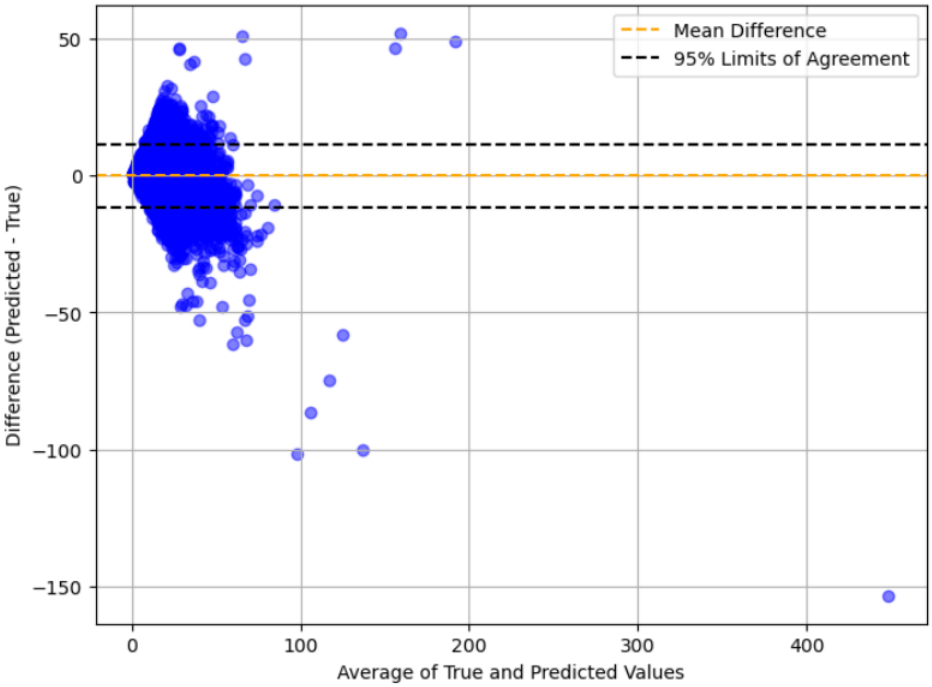


Figure 5.16: Bland–Altman plot for the merged dataset

However, upon analyzing the Bland-Altman plot in Figure 5.16, it becomes apparent that there exist relatively high absolute differences between the predicted and actual values, particularly within the first range of values where the majority of records are concentrated. This discrepancy implies that while the predictions generally fall within the desired range considering the wide scope of values (over 87k records), the model's precision in predicting exact values is suboptimal.
One possible explanation for this phenomenon is the variability of PM values across different geographical areas attributable to diverse environmental conditions. Without incorporating a feature that delineates between the two areas, the model treats the PM range as a unified domain for both datasets, endeavoring to predict within that domain without differentiation due to the absence of pertinent information. These findings underscore the original hypothesis, emphasizing the necessity to either incorporate features that encapsulate environmental conditions or devise distinct models for different areas, as the available features alone are insufficient to infer such information.

To conclude this discussion and affirm the thesis, a final test was conducted by creating a new independent model using only the Southampton data.

Metric	**MAE**	**RMSE**	**MdAE**	$\mathbf{R^2}$
Southampton Model	1.73	3.04	1.01	0.88

Table 5.10: Performance metrics for Southampton model

The latest results presented in Table 5.10 serve to reinforce the thesis that tailoring a model to a specific geographical area yields superior outcomes in accurately capturing

and predicting PM levels using machine learning techniques. The model trained exclusively on Southampton data demonstrates excellent performance across all metrics utilized, consolidating the argument for geographic specialization in PM forecasting models.

5.3.3 Multilayer Perceptron

In this part of the study, I utilized a Multilayer Perceptron architecture, a type of feed-forward neural network, to model complex relationships in the data.

An MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of interconnected nodes, or neurons, where each neuron in a layer receives input from the neurons in the previous layer and passes its output to the neurons in the subsequent layer. The information flow is unidirectional, from input to output, with no cycles or feedback loops, characterizing it as a feed-forward network.

Each neuron performs a weighted sum of its inputs, adds a bias term, and passes the result through a non-linear activation function, such as the Rectified Linear Unit (ReLU), sigmoid, or hyperbolic tangent (tanh). This non-linearity enables the network to approximate complex functions and learn intricate patterns in the data.

A critical aspect of applying MLPs is determining the optimal network configuration, including the number of hidden layers, the number of neurons per layer, the choice of activation functions, and regularization strategies such as dropout or weight decay. These choices directly influence the model's ability to generalize from training data while avoiding overfitting and managing computational complexity. In this study, a tailored MLP architecture was developed and optimized to balance high predictive performance with efficiency, ensuring reliable modeling of the non-linear dependencies within the air quality dataset.

The MLP is well-suited for capturing non-linear patterns and offers significant flexibility through the customization of its architecture and hyperparameters. A critical aspect of this approach was determining the optimal network configuration that achieves a balance between high performance, avoiding overfitting, and managing computational complexity.

A complete pipeline was constructed that facilitated the training of each network with a unique combination of hyperparameters. Specifically, the focus is on two key hyperparameters: the number of consecutive old records used to create the new dataset, the *loopback* number, and the number of neurons per layer. These parameters collectively defined the specific architecture of the model being trained. Furthermore, experiments conducted involve three different batch sizes to assess their impact during training.

The values considered for each hyperparameter were as follows: the number of neurons per layer ranged from $\{300, 500, 700\}$ for the first, second, and third layers respectively; the number of consecutive records varied among $[6, 12, 20]$; and batch sizes encompassed $[64, 256, 512]$. These hyperparameters were evaluated across five different network architectures, resulting in a total of 135 possible configurations, each requiring independent training and evaluation. The Rectified Linear Unit (ReLU) activation function was employed in the hidden layers due to its computational efficiency and ability to mitigate the vanishing gradient problem. Each model was trained for 200 epochs. This comprehensive approach allowed us to gain a deeper understanding of the

network's performance under various conditions.

MLP Results

Model	Neurons	64:6	256:6	512:6	64:12	256:12	512:12	64:20	256:20	512:20
AirMLP6	**300**	0.801	0.772	0.752	0.827	0.793	0.758	0.859	0.812	0.788
	500	0.833	0.801	0.769	0.863	0.830	0.808	0.884	0.841	0.821
	700	0.862	0.821	0.780	0.878	0.853	0.819	**0.901**	0.858	0.849
AirMLP7	**300**	0.811	0.780	0.760	0.847	0.805	0.778	0.865	0.823	0.801
	500	0.844	0.814	0.770	0.883	0.838	0.815	0.888	0.859	0.826
	700	0.857	0.826	0.813	0.886	0.855	0.841	**0.904**	0.871	0.848
AirMLP8	**300**	0.801	0.775	0.751	0.848	0.805	0.773	0.878	0.830	0.812
	500	0.840	0.807	0.792	0.884	0.848	0.809	0.901	0.862	0.842
	700	0.885	0.848	0.807	0.896	0.876	0.826	**0.905**	0.883	0.859
AirMLP7h	**300**	0.797	0.768	0.738	0.836	0.876	0.826	0.856	0.825	0.805
	500	0.833	0.789	0.778	0.872	0.833	0.808	0.887	0.853	0.812
	700	0.848	0.823	0.796	0.883	0.855	0.829	**0.910**	0.860	0.844
AirMLP8h	**300**	0.808	0.760	0.769	0.853	0.819	0.785	0.876	0.836	0.807
	500	0.851	0.822	0.803	0.873	0.836	0.808	0.887	0.859	0.841
	700	0.867	0.832	0.810	0.887	0.867	0.840	**0.916**	0.867	0.848

Table 5.11: R^2 score on models trained differently. The meaning of the first row is [batch size]:[num records]

In Table 5.11, there are shown the R^2 scores corresponding to each architecture evaluated on the test set, alongside different sets of hyperparameters. Each column in the table represents the results obtained for various combinations of batch size and consecutive records, with the column name structured as *"batch size : loopback number"*.

As evident from the results in Table 5.11, the optimal configuration for each model consistently features a higher number of neurons per layer and a batch size of 64. However, in our evaluation, it is crucial to take various factors into account.

To gain a more comprehensive understanding, we should investigate the effects of further increasing the number of neurons per layer and assess whether this leads to the onset of overfitting. This analysis will help us strike a balance between model complexity and performance, ensuring that we do not compromise generalization capabilities while seeking optimal performance.

Analyzing the loss graph provides valuable insights. Firstly, it is apparent that overfitting is not occurring in the initial training attempts, see Figures 5.17, 5.18 and 5.19, which suggests that the increase in network size should be considered without immediately worrying about overfitting.

Secondly, the loss graph indicates two important observations. When employing a larger batch size, the network tends to converge more quickly and smoothly but at a higher loss value, ultimately resulting in poorer performance. Conversely, when using a smaller batch size, the network converges to a lower loss value, even with an identical number of neurons. This underscores the significance of batch size in model convergence

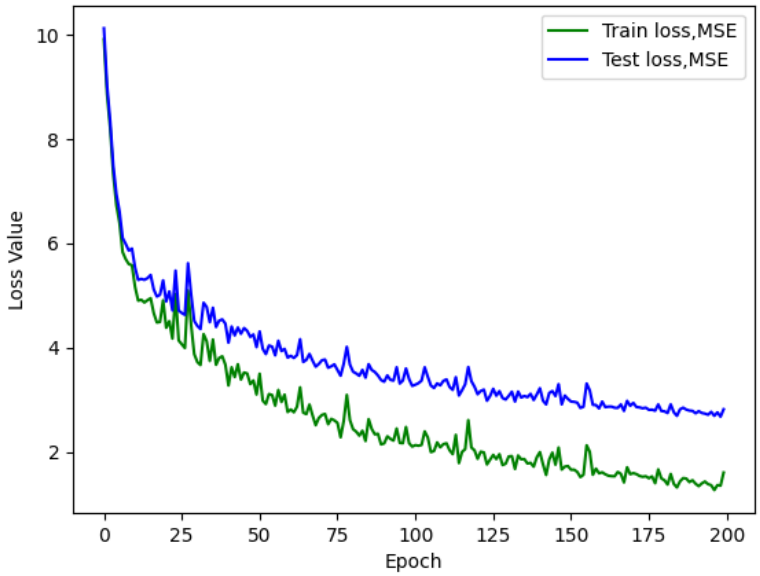


Figure 5.17: AirMLP8-700 loss with a batch size of 64, test L1 value of 2.732.

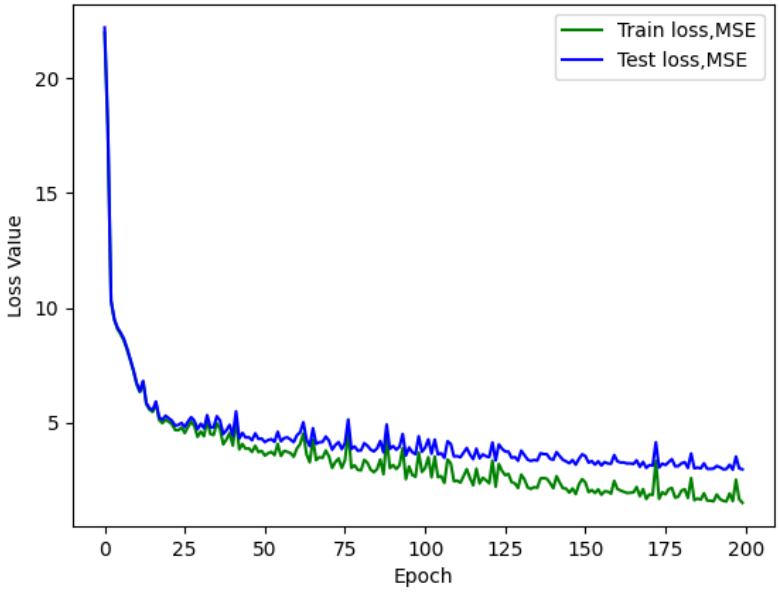


Figure 5.18: AirMLP8-700 loss with a batch size of 256, test L1 value of 2.976.

and its subsequent impact on performance. These insights can guide our decisions when fine-tuning the network architecture for optimal results.

The fact that the loss curves are still descending suggests that there is room for improvement in the model's performance. Instead of selecting a fixed number of epochs for training as in this study, considering the implementation of an early stopping technique could prove to be highly beneficial.

After our initial evaluation, which identified the best set of hyperparameters for each model as a batch size of 64, a *loopback* of 20 consecutive records (including the past 19 observations in addition to the current one to be corrected), and 700 neurons per layer, it is valuable to explore the effects of further increasing the number of neurons per layer.

Model	**Neurons**	$\mathbf{R^2}$
AirMLP6	900	0.901
	1100	0.912
	1500	0.926
AirMLP7	900	0.919
	1100	0.926
	1500	**0.932**
AirMLP8	900	0.917
	1100	0.928
	1500	0.925
AirMLP7h	900	0.915
	1100	0.921
	1500	0.927
AirMLP8h	900	0.917
	1100	0.921
	1500	0.923

Table 5.12: Performance with a batch size of 64, and a *loopback* of 20, changing the number of neurons per layer.

To achieve this, additional training was conducted for all five models, with variations of 900, 1100, and 1500 neurons per layer, while maintaining 20 consecutive records and a batch size of 64, see Table 5.12.

The outcomes depicted in Figure 5.20 are rather remarkable, as they reveal not only enhancements in terms of the R^2 metrics but also significant improvements in the loss reduction process during training. The loss curves now exhibit a smoother descent, and they converge to lower values compared to the previous configurations. These findings strongly indicate that augmenting the number of neurons per layer has exerted a positive influence on both the model's predictive performance and its training stability.

Upon inspecting the figures, it becomes evident that there is still no indication of overfitting throughout the training process. Notably, the AirMLP7-1500 model performs well, potentially serving as a solid foundation for further exploration.

While it is tempting to increase the network's dimensions even further, it is im-

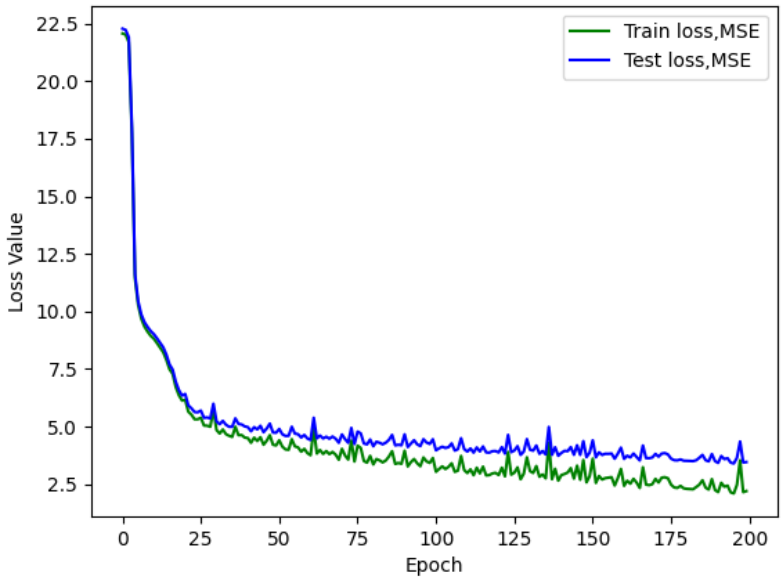


Figure 5.19: AirMLP8-700 loss with a batch size of 512, test L1 value of 3.461.

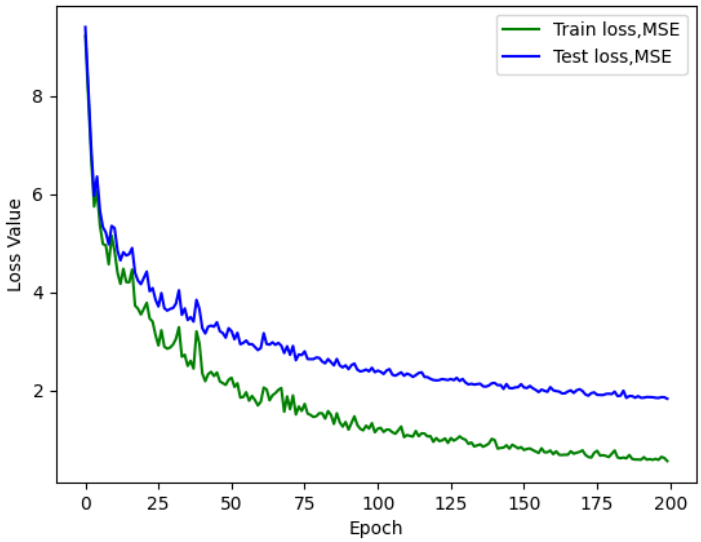


Figure 5.20: AirMLP7-1500 loss. Batch size of 64 and a *loopback* of 20.

perative to carefully weigh the trade-off between potential performance enhancements, which may be modest, and the accompanying escalation in computational requirements.

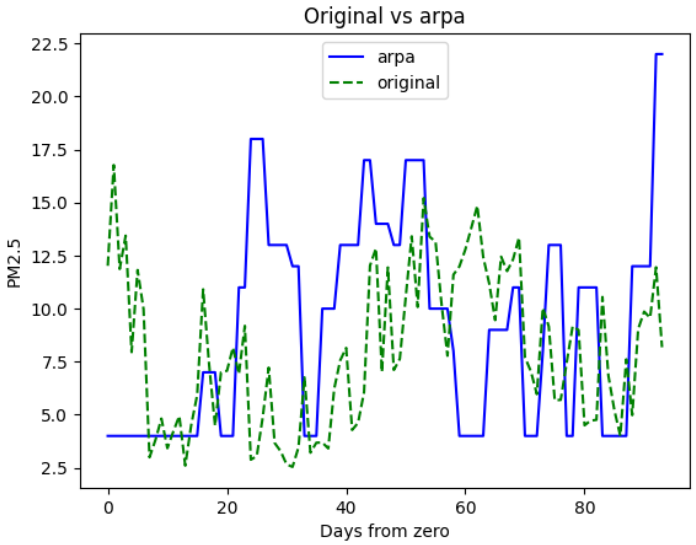


Figure 5.21: Original data vs Arpa ground truth data.

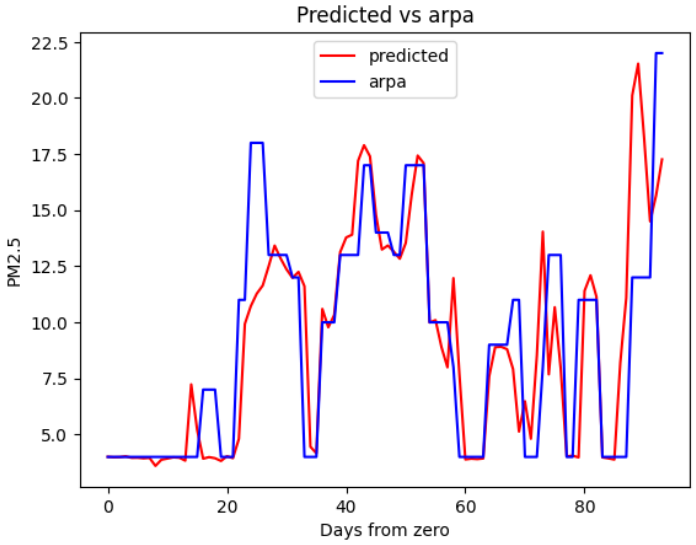


Figure 5.22: AirMLP7-1500's predicted output vs Arpa ground truth data.

Now, let's consider a small dataset consisting of data from a few days, which is not part of the training or test set. We'll compare the original PM 2.5 values from this dataset with the ground truth (Arpa) (Figure 5.21) and also compare the PM 2.5 values produced by the AirMLP7-1500 model concerning the ground truth (Figure 5.22). This comparison will provide insights into how well the model generalizes to new, unseen data and its performance in a real-world scenario outside the training and test sets.

Indeed, the close alignment between the predicted output of the AirMLP7-1500 model and the ground truth, especially when compared to the original data, is a note-

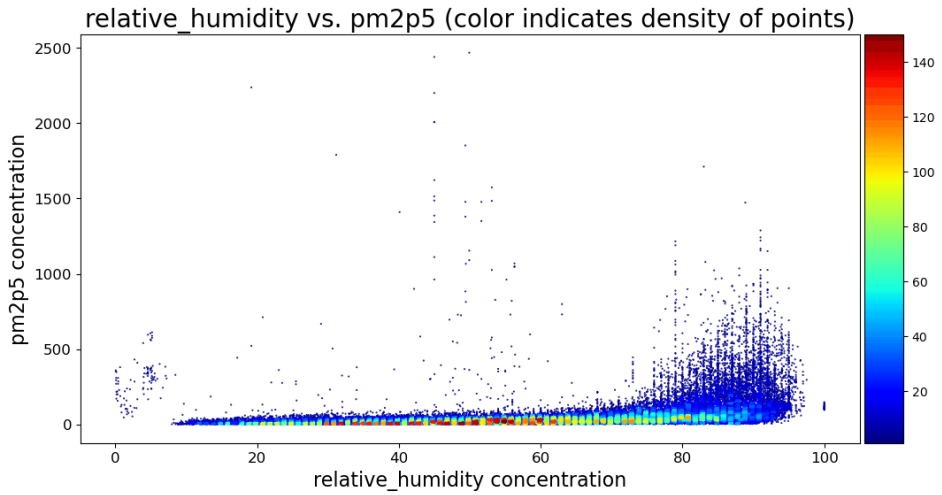


Figure 5.23: Increasing of PM 2.5 mass concentration detected by the LC sensor due to a higher RH.

worthy observation. Importantly, these results were obtained using a small dataset of unseen data.

This outcome serves as an indicator of the model's generalization capabilities and provides further evidence that the network does not suffer from overfitting. It highlights the robustness of the model and its ability to provide accurate PM 2.5 estimates, reinforcing its potential for real-world applications in air quality monitoring.

Regarding the influence of relative humidity, our study emphasizes the significant impact of humidity on the accuracy of low-cost PM 2.5 measurements. As illustrated in Figure 5.23, when relative humidity surpasses a specific threshold, low-cost sensors tend to overestimate PM 2.5 levels. This hygroscopic effect can result in misleading air quality assessments, thus reducing the utility of such data in pollution-related decisions.

To mitigate the adverse impact of humidity and other factors on PM 2.5 measurements, I opted for MLP architectures. The AirMLP7-1500 model, in particular, exhibited impressive performance, achieving an R^2 score of 0.932. This underscores again the potential for AI-based correction methods to enhance the accuracy of low-cost sensor data.

It is important to acknowledge that the proposed correction models are tailored to a specific location and a particular set of low-cost sensors. As previously highlighted, applying these models to different devices or locations may yield suboptimal results. Therefore, the research serves as a blueprint for training similar models tailored to specific contexts, rather than offering a one-size-fits-all solution.

Numerous promising avenues for future research emerge from this study. One immediate direction is to investigate the impact of further increasing the model's complexity, with a focus on balancing improved performance with computational resources. Regularization techniques like dropout could also be explored to enhance model robustness.

An interesting observation regarding the specular pattern between training loss and test loss is shown in the loss figures. This phenomenon could likely be attributed to the initial shuffling of data, which may not have adequately accounted for the presence of very similar data points in both the training and test sets. Further investigation and refinement of the data splitting and shuffling procedures could potentially mitigate this

issue and provide a more accurate representation of the model's performance on unseen data.

Another promising avenue for future research involves considering seasonality when training PM 2.5 correction models. Seasonal variations in air quality are a well-known phenomenon [BBKS20, TZW+21]. By training separate models tailored to different seasons, the potential exists to capture and correct seasonal variations in PM 2.5 measurements. Moreover, incorporating temporal information into the dataset is a strategy that can further enhance the network's ability to differentiate between diverse environmental contexts. This could involve including features such as day of the week, time of day, or even specific holidays or events that might impact air quality. By accounting for both seasonality and temporal dynamics, the models can become more adaptive and capable of providing accurate corrections across various environmental scenarios. This approach aligns with the idea of developing context-aware models that tailor their corrections based on the prevailing conditions [DTS+23], ultimately advancing the effectiveness of low-cost sensor data in air quality monitoring.

Beyond correction, the models developed here could be adapted for anomaly detection, helping to identify unusual PM 2.5 readings that may indicate pollution events or sensor malfunctions. Additionally, transitioning from correction to prediction could enable the forecasting of future PM 2.5 levels based on atmospheric conditions, contributing to proactive pollution management at the installation site.

The final thought is that extending the models' evaluation to new, unseen data from diverse geographical locations is essential to assess their generalization capabilities [NKGC21]. This would help determine the models' adaptability to varying environmental conditions. Incorporating data from low-cost laser scattering sensors manufactured by different companies is also an important consideration. Different sensor models may exhibit unique characteristics and behaviors. Testing the models with data from various sensor manufacturers can help validate their robustness and ensure that they perform effectively across a spectrum of sensor types.

Chapter 6

Cross-Location Calibration

The aim of this analysis is to development a calibration model that does not need direct co-location with a reference station. This approach aims to address two critical needs: to enable the immediate use of LCS data upon installation and to improve data quality in contexts where traditional calibration procedures are not practical. By exploiting contextual and environmental similarities between locations, the proposed solution aims to expand the applicability and reliability of LCSs in air quality monitoring. We will analyze machine learning and deep learning models, including linear regression, ridge regression, lasso regression, elastic net, support vector regression, decision tree, random forest, XGBoost, LightGBM, and a recurrent neural network, in order to evaluate the generalization capabilities of these models across various experimental scenarios.

To the best of my knowledge, no previous work has systematically studied the ability of models to generalize across such geographically and contextually diverse datasets. In our approach, each dataset was standardized by aligning the LCS data with those of the corresponding reference station, resampling the measurements to an hourly frequency and creating homogeneous feature vectors. This procedure ensures comparability and consistency between the datasets, providing a solid basis for exploring the generalization capabilities of calibration models under different environmental and meteorological conditions.

This chapter provides an in-depth overview of the methodologies employed in this study, encompassing data acquisition, pre-processing, feature selection, and model evaluation. The aim is to establish a rigorous framework for the analysis and prediction of air quality across diverse geographical locations. The chapter is organized into three primary sections, each addressing a critical aspect of the methodology:

- **Data Acquisition and Preprocessing (6.1.1)**: This section outlines the sources and characteristics of the datasets used, including sensor types and geographic locations. It also describes the data standardization procedures to ensure homogeneity across datasets and details the anomaly detection methods applied during pre-processing to enhance data reliability.
- **Features Selection (6.1.1)**: This section describes the comprehensive set of features extracted from the datasets, including temporal, geographic, and environmental attributes. It highlights the importance of these features in capturing

the variability of air quality data and their transformation for machine learning compatibility.

- **Model Selection and Experimental Setup (6.1.2)**: This section presents the machine learning and deep learning models utilized in the study, including LightGBM and recurrent neural networks. It elaborates on the experimental procedures, such as baseline evaluation, pairwise dataset analysis, incremental data inclusion, and leave-one-location-out testing, which were designed to assess model performance and generalization capabilities.

- **Generalization Results (6.2)**: Finally, this section provides a comprehensive analysis of the models' performance across diverse datasets, emphasizing the adaptability and robustness of the proposed methodology in varying environmental and geographical contexts.

6.1 Methods

This section describes the methodologies employed in this study, including the dataset acquisition and standardization procedures (6.1.1), the selection of relevant features (6.1.1), and the choice of machine learning models and experimental setups (6.1.2). Additionally, the strategies used for splitting the dataset into training and testing sets to optimize model performance are described in detail (5.1.1).

6.1.1 Data and Standardization Procedure

The datasets utilized in this study were sourced from eight distinct locations: Aosta (Italy), Badajoz (Spain), Lima (Peru), Bangalore (India), Delhi (India), Calgary (Canada), Hamirpur (India) and Southampton (UK). These datasets are publicly available [CM24, CPZ23b].

PM data were collected using low-cost laser scattering sensors, with the specific sensor types listed in Table 3.3. It is important to note that each data set was collected in different time periods, varying in both duration and specific period of collection.

The standardization process aimed to create homogeneous datasets in terms of frequency and included characteristics. In particular, the original data, which had sampling frequencies between 15 minutes and 1 hour, were resampled to an hourly frequency. In addition, the meteorological characteristics of all datasets were harmonised using the Weather section of the Visual Crossing API [Cro], ensuring the consistency of the meteorological data.

To pre-process the data and resolve anomalies caused by technical problems, various techniques were tested, including 3-sigma, interquartile range and DBSCAN methods (see Section 5.1.2). These techniques were applied over different time periods: the whole series, month, three weeks, two weeks and weekly intervals. It was determined that the application of the interquartile correction on a weekly basis outperformed the other methods, effectively minimizing the impact of outlier data.

Features Selection

Each dataset contains a comprehensive set of features aimed at capturing environmental conditions, geographic context, and sensor-specific attributes. Below is a detailed description of the included features:

- **Temporal Features:**
  - `valid_at`: Timestamp indicating the exact date and time of the measurement.
  - `hour`: The hour at which the measurement was taken.
  - `day`: The calendar day of measurement.
  - `day_of_week`: Specifies the day of the week (for example, Monday, Tuesday).
  - `month`: The month of the measurement, is useful for seasonal analysis.
  - `year`: The year of the measurement, providing a long-term temporal perspective.
- **Geographic and Demographic Features:**
  - `city`: The city where the measurements were taken.
  - `altitude`: Elevation above sea level (in meters) of the sensor's location.
  - `latitude`: Latitude of the sensor's location.
  - `longitude`: Longitude of the sensor's location.
  - `density_of_population`: Population density of the area (e.g., people per square kilometre).
- **Sensor Information:**
  - `sensor_id`: A unique identifier for the sensor.
  - `type_sensor`: Type or model of the sensor.
- **Pollution and Weather Features:**
  - `pm2p5_x`: $PM_{2.5}$ concentration (in $\mu g/m^3$) measured by the low-cost sensor.
  - `pm2p5_y`: $PM_{2.5}$ concentration (in $\mu g/m^3$) measured by the reference station.
  - `pressure`: Atmospheric pressure (in hPa).
  - `relative_humidity`: Percentage of relative humidity.
  - `temperature`: Ambient air temperature (in °C).
  - `wind_speed`: Wind speed (in m/s).
  - `rain`: Precipitation levels (e.g., mm/hour).

Categorical features are converted using the *dummies method*, which creates one-hot encoded variables for each category. This ensures that the data is in a suitable format for machine learning models while preserving the categorical information.

6.1.2 Model Selection and Experimental Setup

This study employed a diverse set of machine learning and deep learning models, including linear regression, ridge regression, lasso regression, elastic net, support vector regression, decision tree, random forest, XGBoost, LightGBM, and a recurrent neural network.

The goal was to evaluate the generalization capabilities of these models across various experimental scenarios. Each dataset was divided into training and testing sets based on an optimized split identified as suitable for the problem, as explained in Section 5.1.1. The models were tested through the following procedures:

- **Baseline Evaluation Across All Locations:** All models were evaluated on each location to establish baseline and state-of-the-art performance for the collected datasets. LightGBM performed better than the other methods, even in subsequent procedures, so it was retained as the preferred method.
- **Pairwise Dataset Analysis:** LightGBM was trained and tested on paired datasets to assess how incorporating data from multiple locations impacts performance. To our knowledge, this type of analysis is novel in the literature.
- **Incremental Location Inclusion:** The effect of gradually adding data from one location at a time was tested, evaluating the LightGBM model's performance on the location of interest as additional datasets were included in the training set.
- **Leave-One-Location-Out Testing:** Models were trained on data from all but one location and then tested on the excluded location. Additionally, a variation of this test involved withholding 25% of the data from the target location during training to assess the impact on model performance.
- **PCA-Based Closest Location Testing:** Models were trained on data from external datasets that showed the greatest proximity to the target dataset in a two-dimensional PCA space, based on their overlap in the projection.

These procedures were designed to comprehensively evaluate the ability of the models, with particular focus on LightGBM, to generalize across different environmental and geographical contexts. Details of the model implementation and experimental results are given in Section 6.2.

6.2 Cross-Location Results

This section provides a detailed analysis of the model's performance and adaptability across diverse datasets. Initially, a baseline evaluation of various models is conducted in multiple locations (6.2.1), with a focus on their effectiveness in localized contexts. Following this, pairwise dataset combinations (6.2.2) and incremental dataset integration (6.2.3) are analyzed to assess the impact of incorporating geographically and climatically diverse data. Finally, the model's generalization capabilities are examined using a leave-one-location-out approach (6.2.4), offering insights into its performance in unseen environments.

6.2.1 Baseline Evaluation Across All Locations

To evaluate the performance of the various models on different datasets, a baseline evaluation was conducted using different regressive models. The feature vector consisted of the attributes described in Section 6.1.1 as stand-alone data points, without incorporating temporal dependencies such as past or future data.

The results of this evaluation, presented in Table 6.1, show the effectiveness of each model in adjusting air quality measurements in different cities. The models tested include linear regression (LinR), ridge regression (RR), lasso regression (LR), elastic network (EN), support vector regression (SVR), decision tree (DT), random forest (RF), XGBoost, LightGBM and a recurrent neural network (RNN). For each dataset, the table provides performance metrics for the applied models, highlighting variations between locations and showing the capabilities of each specialized method for each location.

	Dataset									
Method	**Aosta**	**Badajoz**	**Bangalore**	**Calgary**	**Delhi**	**Hamirpur**	**Lima IQAir**	**Lima AIRBEAM**	**UK PMS003**	**UK SPS03**
LinR	0.71	0.59	0.86	0.81	0.85	0.95	0.62	0.58	0.66	0.70
RR	0.71	0.59	0.86	0.81	0.85	0.95	0.62	0.58	0.66	0.70
LR	0.64	0.50	0.82	0.69	0.84	0.95	0.58	0.58	0.62	0.66
EN	0.56	0.48	0.70	0.61	0.71	0.85	0.54	0.55	0.54	0.57
SVR	0.79	0.56	0.82	0.86	0.72	0.68	0.65	0.73	0.69	0.72
DT	0.80	0.72	0.78	0.96	0.82	0.93	0.63	0.60	0.74	0.61
RF	0.89	0.85	0.90	**0.98**	**0.91**	**0.96**	0.77	0.78	0.81	0.81
XGBoost	**0.90**	0.86	**0.91**	**0.98**	**0.91**	**0.96**	**0.81**	0.76	0.80	**0.82**
LightGBM	**0.90**	**0.88**	**0.91**	**0.98**	**0.91**	**0.96**	**0.81**	**0.78**	**0.83**	**0.82**
RNN	0.75	0.73	0.79	0.73	0.82	0.87	0.80	0.65	0.82	0.81

Table 6.1: Results for Baseline Evaluation (R^2 performance values)

As shown in Table 6.1, tree-based methods consistently outperform the other models. While training and testing within the same location is a well-established practice in the literature, it serves as a valuable baseline for assessing the effectiveness of the subsequent generalization procedures. Notably, LightGBM demonstrates superior performance compared to other models across all setups, although the results for some configurations are not explicitly presented in the other sections.

6.2.2 Pairwise Dataset Analysis

This section presents the results of a performance analysis in which datasets were combined iteratively during the training phase. Specifically, at each stage, an additional dataset was incorporated into the basic training set.

In Table 6.2, the first column, labelled *Training*, indicates the dataset's location used as the base training set. The rows represent the R^2 performance values of the test set when combining the base training set with the dataset from the specified location.

Comparing the performance values obtained by training on a single dataset (Table 6.1) with those obtained by iteratively combining additional datasets (Table 6.2), it is evident that in some cases, dataset combination improves performance, such as for the UK datasets. However, in other cases, it results in decreased performance, such as for the Aosta dataset.

These performance variations can be attributed to the inherent characteristics of the datasets. The datasets are sourced from geographically and climatically diverse locations, use different types of sensors, and vary in terms of data collection periods,

both in length and temporal scope. This highlights the complexity of generalization across heterogeneous datasets.

	Test									
Training	**Aosta**	**Badajoz**	**Bangalore**	**Calgary**	**Delhi**	**Hamirpur**	**Lima IQAir**	**Lima AIRBEAM**	**UK PMS003**	**UK SPS03**
Aosta	-	0.87	0.82	0.79	0.72	0.81	-1.81	0.67	0.66	0.76
Badajoz	0.60	-	0.54	0.63	0.55	0.64	0.25	-0.71	0.53	0.64
Bangalore	**0.91**	0.90	-	**0.91**	0.90	0.90	0.88	0.90	0.87	0.89
Calgary	0.97	0.90	0.95	-	0.80	0.96	0.94	0.96	0.90	0.90
Delhi	0.81	0.86	**0.92**	**0.92**	-	0.86	0.84	0.83	0.86	0.83
Hamirpur	**0.97**	**0.97**	**0.96**	**0.97**	**0.96**	-	**0.96**	**0.97**	0.94	0.94
LimaIQAir	**0.84**	**0.83**	**0.83**	**0.83**	0.79	**0.82**	-	**0.84**	0.78	0.75
Lima AIRBEAM	**0.81**	**0.80**	**0.80**	**0.78**	0.75	**0.81**	0.63	-	0.74	0.79
UK PMS003	**0.89**	**0.85**	**0.89**	**0.89**	**0.88**	**0.88**	**0.84**	**0.85**	-	**0.92**
UK SPS30	**0.90**	**0.88**	**0.90**	**0.90**	**0.90**	**0.90**	**0.88**	**0.85**	0.60	-

Table 6.2: Results for Pairwise Dataset Analysis (R^2 performance values).

The improved performance observed for the UK datasets can be attributed to site-specific characteristics, in particular high levels of rainfall. This could result in an uneven distribution of data points between rainy and non-rainy conditions. To further explore this hypothesis, A visualization based on a clustering perspective using Principal Component Analysis (PCA) with two components applied to the datasets is provided in Section 6.2.5.

6.2.3 Incremental Location Inclusion

In these tests, the behavior of the model is evaluated by progressively combining data sets from different locations in the training set. For each experiment, the training set started with the data set of one location and progressively data sets of other locations were added. After each addition, the model was retrained and evaluated using the specific test set of the initial location.

The integration of the datasets is based on Table 6.2, which shows the ranking of each dataset in terms of increasing performance. The first datasets included are those that contribute to the increasing performance improvement. The results are shown in Table 6.3 and in Figure 6.1 for more immediate analysis.

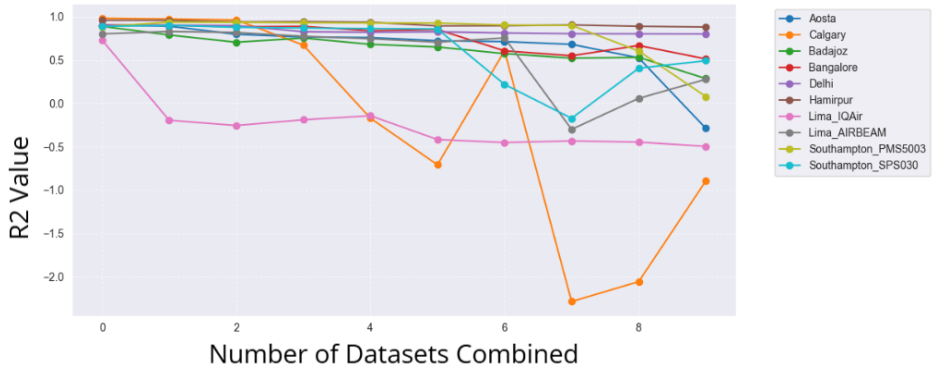


Figure 6.1: Iterative training results: R^2 values for each training dataset with incrementally added datasets.

Cities (Training Set)	**Number of Datasets Combined**	R^2	**Cities (Training Set)**	**Number of Datasets Combined**	R^2	**Cities (Training Set)**	**Number of Datasets Combined**	R^2
Aosta	0	0.9046	Badajoz	0	0.8903	Bangalore	0	0.9051
	1	0.8926		1	0.7905		1	0.9036
	2	0.7981		2	0.7067		2	0.8847
	3	0.7668		3	0.7550		3	0.8898
	4	0.7609		4	0.6827		4	0.8388
	5	0.7242		5	0.6522		5	0.8536
	6	0.7140		6	0.5752		6	0.6078
	7	0.6827		7	0.5238		7	0.5507
	8	0.5262		8	0.5323		8	0.6692
	9	-0.2862		9	0.2876		9	0.5144
Calgary	0	0.9846	Delhi	0	0.9023	Hamirpur	0	0.9598
	1	0.9761		1	0.9042		1	0.9590
	2	0.9638		2	0.8993		2	0.9408
	3	0.6745		3	0.8280		3	0.9441
	4	-0.1688		4	0.8223		4	0.9378
	5	-0.7081		5	0.8273		5	0.8956
	6	0.6078		6	0.8145		6	0.8972
	7	-2.2869		7	0.8062		7	0.9070
	8	-2.0571		8	0.8045		8	0.8907
	9	-0.8954		9	0.8029		9	0.8828
Lima IQAir	0	0.7318	Lima AIRBEAM	0	0.8041	Southampton PMS5003	0	0.8851
	1	-0.1936		1	0.8307		1	0.9399
	2	-0.2543		2	0.8201		2	0.9377
	3	-0.1882		3	0.7759		3	0.9334
	4	-0.1417		4	0.7469		4	0.9298
	5	-0.4159		5	0.7074		5	0.9294
	6	-0.4512		6	0.7568		6	0.9056
	7	-0.4322		7	-0.2989		7	0.9016
	8	-0.4463		8	0.0590		8	0.6052
	9	-0.4951		9	0.2788		9	0.0803
Southampton SPS30	0	0.8895						
	1	0.9021						
	2	0.8793						
	3	0.8712						
	4	0.8644						
	5	0.8603						
	6	0.2201						
	7	-0.1713						
	8	0.4067						
	9	0.4933						

Table 6.3: Iterative training results: R^2 values for each training dataset with incrementally added datasets.

6.2.4 Leave-One-Location-Out Testing

This section presents the results regarding the model's ability to generalize across datasets from different geographic regions and characteristics.

Two analyses were conducted:

1. **External Datasets as Testing Set (GM_1):** An entire dataset corresponding to a specific location is used as the testing set, while all other datasets are used for training.

2. **External Datasets Plus a Portion of the Local Dataset as Testing Set (GM_2):** The dataset from a particular location is split into training and testing subsets using the RMS. The training subset, combined with all other datasets, is used to train the model, while the testing subset from the same location is used for evaluation.

Dataset	LM	GM_1	GM_2	Exp_1	Exp_2	**Intersected Datasets**
Aosta	0.88	0.48	0.56	0.53	0.63	Delhi, Calgary, Bangalore, Hamirpur, Badajoz, $Lima_1$, $Lima_2$
Badajoz	0.83	0.16	0.35	0.17	0.64	Delhi, Calgary, Bangalore, Hamirpur, $Lima_1$, $Lima_2$, Aosta
Bangalore	0.90	0.40	0.85	0.66	0.89	Delhi, Calgary, Hamirpur, Badajoz, $Lima_1$, $Lima_2$, Aosta
Calgary	0.98	-4.34	-0.03	0.55	0.76	Delhi, Bangalore, Hamirpur, Badajoz, $Lima_1$, $Lima_2$, Aosta
Delhi	0.91	-0.75	0.83	0.71	0.82	Calgary, Bangalore, Hamirpur, Badajoz, $Lima_1$, $Lima_2$, Aosta
Hamirpur	0.96	0.05	0.94	0.78	0.95	Delhi, Calgary, Bangalore, Badajoz, $Lima_1$, $Lima_2$, Aosta
$Lima_2$	0.80	0.09	0.71	-0.04	0.79	Delhi, Calgary, Bangalore, Hamirpur, Badajoz, $Lima_1$, Aosta
$Lima_1$	0.81	-1.03	0.42	0.16	0.55	Delhi, Calgary, Bangalore, Hamirpur, Badajoz, $Lima_2$, Aosta
$Southampton_1$	0.84	-17.36	0.54	-0.11	0.93	$Southampton_2$
$Southampton_2$	0.82	0.23	0.33	0.61	0.60	$Southampton_1$

Table 6.4: Performance R^2 for Leave-One-Location-Out Testing (GM_1 and GM_2) and PCA-Based Closest Location Testing (Exp_1 and Exp_2), including a comparison with the localized model LM. For brevity, $Lima_1$ refers to Lima_AIRBEAM, $Lima_2$ refers to Lima_IQAir, $Southampton_1$ refers to Southampton_PMS5003, and $Southampton_2$ refers to Southampton_SPS030

Table 6.4 presents the results of the two analyses using a generalized model (GM) in the GM_1 and GM_2 columns, compared to the localized model (LM) in the first column (localized model performance is based on the Instance RMS split from Table 5.1).

As shown, the results of the generalized model (GM) never exceed and are far from those of the localized model (LM), which remains the best-performing approach. However, the localized model requires a substantial amount of training data to achieve

its performance. The first column shows the results of the GM_1, which can be applied immediately after the sensor is installed. Given this, it is reasonable for its R^2 values to be much lower than those of LM, as no local training data are yet available for comparison. The second column displays the results of the GM_2, which incorporates a portion of local training data. These results highlight the significant impact of including local data, as performance improves consistently compared to GM_1.

From this preliminary study, we can infer that the generalized model GM_1 performs adequately for initial deployment in locations like Aosta and Bangalore, where R^2 greater than 0.4. However, there are several datasets with R^2 values between 0 and 0.3, indicating subpar performance of the generalized model in these cases. Lastly, for some datasets such as Calgary, Lima AIRBEAM, and Southampton PMS5003, the generalized model (GM_1) shows negative R^2 values. This suggests that the external datasets used for training in these cases are highly dissimilar to the data from the target location.

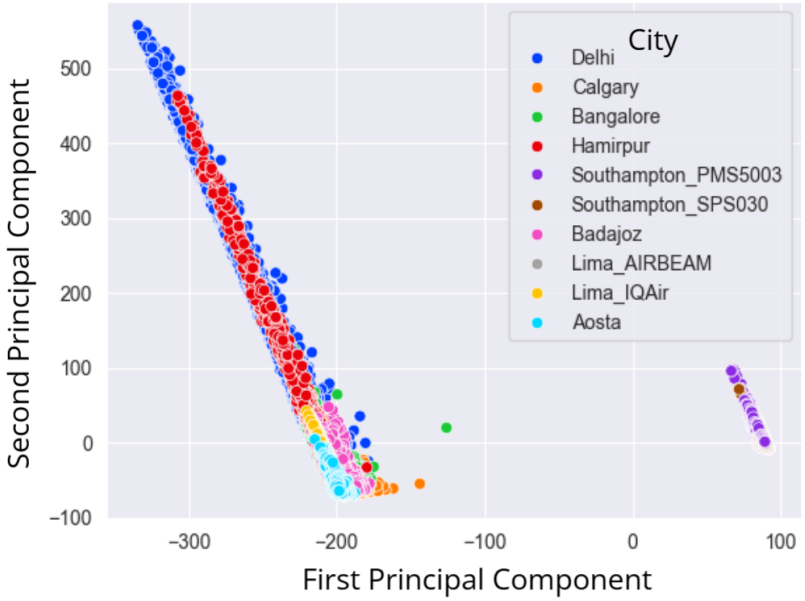


Figure 6.2: Dataset distribution over *PCA2*.

6.2.5 PCA-Based Closest Location Testing:

Principal Component Analysis (PCA) is a statistical technique used for **dimensionality reduction**. It transforms a dataset with potentially correlated features into a set of **linearly uncorrelated variables** called **principal components**, ordered by the amount of variance they explain in the data.

Let $X \in R^{n \times p}$ be your data matrix, where:

- n = number of observations (samples)
- p = number of features (dimensions)

1. Standardize the Data: PCA is sensitive to scale. Standardize each feature to have mean zero and unit variance:

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation vectors.

2. Compute the Covariance Matrix: Compute the $p \times p$ covariance matrix Σ:

$$\Sigma = \frac{1}{n-1} X_{\text{standardized}}^{\top} X_{\text{standardized}}$$

3. Compute Eigenvalues and Eigenvectors: Solve the eigenvalue problem:

$$\Sigma \mathbf{v}_i = \lambda_i \mathbf{v}_i$$

where λ_i are the eigenvalues (variance explained by each component), and $\mathbf{v}_i$ are the eigenvectors (principal directions).

4. Select Principal Components: Sort eigenvalues in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and choose the top k eigenvectors to form the projection matrix V_k.

5. Project the Data: Transform the data into the new subspace:

$$X_{\text{PCA}} = X_{\text{standardized}} \cdot V_k$$

PCA is fundamentally based on the covariance matrix, which measures **linear relationships** between features. If features are linearly correlated, the covariance matrix will have off-diagonal entries significantly different from zero. PCA identifies directions (principal components) that capture the maximal variance caused by this linear dependency.

- If two features are strongly positively correlated (e.g., correlation $r \approx 0.9$), PCA can combine them into one component with minimal information loss.
- PCA works by decorrelating the features—transforming them into new axes with zero correlation.

The following variables were used in the PCA: `relative_humidity`, `temperature`, `wind_speed`, and `rain`, `pm2p5_x` from LCS, `pm2p5_y` from RS. These variables were chosen because they encapsulate key meteorological factors.

In Figure 6.3 the correlations between the features used in this study are presented. Correlation of $r \approx 0.4$ is considered a moderate linear correlation. In this case:

- PCA can still reduce redundancy, but not as effectively as with stronger correlations.
- Variance will be less concentrated in the first few principal components.

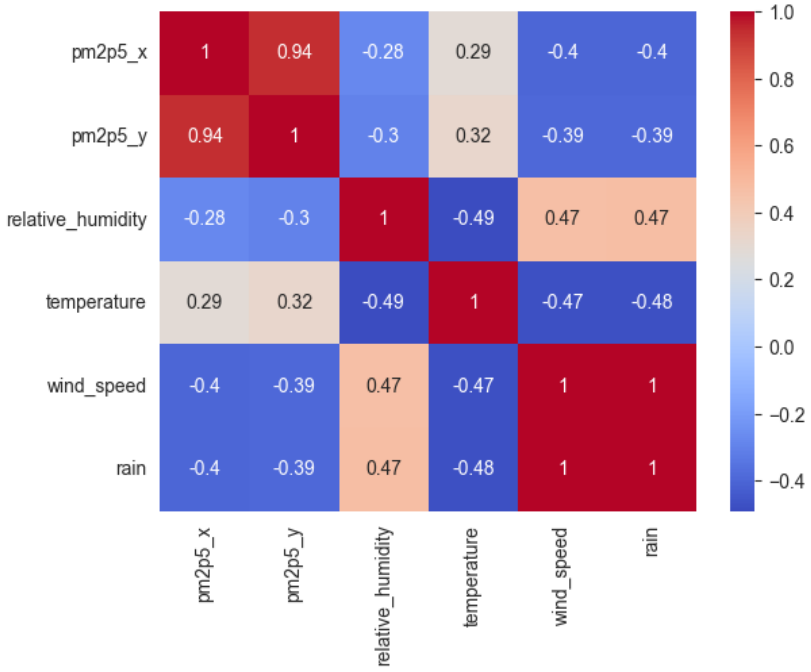


Figure 6.3: Features correlation heatmap.

- It is still worthwhile to try PCA and evaluate the explained variance ratio:

$$\text{Explained Variance Ratio} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

  If the first few components explain a high percentage (e.g., 80%), PCA is effective.

Based on the PCA results:

- **PCA with 2 components (PCA2):** Retains 97.8% of the total variance (explained variance ratio: [0.8461, 0.1322]) with eigenvalues [48830.35, 19298.23].
- **PCA with 3 components (PCA3):** Retains 99.3% of the total variance (explained variance ratio: [0.8461, 0.1322, 0.0153]) with eigenvalues [48830.35, 19298.23, 6562.29].

To investigate the advantages of using PCA, we present a visualization from a clustering perspective using PCA2.

In Figure 6.4, two distinct clusters are observed. The cluster highlighted in dark red, located in the lower right corner, represents the Southampton datasets, which only differ in the type of sensor used. This finding further confirms the hypothesis proposed at the end of Section 6.2.1, where we noted a unique characteristic of the Southampton datasets, specifically their strong association with rainfall patterns.

PCA2 is the preferred approach to PCA3. Indeed, with PCA3, the addition of a third component increases the retained variance by only 1.5%, a relatively low value that does not justify the added complexity. Furthermore, the third component has a much smaller eigenvalue, indicating that it captures less meaningful information and may introduce noise or redundancy.

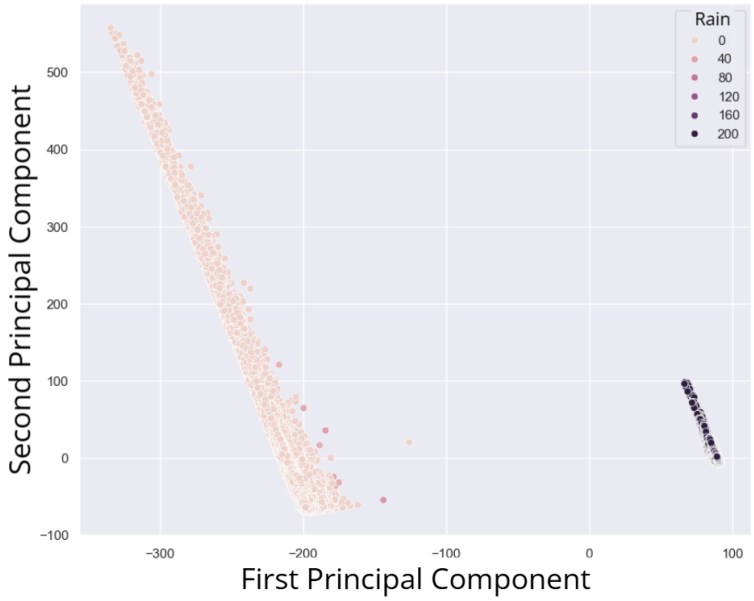


Figure 6.4: Rain feature highlighted over a PCA performed with two components.

Figure 6.2 presents the PCA2 representation, where 10 datasets from different cities are distinguished using different colors.

For each dataset d, a list of intersecting datasets is determined. These intersecting datasets consist of those that overlap with the original dataset d. Specifically, we consider the original dataset d along with all the data points from external datasets d_{ext} that fall within the region of the PCA space defined by the minimum and maximum values of each principal component for the original dataset d.

To assess how incorporating data from nearby regions in the PCA space can improve model performance and ensure reliable predictions across different datasets, two experimental strategies were implemented:

1. **Experiment 1 (Exp_1):** LightGBM is trained exclusively on data from d_{ext}, with no data from d included.

2. **Experiment 2 (Exp_2):** LightGBM is trained on data from d_{ext}, augmented with a randomly selected 25% subset of data from d.

The results of the two experiments are summarized in the right-hand section of Table 6.4, while the left-hand section presents the performance of the localized model (LM) and the Leave-One-Location-Out testing (GM_1 and GM_2).

By consolidating the results of the Leave-One-Location-Out testing and PCA-based Closest Location testing into a single table, it becomes easier to compare GM_1 with Ext_1 and GM_2 with Ext_2.

Exp_2 reveals that incorporating even a small fraction (25%) of local data into the training process significantly enhances the model's predictive performance. This underscores the importance of local data in providing contextual information that external datasets alone cannot fully capture. A notable case is observed in Southampton. In the PCA space, there is a single cluster where only Southampton_PMS5003 and Southampton_SPS030 datasets intersect. Consequently, in these cases, Exp_1 is trained using data

from only one external dataset, unlike other cases where multiple intersected datasets are used. This may explain why datasets belonging to distinct clusters struggle in Exp_1. Additionally, the type of sensor appears to influence the performance of Exp_1. While Exp_1 yields a negative R^2 for Southampton_PMS5003, it achieves a good R^2 for Southampton_SPS030. The performance improvement observed in Exp_2 suggests that intersected datasets provide a reasonable baseline, but their contribution is limited when the test city exhibits unique data characteristics.

Exp_1 can be compared to GM_1: while GM_1 includes all data from external datasets, Exp_1 uses only data from nearby regions in the PCA space. Consequently, the model in Exp_1 is trained on a smaller subset of data. However, as shown in the table, this approach leads to improved performance (R^2). Similarly, Exp_2 can be compared to GM_2: in both cases, data from external datasets is augmented with a portion of local data. The comparability of GM_2 with Ext_2 is maintained due to the same test set being used. Notably, leveraging the training data as outlined in this section often results in improved performance for both comparisons, GM_1/Ext_1 and GM_2/Ext_2.

6.2.6 Limitations

While this study provides valuable insights into the generalization of machine learning calibration models for low-cost air quality sensors, several limitations must be acknowledged:

- **Limited Number and Duration of Datasets**: Although we collected datasets from various global locations, the number of datasets remains relatively limited. Furthermore, data collection periods for many datasets do not span an entire year, which restricts their ability to capture a full range of meteorological and seasonal variations. This limitation may affect the robustness of the proposed models under various weather conditions and long-term deployments.

- **Variability in Sensor Specifications**: The datasets were collected using various LCSs with differing specifications and measurement accuracies. Although this diversity reflects real-world scenarios, it also introduces variability that can complicate model generalization and calibration consistency.

- **Assumptions on Data Quality**: The training and testing procedures assume homogeneous pre-processed datasets. However, in practical applications, the quality of LCS data may vary significantly due to hardware issues, installation environments, and maintenance practices. These factors were not explicitly addressed in this study.

- **Simplifications in PCA-Based Dataset Selection**: In our PCA-based dataset selection methodology, we assume that the reduced components adequately represent the environmental and sensor-specific contexts of the datasets. This simplification may overlook subtler location-specific factors that influence air quality measurements.

Chapter 7

Conclusion

This thesis has focused on addressing the challenges associated with improving the accuracy of air quality measurements obtained from low-cost sensors, with a specific emphasis on $PM_{2.5}$ data. The work presented herein contributes to the broader field of air quality monitoring and data-driven environmental research through advancements in sensor calibration, data pre-processing, clustering techniques, and generalization methodologies.

The key contributions of this thesis can be summarized as follows:

- **Data Collection**: A major focus of this research was the collection and standardization of air quality datasets from a wide range of global locations. This includes data obtained from open sources as well as collaborative efforts with Italian Environmental Agencies. The creation of the *Air Quality Datasets Repository (AQDR)* [CM24] has provided a significant resource for the research community, promoting further studies on low-cost sensors.

- **Datasets Standardization**: A dedicated pipeline for dataset standardization was developed, which handles resampling to hourly frequency and removing gross anomalies. Various anomaly detection methods were evaluated, with the *interquartile range (IQR)* method emerging as the most effective. The relevant code is available in the code directory of AQDR [CM24].

- **Datasets Splitting**: Various splitting techniques were tested using LightGBM, with the *Random Month Split (RMS)* method showing the best performance. In this approach, the dataset is grouped by month, and for each month, 75% of the instances are randomly assigned to the training set, while the remaining 25% are used for testing. This method showed more consistent and higher results compared to others.

- **MitH Framework for Hygroscopicity Mitigation**: The Mitigate Hygroscopicity (MitH) framework [CP24] was developed to mitigate the impact of high relative humidity on PM measurements. This framework can function independently or as a pre-processing step for calibration models, significantly enhancing the reliability of low-cost sensor measurements under humid conditions. It has been tested under various meteorological conditions and in different modes, either

correcting one observation at a time or in batches. When compared to existing correction functions, the proposed approach outperforms them, improving R^2, RMSE, and NRMSE. A key advantage of this framework is its applicability even in locations lacking reference stations, improving data accuracy without requiring locally trained calibration models. MitH has also been tested as a pre-processing step followed by different models, showing improved results over models applied to raw data. Specifically, R^2 increased by 0.14 for linear regression, 0.03 for LightGBM, and 0.6 for AirMLP.

- **ANFIS Method**: The Adaptive Neuro-Fuzzy Inference System was applied to adjust low-cost sensor PM concentrations, showcasing both its advantages and limitations. Unlike deterministic methods, AI approaches like ANFIS rely on test procedures with random selection and validation across datasets, introducing inherent uncertainty due to the lack of formal proof. Despite this, ANFIS benefits from fuzzy logic, which bridges complex AI computations with human interpretability. In comparison, linear regression showed limited performance ($R^2 < 0.5$), while random forest performed better. The Multilayer Perceptron outperformed ANFIS in some cases, particularly when using the full feature set, though its RMSE was slightly higher. ANFIS was ultimately preferred for its interpretability and explainability, which can be valuable in certain contexts, despite the higher performance of neural networks. Code and data are available [CPZ23b, CK24].

- **Calibration Models**: A variety of advanced calibration techniques were explored, including linear regression, ridge regression, lasso regression, elastic network, support vector regression, decision tree, random forest, XGBoost, LightGBM, and recurrent neural networks. Among these, random forest, XGBoost, and LightGBM consistently achieved the highest performance across most datasets, thanks to their ability to model non-linear relationships and capture complex patterns in the data. For example, LightGBM achieved R^2 values of 0.98 and 0.91 for the Calgary and Delhi datasets, respectively. Further analysis of LightGBM [CAP24] and AirMLP, [CPZ23a], an MLP tailored for this application, was conducted in dedicated sections, focusing on specific improvements such as skewness transformation and the use of past data. The AirMLP7-1500 model, in particular, delivered impressive results, achieving an R^2 score of 0.932.

- **Generalization Techniques**: Efforts were made to improve the generalization of calibration models to unseen locations. This aspect is critical for deploying LCSs in regions where localized calibration data are unavailable. Results highlighted the effectiveness of LightGBM as the most reliable model for generalization purposes. Additionally, the study provided new insights into the potential for further analyzing and implementing generalization techniques on air quality datasets, challenging current perspectives in the literature. Moreover, including a portion of the local dataset in the training phase consistently outperforms models trained solely on external datasets. This is evident in Table 6.4, where GM_2 achieves higher $\mathbf{R^2}$ values than GM_1, and Exp_2 outperforms Exp_1.

  The PCA-Based Closest Location Testing further highlights the importance of leveraging data from contextually similar PCA regions. This targeted approach,

which uses datasets with comparable characteristics, improves model performance relative to Leave-One-Location-Out Testing. Specifically, Exp_1 surpasses GM_1, and Exp_2 achieves better results than GM_2, reinforcing the value of selecting relevant external data.

Building on the findings of this thesis, several avenues for future research are proposed:

- **Integration of Advanced Sensor Technologies**: Exploring enhanced sensors that combine low-cost sensors (LCSs) with other technologies under the AIQS project, within the ECOSISTER framework.
- **Improvement of the MitH Framework**: Further development of the MitH framework is underway within the AIQS project, with the goal of implementing it in real-world scenarios.
- **Expanding Data Collection**: Ongoing efforts are focused on gathering new data, with additional collaborations being formed. The launch of AQDR facilitates the acquisition of more data from various research projects.
- **Advancing Clustering Approaches**: Investigating more sophisticated clustering techniques, such as hierarchical clustering and density-based methods, to better capture subtle environmental variations.
- **Comprehensive Model Validation**: Extending the evaluation of generalization techniques to a wider variety of datasets, particularly those from regions with extreme environmental conditions.
- **Digital Twin Development**: As part of the broader research trajectory, efforts will be made to integrate findings into the Digital Twin framework, enabling predictive, personalized air quality monitoring solutions.
- **Citizen Science Initiatives**: Leveraging LCS networks to empower communities with actionable air quality data, fostering environmental awareness, and improving public health. This includes integrating air quality analysis with traffic flow and pollution source data to optimize pedestrian pathways.

This thesis represents a step forward in harnessing the potential of low-cost sensors for air quality monitoring. By addressing challenges in calibration, data pre-processing, and generalization, it paves the way for more reliable and accessible air quality solutions. These advancements have the potential to support policymakers, researchers, and communities in mitigating the impacts of air pollution, ultimately contributing to a healthier and more sustainable future.

Bibliography

[100] 10000 ambient air monitor datasheet. https://particlesplus.com/wp-content/datasheets/10000/Particles%20Plus%2010000%20Datasheet.pdf. Accessed: 2023-09-20.

[120] 12000 ambient air monitor datasheet. https://particlesplus.com/wp-content/datasheets/12000/Particles%20Plus%2012000%20Datasheet.pdf. Accessed: 2023-09-20.

[ABDG+20] Brigida Alfano, Luigi Barretta, Antonio Del Giudice, Saverio De Vito, Girolamo Di Francia, Elena Esposito, Fabrizio Formisano, Ettore Massera, Maria Lucia Miglietta, and Tiziana Polichetti. A review of low-cost particulate matter sensors from the developers' perspectives. *Sensors*, 20(23):6819, 2020.

[ABOS22] Priscilla Adong, Engineer Bainomugisha, Deo Okure, and Richard Sserunjogi. Applying machine learning for large scale field calibration of low-cost pm2. 5 and pm10 air pollution sensors. *Applied AI Letters*, 3(3):e76, 2022.

[ADMJ22] A. Kofi Amegah, Gordon Dakuu, Pierpaolo Mudu, and Jouni J. K. Jaakkola. Particulate matter pollution at traffic hotspots of accra, ghana: levels, exposure experiences of street traders, and associated respiratory and cardiovascular symptoms. *Journal of Exposure Science and Environmental Epidemiology*, 32(2):333 – 342, 2022.

[AGSSL21] Patricia Arroyo, Jaime Gómez-Suárez, José Ignacio Suárez, and Jesús Lozano. Low-cost air quality measurement system based on electrochemical and pm sensors with cloud connection. *Sensors*, 21(18), 2021.

[Alt23] Astrid Altamar. Data pm 2.5, Feb 2023.

[AM5] Am520 datasheet. https://tsi.com/getmedia/3b6a2fdc-b348-466f-b6f6-b2014be9a0d5/SidePak_AM520-AM520i_A4_5001738_RevC_Web?ext=.pdf. Accessed: 2023-09-20.

[Ame18] A. Kofi Amegah. Proliferation of low-cost sensors. what prospects for air pollution epidemiologic research in sub-saharan africa? *Environmental Pollution*, 241:1132 – 1137, 2018.

[AQM] AQMESH. Aqmesh technical documentation. https://d3pcsg2wjq9izr.cloudfront.net/files/84570/download/

667711/10reasonswhyyoushouldchooseAQMesh.pdf. Accessed: 2023-09-20.

[AS] AQ-SPEC. Aiq quality sensor performance evaluation center. https://www.aqmd.gov/aq-spec/evaluations/criteria-pollutants/field. Accessed: July 13, 2024.

[AT] ARPA-Torino. Torino rubino air quality monitoring station. Accessed: December 5, 2023.

[BBKS20] Zsolt Bodor, Katalin Bodor, Ágnes Keresztesi, and Róbert Szép. Major air pollutants seasonal variation analysis and long-range transport of pm 10 in an urban environment with specific climate condition in transylvania (romania). *Environmental Science and Pollution Research*, 27:38181–38199, 2020.

[BCA+21] Lorenzo Brilli, Federico Carotenuto, Bianca Patrizia Andreini, Alice Cavaliere, Andrea Esposito, Beniamino Gioli, Francesca Martelli, Marco Stefanelli, Carolina Vagnoli, Stefania Venturi, Alessandro Zaldei, and Giovanni Gualtieri. Low-cost air quality stations' capability to integrate reference stations in particulate matter dynamics assessment. *Atmosphere*, 12(8), 2021.

[BCC98] Dino Borri, Grazia Concilio, and Emilia Conte. A fuzzy approach for modelling knowledge in environmental systems evaluation. *Computers, Environment and Urban Systems*, 22(3):299–313, 1998.

[Bel24] Belgian Interregional Environment Agency. What are the causes of particulate matter?, 2024. Accessed: 2024-12-29.

[BGBD23] Yas Barzegar, Irina Gorelova, Francesco Bellini, and Fabrizio D'Ascenzo. Drinking water quality assessment using a fuzzy inference system method: A case study of rome (italy). *International Journal of Environmental Research and Public Health*, 20(15), 2023. All Open Access, Gold Open Access, Green Open Access.

[BGC21] Karoline K. Barkjohn, Brett Gantt, and Andrea L. Clements. Development and application of a united states-wide correction for pm2.5 data collected with the purpleair sensor. *Atmospheric Measurement Techniques*, 14(6):4617 – 4637, 2021. All Open Access, Gold Open Access, Green Open Access.

[BNV+23] Leen Brusseleers, Vu Giang Nguyen, Kim Chi Vu, Han Huy Dung, Ben Somers, and Bruno Verbist. Assessment of the impact of local climate zones on fine dust concentrations: A case study from hanoi, vietnam. *Building and Environment*, 242, 2023.

[BOM+23a] Florentin MJ Bulot, Steven J Ossont, Andrew KR Morris, Philip J Basford, Natasha HC Easton, Hazel L Mitchell, Gavin L Foster, Simon J Cox, and Matthew Loxham. Characterisation and calibration of low-cost

pm sensors at high temporal resolution to reference-grade performance. *Heliyon*, 9(5), 2023.

[BOM+23b] Florentin M.J. Bulot, Steven J. Ossont, Andrew K.R. Morris, Philip J. Basford, Natasha H.C. Easton, Hazel L. Mitchell, Gavin L. Foster, Simon J. Cox, and Matthew Loxham. Characterisation and calibration of low-cost pm sensors at high temporal resolution to reference-grade performance. *Heliyon*, 9(5), 2023. All Open Access, Gold Open Access.

[BOM+23c] Florentin M.J. Bulot, Steven J. Ossont, Andrew K.R. Morris, Philip J. Basford, Natasha H.C. Easton, Hazel L. Mitchell, Gavin L. Foster, Simon J. Cox, and Matthew Loxham. Characterisation and calibration of low-cost pm sensors at high temporal resolution to reference-grade performance. *Heliyon*, 9(5), May 2023.

[BPL+22] T. Bush, N. Papaioannou, F. Leach, F. D. Pope, A. Singh, G. N. Thomas, B. Stacey, and S. Bartington. Machine learning techniques to improve the field performance of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 15(10):3261–3278, 2022.

[BRP24] Chiara Bachechi, Federica Rollo, and Laura Po. Hypeair: A novel framework for real-time low-cost sensor calibration for air quality monitoring in smart cities. *Ecological Informatics*, 81:102568, 2024.

[BSPP21] Peter Biber, Fabian Schwaiger, Werner Poschenrieder, and Hans Pretzsch. A fuzzy logic-based approach for evaluating forest ecosystem service provision and biodiversity applied to a case study landscape in southern germany. *European Journal of Forest Research*, 140(6):1559 – 1586, 2021. All Open Access, Green Open Access, Hybrid Gold Open Access.

[Bul22a] Florentin M. J. Bulot. Characterisation and calibration of low-cost pm sensors at high temporal resolution to reference grade performances - dataset, October 14 2022.

[Bul22b] Florentin M. J. Bulot. Characterisation and calibration of low-cost pm sensors at high temporal resolution to reference grade performances - dataset, 10 2022. https://doi.org/10.5281/zenodo.7198378.

[CAP24] Martina Casari, Andrea Arigliano, and Laura Po. A comparative study of lightgbm on air quality data across multiple locations. volume 3762, page 505 – 509, 2024.

[Car22] Ken S Carslaw. Aerosol processes. In *Aerosols and Climate*, pages 135–185. Elsevier, 2022.

[CBM+22] Eric S. Coker, Rafael Buralli, Andres Felipe Manrique, Claudio Makoto Kanai, A. Kofi Amegah, and Nelson Gouveia. Association between pm2.5 and respiratory hospitalization in rio branco, brazil: Demonstrating the potential of low-cost air quality sensor for epidemiologic research. *Environmental Research*, 214, 2022.

[CFd+23] Fernando Campo, Davide Franco, Felipe de Campos Santos, Andy Blanco-Rodríguez, Alejandro Rafael Garcia-Ramirez, Gabriel Ratão, and Leonardo Hoinaski. Clean - collaborative low-cost environmental and air-quality network. *Environmental Modelling & Software*, 163:105664, 2023.

[CFDS04] Bhabesh Chakrabarti, Philip M. Fine, Ralph Delfino, and Constantinos Sioutas. Performance evaluation of the active-flow personal dataram pm2.5 mass monitor (thermo anderson pdr-1200) designed for continuous personal exposure measurements. *Atmospheric Environment*, 38(20):3329 – 3340, 2004.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[CGS+23a] M. J. Campmier, J. Gingrich, S. Singh, N. Baig, S. Gani, A. Upadhya, P. Agrawal, M. Kushwaha, H. R. Mishra, A. Pillarisetti, S. Vakacherla, R. K. Pathak, and J. S. Apte. Seasonally optimized calibrations improve low-cost sensor performance: long-term field evaluation of purpleair sensors in urban and rural india. *Atmospheric Measurement Techniques*, 16(19):4357–4374, 2023.

[CGS+23b] Mark Campmier, Jonathan Gingrich, Saumya Singh, et al. Seasonally optimized calibrations improve low-cost sensor performance: Long-term field evaluation of purpleair sensors in urban and rural india [dataset], 2023.

[Che13] Mu-Yen Chen. A hybrid anfis model for business failure prediction utilizing particle swarm optimization and subtractive clustering. *Information Sciences*, 220:180–195, 2013.

[CK24] Martina Casari and Piotr A. Kowalski. Anfis low-cost pm adjustment, 3 2024.

[CKC+18] Chen-Chia Chen, Chih-Ting Kuo, Ssu-Ying Chen, Chih-Hsing Lin, Jin-Ju Chue, Yi-Jie Hsieh, Chun-Wen Cheng, Chieh-Ming Wu, and Chun-Ming Huang. Calibration of low-cost particle sensors by using machine-learning method. In *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 111–114, 2018.

[CKP24] Martina Casari, Piotr A. Kowalski, and Laura Po. Optimisation of the adaptive neuro-fuzzy inference system for adjusting low-cost sensors pm concentrations. *Ecological Informatics*, 83:102781, 2024.

[CL07] Shuenn-Chin Chang and Chung-Te Lee. Secondary aerosol formation through photochemical reactions estimated by using air quality monitoring data in taipei city from 1994 to 2003. *Atmospheric Environment*, 41(19):4002 – 4017, 2007.

[CLL+16] Donatella Caniani, Alessandro Labella, Donata Serafina Lioi, Ignazio M. Mancini, and Salvatore Masi. Habitat ecological integrity and environmental impact assessment of anthropic activities: A gis-based fuzzy logic model for sites of high biodiversity conservation interest. *Ecological Indicators*, 67:238–249, 2016.

[CM04] Vladimir Cherkassky and Yunqian Ma. Practical selection of svm parameters and noise estimation for svm regression. *Neural networks*, 17(1):113–126, 2004.

[CM24] Martina Casari and Eleonora Montorsi. Martinacasari/airqualitydatasetsrepository: Aqdr - v1.0.0 (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.13982208, 2024.

[CML+21] Francesco Concas, Julien Mineraud, Eemil Lagerspetz, Samu Varjonen, Xiaoli Liu, Kai Puolamäki, Petteri Nurmi, and Sasu Tarkoma. Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Transactions on Sensor Networks (TOSN)*, 17(2):1–44, 2021.

[CMLVS22] Miriam Chacón-Mateos, Bernd Laquai, Ulrich Vogt, and Cosima Stubenrauch. Evaluation of a low-cost dryer for a low-cost optical particle counter. *Atmospheric Measurement Techniques*, 15(24):7395–7410, 2022.

[CMTS22] I. Christakis, K. Moutzouris, O. Tsakiridis, and I. Stavrakas. Barometric pressure as a correction factor for low-cost particulate matter sensors. In *IOP Conference Series: Earth and Environmental Science*, volume 1123, page 012068, 2022. All Open Access, Gold Open Access.

[CP23a] M. Casari and L. Po. Airmlp - sps30 low-cost sensors and tecora reference station pm 2.5 data, October 24 2023.

[CP23b] Martina Casari and Laura Po. Mitigating the impact of humidity on low-cost pm sensors. volume 3486, page 599 – 604, 2023.

[CP24] Martina Casari and Laura Po. Mith: A framework for mitigating hygroscopicity in low-cost pm sensors. *Environmental Modelling & Software*, 173:105955, 2024.

[CPZ23a] Martina Casari, Laura Po, and Leonardo Zini. Airmlp: A multilayer perceptron neural network for temporal correction of pm2.5 values in turin. *Sensors*, 23(23), 2023.

[CPZ23b] Martina Casari, Laura Po, and Leonardo Zini. Low-cost pm data, 10 2023. https://doi.org/10.5281/zenodo.10037781.

[Cro] Visual Crossing. Visual crossing weather api. https://www.visualcrossing.com/weather-api. Accessed: 2024-10-23.

[CROD21] Ellen M. Considine, Colleen E. Reid, Michael R. Ogletree, and Timothy Dye. Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network. *Environmental Pollution*, 268:115833, 2021.

[CSK+20] L. R. Crilley, A. Singh, L. J. Kramer, M. D. Shaw, M. S. Alam, J. S. Apte, W. J. Bloss, L. Hildebrandt Ruiz, P. Fu, W. Fu, S. Gani, M. Gatari, E. Ilyinskaya, A. C. Lewis, D. Ng'ang'a, Y. Sun, R. C. W. Whitty, S. Yue, S. Young, and F. D. Pope. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. *Atmospheric Measurement Techniques*, 13(3):1181–1193, 2020.

[CSP+18] Leigh R Crilley, Marvin Shaw, Ryan Pound, Louisa J Kramer, Robin Price, Stuart Young, Alastair C Lewis, and Francis D Pope. Evaluation of a low-cost optical particle counter (alphasense opc-n2) for ambient air monitoring. *Atmospheric Measurement Techniques*, 11(2):709–720, 2018.

[CYP+22] Damien Chanal, Nadia Yousfi Steiner, Raffaele Petrone, Didier Chamagne, and Marie-Cécile Péra. Online diagnosis of pem fuel cell by fuzzy c-means clustering. In Luisa F. Cabeza, editor, *Encyclopedia of Energy Storage*, pages 359–393. Elsevier, Oxford, 2022.

[DA23] Giuseppe D Aniello. Fuzzy logic for situation awareness: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):4419 – 4438, 2023. All Open Access, Hybrid Gold Open Access.

[DAPO+18] Andrea Di Antonio, Olalekan A. M. Popoola, Bin Ouyang, John Saffell, and Roderic L. Jones. Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter. *Sensors (Switzerland)*, 18(9), 2018. All Open Access, Gold Open Access, Green Open Access.

[deS22] Priyanka Nadia deSouza. Key concerns and drivers of low-cost air quality sensor use. *Sustainability*, 14(1), 2022.

[DKS+22] Priyanka DeSouza, Ralph Kahn, Tehya Stockman, William Obermann, Ben Crawford, An Wang, James Crooks, Jing Li, and Patrick Kinney. Calibrating networks of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 15(21):6309–6328, 2022.

[DM01] Derek E Day and William C Malm. Aerosol light scattering measurements as a function of relative humidity: a comparison between measurements made at three different sites. *Atmospheric Environment*, 35(30):5169–5176, 2001. Visibility, Aerosol and Atmospheric Optics.

[DSG+23] Konstantinos Dimitriou, Iasonas Stavroulas, Georgios Grivas, Charalampos Chatzidiakos, Georgios Kosmopoulos, Andreas Kazantzidis,

Konstantinos Kourtidis, Athanasios Karagioras, Nikolaos Hatzianastassiou, Spyros N Pandis, Nikolaos Mihalopoulos, and Evangelos Gerasopoulos. Intra- and inter-city variability of pm2.5 concentrations in greece as determined with a low-cost sensor network. *Atmospheric Environment*, 301, 2023.

[DSM] Dsm501 datasheet. https://www.elecrow.com/download/DSM501.pdf. Accessed: 2023-09-20.

[DTA22] Anh Ngoc Thi Do, Hau Duc Tran, and Matthew Ashley. Employing a novel hybrid of ga-anfis model to predict distribution of whiting fish larvae and juveniles from tropical estuaries in the context of climate change. *Ecological Informatics*, 71:101780, 2022.

[DTS+23] Racha Dejchanchaiwong, Perapong Tekasakul, Apichat Saejio, Thanathip Limna, Thi-Cuc Le, Chuen-Jinn Tsai, Guan-Yu Lin, and John Morris. Seasonal field calibration of low-cost pm2. 5 sensors in different locations with different sources in thailand. *Atmosphere*, 14(3):496, 2023.

[EEA21] EEA. Eu air quality directives, 2021.

[Eur08] European Parliament and Council. Directive 2008/50/ec of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for europe. https://eur-lex.europa.eu/eli/dir/2008/50/oj/eng, 2008. Accessed: 2025-01-07.

[FZZ+23] Zikang Feng, Lina Zheng, Xuehan Zhang, Jia Liu, Ning Xue, and Mengmeng Wang. Evaluation and calibration of low-cost particulate matter sensors for respirable coal mine dust monitoring. *Aerosol Science and Technology*, 0(0):1–12, 2023.

[GA18] Nevin Guler Dincer and Ozge Akkuş. A new fuzzy time series model based on robust clustering for forecasting of air pollution. *Ecological Informatics*, 43:157–164, 2018.

[GMP+21a] Michael R Giordano, Carl Malings, Spyros N Pandis, Albert A Presto, VF McNeill, Daniel M Westervelt, Matthias Beekmann, and R Subramanian. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, 158:105833, 2021.

[GMP+21b] Michael R. Giordano, Carl Malings, Spyros N Pandis, Albert A. Presto, V.F. McNeill, Daniel M. Westervelt, Matthias Beekmann, and R. Subramanian. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, 158, 2021. All Open Access, Green Open Access, Hybrid Gold Open Access.

[GP2] Gp2y1010au0f datasheet. https://4donline.ihs.com/images/VipMasterIC/IC/SHCO/SHCOS00939/SHCOS00939-1.pdf?hkey=6D3A4C79FDBF58556ACFDE234799DDF0. Accessed: 2023-09-20.

[GRS03] Gregory C. Pratt Gurumurthy Ramachandran, John L. Adgate and Ken Sexton. Characterizing indoor and outdoor 15 minute average pm 2.5 concentrations in urban neighborhoods. *Aerosol Science and Technology*, 37(1):33–45, 2003.

[GSB17] Michel Gerboles, Laurent Spinelle, and Annette Borowiak. *Measuring air pollution with low-cost sensors*, 2017.

[GSV17] Alexander Gegov, David Sanders, and Boriana Vatchova. Aggregation of inconsistent rules for fuzzy rule base simplification. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 21(3):135 – 145, 2017. All Open Access, Green Open Access.

[GTM+16] RS Gao, H Telg, RJ McLaughlin, SJ Ciciora, LA Watts, MS Richardson, JP Schwarz, AE Perring, TD Thornberry, AW Rollins, et al. A light-weight, high-sensitivity particle spectrometer for pm2. 5 aerosol measurements. *Aerosol Science and Technology*, 50(1):88–99, 2016.

[HDQ+22] Jelle Hofman, Tien Huu Do, Xuening Qin, Esther Rodrigo Bonet, Wilfried Philips, Nikos Deligiannis, and Valerio Panzica La Manna. Spatiotemporal air quality inference of low-cost sensor data: Evidence from multiple sensor testbeds. *Environmental Modelling and Software*, 149, 2022.

[HK20] David H Hagan and Jesse H Kroll. Assessing the accuracy of low-cost optical particle sensors using a physics-based approach. *Atmospheric measurement techniques*, 13(11):6343–6355, 2020.

[HLT+21] Gung-Hwa Hong, Thi-Cuc Le, Jing-Wei Tu, Chieh Wang, Shuenn-Chin Chang, Jhih-Yuan Yu, Guan-Yu Lin, Shankar G Aggarwal, and Chuen-Jinn Tsai. Long-term evaluation and calibration of three types of low-cost pm2. 5 sensors at different air quality monitoring stations. *Journal of Aerosol Science*, 157:105829, 2021.

[HNS+22] Jelle Hofman, Mania Nikolaou, Sharada Prasad Shantharam, Christophe Stroobants, Sander Weijs, and Valerio Panzica La Manna. Distant calibration of low-cost pm and no2 sensors; evidence from multiple sensor testbeds. *Atmospheric Pollution Research*, 13(1), 2022.

[HPM] Hpma115c0 datasheet. https://media.distrelec.com/Web/Downloads/_t/ds/HPMA115C0-003_eng_tds.pdf. Accessed: 2023-09-20.

[HRJM15] Jeffery S. Horsburgh, Stephanie L. Reeder, Amber Spackman Jones, and Jacob Meline. Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental*

Modelling and Software, 70:32 – 44, 2015. All Open Access, Hybrid Gold Open Access.

[HZdF+21a] Jinxi Hua, Yuanxun Zhang, Benjamin de Foy, Xiaodong Mei, Jing Shang, Yang Zhang, Ishaq Dimeji Sulaymon, and Dandan Zhou. Improved pm2. 5 concentration estimates from low-cost sensors using calibration models categorized by relative humidity. *Aerosol Science and Technology*, 55(5):600–613, 2021.

[HZdF+21b] Jinxi Hua, Yuanxun Zhang, Benjamin de Foy, Xiaodong Mei, Jing Shang, Yang Zhang, Ishaq Dimeji Sulaymon, and Dandan Zhou. Improved pm2.5 concentration estimates from low-cost sensors using calibration models categorized by relative humidity. *Aerosol Science and Technology*, 55(5):600 – 613, 2021.

[IL67] A.G. Ivakhnenko and V.G. Lapa. *Cybernetics and Forecasting Techniques*. Modern analytic and computational methods in science and mathematics. American Elsevier Publishing Company, 1967.

[IPdGV+10] Luc Int Panis, Bas de Geus, Grégory Vandenbulcke, Hanny Willems, Bart Degraeuwe, Nico Bleux, Vinit Mishra, Isabelle Thomas, and Romain Meeusen. Exposure to particulate matter in traffic: A comparison of cyclists and car passengers. *Atmospheric Environment*, 44(19):2263 – 2270, 2010.

[IQA24] IQAir. IQAir - Air Quality Information and Pollution Data, 2024. Accessed: 2024-10-28.

[JBLL14] Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine learning*, 97:155–176, 2014.

[JLA+20] Rohan Jayaratne, Xiaoting Liu, Kang Ho Ahn, Akwasi Asumadu-Sakyi, Gavin Fisher, Jian Gao, Adrian Mabon, Mandana Mazaheri, Benjamin Mullins, Mawutorli Nyaku, et al. Low-cost pm2. 5 sensors: An assessment of their suitability for various applications. *Aerosol and Air Quality Research*, 20(3):520–532, 2020.

[JLT+18] Rohan Jayaratne, Xiaoting Liu, Phong Thai, Matthew Dunbabin, and Lidia Morawska. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmospheric Measurement Techniques*, 11(8):4883 – 4890, 2018. All Open Access, Gold Open Access, Green Open Access.

[JLW+22] Xiaoai Jin, Zhanqing Li, Tong Wu, Yuying Wang, Yafang Cheng, Tianning Su, Jing Wei, Rongmin Ren, Hao Wu, Shangze Li, Dongmei Zhang, and Maureen Cribb. The different sensitivities of aerosol optical properties to particle concentration, humidity, and hygroscopicity between the surface level and the upper boundary layer in guangzhou, china. *Science of the Total Environment*, 803, 2022.

[JXF+24] Weaam Jaafar, Junshi Xu, Emily Farrar, Cheol-Heon Jeong, Arman Ganji, Greg Evans, and Marianne Hatzopoulou. Challenges and opportunities of low-cost sensors in capturing the impacts of construction activities on neighborhood air quality. *Building and Environment*, 254:111363, 2024.

[KANZ22] Ye Kang, Lu Aye, Tuan Duc Ngo, and Jin Zhou. Performance evaluation of low-cost air quality sensors: A review. *Science of The Total Environment*, 818:151769, 2022.

[KFH+23] Chris Kelly, Julian Fawkes, Rachel Habermehl, Davi de Ferreyro Monticelli, and Naomi Zimmerman. Plume dashboard: A free and open-source mobile air quality monitoring dashboard. *Environmental Modelling and Software*, 160, 2023. All Open Access, Hybrid Gold Open Access.

[KJK20] Anupam Khatua, Soovoojeet Jana, and Tapan Kumar Kar. A fuzzy rule-based model to assess the effects of global warming, pollution and harvesting on the production of hilsa fishes. *Ecological Informatics*, 57:101070, 2020.

[KK19] Dervis Karaboga and Ebubekir Kaya. Adaptive network based fuzzy inference system (anfis) training approaches: a comprehensive survey. *Artificial Intelligence Review*, 52:2263–2293, 2019.

[KMF+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[KNS+21] Nareg Karaoghlanian, Batoul Noureddine, Najat Saliba, Alan Shihadeh, and Issam Lakkis. Low cost air quality sensors "purpleair" calibration and inter-calibration dataset in the context of beirut, lebanon, 2021.

[KPDWP24a] Slawomir Koziel, Anna Pietrenko-Dabrowska, Marek Wojcikowski, and Bogdan Pankiewicz. Efficient calibration of cost-efficient particulate matter sensors using machine learning and time-series alignment. *Knowledge-Based Systems*, 295:111879, 2024.

[KPDWP24b] Slawomir Koziel, Anna Pietrenko-Dabrowska, Marek Wojcikowski, and Bogdan Pankiewicz. Field calibration of low-cost particulate matter sensors using artificial neural networks and affine response correction. *Measurement*, 230:114529, 2024.

[KS21a] Vikas Kumar and Manoranjan Sahu. Evaluation of nine machine learning regression algorithms for calibration of low-cost pm2. 5 sensor. *Journal of Aerosol Science*, 157:105809, 2021.

[KS21b] Vikas Kumar and Manoranjan Sahu. Evaluation of nine machine learning regression algorithms for calibration of low-cost pm2.5 sensor. *Journal of Aerosol Science*, 157:105809, 2021.

[KSHA15] Gregor Kiesewetter, Wolfgang Schoepp, Chris Heyes, and Markus Amann. Modelling pm2.5 impact indicators in europe: Health effects and legal compliance. *Environmental Modelling and Software*, 74:201 – 211, 2015. All Open Access, Hybrid Gold Open Access.

[KST+18] Evangelos Kosmidis, Panagiota Syropoulou, Stavros Tekes, Philipp Schneider, Eleftherios Spyromitros-Xioufis, Marina Riga, Polychronis Charitidis, Anastasia Moumtzidou, Symeon Papadopoulos, Stefanos Vrochidis, Ioannis Kompatsiaris, Ilias Stavrakas, George Hloupis, Andronikos Loukidis, Konstantinos Kourtidis, Aristeidis K. Georgoulias, and Georgia Alexandri. Hackair: Towards raising awareness about air quality in europe by developing a collective online platform. *ISPRS International Journal of Geo-Information*, 7(5), 2018. All Open Access, Gold Open Access, Green Open Access.

[LGD+17] Zhanqing Li, Jianping Guo, Aijun Ding, Hong Liao, Jianjun Liu, Yele Sun, Tijian Wang, Huiwen Xue, Hongsheng Zhang, and Bin Zhu. Aerosol and boundary-layer interactions and impact on air quality. *National Science Review*, 4(6):810–833, 09 2017.

[LK] Bernd Laquai and Bianca Kroseberg. Comparison of a computational method for correcting the humidity influence with the use of a low-cost aerosol dryer on a sds011 low-cost pm-sensor. *researchgate*.

[LKV+19] Chris C. Lim, Ho Kim, M.J. Ruzmyn Vilcassim, George D. Thurston, Terry Gordon, Lung-Chi Chen, Kiyoung Lee, Michael Heimbinder, and Sun-Young Kim. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in seoul, south korea. *Environment International*, 131:105022, 2019.

[Loh11] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

[LSHV19] Hai-Ying Liu, Philipp Schneider, Rolf Haugen, and Matthias Vogt. Performance assessment of a low-cost pm2.5 sensor for a near four-month period in oslo, norway. *Atmosphere*, 10(2), 2019.

[LZZ+20] Hongyong Li, Yujiao Zhu, Yong Zhao, Tianshu Chen, Ying Jiang, Ye Shan, Yuhong Liu, Jiangshan Mu, Xiangkun Yin, Di Wu, Cheng Zhang, Shuchun Si, Xinfeng Wang, Wenxing Wang, and Likun Xue. Evaluation of the performance of low-cost air quality sensors at a high mountain station with complex meteorological conditions. *Atmosphere*, 11(2), 2020.

[MA75] Ebrahim H Mamdani and Saïd Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1):1–13, 1975.

[MBC+22] Mohammad Saood Manzar, Mohammed Benaafi, Romulus Costache, Omar Alagha, Nuhu Dalhat Mu'azu, Mukarram Zubair, Jazuli Abdullahi, and S.I. Abba. New generation neurocomputing learning coupled with a hybrid neuro-fuzzy model for quantifying water quality index variable: A case study from saudi arabia. *Ecological Informatics*, 70:101696, 2022.

[MIF+20] Agnes Molnár, Kornelia Imre, Zita Ferenczi, Gyula Kiss, and Andras Gelencsér. Aerosol hygroscopicity: Hygroscopic growth proxy based on visibility for low-cost pm monitoring. *Atmospheric Research*, 236, 2020. All Open Access, Hybrid Gold Open Access.

[MTH+20] Carl Malings, Rebecca Tanzer, Aliaksei Hauryliuk, Provat K Saha, Allen L Robinson, Albert A Presto, and R Subramanian. Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Science and Technology*, 54(2):160–174, 2020.

[MZHF16] Abhisek Manikonda, Naděžda Zíková, Philip K. Hopke, and Andrea R. Ferro. Laboratory assessment of low-cost pm monitors. *Journal of Aerosol Science*, 102:29–40, 2016.

[NK24] Armin Nakhjiri and Ata Abdollahi Kakroodi. Air pollution in industrial clusters: A comprehensive analysis and prediction using multi-source data. *Ecological Informatics*, 80:102504, 2024.

[NKGC21] Peer Nowack, Lev Konstantinovskiy, Hannah Gardiner, and John Cant. Machine learning calibration of low-cost no 2 and pm 10 sensors: Non-linear algorithms and their impact on site transferability. *Atmospheric Measurement Techniques*, 14(8):5637–5655, 2021.

[NNLV21] Nam H Nguyen, Huy X Nguyen, Thuan TB Le, and Chinh D Vu. Evaluating low-cost commercially available sensors for air quality monitoring and application of sensor calibration methods for improving accuracy. *Open Journal of Air Pollution*, 10(1), 2021.

[NPM] Npm 2 datasheet. https://metone.com/wp-content/uploads/pdfs/npm-2-network-particulate-monitor.pdf. Accessed: 2023-09-20.

[ODFHDS06] William Ocampo-Duque, Núria Ferré-Huguet, José L. Domingo, and Marta Schuhmacher. Assessing water quality in rivers with fuzzy inference systems: A case study. *Environment International*, 32(6):733–742, 2006.

[OFO+20] Opeyemi Omokungbe, Olusegun Gabriel Fawole, Oyediran Owoade, Olalekan Popoola, Roderic Jones, Felix Olise, Muritala Ayoola, Abiodun Olaitan, Adekunle Toyeje, Ayodele Olufemi, Sunmonu Lukman Ayobami, and Olawale Abiye. Analysis of the variability of airborne particulate matter with prevailing meteorological conditions across a semi-urban environment using a network of low-cost air quality sensors. *Heliyon*, 6:e04207, 06 2020.

[OPCa] Opc-n2 datasheet. https://parmex.com.mx/show_catalogue_pdf/142183/1. Accessed: 2023-09-20.

[OPCb] Opc-n3 datasheet. https://www.alphasense.com/wp-content/uploads/2019/03/OPC-N3.pdf. Accessed: 2023-09-20.

[ORC20] Tomasz Owczarek, Mariusz Rogulski, and Piotr O. Czechowski. Assessment of the equivalence of low-cost sensors with the reference method in measuring pm10 concentration using selected correction functions. *Sustainability (Switzerland)*, 12(13), 2020. All Open Access, Gold Open Access, Green Open Access.

[Org21] World Health Organization. What are the who air quality guidelines?, 2021. Accessed: 2024-11-23.

[PAB+16] Fabrício José Pontes, GF Amorim, Pedro Paulo Balestrassi, AP Paiva, and João Roberto Ferreira. Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing*, 186:22–34, 2016.

[PC24] European Parliament and Council. Directive (eu) 2024/2881 of the european parliament and of the council of 23 october 2024 on ambient air quality and cleaner air for europe (recast), 2024. Accessed: 2025-01-07.

[PDTD24] Van The Pham, Tuyet Anh Thi Do, Hau Duc Tran, and Anh Ngoc Thi Do. Classifying forest cover and mapping forest fire susceptibility in dak nong province, vietnam utilizing remote sensing and machine learning. *Ecological Informatics*, 79:102392, 2024.

[PGG16] Kanchan Prasad, Amit Kumar Gorai, and Pramila Goyal. Development of anfis models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmospheric Environment*, 128:246–262, 2016.

[PK13] Frank A Pouw and Mila Kwiatkowska. An overview of fuzzy-logic based approaches to ecology: Addressing uncertainty. In *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, pages 540–545. IEEE, 2013.

[PMSa] Pms1 datasheet. https://www.shinyei.co.jp/stc/eng/products/optical/pm2.html. Accessed: 2023-09-20.

[PMSb] Pms1003 datasheet. https://www.aqmd.gov/docs/default-source/aq-spec/resources-page/plantower-pms1003-manual_v2-5.pdf. Accessed: 2023-09-20.

[PMSc] Pms3003 datasheet. https://download.kamami.pl/p563980-PMS3003%20series%20data%20manual_English_V2.5.pdf. Accessed: 2023-09-20.

[PMSd] Pms5003 datasheet. https://www.digikey.jp/htmldatasheets/production/2903006/0/0/1/PMS5003-Series-Manual.pdf. Accessed: 2023-09-20.

[PMSe] Pms7003 datasheet. https://www.espruino.com/datasheets/PMS7003.pdf. Accessed: 2023-09-20.

[POP] Pops. https://handixscientific.com/pops/. Accessed: 2023-09-20.

[PVK+23] M. Y. Patel, P. F. Vannucci, J. Kim, W. M. Berelson, and R. C. Cohen. Towards a universal hygroscopic growth calibration for low-cost $pm_{2.5}$ sensors. *EGUsphere*, 2023:1–14, 2023.

[PVK+24] M. Y. Patel, P. F. Vannucci, J. Kim, W. M. Berelson, and R. C. Cohen. Towards a hygroscopic growth calibration for low-cost $pm_{2.5}$ sensors. *Atmospheric Measurement Techniques*, 17(3):1051–1060, 2024.

[PYPL21a] Donggeun Park, Geon-Woo Yoo, Seong-Ho Park, and Jong-Hyeon Lee. Assessment and calibration of a low-cost pm2. 5 sensor using machine learning (hybridlstm neural network): Feasibility study to build an air quality monitoring system. *Atmosphere*, 12(10):1306, 2021.

[PYPL21b] Donggeun Park, Geon-Woo Yoo, Seong-Ho Park, and Jong-Hyeon Lee. Assessment and calibration of a low-cost pm2.5 sensor using machine learning (hybridlstm neural network): Feasibility study to build an air quality monitoring system. *Atmosphere*, 12(10), 2021.

[PYPL21c] Donggeun Park, Geon-Woo Yoo, Seong-Ho Park, and Jong-Hyeon Lee. Assessment and calibration of a low-cost pm2.5 sensor using machine learning (hybridlstm neural network): Feasibility study to build an air quality monitoring system. *Atmosphere*, 12(10), 2021.

[Rah20] Md Anisur Rahman. Improvement of rainfall prediction model by using fuzzy logic. *American Journal of Climate Change*, 9(4):391–399, 2020.

[RAKB18] Sanjay Rajagopalan, Sadeer G. Al-Kindi, and Robert D. Brook. Air pollution and cardiovascular disease: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 72(17):2054 – 2070, 2018. All Open Access, Bronze Open Access.

[RAM+99] L. Willard Richards, Siana H. Alcorn, Charles McDade, Tiina Couture, Douglas Lowenthal, Judith C. Chow, and John C. Watson. Optical properties of the san joaquin valley aerosol collected during the 1995 integrated monitoring study. *Atmospheric Environment*, 33(29):4787 – 4795, 1999.

[RB20] Mariusz Rogulski and Artur Badyda. Investigation of low-cost and optical particulate matter sensors for ambient monitoring. *Atmosphere*, 11(10), 2020. All Open Access, Gold Open Access.

[Say21] Hoseyn Sayyaadi. Chapter 8 - real-time optimization of energy systems using the soft-computing approaches. In Hoseyn Sayyaadi, editor, *Modeling, Assessment, and Optimization of Energy Systems*, pages 479–527. Academic Press, 2021.

[SCC+17] Jingjin Shi, Feier Chen, Yunfei Cai, Shichen Fan, Jing Cai, Renjie Chen, Haidong Kan, Yihan Lu, and Zhuohui Zhao. Validation of a light-scattering pm2.5 sensor monitor based on the long-term gravimetric measurements in field tests. *PLoS ONE*, 12(11), 2017. All Open Access, Gold Open Access, Green Open Access.

[SCT+14] Sutyajeet Soneja, Chen Chen, James M. Tielsch, Joanne Katz, Scott L. Zeger, William Checkley, Frank C. Curriero, and Patrick N. Breysse. Humidity and gravimetric equivalency adjustments for nephelometer-based particulate matter measurements of emissions from solid biomass fuel use in cookstoves. *International Journal of Environmental Research and Public Health*, 11(6):6400 – 6416, 2014. All Open Access, Gold Open Access, Green Open Access.

[SDS] Sds011 datasheet. https://cdn-reichelt.de/documents/datenblatt/X200/SDS011-DATASHEET.pdf. Accessed: 2023-09-20.

[SG16] T. Sheehan and M. Gough. A platform-independent fuzzy logic modeling framework for environmental decision support. *Ecological Informatics*, 34:92–101, 2016.

[SGS+21] T. Siciliano, R. Giua, M. Siciliano, S. Di Giulio, and A. Genga. The morphology and chemical composition of the urban pm10 near a steel plant in apulia determined by scanning electron microscopy. source apportionment. *Atmospheric Research*, 251, 2021.

[Si19] Minxing Si. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine learning methods, October 5 2019.

[Sku13] Annett Skupin. *Optische und mikrophysikalische Charakterisierung von urbanem Aerosol bei (hoher) Umgebungsfeuchte*. PhD thesis, Verlag nicht ermittelbar, 2013.

[SPS] Sps30 datasheet. https://sensirion.com/media/documents/8600FF88/616542B5/Sensirion_PM_Sensors_Datasheet_SPS30.pdf. Accessed: 2023-09-20.

[SSC19] Shwetank, Suhas, and Jitendra Kumar Chaudhary. Estimation of groundwater contamination using fuzzy logic: A case study of haridwar, india. *Groundwater for Sustainable Development*, 8:644–653, 2019.

[Str17] Norbert Streibl. Influence of humidity on the accuracy of low-cost particulate matter sensors. *Techn. Ber. Technical report. DOI*, 10, 2017.

[SXDD20] M. Si, Y. Xiong, S. Du, and K. Du. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. *Atmospheric Measurement Techniques*, 13(4):1693–1707, 2020.

[TTK+22] Roman Trach, Yuliia Trach, Agnieszka Kiersnowska, Anna Markiewicz, Marzena Lendo-Siwicka, and Konstantin Rusakov. A study of assessment and prediction of water quality index using fuzzy logic and ann models. *Sustainability*, 14(9), 2022.

[TZW+21] Yulu Tian, Lingnan Zhang, Yang Wang, Jinxi Song, and Haotian Sun. Temporal and spatial trends in particulate matter and the responses to meteorological conditions and environmental management in xi'an, china. *Atmosphere*, 12(9):1112, 2021.

[Uni24] United States Environmental Protection Agency. Particulate matter (pm) basics, 2024. Accessed: 2024-12-29.

[UOY+20] Sayako Ueda, Kazuo Osada, Makiko Yamagami, Fumikazu Ikemori, and Kunihiro Hisatsune. Estimating mass concentration using a low-cost portable particle counter based on full-year observations: Issues to obtain reliable atmospheric pm 2.5 data. *Asian Journal of Atmospheric Environment (AJAE)*, 14(2), 2020.

[VAV+21] Jamie I Verhoeven, Youssra Allach, Ilonca C H Vaartjes, Catharina J M Klijn, and Frank-Erik de Leeuw. Ambient air pollution and the risk of ischaemic and haemorrhagic stroke. *The Lancet Planetary Health*, 5(8):e542–e552, 2021.

[VEC+23] Edwin Villanueva, Soledad Espezua, George Castelar, Kyara Diaz, and Erick Ingaroca. Smart multi-sensor calibration of low-cost particulate matter monitors. *Sensors*, 23(7), 2023.

[VPSP+23] Martine Van Poppel, Philipp Schneider, Jan Peters, Sinan Yatkin, Michel Gerboles, Christina Matheeussen, Alena Bartonova, Silvije Davila, Marco Signorini, Matthias Vogt, Franck René Dauge, Jøran Solnes Skaar, and Rolf Haugen. Senseurcity: A multi-city air quality dataset collected for 2020/2021 using open low-cost sensor systems. *Scientific Data*, 10(1), 2023. All Open Access, Gold Open Access.

[W+21] WHO et al. Who global air quality guidelines: particulate matter (pm2. 5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary. *World Health Organization*, 2021.

[WDWL19] Yanwen Wang, Yanjun Du, Jiaonan Wang, and Tiantian Li. Calibration of a low-cost pm2. 5 monitor using a random forest model. *Environment international*, 133:105161, 2019.

[Wei05] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[WOL+21a] Wan-Sik Won, Rosy Oh, Woojoo Lee, Sungkwan Ku, Pei-Chen Su, and Yong-Jin Yoon. Hygroscopic properties of particulate matter and effects of their interactions with weather on visibility. *Scientific reports*, 11(1):16401, 2021.

[WOL+21b] Wan-Sik Won, Rosy Oh, Woojoo Lee, Sungkwan Ku, Pei-Chen Su, and Yong-Jin Yoon. Hygroscopic properties of particulate matter and effects of their interactions with weather on visibility. *Scientific Reports*, 11(1), 2021. All Open Access, Gold Open Access.

[Wor24] World Health Organization. Sustainable development goals and air pollution, 2024. Accessed: 2024-10-23.

[XWZ+22] Yueguang Xue, Liuxiang Wang, Yiming Zhang, Yuliang Zhao, and Ying Liu. Air pollution: A culprit of lung cancer. *Journal of Hazardous Materials*, 434:128937, 2022.

[Zad65] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[Zad78] L.A Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28, 1978.

[ZDL+19] Furong Zhu, Rui Ding, Ruoqian Lei, Han Cheng, Jie Liu, Chaowei Shen, Chao Zhang, Yachun Xu, Changchun Xiao, Xiaoru Li, Junqing Zhang, and Jiyu Cao. The short-term effects of air pollution on respiratory diseases and lung cancer mortality in hefei: A time-series analysis. *Respiratory Medicine*, 146:57–65, 2019.

[ZFS96] Lotfi A Zadeh, K.S. Fu, and M. Shimura. An introduction to fuzzy logic applications in intelligent systems. *Proceedings of the IEEE*, 83(3):345–377, 1996.

[Zhi21] Petar Zhivkov. Optimization and evaluation of calibration for low-cost air quality sensors: Supervised and unsupervised machine learning models. page 255 – 258, 2021. All Open Access, Gold Open Access.

[ZPBdlC+23] A. Zafra-Pérez, C. Boente, A. Sánchez de la Campa, J.A. Gómez-Galán, and J.D. de la Rosa. A novel application of mobile low-cost sensors for atmospheric particulate matter monitoring in open-pit mines. *Environmental Technology and Innovation*, 29, 2023. All Open Access, Gold Open Access, Green Open Access.

[ZSG+20a] Marina Zusman, Cooper S. Schumacher, Amanda J. Gassett, Elizabeth W. Spalt, Elena Austin, Timothy V. Larson, Graeme Carvlin, Edmund Seto, Joel D. Kaufman, and Lianne Sheppard. Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. *Environment International*, 134:105329, 2020.

[ZSG+20b] Marina Zusman, Cooper S. Schumacher, Amanda J. Gassett, Elizabeth W. Spalt, Elena Austin, Timothy V. Larson, Graeme Carvlin, Edmund Seto, Joel D. Kaufman, and Lianne Sheppard. Calibration of

low-cost particulate matter sensors: Model development for a multi-city epidemiological study. *Environment International*, 134, 2020. All Open Access, Gold Open Access, Green Open Access.

[ZWZK24] Agnieszka Ziernicka-Wojtaszek, Zbigniew Zuśka, and Joanna Kopcińska. Assessment of the effect of meteorological conditions on the concentration of suspended pm2.5 particulate matter in central europe. *Sustainability*, 16(11), 2024.