Politecnico di Torino

Master's degree in environmental and Land Management Engineering

Climate Change path

Data management and optimization in air pollutions field

Supervisor:
Professor Marina Clerico

Co-Supervisor:
Davide Gallione

Candidate:
Ghazaleh Janfada

2025-2026

Abstract

Air pollution is a significant environmental problem that impacts human health and the environment. A large amount of environmental and meteorological data is being generated on a continuous basis through monitoring systems and analysis software. However, this data is often hard to handle and analyze because of a lack of clarity on variable names and inefficient update mechanisms. The current thesis is concerned with data handling and optimization in the area of air pollution, with the objective of converting unorganized exported data into a systematically organized database.

The data set used for this research was extracted from the PDAnalyze software and contains air pollution as well as meteorological variables from 2019 to 2026. One of the initial aims of this research was to provide a clear definition of certain variables that are particle-related, specifically the X and M variables. The X variables were found to be particle size percentiles, while the M values correspond to the statistical mean particle size according to number distribution.

In an effort to optimize the process of updating the database, a MATLAB script was created to enable the automatic merging of new data into an existing Excel-based database. This optimized process ensured that there was less human error involved. The organized database also allowed for comparisons to be made between different seasons and years, which aided in the analysis of air pollution trends over time.

Lastly, this thesis concludes with a database design that is user-friendly and can be easily interpreted by other researchers. The findings of this study emphasize the significance of effective data management as an essential step in conducting accurate air quality analysis and future studies.

Acknowledgments

To the brave people of my homeland, Iran, who struggle for freedom with courage, hope, and resilience in a glorious sadness.

To all the hardworking women around the world who have been a source of inspiration for my strength and perseverance.

And to my family, whose unwavering support sustained me through the hardest times.

Table of Contents

Table of Figures

Chapter 1. Introduction

Air pollution is one of the most significant issues in the modern environment. It impacts human health, the environment, and climate. Various studies have confirmed that air pollutants can lead to respiratory diseases, cardiovascular diseases, and other severe health issues. In urban and industrial regions, air pollution is affected by traffic emissions, industrial processes, and atmospheric factors such as temperature, wind speed, and humidity. Therefore, it is necessary to monitor air pollution and atmospheric factors continuously to understand air pollution behavior and aid in the development of environmental protection strategies.

Technological advancements have increased the volume of environmental data obtained from monitoring stations and analysis software in recent years. These tools produce massive amounts of data regarding air pollutant concentration and particle size distribution. However, it is not sufficient to have the data. Data management and interpretation have a significant role in the scientific significance of the data. Unorganized data can result in confusion, inefficiency, and difficulties in future analysis. Hence, data management has become a basic need in environmental studies, particularly in air pollution.

One of the biggest difficulties in air pollution research is the complexity of particle-related data. Contemporary measuring equipment is capable not only of measuring mass concentrations like PM10 and PM2.5, but also particle size distribution and statistical parameters calculated from it. These parameters are usually exported from dedicated software, and their interpretation is not always clearly described. This can cause problems for researchers and practitioners in understanding and properly using these parameters. This can make the reusability of data and its reliability for long-term research questionable.

This thesis is dedicated to data management and optimization in air pollution research, using data exported from PDAnalyze software. This data set contains both meteorological and air pollution-related parameters and spans a time period from 2019 to 2025. The primary aim of this thesis is to turn raw exported data into a structured, understandable, and reusable database that can serve as a basis for future research and analysis. Rather than focusing solely on pollutant trends, the significance of data organization and clarification is highlighted in this thesis.

The first significant issue dealt with in this research is the explanation of certain parameters related to particles, called X and M values. These parameters were not well understood and could not be easily explained in the exported data set. By analytical analysis, the X parameters were found to be particle size percentiles. Percentiles are a measure of the diameter below which a given percentage of particles is present and are significant in understanding particle size distribution. The M parameters were found to be the mean particle size, statistically calculated from the number distribution. These parameters are crucial for proper scientific analysis, as particle size is a significant factor in determining atmospheric and health-related phenomena.

Another significant part of this thesis is the optimization of the database update process. Environmental monitoring is a continuous process that generates data over a long period of time. It is a time-consuming process to manually enter new data points into an existing database. To address this problem, a MATLAB script was created to automatically merge the new exported data from the software into the existing Excel database. Automation is particularly important for long-term monitoring systems, where data volume increases continuously.

The structured database also enables temporal analysis, and it is thus possible to compare air pollution data from different seasons and years. Seasonal variations of air quality are sometimes associated with variations in meteorological factors and emissions. For instance, winter seasons may have higher levels of air pollution due to the use of heating devices and a stable atmosphere, while summer seasons may be affected by photochemical reactions and higher temperatures. A well-structured database makes it possible to conduct such comparisons and enables the determination of trends between 2019 and 2025. Even though this thesis does not involve sophisticated statistical analysis, it offers the required database to facilitate such analysis in future studies.

Another aim of this research is to develop a database that can be easily utilized by other researchers or experts. Data sharing and reproducibility in environmental studies are becoming increasingly important. A dataset that is difficult to interpret or not well-documented loses much of its scientific significance. For this reason, particular emphasis was placed on the definition of variables, data consistency, and logical structuring. The parameters in the database were explained and structured in a manner that enables users to easily interpret them and understand their significance.

The overall aims of this thesis can be listed as follows:

- to interpret and explain the meaning of particle-related parameters (X and M values),
- to create an optimized and automated system for updating the database,
- to allow comparisons between different seasons and years,
- to create a structured and reusable database for future use.

The aims of this thesis cover both technical and application aspects of environmental data management. This thesis does not solely concentrate on the level of pollution but also emphasizes the significance of data quality, understanding, and accessibility.

This thesis consists of several chapters. The first chapter explains the background of the thesis and the motivation and aims of the research. The second chapter explains the data sources and the overall methodology for constructing and managing the database. The third chapter explains the database structure and optimization, including the automated update system. The fourth chapter explains the comparisons between seasons and years based on the structured dataset. Finally, the final chapter concludes with the thesis and provides recommendations for future improvements.

In conclusion, this research intends to show how effective data management can turn raw environmental data into a valuable resource for air quality research. By clarifying important variables, improving the update process, and creating a friendly database, this thesis will help make air pollution data more useful. As environmental issues continue to escalate, organized and well-structured data will be a crucial component in facilitating scientific research and informed decision-making in air quality management.

Chapter 2. Literature review

2.1. Introduction

The practice of managing data has evolved into a pillar of contemporary science and industry. Amidst large volumes of data, the ability to store, organize and retrieve data seamlessly and efficiently is a necessity. As per Dhudasia et al. (2021), managing data entails collecting, checking, storing, data protecting, and processing, involving making data accessible, reliable, and timely available to the user. In present day research, effective data management is crucial not only for reproducibility of a study, but for the transparency and accountability the study entails (Bernardo, 2024; Kanza, 2022).

Data has become one of the most important resources for science, industry and government in the last few decades. The collection, management and interpretation of large sets of data have transformed the creation of knowledge and the making of decisions (Dhudasia et al., 2021). Today's data management and controlling have evolved beyond simplistic record keeping to sophisticated management that guarantees the accuracy, accessibility and security of the data. This is of particular importance in the environmental monitoring, health care and financial sectors where large, varied and rapidly changing datasets are produced. The management of data, for the most part, is the driver of progress in the world today, tackling innovations, and shaping policies in multiple sectors.

More scholars acknowledge that data management involves more than a technical task. As Bernardo (2024) explains, data practices are instrumental to the governance, accountability, and enduring sustainability of information systems. Advanced databases can provide little dependable information to inform decisions if structural clarity and methodological precise are lacking. On top of organizational advantages, effective management of data promotes research reproducibility, helping to underpin the trust of the broader scientific community in the findings. Thus, the broadest and most positive consequences of managing data are the credibility, accountability, and positive impact that come from transparent information flows.

Meanwhile, the FAIR principles Findable, Accessible, Interoperable, and Reusable have become one of the most recognized international standards for the organization of datasets (Mitchell et al., 2021). These principles are concerned not just with the technical aspects of storage but also with the means of cross-discipline sharing, integration, and reusage of data. For example, environmental studies frequently require the integration of air quality data, weather records, and satellite imagery. In such cases, compliance with the FAIR principles becomes critical for effective joint work and large-scale research (Rosales et al., 2025). In practice, FAIR principles help transform data from isolated records into resources that can drive innovation and solve complex, cross-disciplinary challenges.

Technological progress has dramatically reshaped how data is handled. The growth of cloud computing, artificial intelligence, and the Internet of Things has given researchers powerful tools for real-time storage, analysis, and optimization of data streams (Zhang et al., 2024; Guo et al., 2024). While these technologies greatly expand the possibilities of data management, they also raise pressing challenges around privacy, interoperability, and computational efficiency. As a result, recent scholarships increasingly emphasize the need to balance innovation with ethical responsibility and practical feasibility. In many cases, the success of modern data systems depends not only on technical capacity but also on the frameworks put in place to ensure trust, fairness, and sustainability.

In summary, strong data management practices have become a fundamental requirement for modern research and policymaking. Whether it is ensuring transparency in academic studies or supporting predictive modelling in environmental sciences, the importance of managing data effectively continues to grow. The different approaches, coupled with the intricacy of datasets in the real world, offer both opportunities and challenges which require constant investigation and improvement (Huang, 2021; Zhou et al., 2025). Having recognized this, the next parts focusse on the fundamental ideas and principles with new methodological approaches and the applications of data management in a particular area, primarily environmental and air quality data. These viewpoints underscore the dual nature of data management: as a specialized discipline and as an important catalyst for advancement.

2.2. Data Management Concepts and Frameworks

2.2.1. FAIR Principles

Regarding to Maintainable principles, there is only one point to consider the dataset must be 'reusable', even for potential use cases, while 'interoperable' is tied to 'integrate'; that is, the system must 'integrate' with other systems with or without the other independent systems. Not 'interoperability' as the dataset structure. The explicit structure must 'inter operate' with other data sets. The sets must have some explicit connections that are 'interoperable across borders' not just 'interoperable'. It is a fine distinction to make the dataset usable when it is independently applied to other structured sets. Implementing the FAIR principles requires technical infrastructure as much as cultural change in institutions. Metadata standards, for example, are the foundation upon which datasets become findable and interoperable across platforms (Zhang et al., 2024). In environmental science, where data sources may be satellite imagery, IoT sensors, and ground stations, common metadata standards enable smooth integration and meaningful comparison. But reaching FAIR compliance is not a simple matter, as most organizations lack the technical acumen or governance in place to mandate interoperability at scale (Bernardo, 2024). Filling such gaps will require ongoing investment, collaboration, and learning so that FAIR becomes a reality and not an ideal.

Singh and Pandey (2023) introduce DMPFrame, a theoretical framework for building data management plans (DMPs) that are closely aligned with the FAIR principles. The framework strongly emphasizes the documentation of data by using standardized metadata from the outset, which supports better discoverability and accessibility. The authors note that many projects forget FAIR at the initial stage, which in most instances causes recurring errors and limited potential for reusing data. By applying DMPFrame, scientists are able to design the arrangement and movement of their data in a meticulous manner so that it is transparent, well-structured, and effective throughout the whole research cycle. Ultimately, this process enables data to be converted from stand-alone records into an assured, reusable asset that informs the broader scientific community.

2.2.2. Data Governance

Data governance refers to the processes, policies, and duties that guide management, security, and sharing of data in organizations (Bernardo, 2024). Effective governance guarantees data quality and accountability via precisely defined roles and methods for monitoring. If no governance exists, datasets may end up being inconsistent, replicated, or inaccurate, thereby hindering scientific advancement and organizational decision-making. As such, governance has become the cornerstone of sustainable information ecosystems, providing the basis for maintaining trust and maximizing value of information in the long term.

Current research shows that data governance is increasingly linked to automation and advanced analytics. For instance, Zhang et al. (2024) observes that machine learning techniques can automate some of the governance aspects, for instance, detecting data anomalies within data pipelines or automatically generating metadata. Similarly, Guo et al. (2024) illustrate that governance models integrated with optimization models allow organizations to balance the accessibility of data against quality assurance. These breakthroughs represent a step from rigid, rule-based governance towards adaptive, intelligent systems that can dynamically respond to changing data environments. Ultimately, this union of governance and automation improves reliability while making more responsive and scalable data management methods possible.

Marcucci, Gonzalez Alarcon, Verhulst, and Wullhorst (2023) outline a comparative review of some of the data governance models. In their analysis, they recognize some significant differences in the definition of roles, responsibilities, and data policies across industries and sectors. The study highlights that having clear policies in place regarding data ownership, access, and privacy is critical. These comparative observations are especially effective for environmental researchers because they can describe governance frameworks best suited for complex, multi-source projects like air quality monitoring. Organizations can apply these tailored measures in order to enhance compliance and efficiency in operations through an understanding of these differences.

Governance also plays a crucial role in addressing moral concerns in dealing with data management. Kanza (2022) asserts that every research undertaking relies not only on technical infrastructure but also on social responsibility to deal with information openly and ethically. The situation is particularly critical in environmental research where information directly influences public policy, and strong governance mechanisms provide confidence and accountability that will allow stakeholders to make decisions informed by results. Therefore, governance stands at the intersection of technology, policy, and ethics in a manner that ensures data practices are supportive of innovation and societal responsibility.

2.2.3. Data Management and Database Creation

The management of data has become an integral part of research on air pollution due to the constant generation of large quantities of complex data. An open air quality data platform has been developed to promote data sharing to support environmental research and public health studies (Rosales et al., 2025). This shows that there is a need to have a system of storing data to ensure consistency, accuracy, and accessibility of information. This also indicates that the use of standardized formats and metadata descriptions is vital to allow the sharing of data between different research groups. Data management also helps to ensure the transparency of research outcomes. Without proper management of data, the interpretation of pollution data may become fragmented and inaccurate. Effective database creation has been recognized as a prerequisite to ensure high-quality environmental research (Katzenstein & Etcheverry, 2025).

Recent literature also emphasized the application of FAIR principles— “Findable, Accessible, Interoperable, and Reusable”—as a framework for managing air quality data (Katzenstein & Etcheverry, 2025). This can ensure the long-term usability of data sets and foster cooperation between institutions and researchers. However, most studies have emphasized conceptual guidelines rather than practical approaches to creating databases from raw data sets produced by monitoring tools. In fact, the process of converting software-generated files to database format may not be explicitly addressed in some studies. This indicates the need to develop methodologies to create databases that specifically address the process of database construction in detail. This can also enhance the quality of data sets to some extent. Therefore, database construction should be an integral part of air pollution studies rather than an auxiliary process (Rosales et al., 2025).

2.2.4. Provenance and Traceability

Provenance is the meticulous record of a dataset's history and processing, tracing it from start to finish (Mitchell et al., 2021). This encompasses information regarding how data were gathered, processed, and analyzed, and what algorithms or models were used along the way. Provenance enhances traceability, making it possible for researchers to replicate results and validate conclusions with certainty. In large science endeavors, where datasets are reused across many investigations, provenance assures consistency and transparency in research results. Lastly, careful tracking of provenance not only builds trust in data but also facilitates more secure, stable, and responsible science.

Provenance is especially critical in computational workflows. In air pollution modelling, for instance, minor differences in pre-processing data or parameter choices can have a profound impact on final predictions (Lu et al., 2020; Huang, 2021). Capturing high-fidelity provenance allows researchers to track differences, determine the causes of error, and improve reproducibility. Provenance also facilitates collaboration, with groups within institutions able to understand how datasets were produced and modified before they were combined. In this case, provenance and traceability are not only technical guardians but rather integral tools in the building of scientific credibility and ascertaining that research findings are accurate and intelligible in context.

Marcucci et al. (2023) highlights the importance of standardization of data traceability. While provenance is an integral component of data governance in most models, its implementation varies widely across disciplines. Social science research prefers confidentiality, environmental studies reproduce and are resilient, and so on. The variety highlights an essential gap: the need for inter-disciplinary standards that allow for uniform and reliable traceability. Having such standards in place would not only facilitate comparability and integration of data across fields but also raise the general credibility and usefulness of science data.

Data management is one of the cornerstones of science and business today. In the age of big data, being able to store, organize, and access information efficiently is no longer a luxury it's a necessity. Data management, according to Dhudasia et al. (2021), involves activities like collecting, validating, storing, protecting, and processing data so that it will be available, reliable, and on hand when it's needed. In modern research, good data management is not just vital to allow studies to be reproducible, but also for accountability. Transparency (Bernardo, 2024; Kanza, 2022). In recent decades, data has become one of the most valuable resources for science, industry, and government. The ability to collect, structure, and analyze large datasets has transformed the way knowledge is generated, and decisions are made (Dhudasia et al., 2021). Data management is far from the old days of simple record-keeping, and it is now underpinned by high-end processes that guarantee accuracy, accessibility, and security. Such a transition is especially critical in domains where enormous, diverse, and constantly changing datasets are produced, such as environmental monitoring, medicine, and finance. In many ways, effective data management is now the backbone of progress, driving innovation and policy across sectors.

2.2.5. Integrated Frameworks

Combining conceptual frameworks like DMPFrame with data governance frameworks offers a more holistic approach towards data management. Singh and Pandey (2023) identify that DMPFrame can be utilized to integrate governance policies specific to specific organizations, ensuring workflows to be structured and aligned. Marcucci et al. (2023) suggest in this regard that tools for comparison and benchmarking must be employed to enable organizations to select the most suitable framework. Through this synergy of strategies, organizations may achieve transparent governance and FAIR-aligned data practices at the same time, thereby enhancing their data systems' long-term sustainability, reliability, and quality. This synergy emphasizes bridging the gap between organizational strategy and technical design to maximize the impact of data management.

2.3. Data Management Methods

2.3.1. Database Systems

Database systems continue to be the backbone of data management infrastructures, with organized storage and querying efficiency. Relational databases, founded on SQL platforms, are defined by consistency, reliability, and widespread application in business and research environments (Smith & Kumar, 2022). However, increasing diversity and volume of datasets have created the need to apply NoSQL systems, which enable the management of unstructured and semi-structured data. This flexibility is especially crucial in fields like air quality monitoring, where data can range from text metadata to streams of real-time sensor inputs (Zhang et al., 2024). As datasets become more complex, hybrid database solutions that unify relational and NoSQL approaches are proving to be the solution to scalable, efficient, and adaptable data management.

Environmental data management usually relies on databases with capabilities to manage and process massive amounts of spatial data. Environmental data management systems (EDMS), according to the ITRC (2022), can be characterized by client-server databases, commercial off-the-shelf (COTS) products, or cloud models, each offering different levels of security, adoption, and cost. Client-server databases provide good data integrity and multiple-user support on a parallel basis, hence suitable for large projects. Desktop or spreadsheet systems are perfect for small projects but limited in scalability and data quality. Selecting the right database system requires serious consideration of project specifications, including the spatial data structure, size, and functional requirements, to ensure that it suits present and future needs in an effective manner (ITRC, 2022).

Whether relational or non-relational databases are used often depends on project requirements. Chen and Dubois (2021) note that relational systems are ideal for maintaining data integrity, while NoSQL systems offer the scalability needed in big data environments. Hybrid patterns of design are now common, combining structured relational centres with unstructured extras to handle mixed data types (Miller & Rossi, 2023). For instance, environmental data platforms can make use of relational databases to store official government records while using NoSQL systems to process large-scale sensor feeds. Hybrid design of this sort allows organizations to draw benefits from both database types while ensuring reliability, flexibility, and scalability in data environments that are complex.

2.3.2. Data Pipelines

Data pipelines are among the major systems to enable data to move smoothly from acquisition into transformation, storage, and analysis. Dhudasia et al. (2021) note that pipelines reduce manual effort, further preventing errors and enhancing reproducibility. Pipelines can, in environmental studies, automatically ingest air pollution data from IoT sensors, clean it out for missing values, and transfer it into cloud databases for further modeling (Temkov, 2025). This level of automation is especially important in real-time monitoring systems, where inconsistency or lag can nullify predictive accuracy. Pipelines also free researchers from wasting too much time on data movement and processing by minimizing data movement and processing.

Data pipeline design is a key part of ecological research in managing the flow of incoming data from sensors, laboratories, and government systems. Rosini et al. (2025) note that the use of standardized pipelines enhances data quality at reduced processing costs. Raw data from different sources are collected by these systems, cleaned, checked for consistency, and normalized and ready for analysis in databases or analytical models. Essentially, pipelines act as an intermediary layer, ensuring that data transmission from where it is created to final use without loss or unauthorized modification. Such structured workflows are most important in air quality projects, where harmonizing remote sensing data, IoT sensor streams, and lab measurements must be conducted for accurate modelling and decision-making.

Modern data pipelines increasingly integrate optimization and monitoring features to achieve peak efficiency as well as dependability. Guo et al. (2024) document pipelines that not only transfer data but also perform real-time predictive testing on air quality standards. Likewise, Kanza and Clark (2022) emphasize that provenance tracking can be incorporated into pipelines such that transparency and traceability are realized during all processing phases. With integration of automation, optimization, and governance, pipelines become smart workflows with FAIR principles and enabling seamless collaboration across institutions. These combined systems are becoming the standard for environmental research where accurate, timely, and reproducible data forms the basis for modelling as well as policy-making decisions.

2.3.3. Ontology-driven Methods

Ontology-based strategies are crucial for enabling interoperability across disparate datasets. Bernardo (2024) explains that ontologies provide shared vocabularies and semantic relations whereby disparate datasets can "speak the same language." In environmental science, such a strategy is most beneficial when merging government air quality reports, satellite, and crowdsourced IoT measurements (Rosales & Fernandez, 2025). Using ontologies, researchers are able to harmonize otherwise incommensurable data sources, improving the quality and interpretability of analyses. Eventually, ontology-based integration enables more holistic and realistic environmental evaluations, promoting collaboration within institutions and disciplines.

Zhang, Lin, and Wang (2024) illustrate how ontology-based frameworks not only improve dataset integration but also uncover hidden relationships in data. To illustrate, comparing meteorological and pollution data can add new insights to air quality seasonality patterns. Springer Insights (2021) describe how ontologies also facilitate advanced analytics by encapsulating domain knowledge in machine-processable forms. This avoids ambiguity and makes computational models built from integrated datasets more robust. By combining semantic alignment with analytical accuracy, ontology-guided methods enable scientists to extract more accurate, interpretable, and actionable conclusions from complex environmental data.

One of the solutions currently emerging in data management includes ontology-based models to organize and consolidate diversified sets of data. The ITRC (2022) further points out that challenges with unstructured data images, PDFs, or readings from the environment—can be addressed by conceptual structures. Such structures provide shared terms and relationships between data elements, making data exchange, querying, and interpretation more efficient. For example, employing an ontology for biodiversity or pollution data avoids confusion in terms such as "sample" or "site," which can have different meanings across research groups. Accordingly, ontologies have been found to be essential tools for providing consistency, traceability, and interoperability of data in environmental research so that datasets are merged and analyzed consistently across studies.

2.3.4. Machine Learning for Data Quality

Machine learning (ML) algorithms are being employed to improve data quality management. Guo and Li (2024) demonstrate how predictive models can be used to detect anomalies in environmental data sets automatically, so that any further analysis can be made more reliable. In the case of noisy or missing sensor data, ML-based imputation methods perform better than traditional statistical approaches (Huang, 2021). These methods not only achieve better accuracy but also significantly reduce the number of human hours required for data cleaning. Through the integration of ML in data management workflows, researchers can deliver quality datasets with additional time for analysis and decision-making.

Machine learning (ML) algorithms have increasingly been recognized as valuable tools for enhancing data quality in environmental programs. According to the EPA (2025) Guide to Air Quality Data Management, ML algorithms can detect measurement errors and estimate missing data with accuracy, increasing the validity of subsequent analysis. The methods also enable the establishment of anomalies and unusual patterns in datasets that would otherwise go unnoticed. By combining ML with traditional data management methods, researchers can process large and heterogeneous data streams with greater speed and certainty. Such applications are particularly important for urban-scale air quality monitoring, where timely and reliable forecasting is worth a lot for policy and public health intervention.

Beyond anomaly detection, machine learning (ML) techniques are also increasingly being used to automate data validation and standardization. Lu, Zhang, and Chen (2020) survey computational methods for estimating pollutant concentrations and conclude that ML models can more easily adapt to diverse and complex data inputs. Smith and Kumar (2022) note that ML-based solutions are now embedded directly in the majority of commercial and open-source data management systems. By delivering data quality at the earliest possible processing stages, ML not only boosts reliability but also improves the integrity and reproducibility of research workflows overall (Mitchell & Johnson, 2021). These capabilities make ML a vital component of modern environmental data management solutions.

2.3.5. Automated Database Management

Automation is becoming more significant in managing environmental data, particularly considering the increase in data volume and frequency. Programming tools, such as MATLAB and Python, are being utilized to process and organize data more efficiently. A near real-time system for monitoring data is developed to integrate data from different sources and update concentrations of pollutants, including PM2.5, automatically (Geng et al., 2021). This system reduces human intervention, minimizing the possibility of human error during data handling. The literature also indicates that one of the essential aspects of maintaining a reliable air quality database is automation, considering its importance for providing data updates more quickly. Scripting-based approaches are being considered more for environmental monitoring research studies (Gianquintieri et al., 2025).

Despite these advancements, the majority of existing studies emphasize the aspects of real-time data recovery and visualization rather than optimizing local database update procedures. Less emphasis has been placed on the process of systematically integrating exported data from monitoring software into a structured database. While automated pipelines are recognized, descriptions of the algorithms for database updates remain limited in the literature (Geng et al., 2021). This highlights a gap in the development of methods for fast database updating. Efficient database update strategies play a critical role in maintaining consistency between new acquired data and existing data. Without such methods, the database may become fragmented. Therefore, more research should be conducted to formalize automated database update procedures as part of data management frameworks (Katzenstein & Etcheverry, 2025).

2.4. Data Management in Environmental Studies

2.4.1. Air Quality Data Platforms

Air quality data platforms act as hubs for receiving, storing, and disseminating environmental data to scientists, policymakers, and the public. Rosales et al. (2025) note that these platforms give easier access to data and improve collaboration by aggregating data from different sources like government monitoring stations and research institutions. Zhou, Wang, and Liu (2025) point out that open-access platforms enhance transparency and ease comparative analysis between regions. Interoperable and long-term use require standardized data schemas and metadata frameworks, as the OpenAir Consortium (2023) posits. Beyond support for research, these platforms also play a critical role in informing decision-making at the regulatory level and for public health interventions, making high-quality, integrated data actionable across different levels.

Lu et al. (2020) illustrate how integrated data platforms improve air pollution modeling accuracy by combining historical observations with real-time data. Cloud storage systems enable the platforms to handle significant volumes of heterogeneous data affordably (Smith & Kumar, 2022). Rosales and Fernandez (2025) also highlight the importance of simple-to-use interfaces and APIs, which allow outside researchers and stakeholders to interact and access the data with ease. Despite these advantages, challenges remain in how data consistency and quality are maintained when integrating large numbers of heterogeneous datasets. Reducing these challenges is crucial in ascertaining that integrated platforms provide correct insights for research, policy-making, and public health utilization.

2.4.2. IoT and Sensor Networks

The rapid growth of Internet of Things (IoT) sensors has transformed environmental monitoring, especially in air quality monitoring. Temkov (2025) describes collections of distributed sensors that provide high-frequency samples of pollutants, temperature, and humidity in urban and rural spaces. These high-density deployments radically enhance spatial and temporal resolution, allowing detailed analysis of local pollution dynamics. Zhang et al. (2024) note that integrating IoT data into centralized systems must have standard protocols for data transmission, synchronization, and storage. Through seamless and trustworthy integration, IoT-enabled monitoring networks provide real-time and high-granular information to support study and policy intervention.

Crowdsensing methods, in which citizens report data using low-cost sensors or smartphones, are a valuable addition to traditional monitoring networks (Guo et al., 2024). Crowdsensing increases spatial coverage and provides near real-time data on variability of air quality. Kanza and Clark (2022) warn that variability and potential unreliability of measurements gathered from crowds require robust validation and filtering mechanisms. By using a blend of data from official monitoring sites, IoT networks, and citizen reports, platforms can establish a more comprehensive and detailed understanding of the environment. This hybrid approach both enhances the resolution and credibility of air quality determinations and enables more informed decisions to be made at both local and regional levels.

2.4.3. Particle Size Distribution

Particle size distribution (PSD) is frequently used in the characterization of airborne particulate matter, particularly in the study of the physical and environmental characteristics of the same. Percentile-based indicators such as X10, X50, and X90 are frequently used in the characterization of the distribution of particle sizes, particularly in the differentiation between fine and coarse particle fractions (Wang et al., 2024). These indicators have been found to offer a statistical representation of the distribution of the particles in the sample, thereby making it possible to make meaningful comparisons between the data obtained in the study. Furthermore, the use of the mean particle size (M) is frequently used in the representation of the characteristics of the particle population in the study of air pollution (Gianquintieri et al., 2025). The use of the indicators, particularly in the study of the characteristics of the particles, is found to be essential in the study of long-term monitoring, particularly in the organization of the data in the study (Gianquintieri et al., 2025).

Moreover, a number of research studies have revealed that PSD parameters show significant temporal variability, which is influenced by seasonal and meteorological factors. For instance, a greater number of smaller-sized particles are usually present during the winter period, as this is a result of increased combustion and reduced atmospheric dispersion (Wang et al., 2024). On the contrary, a greater particle size is usually present during the warmer period of the year, which is a result of resuspension and natural sources. The use of percentile-based parameters is more accurate for detecting this variability than when using mass concentration parameters. The mean particle size (M) also indicates a change in particle size, as revealed by previous research studies, which state that when PSD parameters are analyzed, a better understanding of pollution dynamics is acquired, making assessments more reliable (Wójcik-Gront & Gozdowski, 2025). This confirms the significance of clearly understanding and defining X and M parameters, as revealed by environmental data.

2.4.4. Integration with Models

The integration of environmental datasets and predictive models is the most important step towards providing actionable information. Lu, Zhang, and Chen (2020) show that integrating ground-based measurements and statistical and machine learning models enhances the accuracy of predictions of pollutant concentrations. Mitchell and Johnson (2021) highlight that it is important to maintain provenance during the integration process, tracking the complete processing history for each dataset. This traceability enables scientists to assess the credibility of predictions and to diagnose sources of uncertainty in model outputs. By linking quality data to robust predictive frameworks, environmental science can deliver more precise, transparent, and credible information for decision-making and policy interventions.

Huang (2021) notes that complex models, including hybrid physical-statistical schemes, rely on consolidated, high-quality data from diverse sensors and platforms. Errors or data gaps can be propagated via models, which can lead to imperfect predictions. Chen and Dubois (2021) emphasize that satellite imagery merged with ground-based sensors improves spatial resolution along with predictive capability. Optimizing data processing workflows ensures that real-time or near-real-time predictions are still achievable, even in the face of massive, heterogeneous datasets. Taken together, these approaches illustrate the critical function of rigorous data management in reliable and actionable environmental modelling.

Collectively, the literature suggests that environmental data management extends far beyond data collection to include integration, validation, and analysis. Air quality platforms, IoT networks, and predictive modelling are interdependent components that, in combination, support accurate monitoring and informed policy development through decision-making. By leveraging heterogeneous datasets, harmonized metadata schemes, and computationally efficient methods, researchers can extract actionable intelligence on air pollution processes. Such integrated approaches end up strengthening evidence-based decision-making for environmental and public health protection, rendering data-driven interventions both effective and reliable (Rosales et al., 2025; Zhou et al., 2025; Temkov, 2025).

2.5. Data Optimization

2.5.1. Workflow and Pipeline Optimization

Data optimization focuses on enhancement of efficiency, accuracy, and usability of datasets at levels of storage, processing, and analysis. Guo et al. (2024) highlight that optimization techniques reduce computational costs without decreasing predictive performance in environmental models. Optimization becomes important in large datasets, such as those generated by air quality monitoring networks, to deal with high-frequency measurements and heterogeneous streams of data. Through workflow and data structure optimization, businesses can minimize redundancy and improve access times for real-time as well as retrograde analyses. Workflow optimization, in particular, ensures smooth linear flow of data from acquisition, cleaning, transforming, and modelling processes, avoiding bottlenecks and allowing quicker, more precise insights.

Zhang et al. (2024) note that pipelines for data could be designed to provide high priority to computationally intensive activities, thread processing, and dynamic allocation of computing resources. These methods enable effective management of complex environmental data sets, delivering timely access to reliable information to decision-makers. Similarly, Huang (2021) illustrates how automated workflow re-tuning using real-time performance metrics can highly boost data throughputs in monitoring systems. Collectively, these optimization techniques enhance the responsiveness and scalability of environmental data management infrastructures, enabling faster, more trustworthy analysis and decision-enabling intelligence.

Real-time optimization is becoming ever-more vital for managing dynamic data sets. Zhou, Wang, and Liu (2025) demonstrate how continuous analysis of input data streams allows real-time automated realignment of processing pipelines. For example, anomalies or sudden spikes in pollutant concentrations can trigger high-priority execution of critical tasks or alerts to be raised with respective authorities. Installation comes with complex algorithms that balance response speed against computation efficiency. By enabling proactive management of environmental data, real-time optimization ensures that monitoring systems can enable timely, accurate, and actionable decision-making.

Mekouar, Lahmer, and Karim (2025) examine how the choice of data processing tools in technologies such as Pandas, Polars, and PySpark can affect the efficiency of data pipelines. Their work shows that properly conducted pipeline optimization not only saves time for processing but also minimizes energy consumption and maximizes the use of computing resources. This is particularly so in environmental projects, where air quality or climate change data are large, complicated, and diverse. The authors note that the selection of the most suited framework depends on the type of the data (structured or unstructured) and on the aims of the analysis. Pipeline optimization is thus a key tactic for attaining efficient, sustainable, and reliable environmental data management.

One of the significant conclusions of Mekouar et al. (2025) is that columnar data is handled efficiently rapidly by Polars, whereas PySpark is very appropriate for big data applications with scalability. These analyses prove that one tool is not suitable for all scenarios, and the best approach usually results from applying an ensemble of tools. Referring to air quality data optimization, this would imply that the consideration of platform selection should take data volume, real-time processing needs, and computational capacity into account. Hence, data pipelines must be dynamic and versatile in character so they could adapt to varying states of data and ensure efficient, reliable, and scalable environmental data handling.

Mekouar et al. (2025) ended up finding that Polars processes columnar data much faster, and PySpark is suitable for scalability on big data projects at a large scale. This finding reveals that no single tool is optimal for every scenario, and having multiple tools combined typically results in the overall best performance. In air quality data optimization, platform selection should consider the amount of data, whether real-time handling is required, and how much computer resource is available. Thus, pipelines should be elastic and dynamic in nature, capable of changing to accommodate changing data conditions without compromising on efficiency, reliability, and scalability.

2.5.2. Storage Optimization

Data compression and optimization are critical for efficient management of large and complex datasets. Lu, Zhang, and Chen (2020) demonstrate that through appropriate encoding and indexing techniques, storage cost can be significantly reduced without compromising data integrity. In environmental research, where historic data often spans decades, such techniques facilitate long-term storage of data while making it available for analysis. Additionally, Rosales and Fernandez (2025) note that cloud-based architectures also optimize storage efficiency by scaling computational and storage resources elastically. Collectively, these strategies ensure environmental datasets are both manageable and readily accessible for research, modelling, and decision-making.

2.5.3. Model Optimization

Predictive model optimization is an integral component of overall data optimization. Guo and Li (2024) illustrate how techniques such as hyperparameter tuning and feature selection can increase model precision while reducing computational demands. The integration of these optimization procedures with data preprocessing monitoring guarantees that models are fed with consistent, quality inputs, minimizing the propagation of errors. Mitchell and Johnson (2021) highlight that keeping track of provenance in optimizing models ensures transparency, with every modification being traced, recorded, and verified. These integrated optimization methods ensure the reliability, efficiency, and interpretability of predictive environmental research analysis.

A paper by Application of AI in Air Pollution Monitoring and Forecasting (2025) points out that deep learning and machine learning techniques have greatly enhanced the accuracy of air quality forecasting models. The authors note that large-scale, multi-source optimized models have the capability to replicate pollutant behavior at greater temporal and spatial resolution. One very effective approach is the use of hybrid neural networks that integrate ground sensor data, satellite observations, and other environmental readings. Optimization of models not only improves forecast accuracy but also facilitates sustainable environmental management by allowing stronger information to be employed in policymaking and pollution mitigation programs.

The same systematic review (2025) points out that model optimization is more than hyperparameter tuning to include feature selection and dimensionality reduction. These processes not only accelerate model training but also enhance models' generalizability in new and unseen situations. In air quality research, these techniques have boosted the accuracy of forecasts for pollutants like ozone and PM2.5. This stresses that model-level data optimization is a central component of modern environmental data management, enabling predictive models to be effective and robust in facilitating research and policy formulation.

2.5.4. Sensor Network Optimization

Network topology and sensor deployment are critical optimization issues in environmental monitoring. Temkov (2025) emphasizes that strategically deploying sensors based on patterns of pollutant dispersion will improve coverage space and reduce the cost of deployment and operation. Zhang, Lin, and Wang (2024) also point out the benefits of adaptive sensor networks, which change sampling frequency dynamically according to environmental conditions and thereby improve data quality and resource use. These approaches illustrate the necessity to integrate optimization strategies at data and hardware levels such that monitoring systems are not only accurate and reliable but also economical and responsive to evolving conditions.

Sensor network optimization is a critical element in environmental data management. Temkov (2025) demonstrates that careful selection of sensor locations based on pollutant distribution patterns can achieve maximal space coverage at reduced installation and maintenance costs. Spatial optimization methods can identify the best location for sensors, enhancing data quality and model accuracy of air quality forecast models. Zhang, Lin, and Wang (2024) also highlight the benefits of adaptive sensor networks with varying sampling frequency as a function of environmental settings, further improving data quality and resource efficiency. These mechanisms underscore the importance of integrating optimization at both hardware and data levels to ensure monitoring systems are accurate, efficient, and cost saving.

Sensor placement and network configuration are major optimization issues in environmental monitoring. Temkov (2025) suggests that sensor placement based on pollutant dispersion patterns has the capability to increase spatial coverage without increasing installation and operating expenses. Zhang, Lin, and Wang (2024) mention employing dynamic and adaptive sensor networks that modify sampling rates according to environmental conditions. For instance, during sudden bursts in the concentration of pollutants, the sensors can generate data at higher frequency, providing more accurate inputs for modelling. This frequency-based optimization improves data accuracy with optimally maintained resource utilization and thus sensor network optimization is an integral aspect of modern air quality monitoring systems along with other advanced data management strategies.

2.5.5. Data Integration Optimization

Environmental data management optimization also encompasses the integration of different datasets. Lu et al. (2020) demonstrate that data coming from disparate sources need to be integrated using harmonization procedures to reduce inconsistency, duplication, and errors. Kanza and Clark (2022) note that ontology-based alignment and semantic standardization boost interoperability even further so that datasets from different platforms and formats can be used together seamlessly. Optimizing data integration processes helps researchers attain both reliable and scalable large-scale analyses, which can support more precise modelling, forecasting, and decision-making in environmental monitoring.

The study Big Data in Environmental Quality Monitoring and Policy Development (2025) applied bibliometric analysis to show that growing use of big data is revolutionizing environmental policymaking. The authors highlight that streamlined and well-analyzed datasets allow policymakers to make data-based decisions with more confidence. This is a testimony that data optimization is not just a technical step but a strategic step in transforming raw data into useful insights. With bridging the gap between research and policy, effective data optimization ensures environmental monitoring becomes feasible, evidence-based, and timely action.

Briefly, environmental data optimization encompasses a broad spectrum of techniques from workflow and storage improvement to predictive modelling, network and sensor placement, and setup. These techniques enhance efficiency in operations, reduce computational and resource costs, and optimize the accuracy and reliability of environmental data. On all applications, optimization strengthens the connection among raw data and valuable information, enabling researchers, policymakers, and public health practitioners to make improved, timely, and effective decisions (Guo et al., 2024; Zhang et al., 2024; Huang, 2021; Temkov, 2025).

2.6. Data on Air: Organization and Access

2.6.1. Data Accessibility and Infrastructure

Accessible data infrastructures are key to effective environmental dataset management. The OpenAir Consortium (2023) elucidates that well-structured systems with uniform access protocols make it easy for researchers and policymakers to access air quality data, enabling them to easily get, analyses, and utilize the data quickly. Rosales et al. (2025) further explain that central platforms with uniform APIs enhance interoperability, with diverse datasets from monitoring stations, satellites, and IoT sensors readily integrated. Through the provision of an organized and accessible data space, such infrastructures enable collaborative research as well as evidence-based policy making in environmental monitoring.

Smith and Kumar (2022) highlight that cloud storage systems enable elastic management of environmental data sets so that data size can expand without compromising access. This guaranteed access not only optimizes research productivity but also promotes transparency and collaboration between institutions. By providing real-time access to heterogeneous data sets, cloud infrastructures support analyses integrating disparate data sets, enhance cross-institutional collaboration, and enable evidence-based decision-making in environmental monitoring and policy formulation.

Metadata and documentation are required to make datasets findable, accessible, and reusable. Zhou, Wang, and Liu (2025) argue that well-structured metadata allows users to understand the extent, limitation, and quality of datasets at the moment before using them. Guo et al. (2024) prove that platforms with automated metadata generation can reduce errors, enhance traceability, and ease data management processes. By providing accurate, consistent documentation, metadata enables environmental datasets to be comprehensible, reliable, and suitable for research, policy, and application.

Rosales and Fernandez (2025) point out that standard documentation practices are needed to support the strengthening of FAIR principles, making data findable, accessible, interoperable, and reusable in the long run. Together with good infrastructure, high-quality metadata is the keystone of an organized and efficient system of air quality data. Such combined practice not only preserves the long-term usefulness of datasets but facilitates collaboration, data sharing, and reliable analyses among research centres and policy agencies.

ITRC (2022) emphasizes the need for robust and modern technical infrastructure as a cornerstone to maintaining access to environmental data. In their report, they state that it is still common for most organizations to employ centralized and legacy-based architectures that limit scalability, remote access, and interoperability. These issues can be resolved by embracing distributed and cloud-based platforms that will provide data to more people. Such infrastructures not only increase technical efficiency but also enable higher transparency, reliability, and trust within environmental data management and use.

Rosales et al. (2025) highlight the significant role open data platforms play as air quality management. They demonstrate how it becomes feasible for researchers, policymakers, and even citizens to utilize and access environmental data efficiently by creating platforms with user-friendly interfaces and standardized APIs. These platforms facilitate a cooperative ecosystem that supports rapid information exchange, inter-institutional partnerships, and scientific innovation. The authors suggest that establishing open, trusted, and reliable infrastructures is a key requirement for effective air pollution monitoring, management, and evidence-based decision-making.

2.6.2. Interoperability and Policy Integration

Interoperability of heterogeneous environmental datasets is required for integrated analysis and policymaking. Temkov (2025) demonstrates that aligning IoT sensor data with official monitoring networks requires standard formats and protocols to prevent inconsistencies and facilitate trustworthy integration. Chen and Dubois (2021) highlight that semantic frameworks and ontology-based approaches further improve the compatibility of heterogeneous datasets, allowing integration of data from diverse sources in a seamless manner. Through enabling large-scale modelling and decision-making collaboration, these methods improve the reliability, scalability, and actionable usefulness of environmental data systems.

Interoperability and accessibility of data have a direct impact on environmental policy effectiveness. Huang (2021) illustrates that real-time access to standardized datasets enables authorities to respond quickly to environmental pollution events, simplify regulatory interventions, and issue early public warnings. The OpenAir Consortium (2023) also emphasizes that open data sharing also promotes public trust and evidence-based policy. By linking policy structures to properly structured and interoperable data sets, environmental management becomes proactive rather than reactive, showing the critical necessity for structured, accessible, and interoperable air quality data systems to support decision-making.

Moumtzidou, Vrochidis, and Kompatsiaris (2016) speak of combining multimodal data sources—like satellite imaging, IoT sensor data, and meteorological data—to develop improved air quality estimation. One important challenge in doing this, they note, lies in the heterogeneity of data formats, resolutions, and reporting standards. As a countermeasure, the authors propose an integration framework that is capable of fusing both structured and unstructured datasets. Not only does this system enhance the accuracy and resolution of air quality predictions, but it also allows the functional operability of the information for use in environmental policy development, enabling more informed and effective decision-making.

Moumtzidou et al. (2016) demonstrate that the availability of multi-source data can overcome the limitations of traditional air pollution monitoring networks. They explain that data fusion from satellites, IoT sensors, and meteorology provides much improved spatial and temporal coverage, particularly in regions with limited monitoring stations. The authors also emphasize that the establishment of common standards for documentation, data sharing, and interoperability is of paramount significance to the effectiveness of these integrated systems. Ultimately, improving the accessibility, structuring, and harmonization of multi-source datasets enables policymakers and environmental managers to adopt more targeted and effective preventive policy and regulation measures.

Moumtzidou, Vrochidis, and Kompatsiaris (2016) note that a primary obstacle in exploiting multi-source air quality data is the lack of standardization in data exchange. They note that IoT sensor data, satellite data, and monitoring station data are typically maintained in heterogeneous formats, and their fusion becomes challenging without the support of an interface infrastructure. To surpass this, ontology-based and standard metadata solutions have been proposed to enable greater interoperability and simple data fusion. Such harmonization not only renders environmental datasets more homogeneous and simpler to utilize but also provides a solid foundation for evidence-based policymaking and more effective air quality management.

In democratizing access to air pollution data, Berrisford and Menezes (2024) point out that interoperability of data ought to be accompanied by open data policy. They suggest an open-source Python package that aggregates air pollution data from diverse sources and presents it in a unified, standardized format. The software provides policymakers and researchers with easy access to relevant datasets, comparative analysis, and actionable findings. The authors conclude that the coordination of interoperability with open data policies promotes transparency, increases public confidence, and enables evidence-based decision-making in air quality management.

2.6.3. Seasonal and Interannual Comparison of Air Pollution Data

For instance, seasonal comparison is often applied to examine temporal variations of particle concentrations and size distributions. Many research studies have shown that there is an increase in pollution concentrations during the winter period, attributed to increased heating activities and atmospheric stability (Wang et al., 2024). Conversely, summer months are characterized by lower concentrations, yet greater influences of meteorological variables, such as temperature and wind speed, are observed. The seasonal comparison is essential for determining the main emission source and environmental factors of air pollution. This technique is also applied to examine short-term variations of air pollution. Through seasonal trends, more understanding is gained about the processes governing particle formation and dispersion. As a result, seasonal comparison is considered a vital technique for evaluating air pollution (Wójcik-Gront & Gozdowski, 2025).

Moreover, interannual analysis is a further extension of time-series analysis, as this technique is concerned with the investigation of long-term variations in air quality over a period of time. This technique is also helpful for the identification of trends regarding various regulatory policies, urbanization, and climate change, as discussed by Wójcik-Gront & Gozdowski (2025). This technique is helpful for obtaining a better understanding of whether air pollution is improving or deteriorating over a period of time. The combination of seasonal and yearly analyses is quite helpful for obtaining more reliable results for time-series analysis. As discussed in previous studies, multi-year analyses are quite helpful for providing a scientific basis for environmental decision-making. Therefore, temporal analysis is quite essential for evaluating air quality, as discussed by Gianquintieri et al. (2025).

2.7. Challenges and Gaps in Current Literature

Despite progress, several gaps remain:

2.7.1. Integration of heterogeneous data sources

The combination of data from heterogenous different sources remains one of the significant challenges in environmental studies. Lu et al. (2020) note that the integration of ground-based monitor station measurements, IoT sensor measurements, and satellite imagery requires strong harmonization of time resolution, units, and types of data. Temkov (2025) finds that the addition of crowdsourced data adds more variability, and strong validation and filtering processes are required. Zhang et al. (2024) recommend ontology-based frameworks as the optimal solution for harmonizing heterogenous data while ensuring semantic consistency. Despite such methodological advancements, inconsistencies, missing values, and gaps continue to limit the usability and reliability of integrated environmental data in research and policy implementation.

In a systematic review, Za'al Alma'aitah et al. (2024) explain the challenges of merging disparate environmental data sets. They note that variability in data types, spatial and temporal resolutions, and quality is a daunting obstacle to building efficient data management systems. Environmental initiatives have a tendency to collect data from diverse sources IoT sensors, satellite images, and local databases—frequently incompatible with one another. Lack of international standards and adaptation tools for harmonization hinders replication, large-scale analysis, and comparability across studies. These issues can be solved, the authors argue, by using data integration frameworks from distributed architectures to enable more dynamic integration of heterogeneous datasets and improve the dependability and usability of environmental information.

2.7.2. Balancing openness and quality

Open-access data platforms enhance transparency, collaboration, and scientific innovation but also become potential sources of impediment to data quality if not well managed. Zhou, Wang, and Liu (2025) recommend that datasets that are publicly accessible may have incomplete, inconsistent, or unverified entries that lead to misleading analysis. Rosales and Fernandez (2025) suggest these risks can be met through the implementation of standardized metadata practices and automated quality control by enabling contextual information, traceability, and validation controls. Smith and Kumar (2022) further note that achieving a balance between openness and rigorous quality control remains a challenge, particularly where large-scale, heterogeneous environmental data sets are being combined for research and policy purposes.

2.7.3. Scalability

The tension between sharing data as openly as possible and maintaining high quality standards has significant implications for reproducibility. Guo et al. (2024) show how predictive models are highly susceptible to inaccuracies in input data, thereby identifying the necessity for rigorous quality assurance in open-access data sets. Bernardo (2024) further maintains that transparent policies of data curation, verification, and validation should be used by institutions to prevent the dissemination of wrong or misinformed information in scientific research and policy enforcement. Both accessibility and reliability are therefore important in ensuring reproducibility and trust in environmental data management.

Scalability is a serious concern in managing big environmental data sets. Huang (2021) notes that the real-time monitoring generates large volumes of data, which tend to overload usual storage and processing capabilities in a matter of time. Zhang, Lin, and Wang (2024) indicate that distributed processing and cloud computing somewhat reduce some of these pressures, but fall short of fully addressing computational bottlenecks in advanced analytical pipelines. Guo and Li (2024) argue that one must optimize data pipelines, storage structures, and workflow management in the interest of responsiveness and efficiency as datasets get larger. Appropriate scalability strategies are therefore crucial to obtaining timely, consistent, and actionable environmental intelligence.

Ranatunga, Ødegård, Jetlund, and Onstein (2025) write about the scalability problems of environmental data and propose Semantic Web technologies as a solution. The authors note that the exponential growth of geographic and environmental data typically surpasses traditional database management systems. Their Ontology-Based Data Access (OBDA) model provides greater integration of big, heterogeneous datasets and improved scalability via the provision of a semantic layer for freer querying and interoperability. But these techniques remain in infancy stages, and there is more testing and validation under real environmental monitoring conditions needed to judge their reliability and effectiveness thoroughly.

2.7.4. Limitations and Contributions

Despite extensive research on particle size distribution and air pollution modeling, relatively little research is dedicated to data management process optimization. The majority of available literature is focused on aspects related to pollutants, trends, and modeling (Gianquintieri et al., 2025). The practical application of raw data to structured data, which is suitable for repeated use, is often ignored. Additionally, statistical values such as percentile values and mean particle size are often employed, yet there is a lack of detailed description regarding their interpretation. This may negatively affect results' transparency and clarity. The literature review indicates a significant gap between data analysis and data management, and this is a problem that should be addressed to improve the overall quality of air pollution research (Rosales et al., 2025).

Another limitation, which was also identified by previous studies, is that there is a lack of focus on effective data base updating techniques. Although the aspect of data automation is discussed, particularly regarding its application to real-time monitoring, there is a lack of publications on effective data base updating techniques (Geng et al., 2021). The development of such techniques is likely to improve data usability for future researchers. Additionally, development of databases, which are based on FAIR principles, is likely to improve data longevity (Katzenstein & Etcheverry, 2025). These publications, therefore, demonstrate the need to integrate data interpretation with data base development and optimization techniques. This thesis, therefore, seeks to address this need by integrating data interpretation techniques with data base optimization techniques.

2.7.5. Governance and ethics

Data governance and ethics are still critical loopholes in environmental data management literature. Kanza (2022) identifies that good policies are required to define duties, rights of access, and best practices for stewardship of data. Bernardo (2024) further contributes that robust governance mechanisms uphold accountability, maintain public trust, and ensure data is utilized responsibly, particularly when data is utilized to make policy choices. Inadequate governance can lead to mismanagement or ethical abuse that can lead to abuse of sensitive environmental information, misinterpretation of results, and loss of scientific credibility, emphasizing the importance of incorporating governance and ethical practice into all data management activities.

Ethical concerns are particularly relevant in the integration of crowdsourced or IoT-produced environmental data. Temkov (2025) indicates concerns regarding privacy, data ownership, and the need for transparent mechanisms for consent. Mitchell and Johnson (2021) emphasize having clear tracking of provenance to ensure reproducibility, traceability, and accountability during data processing. Together, good governance and good ethical oversight are not only necessary for compliance with regulations but also are critical as basic building blocks to good, responsible, credible, and effective environmental data management practices in furtherance of scientific integrity and public confidence.

Za'al Alma'aitah et al. (2024) observe that the majority of research on the integration of environmental data focus on technical approaches and overlook policy, legal, and ethical issues. Ranatunga et al. (2025) also add that even advanced semantic web technologies facilitating interoperability cannot be utilized to their full potential unless there are explicit policies on data ownership, privacy, and open access. Consequently, there is a critical demand for end-to-end data governance models addressing technical as well as ethical dimensions, bridging gaps in existing research and enabling more accountable, responsible, and efficient environmental data management.

2.8. Summary

The literature examined identifies that successful environmental data management is a multi-faceted problem, involving technical, organizational, and ethical factors. Fundamental principles such as the FAIR principles, data governance, and provenance serve as building blocks for trustworthy and reproducible datasets, while applied methodologies in the guise of relational and NoSQL database systems, automated data pipelines, ontology-driven frameworks, and machine learning approaches offer concrete tools for ensuring data quality, accessibility, and interoperability (Mitchell et al., 2021; Bernardo, 2024; Zhang et al., 2024). Together, they establish the value of synergizing technical solutions with governance and ethical strategies in building dependable, scalable, and actionable environmental data systems.

Environmental studies, particularly for air quality monitoring, emphasize the critical necessity of fusing heterogenous data from IoT sensors, satellites, and crowdsourced locations (Lu et al., 2020; Rosales et al., 2025; Temkov, 2025). Augmenting the datasets via workflow enhancement, predictive modelling, and sensor network management not only renders them more efficient and credible but also facilitates scalable real-time analysis. These approaches enable researchers and policymakers to generate actionable results at an increased pace, increase spatial and temporal reach, and enable proactive environmental management (Guo et al., 2024; Huang, 2021).

Despite such advancements, there remain substantial challenges, including harmonizing disparate data sources, sacrificing open access for data quality, providing scalability for big and dynamic data, and having governance and ethics (Zhou et al., 2025; Kanza, 2022). Bridging these gaps effectively is critical to generating insights not only reproducible and transparent but also actionable, supporting rigorous scientific inquiry and good environmental policymaking.

The methodology of this study is based on the data management and optimization of air pollution data through software monitoring outputs. Particle size distribution parameters, such as percentile indicators (X values) and the average particle size (M), were used to describe the airborne particles and ensure proper interpretation of the data. A database was created from the exported environmental data and was organized based on standardized formats to increase accessibility and usability of the data in the future. An automatic update process was developed through the use of programming tools to effectively incorporate new measurements into the existing database and minimize the occurrence of human error. Temporal analysis was applied to compare the changes in the characteristics of the particles in different seasons and years. This methodology combines data management, automatic update, and temporal analysis to increase the reliability and usability of air pollution data in future studies.

Overall, the literature highlights that the combination of robust conceptual frameworks with cutting-edge computational methods offers an inclusive and effective approach to modern data management. Such practices are technically indispensable as well as indispensable to support evidence-based policymaking, confer scientific credibility, and build public trust in environmental research. The collective evidence supports that future effort will have to further focus on interoperability, data and workflow optimization, and strong ethical governance in order to enable researchers and decision-makers to access the full power of large-scale, heterogeneous environmental data for action-capable insight and sustainable management.

Chapter 3. Methodology

This chapter describes the workflow used for data extraction, preprocessing, database construction, and subsequent analysis.

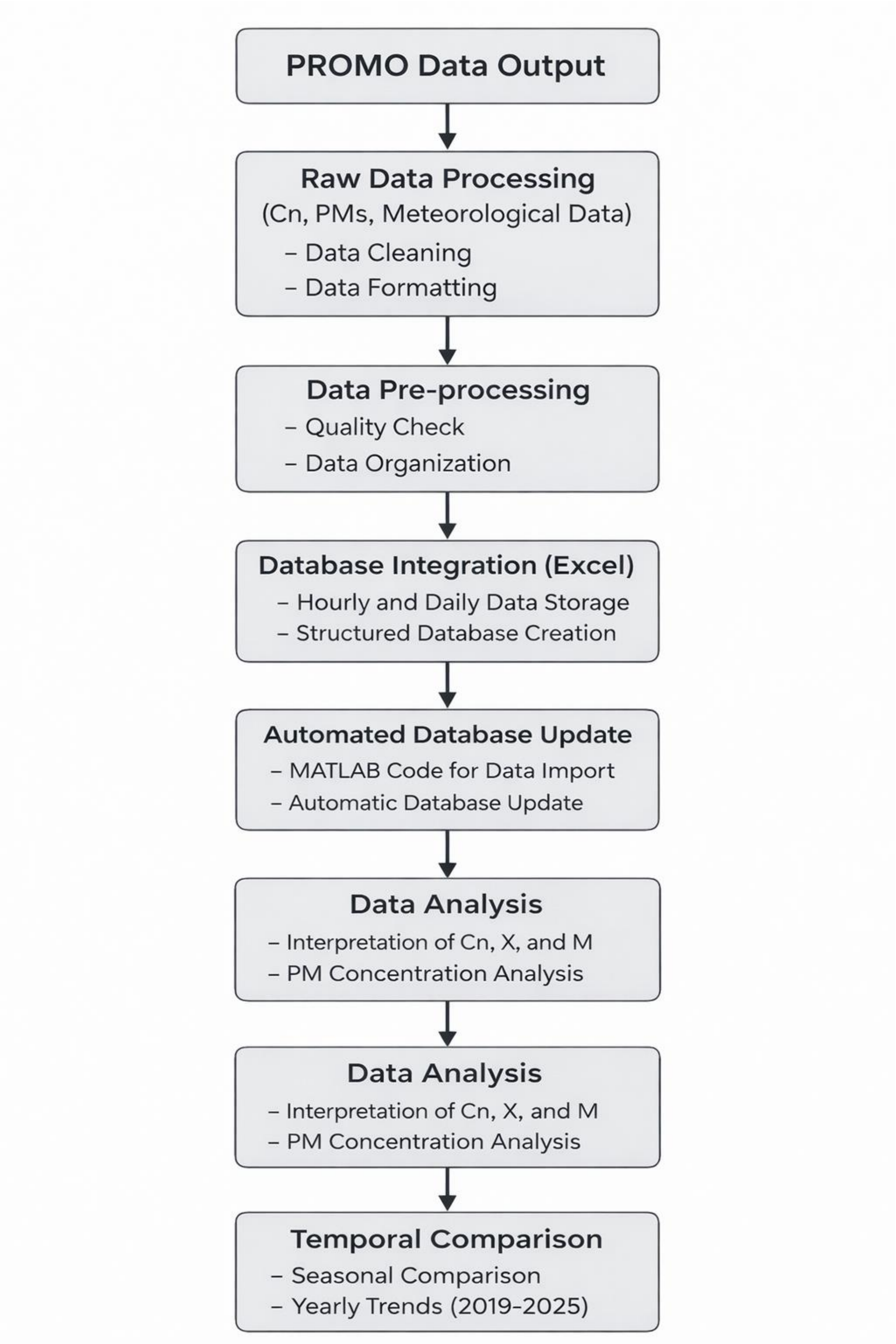

PROMO Data Output
Raw Data Processing
(Cn, PMs, Meteorological Data)
– Data Cleaning
– Data Formatting
Data Pre-processing
– Quality Check
– Data Organization
Database Integration (Excel)
– Hourly and Daily Data Storage
– Structured Database Creation
Automated Database Update
– MATLAB Code for Data Import
– Automatic Database Update
Data Analysis
– Interpretation of Cn, X, and M
– PM Concentration Analysis
Data Analysis
– Interpretation of Cn, X, and M
– PM Concentration Analysis
Temporal Comparison
– Seasonal Comparison
– Yearly Trends (2019–2025)

Figure 3-1 Methodology flow chart

3.1. Data Acquisition

3.1.1. PROMO data extraction

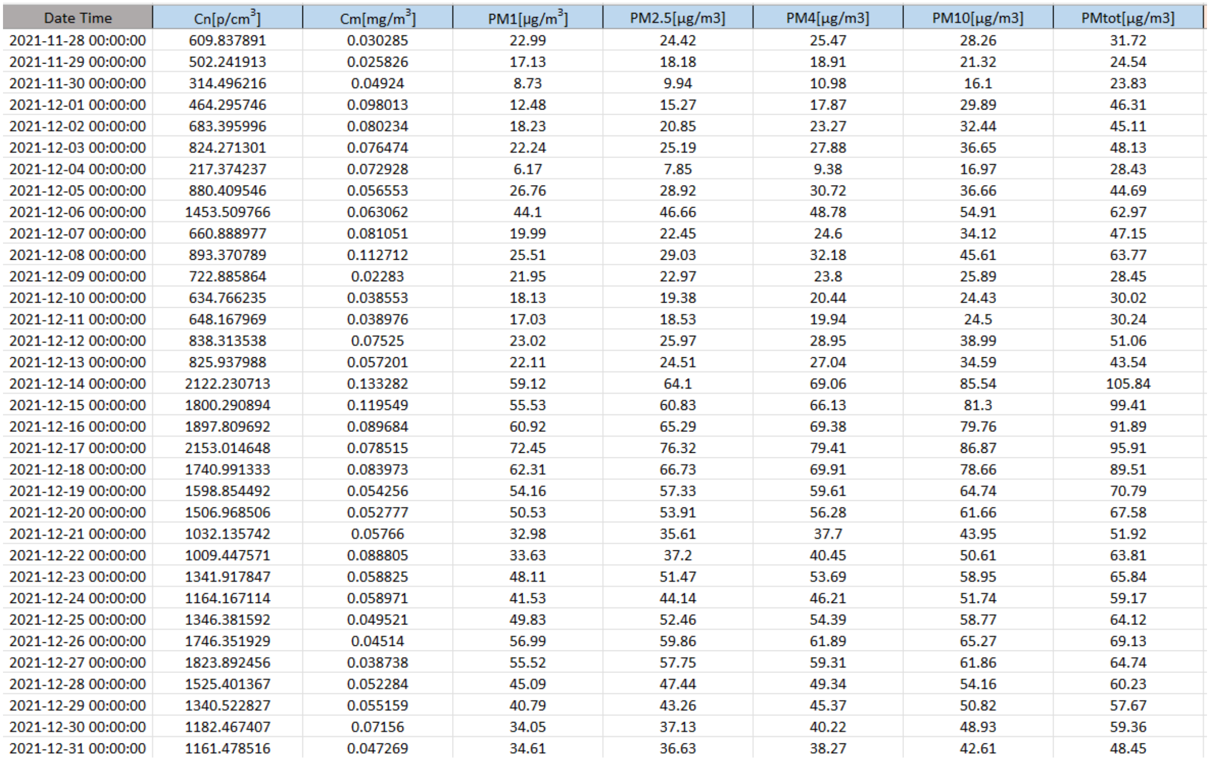
The dataset to be used primarily in this study were extracted from the PROMO system, which is set to continuously measure the concentration of aerosols and related properties of the atmosphere. The PROMO instrument provides high-resolution time series data recorded at short intervals, which are highly important for detailed temporal analysis.The original PROMO output files were extracted via the software interface provided by the Palas GmbH company. All available measurement channels were extracted in their raw format, preserving time stamps and native units. Consequently, the data was stored into different files for further pre-processing. During this phase, no modifications were performed to preserve the integrity of the raw dataset.

3.1.2. Air quality parameters

The PROMO system provides several aerosol metrics that describe the particle number concentration and particulate mass in various size ranges. In this thesis, the following parameters were extracted and analyzed:

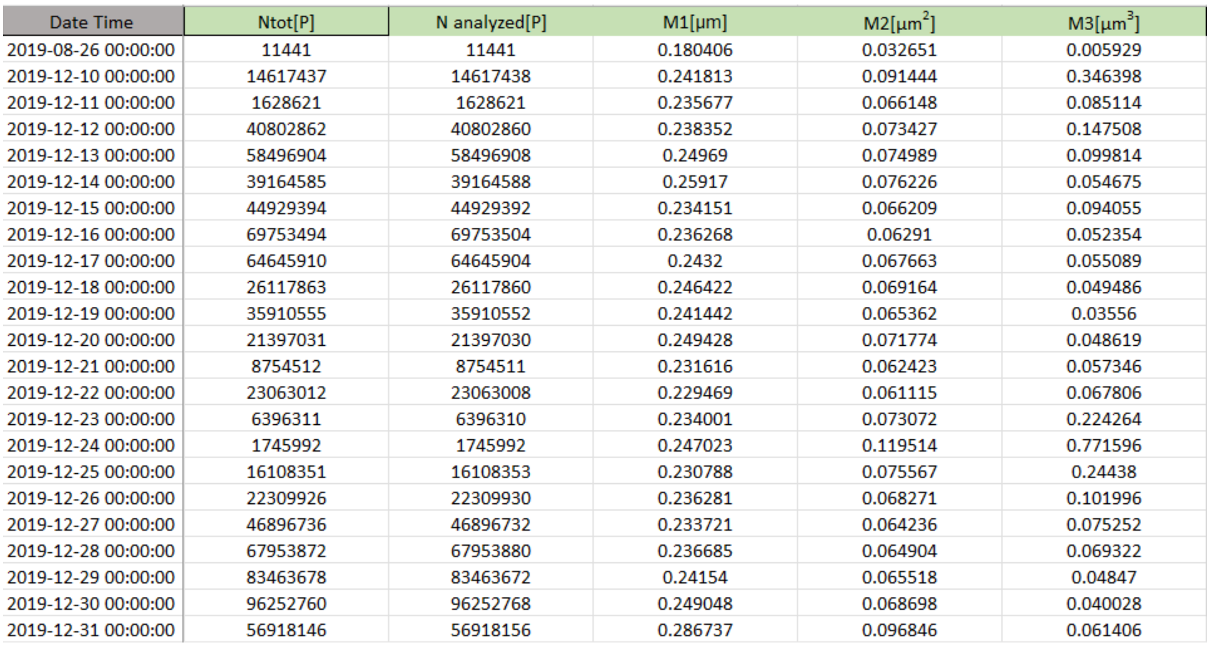
- ✓ CN : Condensation Nuclei: total particle number concentration
- ✓ CM: total mass concentration
- ✓ $N_{Analysed}$: The number of particles included in the statistical analysis within the defined measurement size range
- ✓ N_{tot}: The total particle count recorded by the instrument during each measurement interval
- ✓ PM_1: particulate mass with an aerodynamic diameter less than 1 µm
- ✓ $PM_{2.5}$: particulate mass < 2.5 µm
- ✓ PM_4: particulate mass < 4 µm
- ✓ PM_{10}: particulate mass < 10 µm
- ✓ PM_{tot}: total particulate mass
- ✓ X-Parameters(N): n percent of particles have a size smaller than this value in terms of number
- ✓ X-Parameters(A): n percent of the total particle surface area is from particles smaller than this value
- ✓ X-Parameters(V): n percent of the total particle volume is due to particles with a size smaller than this value
- ✓ M1: Average size based on number of particles
- ✓ M2: Surface-related mean square with emphasis on larger particles
- ✓ M3: Volume-related average with emphasis on largest particles

These parameters enable the characterization of particle behavior for different size fractions. The data reported in this dataset are high-frequency measurements that have been subsequently aggregated to hourly and daily averages to reduce noise and allow for long-term comparative analyses.



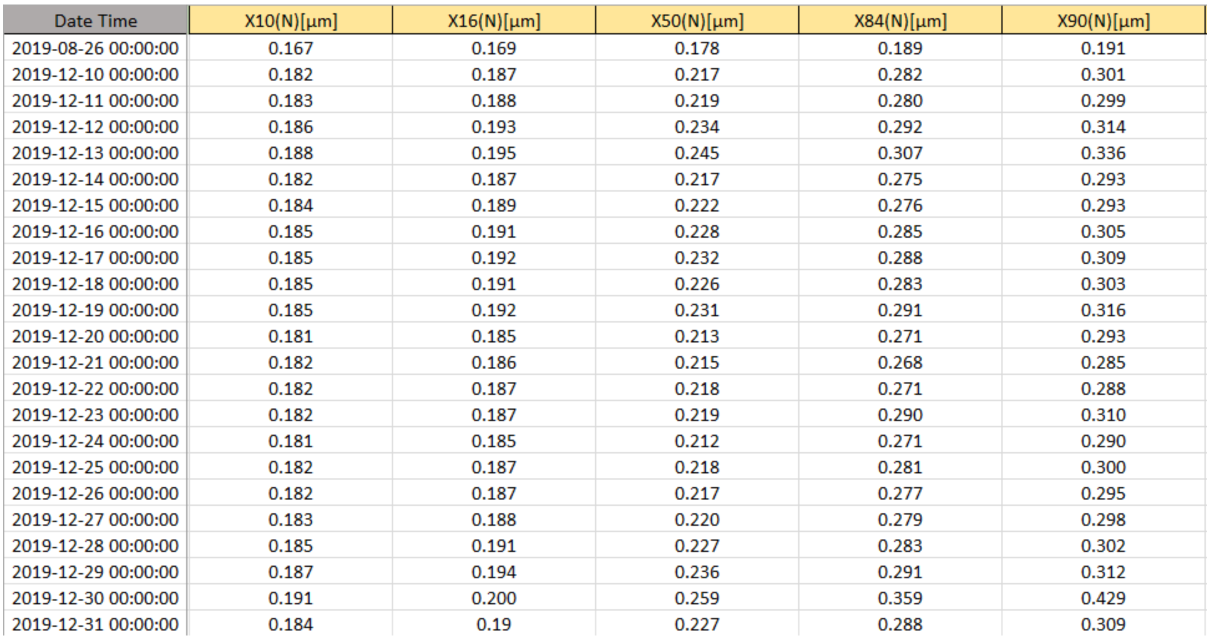
Date Time	Cn[p/cm^3]	Cm[mg/m^3]	PM1[µg/m^3]	PM2.5[µg/m3]	PM4[µg/m3]	PM10[µg/m3]	PMtot[µg/m3]
2021-11-28 00:00:00	609.837891	0.030285	22.99	24.42	25.47	28.26	31.72
2021-11-29 00:00:00	502.241913	0.025826	17.13	18.18	18.91	21.32	24.54
2021-11-30 00:00:00	314.496216	0.04924	8.73	9.94	10.98	16.1	23.83
2021-12-01 00:00:00	464.295746	0.098013	12.48	15.27	17.87	29.89	46.31
2021-12-02 00:00:00	683.395996	0.080234	18.23	20.85	23.27	32.44	45.11
2021-12-03 00:00:00	824.271301	0.076474	22.24	25.19	27.88	36.65	48.13
2021-12-04 00:00:00	217.374237	0.072928	6.17	7.85	9.38	16.97	28.43
2021-12-05 00:00:00	880.409546	0.056553	26.76	28.92	30.72	36.66	44.69
2021-12-06 00:00:00	1453.509766	0.063062	44.1	46.66	48.78	54.91	62.97
2021-12-07 00:00:00	660.888977	0.081051	19.99	22.45	24.6	34.12	47.15
2021-12-08 00:00:00	893.370789	0.112712	25.51	29.03	32.18	45.61	63.77
2021-12-09 00:00:00	722.885864	0.02283	21.95	22.97	23.8	25.89	28.45
2021-12-10 00:00:00	634.766235	0.038553	18.13	19.38	20.44	24.43	30.02
2021-12-11 00:00:00	648.167969	0.038976	17.03	18.53	19.94	24.5	30.24
2021-12-12 00:00:00	838.313538	0.07525	23.02	25.97	28.95	38.99	51.06
2021-12-13 00:00:00	825.937988	0.057201	22.11	24.51	27.04	34.59	43.54
2021-12-14 00:00:00	2122.230713	0.133282	59.12	64.1	69.06	85.54	105.84
2021-12-15 00:00:00	1800.290894	0.119549	55.53	60.83	66.13	81.3	99.41
2021-12-16 00:00:00	1897.809692	0.089684	60.92	65.29	69.38	79.76	91.89
2021-12-17 00:00:00	2153.014648	0.078515	72.45	76.32	79.41	86.87	95.91
2021-12-18 00:00:00	1740.991333	0.083973	62.31	66.73	69.91	78.66	89.51
2021-12-19 00:00:00	1598.854492	0.054256	54.16	57.33	59.61	64.74	70.79
2021-12-20 00:00:00	1506.968506	0.052777	50.53	53.91	56.28	61.66	67.58
2021-12-21 00:00:00	1032.135742	0.05766	32.98	35.61	37.7	43.95	51.92
2021-12-22 00:00:00	1009.447571	0.088805	33.63	37.2	40.45	50.61	63.81
2021-12-23 00:00:00	1341.917847	0.058825	48.11	51.47	53.69	58.95	65.84
2021-12-24 00:00:00	1164.167114	0.058971	41.53	44.14	46.21	51.74	59.17
2021-12-25 00:00:00	1346.381592	0.049521	49.83	52.46	54.39	58.77	64.12
2021-12-26 00:00:00	1746.351929	0.04514	56.99	59.86	61.89	65.27	69.13
2021-12-27 00:00:00	1823.892456	0.038738	55.52	57.75	59.31	61.86	64.74
2021-12-28 00:00:00	1525.401367	0.052284	45.09	47.44	49.34	54.16	60.23
2021-12-29 00:00:00	1340.522827	0.055159	40.79	43.26	45.37	50.82	57.67
2021-12-30 00:00:00	1182.467407	0.07156	34.05	37.13	40.22	48.93	59.36
2021-12-31 00:00:00	1161.478516	0.047269	34.61	36.63	38.27	42.61	48.45

Figure 3-2 Extracted daily air quality parameters in database



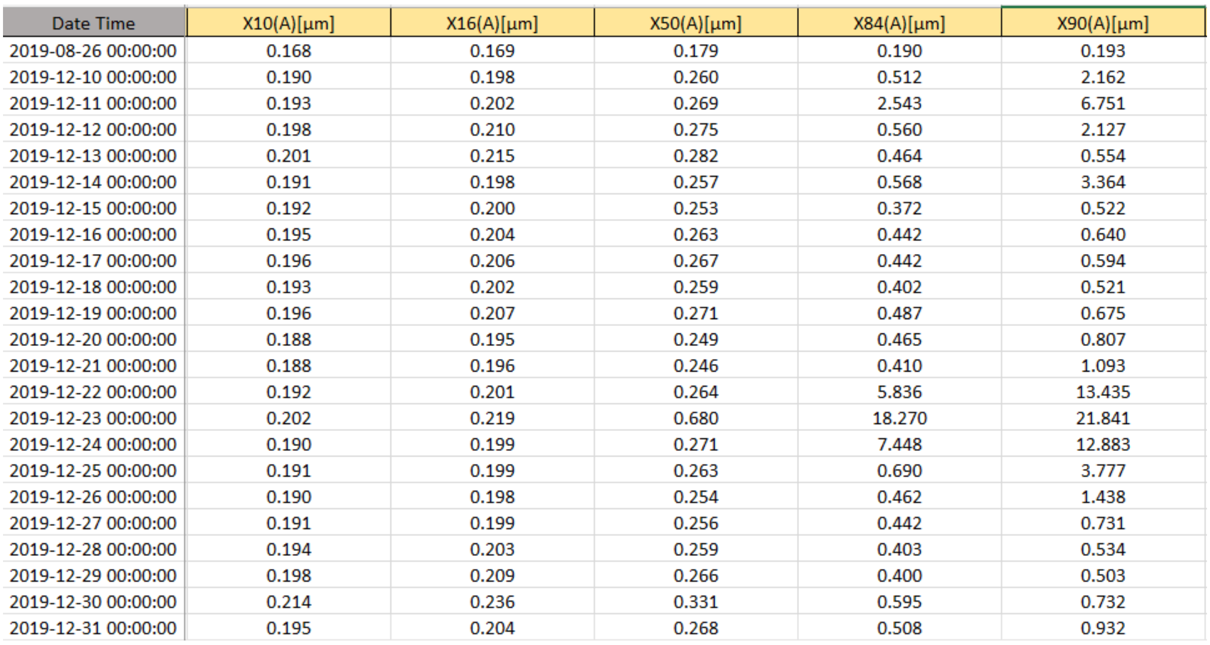
Date Time	Ntot[P]	N analyzed[P]	M1[µm]	M2[µm^2]	M3[µm^3]
2019-08-26 00:00:00	11441	11441	0.180406	0.032651	0.005929
2019-12-10 00:00:00	14617437	14617438	0.241813	0.091444	0.346398
2019-12-11 00:00:00	1628621	1628621	0.235677	0.066148	0.085114
2019-12-12 00:00:00	40802862	40802860	0.238352	0.073427	0.147508
2019-12-13 00:00:00	58496904	58496908	0.24969	0.074989	0.099814
2019-12-14 00:00:00	39164585	39164588	0.25917	0.076226	0.054675
2019-12-15 00:00:00	44929394	44929392	0.234151	0.066209	0.094055
2019-12-16 00:00:00	69753494	69753504	0.236268	0.06291	0.052354
2019-12-17 00:00:00	64645910	64645904	0.2432	0.067663	0.055089
2019-12-18 00:00:00	26117863	26117860	0.246422	0.069164	0.049486
2019-12-19 00:00:00	35910555	35910552	0.241442	0.065362	0.03556
2019-12-20 00:00:00	21397031	21397030	0.249428	0.071774	0.048619
2019-12-21 00:00:00	8754512	8754511	0.231616	0.062423	0.057346
2019-12-22 00:00:00	23063012	23063008	0.229469	0.061115	0.067806
2019-12-23 00:00:00	6396311	6396310	0.234001	0.073072	0.224264
2019-12-24 00:00:00	1745992	1745992	0.247023	0.119514	0.771596
2019-12-25 00:00:00	16108351	16108353	0.230788	0.075567	0.24438
2019-12-26 00:00:00	22309926	22309930	0.236281	0.068271	0.101996
2019-12-27 00:00:00	46896736	46896732	0.233721	0.064236	0.075252
2019-12-28 00:00:00	67953872	67953880	0.236685	0.064904	0.069322
2019-12-29 00:00:00	83463678	83463672	0.24154	0.065518	0.04847
2019-12-30 00:00:00	96252760	96252768	0.249048	0.068698	0.040028
2019-12-31 00:00:00	56918146	56918156	0.286737	0.096846	0.061406

Figure 3-3 Extracted daily N and M parameters in database



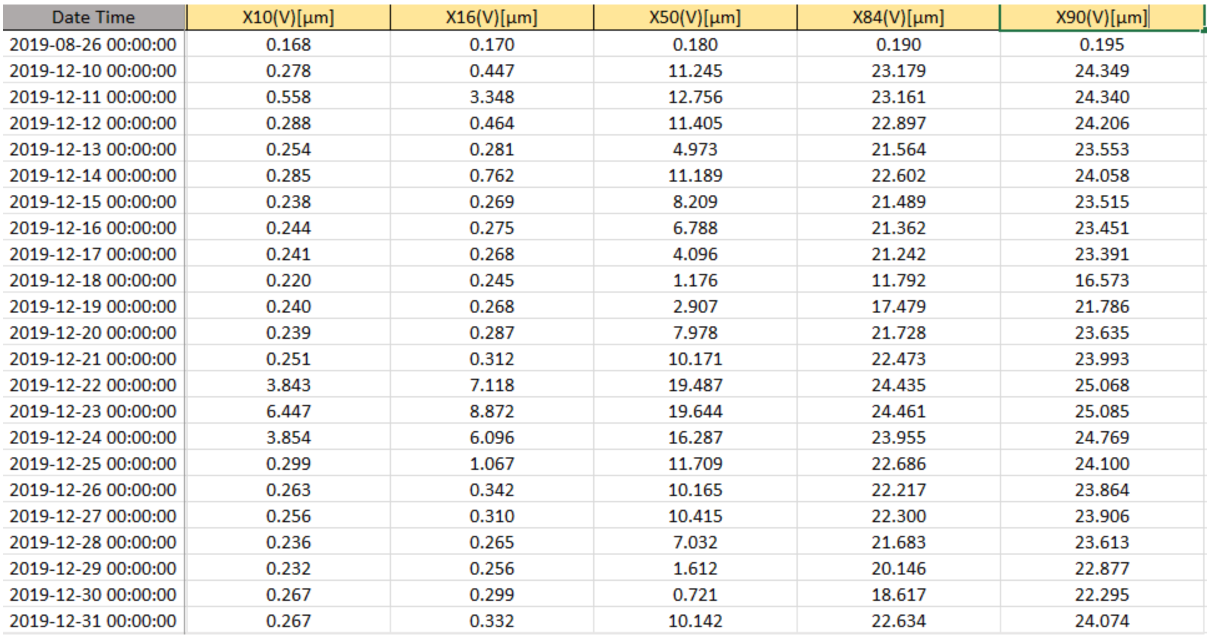
Date Time	X10(N)[μm]	X16(N)[μm]	X50(N)[μm]	X84(N)[μm]	X90(N)[μm]
2019-08-26 00:00:00	0.167	0.169	0.178	0.189	0.191
2019-12-10 00:00:00	0.182	0.187	0.217	0.282	0.301
2019-12-11 00:00:00	0.183	0.188	0.219	0.280	0.299
2019-12-12 00:00:00	0.186	0.193	0.234	0.292	0.314
2019-12-13 00:00:00	0.188	0.195	0.245	0.307	0.336
2019-12-14 00:00:00	0.182	0.187	0.217	0.275	0.293
2019-12-15 00:00:00	0.184	0.189	0.222	0.276	0.293
2019-12-16 00:00:00	0.185	0.191	0.228	0.285	0.305
2019-12-17 00:00:00	0.185	0.192	0.232	0.288	0.309
2019-12-18 00:00:00	0.185	0.191	0.226	0.283	0.303
2019-12-19 00:00:00	0.185	0.192	0.231	0.291	0.316
2019-12-20 00:00:00	0.181	0.185	0.213	0.271	0.293
2019-12-21 00:00:00	0.182	0.186	0.215	0.268	0.285
2019-12-22 00:00:00	0.182	0.187	0.218	0.271	0.288
2019-12-23 00:00:00	0.182	0.187	0.219	0.290	0.310
2019-12-24 00:00:00	0.181	0.185	0.212	0.271	0.290
2019-12-25 00:00:00	0.182	0.187	0.218	0.281	0.300
2019-12-26 00:00:00	0.182	0.187	0.217	0.277	0.295
2019-12-27 00:00:00	0.183	0.188	0.220	0.279	0.298
2019-12-28 00:00:00	0.185	0.191	0.227	0.283	0.302
2019-12-29 00:00:00	0.187	0.194	0.236	0.291	0.312
2019-12-30 00:00:00	0.191	0.200	0.259	0.359	0.429
2019-12-31 00:00:00	0.184	0.19	0.227	0.288	0.309

Figure 3-4 Extracted daily X-Parameters(N) in database



Date Time	X10(A)[μm]	X16(A)[μm]	X50(A)[μm]	X84(A)[μm]	X90(A)[μm]
2019-08-26 00:00:00	0.168	0.169	0.179	0.190	0.193
2019-12-10 00:00:00	0.190	0.198	0.260	0.512	2.162
2019-12-11 00:00:00	0.193	0.202	0.269	2.543	6.751
2019-12-12 00:00:00	0.198	0.210	0.275	0.560	2.127
2019-12-13 00:00:00	0.201	0.215	0.282	0.464	0.554
2019-12-14 00:00:00	0.191	0.198	0.257	0.568	3.364
2019-12-15 00:00:00	0.192	0.200	0.253	0.372	0.522
2019-12-16 00:00:00	0.195	0.204	0.263	0.442	0.640
2019-12-17 00:00:00	0.196	0.206	0.267	0.442	0.594
2019-12-18 00:00:00	0.193	0.202	0.259	0.402	0.521
2019-12-19 00:00:00	0.196	0.207	0.271	0.487	0.675
2019-12-20 00:00:00	0.188	0.195	0.249	0.465	0.807
2019-12-21 00:00:00	0.188	0.196	0.246	0.410	1.093
2019-12-22 00:00:00	0.192	0.201	0.264	5.836	13.435
2019-12-23 00:00:00	0.202	0.219	0.680	18.270	21.841
2019-12-24 00:00:00	0.190	0.199	0.271	7.448	12.883
2019-12-25 00:00:00	0.191	0.199	0.263	0.690	3.777
2019-12-26 00:00:00	0.190	0.198	0.254	0.462	1.438
2019-12-27 00:00:00	0.191	0.199	0.256	0.442	0.731
2019-12-28 00:00:00	0.194	0.203	0.259	0.403	0.534
2019-12-29 00:00:00	0.198	0.209	0.266	0.400	0.503
2019-12-30 00:00:00	0.214	0.236	0.331	0.595	0.732
2019-12-31 00:00:00	0.195	0.204	0.268	0.508	0.932

Figure 3-5 Extracted daily X-Parameters(A) in database



Date Time	X10(V)[µm]	X16(V)[µm]	X50(V)[µm]	X84(V)[µm]	X90(V)[µm]
2019-08-26 00:00:00	0.168	0.170	0.180	0.190	0.195
2019-12-10 00:00:00	0.278	0.447	11.245	23.179	24.349
2019-12-11 00:00:00	0.558	3.348	12.756	23.161	24.340
2019-12-12 00:00:00	0.288	0.464	11.405	22.897	24.206
2019-12-13 00:00:00	0.254	0.281	4.973	21.564	23.553
2019-12-14 00:00:00	0.285	0.762	11.189	22.602	24.058
2019-12-15 00:00:00	0.238	0.269	8.209	21.489	23.515
2019-12-16 00:00:00	0.244	0.275	6.788	21.362	23.451
2019-12-17 00:00:00	0.241	0.268	4.096	21.242	23.391
2019-12-18 00:00:00	0.220	0.245	1.176	11.792	16.573
2019-12-19 00:00:00	0.240	0.268	2.907	17.479	21.786
2019-12-20 00:00:00	0.239	0.287	7.978	21.728	23.635
2019-12-21 00:00:00	0.251	0.312	10.171	22.473	23.993
2019-12-22 00:00:00	3.843	7.118	19.487	24.435	25.068
2019-12-23 00:00:00	6.447	8.872	19.644	24.461	25.085
2019-12-24 00:00:00	3.854	6.096	16.287	23.955	24.769
2019-12-25 00:00:00	0.299	1.067	11.709	22.686	24.100
2019-12-26 00:00:00	0.263	0.342	10.165	22.217	23.864
2019-12-27 00:00:00	0.256	0.310	10.415	22.300	23.906
2019-12-28 00:00:00	0.236	0.265	7.032	21.683	23.613
2019-12-29 00:00:00	0.232	0.256	1.612	20.146	22.877
2019-12-30 00:00:00	0.267	0.299	0.721	18.617	22.295
2019-12-31 00:00:00	0.267	0.332	10.142	22.634	24.074

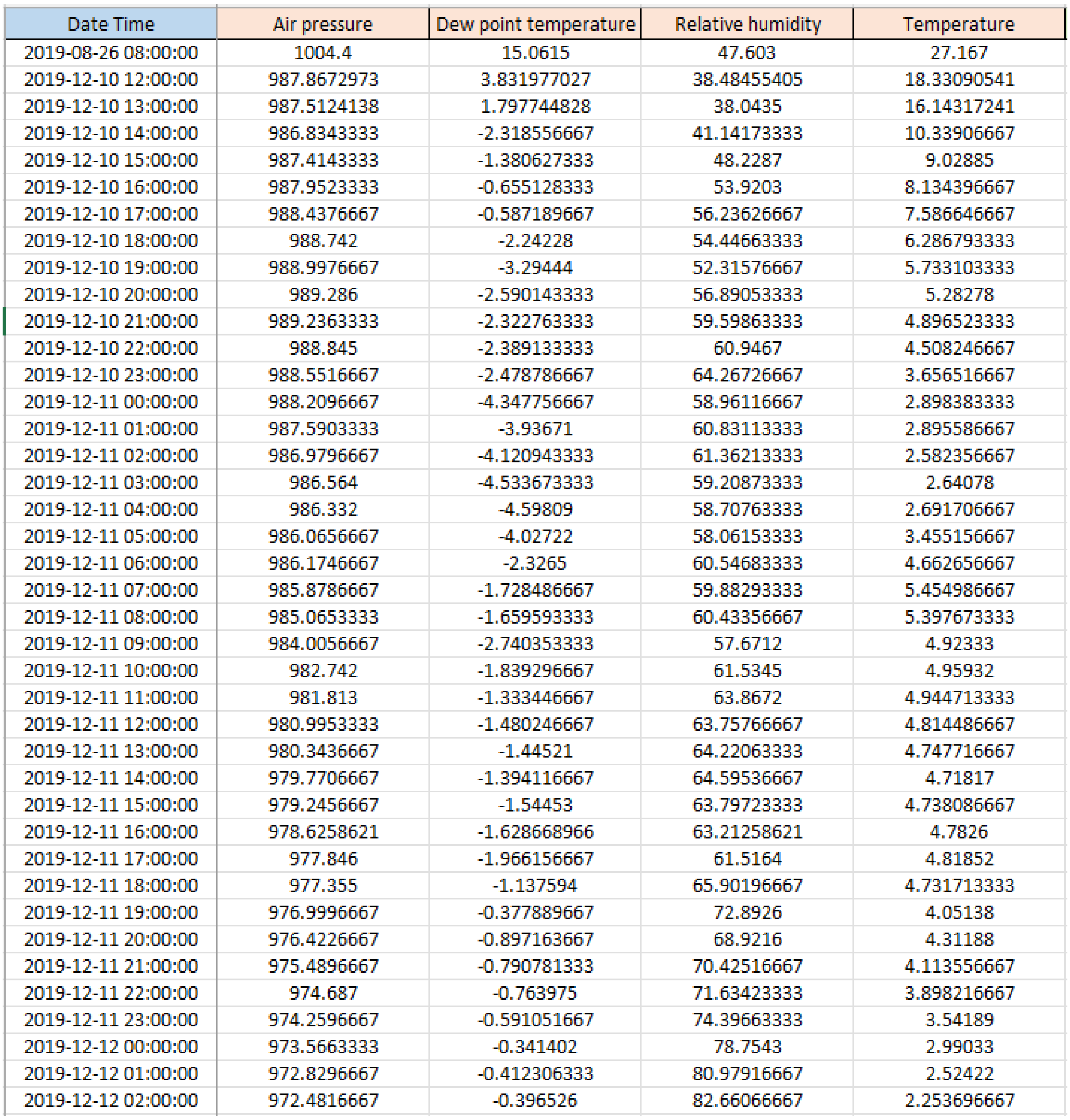
Figure 3-6 Extracted daily X-Parameters(V) in database

3.1.3. Meteorological data collection

Along with aerosol parameters, meteorological variables have also been measured to support the interpretation of air quality patterns. Meteorological data were obtained from the same measurement system and include:

- ✓ Ambient temperature
- ✓ Relative humidity
- ✓ Atmospheric pressure
- ✓ Dew point temperature

All these variables were synchronized to the PROMO timestamps to ensure direct comparability. Meteorological information is necessary to understand the variations in CN and PM concentrations, since atmospheric conditions directly influence the aerosol formation, dispersion, and removal processes.



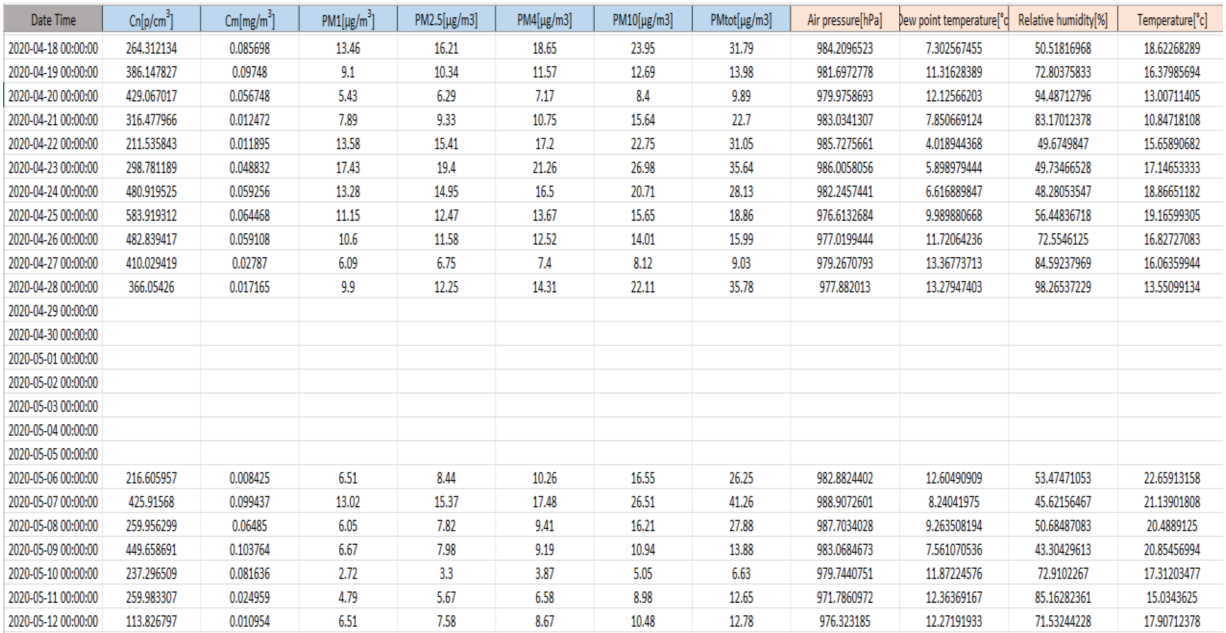
Date Time	Air pressure	Dew point temperature	Relative humidity	Temperature
2019-08-26 08:00:00	1004.4	15.0615	47.603	27.167
2019-12-10 12:00:00	987.8672973	3.831977027	38.48455405	18.33090541
2019-12-10 13:00:00	987.5124138	1.797744828	38.0435	16.14317241
2019-12-10 14:00:00	986.8343333	-2.318556667	41.14173333	10.33906667
2019-12-10 15:00:00	987.4143333	-1.380627333	48.2287	9.02885
2019-12-10 16:00:00	987.9523333	-0.655128333	53.9203	8.134396667
2019-12-10 17:00:00	988.4376667	-0.587189667	56.23626667	7.586646667
2019-12-10 18:00:00	988.742	-2.24228	54.44663333	6.286793333
2019-12-10 19:00:00	988.9976667	-3.29444	52.31576667	5.733103333
2019-12-10 20:00:00	989.286	-2.590143333	56.89053333	5.28278
2019-12-10 21:00:00	989.2363333	-2.322763333	59.59863333	4.896523333
2019-12-10 22:00:00	988.845	-2.389133333	60.9467	4.508246667
2019-12-10 23:00:00	988.5516667	-2.478786667	64.26726667	3.656516667
2019-12-11 00:00:00	988.2096667	-4.347756667	58.96116667	2.898383333
2019-12-11 01:00:00	987.5903333	-3.93671	60.83113333	2.895586667
2019-12-11 02:00:00	986.9796667	-4.120943333	61.36213333	2.582356667
2019-12-11 03:00:00	986.564	-4.533673333	59.20873333	2.64078
2019-12-11 04:00:00	986.332	-4.59809	58.70763333	2.691706667
2019-12-11 05:00:00	986.0656667	-4.02722	58.06153333	3.455156667
2019-12-11 06:00:00	986.1746667	-2.3265	60.54683333	4.662656667
2019-12-11 07:00:00	985.8786667	-1.728486667	59.88293333	5.454986667
2019-12-11 08:00:00	985.0653333	-1.659593333	60.43356667	5.397673333
2019-12-11 09:00:00	984.0056667	-2.740353333	57.6712	4.92333
2019-12-11 10:00:00	982.742	-1.839296667	61.5345	4.95932
2019-12-11 11:00:00	981.813	-1.333446667	63.8672	4.944713333
2019-12-11 12:00:00	980.9953333	-1.480246667	63.75766667	4.814486667
2019-12-11 13:00:00	980.3436667	-1.44521	64.22063333	4.747716667
2019-12-11 14:00:00	979.7706667	-1.394116667	64.59536667	4.71817
2019-12-11 15:00:00	979.2456667	-1.54453	63.79723333	4.738086667
2019-12-11 16:00:00	978.6258621	-1.628668966	63.21258621	4.7826
2019-12-11 17:00:00	977.846	-1.966156667	61.5164	4.81852
2019-12-11 18:00:00	977.355	-1.137594	65.90196667	4.731713333
2019-12-11 19:00:00	976.9996667	-0.377889667	72.8926	4.05138
2019-12-11 20:00:00	976.4226667	-0.897163667	68.9216	4.31188
2019-12-11 21:00:00	975.4896667	-0.790781333	70.42516667	4.113556667
2019-12-11 22:00:00	974.687	-0.763975	71.63423333	3.898216667
2019-12-11 23:00:00	974.2596667	-0.591051667	74.39663333	3.54189
2019-12-12 00:00:00	973.5663333	-0.341402	78.7543	2.99033
2019-12-12 01:00:00	972.8296667	-0.412306333	80.97916667	2.52422
2019-12-12 02:00:00	972.4816667	-0.396526	82.66066667	2.253696667

Figure 3-7 Extracted hourly Meteorological data in database

3.2. Data Pre-processing

3.2.1. Formatting and cleaning raw data

After extracting the raw datasets of PROMO and meteorological data, an initial formatting stage was carried out to put the data into a suitable format for analysis. This included converting the native file structure to a uniform tabular format, ensuring all variables had consistent units and naming conventions. The timestamps were checked for continuity, duplicated entries were removed, and any corrupt rows or incomplete records identified for further treatment. Therefore, during this phase, missing values were flagged by empty cells or removed according to predefined quality-control criteria, ensuring that only reliable observations were preserved for subsequent averaging and integration.



Date Time	Cn[p/cm^{3}]	Cm[mg/m^{3}]	PM1[µg/m^{3}]	PM2.5[µg/m3]	PM4[µg/m3]	PM10[µg/m3]	PMtot[µg/m3]	Air pressure[hPa]	Dew point temperature[°c	Relative humidity[%]	Temperature[°c]
2020-04-18 00:00:00	264.312134	0.085698	13.46	16.21	18.65	23.95	31.79	984.2096523	7.302567455	50.51816968	18.62268289
2020-04-19 00:00:00	386.147827	0.09748	9.1	10.34	11.57	12.69	13.98	981.6972778	11.31628389	72.80375833	16.37985694
2020-04-20 00:00:00	429.067017	0.056748	5.43	6.29	7.17	8.4	9.89	979.9758693	12.12566203	94.48712796	13.00711405
2020-04-21 00:00:00	316.477966	0.012472	7.89	9.33	10.75	15.64	22.7	983.0341307	7.850669124	83.17012378	10.84718108
2020-04-22 00:00:00	211.535843	0.011895	13.58	15.41	17.2	22.75	31.05	985.7275661	4.018944368	49.6749847	15.65890682
2020-04-23 00:00:00	298.781189	0.048832	17.43	19.4	21.26	26.98	35.64	986.0058056	5.898979444	49.73466528	17.14653333
2020-04-24 00:00:00	480.919525	0.059256	13.28	14.95	16.5	20.71	28.13	982.2457441	6.616889847	48.28053547	18.86651182
2020-04-25 00:00:00	583.919312	0.064468	11.15	12.47	13.67	15.65	18.86	976.6132684	9.989880668	56.44836718	19.16599305
2020-04-26 00:00:00	482.839417	0.059108	10.6	11.58	12.52	14.01	15.99	977.0199444	11.72064236	72.5546125	16.82727083
2020-04-27 00:00:00	410.029419	0.02787	6.09	6.75	7.4	8.12	9.03	979.2670793	13.36773713	84.59237969	16.06359944
2020-04-28 00:00:00	366.05426	0.017165	9.9	12.25	14.31	22.11	35.78	977.882013	13.27947403	98.26537229	13.55099134
2020-04-29 00:00:00											
2020-04-30 00:00:00											
2020-05-01 00:00:00											
2020-05-02 00:00:00											
2020-05-03 00:00:00											
2020-05-04 00:00:00											
2020-05-05 00:00:00											
2020-05-06 00:00:00	216.605957	0.008425	6.51	8.44	10.26	16.55	26.25	982.8824402	12.60490909	53.47471053	22.65913158
2020-05-07 00:00:00	425.91568	0.099437	13.02	15.37	17.48	26.51	41.26	988.9072601	8.24041975	45.62156467	21.13901808
2020-05-08 00:00:00	259.956299	0.06485	6.05	7.82	9.41	16.21	27.88	987.7034028	9.263508194	50.68487083	20.4889125
2020-05-09 00:00:00	449.658691	0.103764	6.67	7.98	9.19	10.94	13.88	983.0684673	7.561070536	43.30429613	20.85456994
2020-05-10 00:00:00	237.296509	0.081636	2.72	3.3	3.87	5.05	6.63	979.7440751	11.87224576	72.9102267	17.31203477
2020-05-11 00:00:00	259.983307	0.024959	4.79	5.67	6.58	8.98	12.65	971.7860972	12.36369167	85.16282361	15.0343625
2020-05-12 00:00:00	113.826797	0.010954	6.51	7.58	8.67	10.48	12.78	976.323185	12.27191933	71.53244228	17.90712378

Figure 3-8 Empty cells for days without recorded data

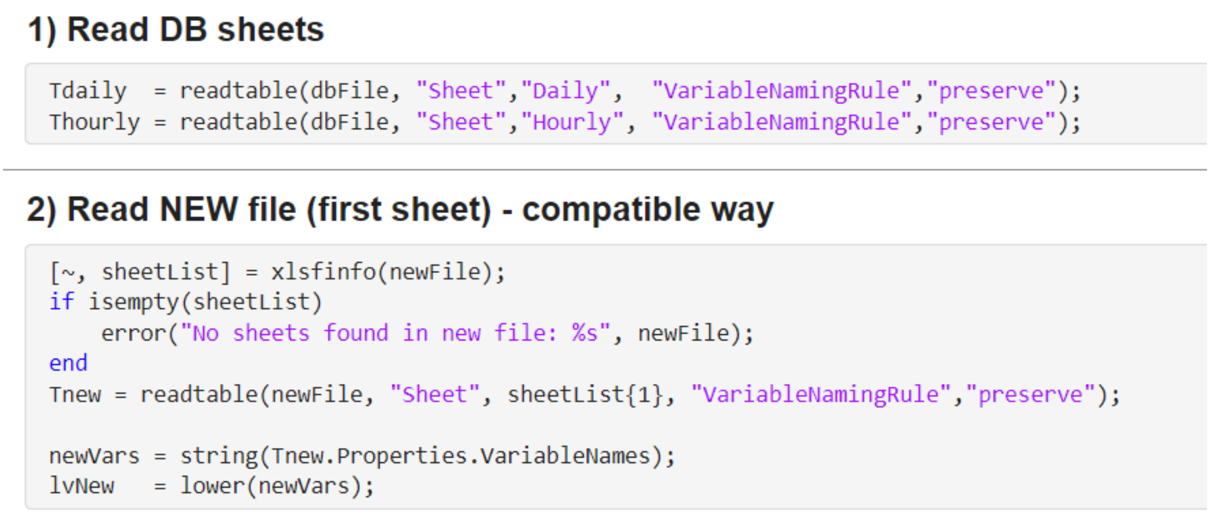
3.2.2. Hourly and daily averaging using MATLAB

The high-frequency PROMO measurements have short-term fluctuations that can mask wider temporal trends. To obtain a more stable representation of aerosol behavior, the dataset was preprocessed in MATLAB to calculate:

Hourly averages, suitable for daily pattern recognition

Daily averages, suitable for long-term and seasonal analysis

Functions were performed through custom MATLAB scripts that grouped data points by timestamp, averaged each parameter, and outputted new time-series tables. This approach automated the accuracy with no chance for manual processing errors. Both of the averaged datasets were then exported into structured form and prepared for entry into the final database.



1) Read DB sheets

```
Tdaily  = readtable(dbFile, "Sheet","Daily",  "VariableNamingRule","preserve");
Thourly = readtable(dbFile, "Sheet","Hourly", "VariableNamingRule","preserve");
```

2) Read NEW file (first sheet) - compatible way

```
[~, sheetList] = xlsfinfo(newFile);
if isempty(sheetList)
    error("No sheets found in new file: %s", newFile);
end
Tnew = readtable(newFile, "Sheet", sheetList{1}, "VariableNamingRule","preserve");

newVars = string(Tnew.Properties.VariableNames);
lvNew   = lower(newVars);
```

Figure 3-9 Hourly and Daily reading codes

3.2.3. Database integration into Excel

All datasets, including raw values, hourly averages, daily averages, and meteorological parameters, were compiled into one single consolidated Excel database after averaging. The database was arranged such that each row reflected a particular time, while each column corresponded to one variable.

This unified structure provides several advantages:

- ✓ Easy comparison of CN, PM fractions, and meteorological indicators.
- ✓ Direct importability in analytical tools and visualization software.
- ✓ Standardized framework for long-term data storage.
- ✓ Reduced risk of mismatched timestamps or fragmented datasets.

The consolidated Excel file is used as the main working database for the analyses made at subsequent stages of the thesis.

3.3. Database Update Framework

3.3.1. Structure for future data insertion

The database was designed in a modular and extendable fashion to incorporate new PROMO measurements that will be collected in the future. Each variable was assigned a dedicated column, while timestamps form the primary indexing parameter. This allows newly acquired data to be directly appended into the existing dataset without requiring adjustment of the structural form.

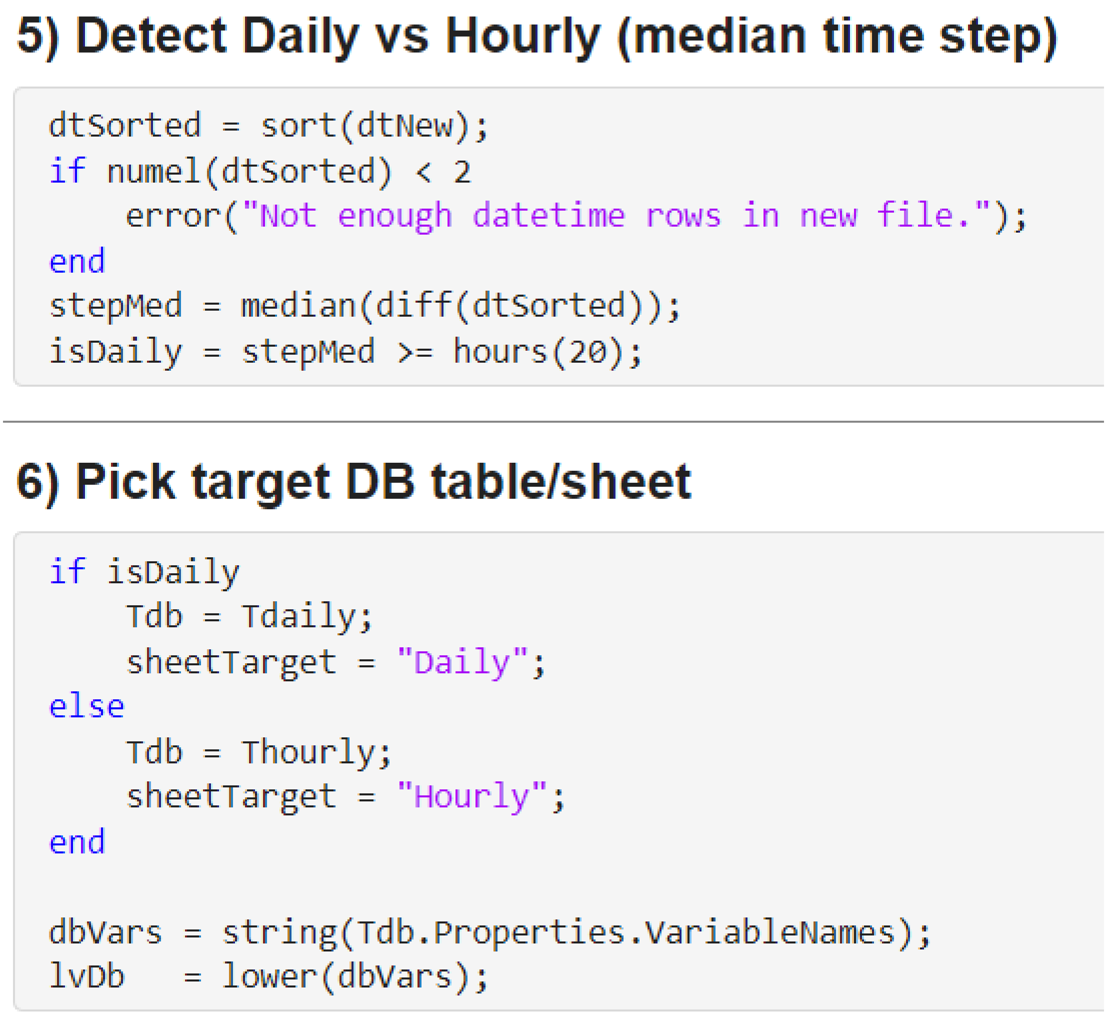
Unique identifiers, consistent variable naming, and standardized units ensure that additional datasets can be merged seamlessly. This modular approach supports long-term monitoring activities and reduces the need for repeated reformatting.

3.3.2. Automation approach to avoid time consumption

An automated workflow has been implemented to minimize manual workload and ensure efficient updates. Scripts developed in MATLAB during the averaging stage can be reused, processing incoming datasets. The scripts automatically execute whenever new PROMO files become available by:

- ✓ Identify and read the excel files
- ✓ Perform formatting and quality checks.
- ✓ Compute hourly and daily averages
- ✓ Append new data to the existing Excel database

This semi-automated pipeline ensures that any future integration of data will be much faster, consistent, and reproducible. The design also allows other users or researchers to update the dataset with minimum technical expertise.



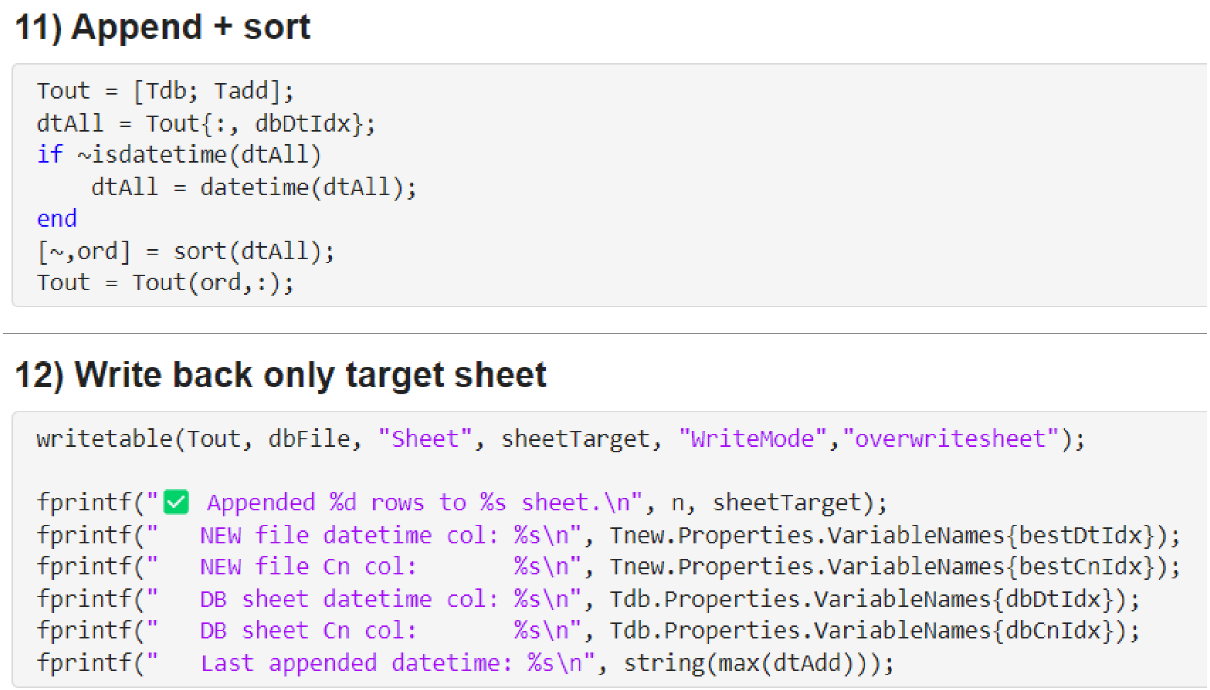
5) Detect Daily vs Hourly (median time step)

```
dtSorted = sort(dtNew);
if numel(dtSorted) < 2
    error("Not enough datetime rows in new file.");
end
stepMed = median(diff(dtSorted));
isDaily = stepMed >= hours(20);
```

6) Pick target DB table/sheet

```
if isDaily
    Tdb = Tdaily;
    sheetTarget = "Daily";
else
    Tdb = Thourly;
    sheetTarget = "Hourly";
end

dbVars = string(Tdb.Properties.VariableNames);
lvDb   = lower(dbVars);
```

Figure 3-10 Detecting daily and hourly data in MATLAB code



11) Append + sort

```
Tout = [Tdb; Tadd];
dtAll = Tout{:, dbDtIdx};
if ~isdatetime(dtAll)
    dtAll = datetime(dtAll);
end
[~,ord] = sort(dtAll);
Tout = Tout(ord,:);
```

12) Write back only target sheet

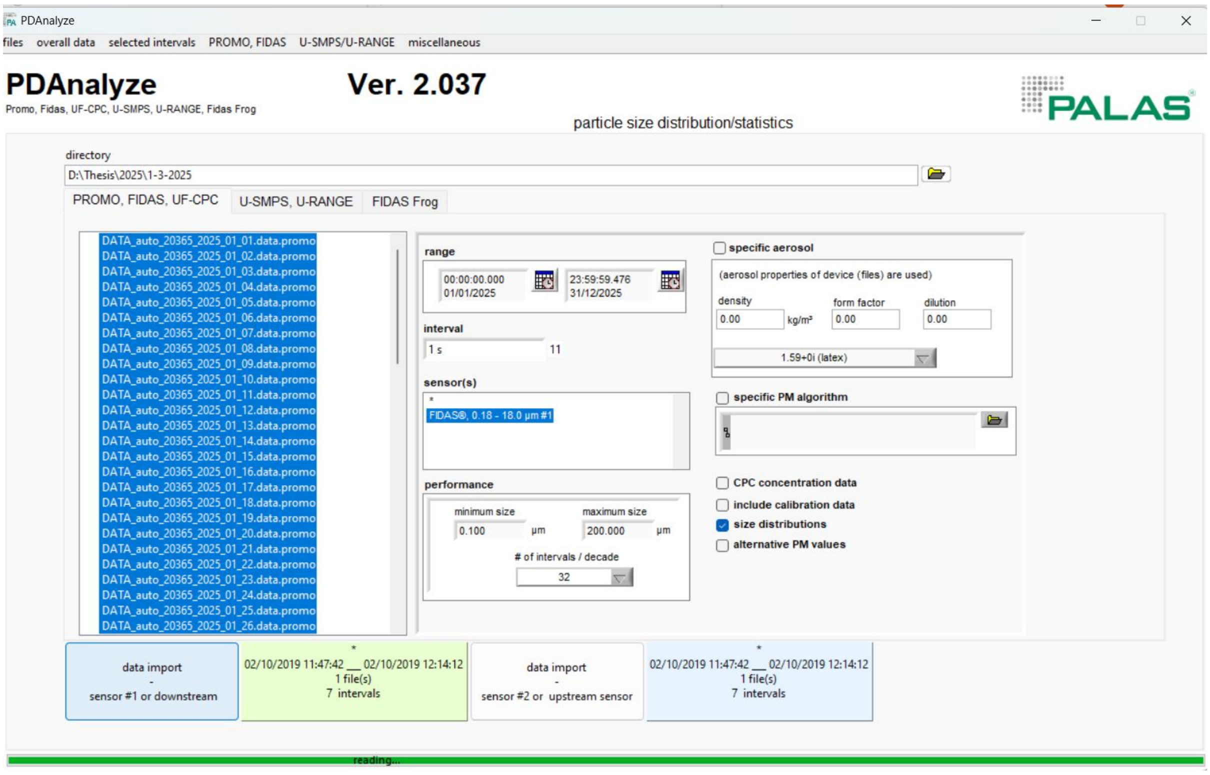
```
writetable(Tout, dbFile, "Sheet", sheetTarget, "WriteMode","overwritesheet");

fprintf("✅ Appended %d rows to %s sheet.\n", n, sheetTarget);
fprintf("   NEW file datetime col: %s\n", Tnew.Properties.VariableNames{bestDtIdx});
fprintf("   NEW file Cn col:       %s\n", Tnew.Properties.VariableNames{bestCnIdx});
fprintf("   DB sheet datetime col: %s\n", Tdb.Properties.VariableNames{dbDtIdx});
fprintf("   DB sheet Cn col:       %s\n", Tdb.Properties.VariableNames{dbCnIdx});
fprintf("   Last appended datetime: %s\n", string(max(dtAdd)));
```

Figure 3-11 Adding new column to main database file

3.4. Data Analysis

3.4.1. Using PDAnalyse software properly

PDAnalyse was used as the main tool for the interpretation of aerosol-related outputs, especially particle number concentrations and particle mass distributions. The software allows visualization of temporal patterns, computation of statistical descriptors, and identification of specific events like pollution peaks. To ensure correct usage, all PROMO data were imported into PDAnalyse in line with recommended formatting. Parameters like CN and PM fractions were analyzed by means of the graphical interfaces already implemented, ensuring uniform interpretation for different time periods. The analysis settings were standardized, to avoid possible variability caused by user-dependent configurations.



directory: D:\Thesis\2025\1-3-2025

DATA_auto_20365_2025_01_01.data.promo
DATA_auto_20365_2025_01_02.data.promo
DATA_auto_20365_2025_01_03.data.promo
DATA_auto_20365_2025_01_04.data.promo
DATA_auto_20365_2025_01_05.data.promo
DATA_auto_20365_2025_01_06.data.promo
DATA_auto_20365_2025_01_07.data.promo
DATA_auto_20365_2025_01_08.data.promo
DATA_auto_20365_2025_01_09.data.promo
DATA_auto_20365_2025_01_10.data.promo
DATA_auto_20365_2025_01_11.data.promo
DATA_auto_20365_2025_01_12.data.promo
DATA_auto_20365_2025_01_13.data.promo
DATA_auto_20365_2025_01_14.data.promo
DATA_auto_20365_2025_01_15.data.promo
DATA_auto_20365_2025_01_16.data.promo
DATA_auto_20365_2025_01_17.data.promo
DATA_auto_20365_2025_01_18.data.promo
DATA_auto_20365_2025_01_19.data.promo
DATA_auto_20365_2025_01_20.data.promo
DATA_auto_20365_2025_01_21.data.promo
DATA_auto_20365_2025_01_22.data.promo
DATA_auto_20365_2025_01_23.data.promo
DATA_auto_20365_2025_01_24.data.promo
DATA_auto_20365_2025_01_25.data.promo
DATA_auto_20365_2025_01_26.data.promo

range: 00:00:00.000 01/01/2025 — 23:59:59.476 31/12/2025

interval: 1 s 11

sensor(s): *
FIDAS®, 0.18 - 18.0 µm #1

performance
minimum size 0.100 µm
maximum size 200.000 µm
# of intervals / decade 32

□ specific aerosol
(aerosol properties of device (files) are used)
density 0.00 kg/m³
form factor 0.00
dilution 0.00
1.59+0i (latex)

□ specific PM algorithm

□ CPC concentration data
□ include calibration data
☑ size distributions
□ alternative PM values

data import
-
sensor #1 or downstream

*
02/10/2019 11:47:42 __ 02/10/2019 12:14:12
1 file(s)
7 intervals

data import
-
sensor #2 or upstream sensor

*
02/10/2019 11:47:42 __ 02/10/2019 12:14:12
1 file(s)
7 intervals

Figure 3-12 Primary setting of PDAnalyze software

3.4.2. Interpretation of CN and PM outputs

Characteristic patterns in the particle behavior were delineated from the processed CN and PM datasets. CN concentrations were studied for completes the behavior of PMs., while PM fractions were analyzed to understand the distribution of coarser particles.

Trends were analyzed with respect to meteorological conditions, which allowed one to assess the impact of atmospheric parameters on aerosol concentration. This integrative interpretation approach makes certain that the observed variations are related to physically meaningful mechanisms.

3.4.3. Visualization and statistical assessment

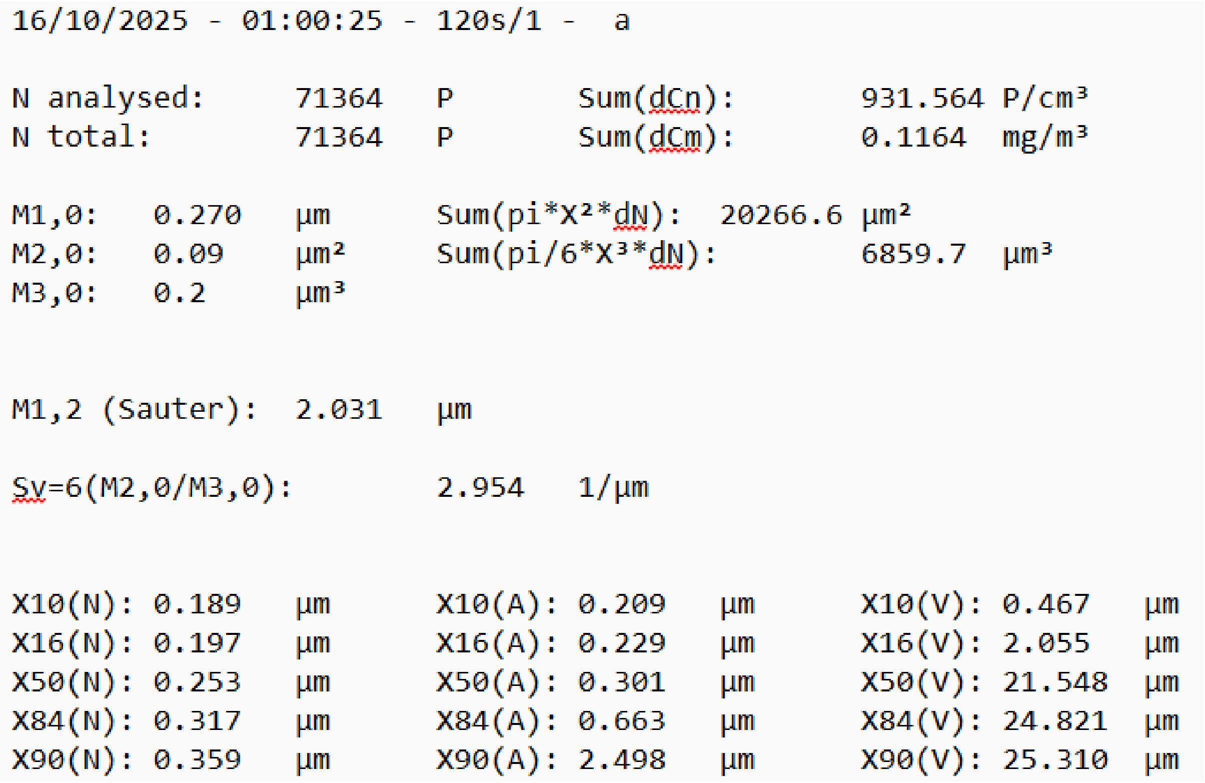
Several visualization techniques were applied to support quantitative interpretation, including:

- ✓ Time-series plots for CN and PM
- ✓ Scatter plots comparing aerosol and meteorological parameters
- ✓ Histograms to explore distribution shapes
- ✓ Correlation analysis to bring out relationships between variables.

These visual and statistical tools provide the underpinning for the temporal comparisons made in the next section of this thesis.

3.4.4. Validation of X-Percentile Particle Size Parameters (X10, X16, X50, X84, X90)

The output data obtained from the particle size analysis software was imported into Microsoft Excel for post-processing and validation purposes. The parameters X10, X16, X50, X84, and X90 were found to be the particle size distribution (PSD) diameters based on the interpretation of the output data from the software, which aligns with the ISO 9276-2 standard that describes the particle size diameters based on the cumulative percentage of the particle size distribution. To validate this assumption, numerical consistency checks were implemented on all the obtained data sets. To begin with, the monotonic ordering of the particle size diameters was evaluated to ensure that the following conditions were met: X10 < X16 < X50 < X84 < X90. Secondly, the median nature of the particle size diameter was evaluated to ensure that the obtained results were consistent with the average of the particle size diameters, which was obtained from the difference between the 84th and 16th particle size diameters. The symmetry of the particle size diameters was also evaluated based on the upper/lower particle size diameters to check for skewness of the particle size distribution.



```
16/10/2025 - 01:00:25 - 120s/1 -  a

N analysed:     71364   P       Sum(dCn):       931.564 P/cm³
N total:        71364   P       Sum(dCm):       0.1164  mg/m³

M1,0:   0.270   µm      Sum(pi*X²*dN):  20266.6 µm²
M2,0:   0.09    µm²     Sum(pi/6*X³*dN):        6859.7  µm³
M3,0:   0.2     µm³


M1,2 (Sauter):  2.031   µm

Sv=6(M2,0/M3,0):        2.954   1/µm


X10(N): 0.189   µm      X10(A): 0.209   µm      X10(V): 0.467   µm
X16(N): 0.197   µm      X16(A): 0.229   µm      X16(V): 2.055   µm
X50(N): 0.253   µm      X50(A): 0.301   µm      X50(V): 21.548  µm
X84(N): 0.317   µm      X84(A): 0.663   µm      X84(V): 24.821  µm
X90(N): 0.359   µm      X90(A): 2.498   µm      X90(V): 25.310  µm
```

Figure 3-13 Extracted values from PDAnalyze software in text file

These values are considered to be the diameters below which 10%, 16%, 50%, 84%, and 90% of the cumulative number of particles are present. This is based on the definition of percentile diameters in ISO 9276-2 for particle size distributions.

3.4.4.1. consistency check

To ensure that the X-parameters are true percentiles, a monotonic order test was performed on the data. For each time series, the condition X10 < X16 < X50 < X84 < X90 was evaluated. This is a necessary condition for cumulative percentiles. If this condition is not met, it would mean that there is inconsistency in the data generated by the software. The data satisfied the condition, and hence it was consistent with the definition of percentiles.

3.4.4.2. Median and symmetrical tests

The median property of X50 was tested by comparing it with the midpoint between X16 and X84 using the formula for normalized deviation:

$$\frac{\left|X_{50}-\left(\frac{X_{16}+X_{84}}{2}\right)\right|}{(X_{84}-X_{16})}$$

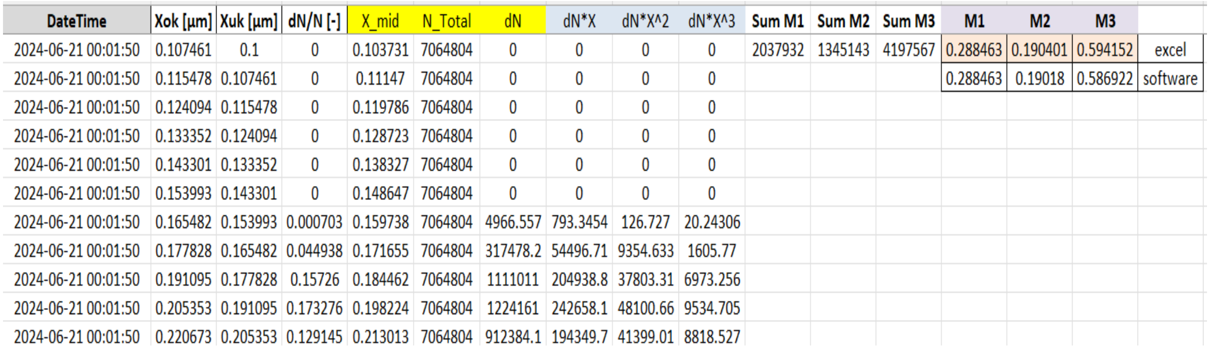
Furthermore, the symmetry of the distribution was tested using the formula $\frac{(X_{84}-X_{16})}{(X_{50}-X_{16})}$. A value close to unity indicates a distribution that is nearly symmetric, while values away from unity indicate skewness.

3.4.4.3. Validation

The results have confirmed that X10, X16, X50, X84, and X90 are true percentile descriptors of the particle size distribution. Despite the fact that the distributions were not strictly normal and right-skewed, this is not uncommon for particle size distributions and does not preclude the percentile description. Thus, the X-parameters calculated by the software were confirmed as cumulative percentile diameters in accordance with ISO 9276-2.

3.4.5. Calculation and Validation of Particle Size Distribution Moments (M1, M2, M3)

To ensure correct calculations, two methods were applied:
1) To understand the meaning of the parameters M1, M2, and M3, the particle size distribution data was exported from the PDAnalyze software and put into an Excel database. Here, the dataset included the values of the particle diameter for each size bin and the number of particles present. For each day, the size bins were constant, and the number of particles varied. Using this, it was possible to understand the particle size distribution without using the software. At the same time, the moment values calculated by the software were noted. This allows for a transparent validation of the results calculated by the software.
Particle size bins were represented by the variable X, where X is the midpoint diameter of the particles extracted directly from software. Using this data, the first three statistical moments of the distribution were determined. The first three statistical moments of the distribution were determined using the following equations in Microsoft Excel. The first moment of the distribution is given by the equation Σ(XdN), where dN is the number of particles in each bin. The second moment of the distribution is given by the equation $\sum(\pi \cdot X^2 \cdot \mathrm{dN})$, while the third moment of the distribution is given by the equation $\sum(\pi/6 \cdot X^3 \cdot \mathrm{dN})$. These equations were used to determine the values for each of the size bins for each of the daily data sets.
Lastly, the calculated values were compared with the values obtained from PDAnalyze software for the corresponding dates. From the comparison, it was evident that the results obtained independently matched closely with the results obtained from the software. Therefore, it can be concluded that the parameters M1, M2, and M3 are indeed related to the first, second, and third normalized moment of particle size distribution. Thus, the validation of results obtained from the software can be said to be reliable, as the parameters are based on their respective statistical definitions.



DateTime	Xok [µm]	Xuk [µm]	dN/N [-]	X_mid	N_Total	dN	dN*X	dN*X^2	dN*X^3	Sum M1	Sum M2	Sum M3	M1	M2	M3	
2024-06-21 00:01:50	0.107461	0.1	0	0.103731	7064804	0	0	0	0	2037932	1345143	4197567	0.288463	0.190401	0.594152	excel
2024-06-21 00:01:50	0.115478	0.107461	0	0.11147	7064804	0	0	0	0				0.288463	0.19018	0.586922	software
2024-06-21 00:01:50	0.124094	0.115478	0	0.119786	7064804	0	0	0	0							
2024-06-21 00:01:50	0.133352	0.124094	0	0.128723	7064804	0	0	0	0							
2024-06-21 00:01:50	0.143301	0.133352	0	0.138327	7064804	0	0	0	0							
2024-06-21 00:01:50	0.153993	0.143301	0	0.148647	7064804	0	0	0	0							
2024-06-21 00:01:50	0.165482	0.153993	0.000703	0.159738	7064804	4966.557	793.3454	126.727	20.24306							
2024-06-21 00:01:50	0.177828	0.165482	0.044938	0.171655	7064804	317478.2	54496.71	9354.633	1605.77							
2024-06-21 00:01:50	0.191095	0.177828	0.15726	0.184462	7064804	1111011	204938.8	37803.31	6973.256							
2024-06-21 00:01:50	0.205353	0.191095	0.173276	0.198224	7064804	1224161	242658.1	48100.66	9534.705							
2024-06-21 00:01:50	0.220673	0.205353	0.129145	0.213013	7064804	912384.1	194349.7	41399.01	8818.527							

Figure 3-14 Extracted and calculated M-values in excel for 21/06/2024

2) The particle size distribution data was exported from the software to Microsoft Excel, and the data included the lower and upper limits of each size bin (Xuk and Xok), the relative number fraction (dN/N), and the total particle number (N_{Total}l) for each time step. The midpoint diameter of each bin (X_{mid}) was calculated.

3.4.5.1. *Calculation of particle counts (dN)*

The number of particles in each size class (dN) was calculated using the formula $dN = \left(\frac{dN}{N}\right) \times N_{Total}$.

This calculation allowed the use of the normalized particle fractions to obtain the actual number of particles. The obtained values of dN were then used to calculate the surface- and volume-related contributions of each size class to the total number of particles.

3.4.5.2. *Calculation of moments M1, M2, and M3*

The first moment (M1), representing the mean particle diameter, was calculated as

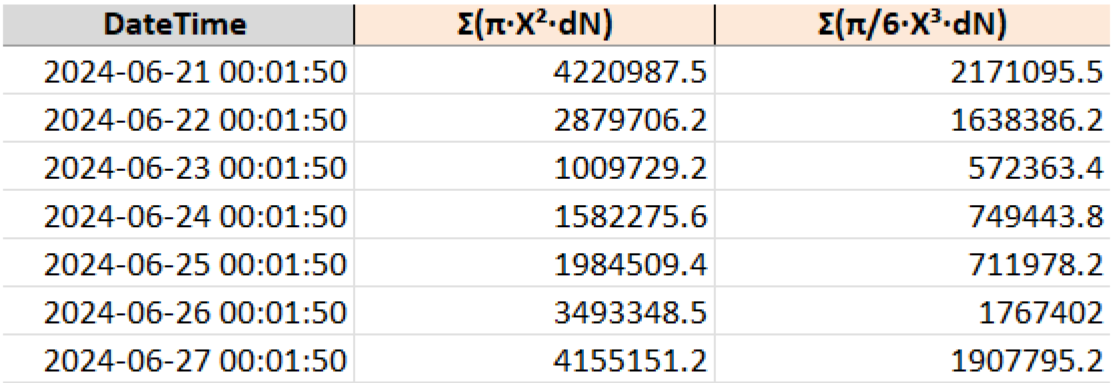
M1=$\sum(X_{mid} \cdot \mathrm{dN/N})$.

The second moment (M2), related to total particle surface area, was calculated as

M2=$\sum(\pi \cdot {X_{mid}}^2 \cdot \mathrm{dN})$.

The third moment (M3), related to total particle volume, was calculated as M3=$\sum(\pi/6 \cdot {X_{mid}}^3 \cdot \mathrm{dN})$.

These summations were performed for each time record in Excel over all size bins.



DateTime	$\Sigma(\pi\cdot X^2\cdot dN)$	$\Sigma(\pi/6\cdot X^3\cdot dN)$
2024-06-21 00:01:50	4220987.5	2171095.5
2024-06-22 00:01:50	2879706.2	1638386.2
2024-06-23 00:01:50	1009729.2	572363.4
2024-06-24 00:01:50	1582275.6	749443.8
2024-06-25 00:01:50	1984509.4	711978.2
2024-06-26 00:01:50	3493348.5	1767402
2024-06-27 00:01:50	4155151.2	1907795.2

Figure 3-15 Extracted moments values from software

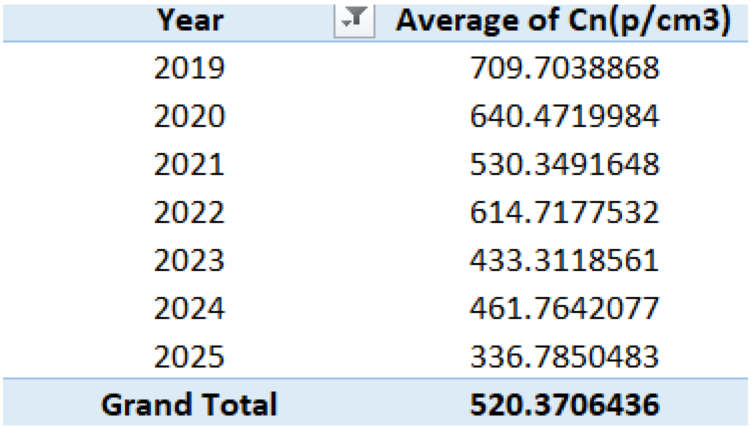
3.4.5.3. Validation against software output

The values of M1, M2, and M3 derived from the Excel calculations were compared with the values provided by the particle size analysis software. The good agreement between the calculated values of the moments and the software values ensured that the software values are indeed the values of the first, second, and third-order moments of the PSD. This ensures that M1, M2, and M3 are indeed the number-weighted mean diameter, surface-related moment, and volume-related moment of the PSD.

3.5. Temporal and Seasonal Comparison

3.5.1. Year-to-year comparison

The processed dataset was divided into yearly segments to assess the long-term variation in aerosol behavior. For each year. time series plots and annual trend lines were used to detect increases, decreases, or stable patterns in air quality indicators.
Comparisons across years also enable the identification of systematic changes, such as the influence of meteorological anomalies, emission variations, or potential environmental events. Multi-year datasets are analyzed in this study to assess whether the observed trends are consistent, transient, or driven by specific external factors.

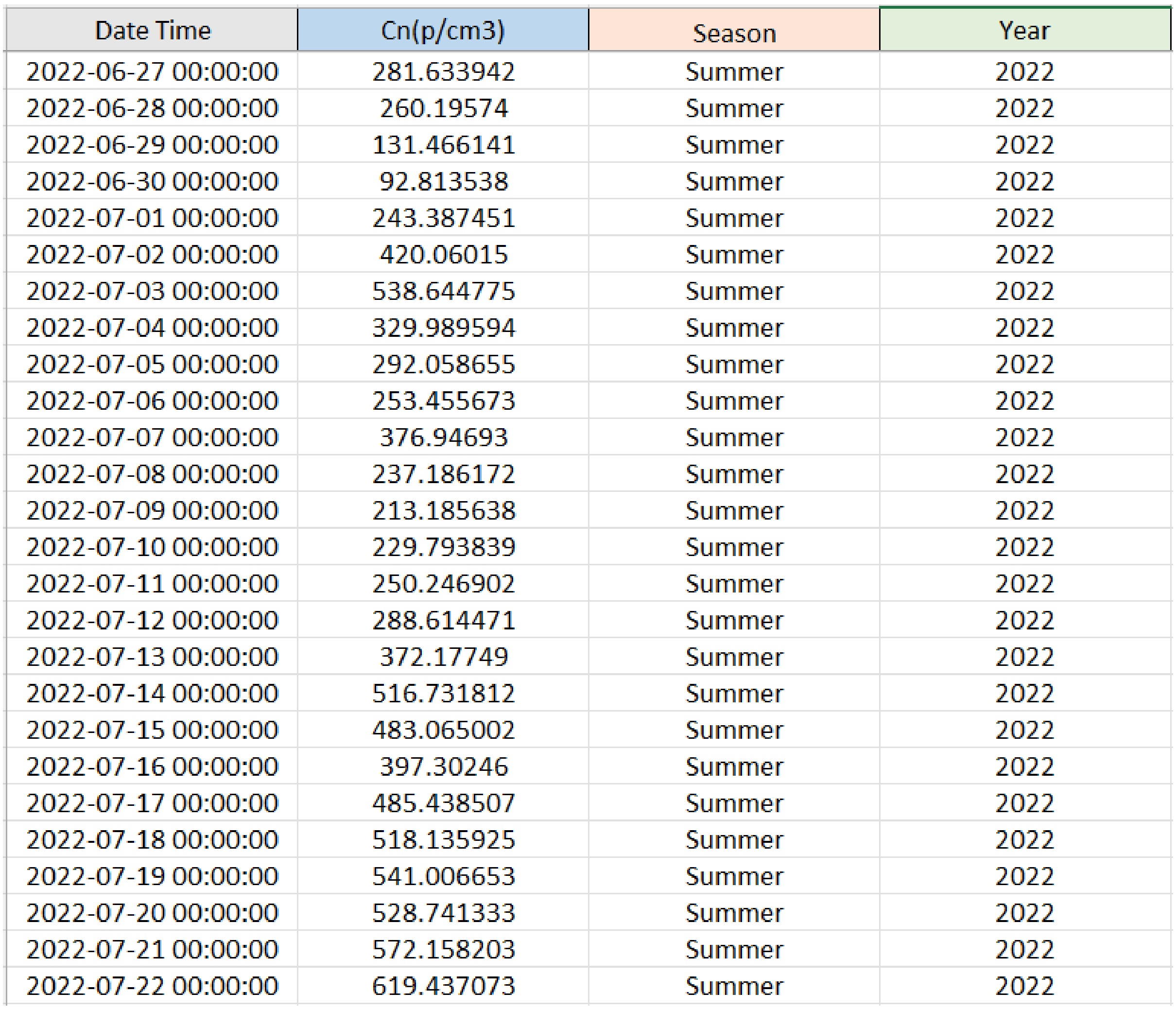


Year	Average of Cn(p/cm3)
2019	709.7038868
2020	640.4719984
2021	530.3491648
2022	614.7177532
2023	433.3118561
2024	461.7642077
2025	336.7850483
Grand Total	**520.3706436**

Figure 3-16 Table of average Cn from 2019 to 2025

3.5.2. Seasonal variation analysis

Besides annual comparisons, the dataset was also divided into four meteorological seasons: winter, spring, summer, and autumn. Seasonal averages of CN and PM concentrations were calculated and compared to highlight typical patterns associated with temperature, humidity, and so on.
Seasonal variations are crucial in understanding the dynamics of particulate matter. Fine particles often present stronger variations during low boundary-layer height periods, while coarse particles may be affected by outdoor activities. The seasonal analysis will give further interpretation of the temporal behavior of the aerosol concentration.



Date Time	Cn(p/cm3)	Season	Year
2022-06-27 00:00:00	281.633942	Summer	2022
2022-06-28 00:00:00	260.19574	Summer	2022
2022-06-29 00:00:00	131.466141	Summer	2022
2022-06-30 00:00:00	92.813538	Summer	2022
2022-07-01 00:00:00	243.387451	Summer	2022
2022-07-02 00:00:00	420.06015	Summer	2022
2022-07-03 00:00:00	538.644775	Summer	2022
2022-07-04 00:00:00	329.989594	Summer	2022
2022-07-05 00:00:00	292.058655	Summer	2022
2022-07-06 00:00:00	253.455673	Summer	2022
2022-07-07 00:00:00	376.94693	Summer	2022
2022-07-08 00:00:00	237.186172	Summer	2022
2022-07-09 00:00:00	213.185638	Summer	2022
2022-07-10 00:00:00	229.793839	Summer	2022
2022-07-11 00:00:00	250.246902	Summer	2022
2022-07-12 00:00:00	288.614471	Summer	2022
2022-07-13 00:00:00	372.17749	Summer	2022
2022-07-14 00:00:00	516.731812	Summer	2022
2022-07-15 00:00:00	483.065002	Summer	2022
2022-07-16 00:00:00	397.30246	Summer	2022
2022-07-17 00:00:00	485.438507	Summer	2022
2022-07-18 00:00:00	518.135925	Summer	2022
2022-07-19 00:00:00	541.006653	Summer	2022
2022-07-20 00:00:00	528.741333	Summer	2022
2022-07-21 00:00:00	572.158203	Summer	2022
2022-07-22 00:00:00	619.437073	Summer	2022

Figure 3-17 Indicating seasons and years for plotting comparison charts in excel file

3.6. Validation and Quality Control

3.6.1. Handling missing and abnormal values

The quality control procedures have been implemented throughout the pipeline of data processing, ensuring that the final dataset is reliable for analysis. During the pre-processing, several problems were identified, such as missing timestamps, incomplete performance of sensors, or physically unrealistic values.

Actions taken varied according to the issue:

- ✓ The isolated missing points remained as empty cells.
- ✓ Long gaps were flagged and excluded from comparative analysis.

These procedures prevent numerical instabilities and ensure that trends detected during the analysis are based on trustworthy data.

3.6.2. Ensuring consistency between PROMO and meteorological data

All PROMO variables were synchronized with the meteorological measurements using the original timestamps to maintain consistency among the datasets. During the integration into the Excel database, any mismatch between the two datasets was corrected.

Cross-checks were carried out to ensure that the number of records in each timestamp was matched, and that all variables fell correctly. This ensured that all multi-variable analyses such as correlations between aerosol values and meteorological were accurate and free from temporal bias.

3.7. Automated Data Integration

3.7.1. Data import and database structure

The output files created by the measurement software were exported in Microsoft Excel format and incorporated into a centralized database using MATLAB. The database was organized into two major tables corresponding to hourly averaged and daily averaged data. Each new data set was treated separately and added to the corresponding table according to its time resolution. This enables the database to be continuously expanded as new data sets become available.

3.7.2. Automatic detection of datetime and measurement variables

For every newly imported Excel file, the script was able to automatically determine the datetime column by trying all possible columns and picking the one that could be converted to a valid datetime format the most. Both numerical serial date formats and text date formats were supported.

The measurement variables (concentration, size-related values, or other calculated values) were determined by looking for column titles and picking the column with the largest number of valid numeric values, excluding columns related to datetime information. This approach can be generally used for other variables besides Cn.

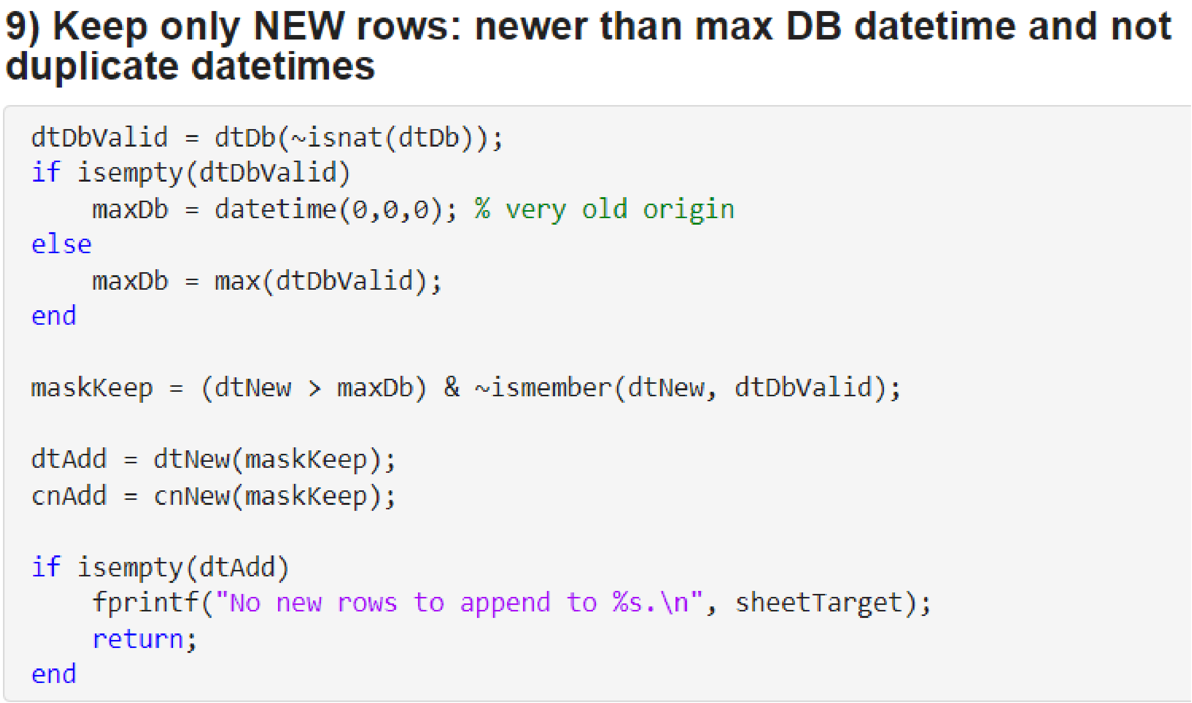
3.7.3. Temporal resolution classification

The temporal resolution of each dataset was calculated by finding the median time difference between consecutive timestamps. A dataset was considered to be daily data if the median time difference was greater than 20 hours, and it was considered to be hourly data otherwise. Based on this classification, new records were assigned to the appropriate table (hourly or daily) in the main database.

3.7.4. Data validation and database update

Before adding new records, the invalid timestamps were eliminated, and all data were formatted in a uniform numeric and datetime format. Only the records with timestamps more recent than the highest timestamp in the database were kept. Duplicates were also eliminated to avoid repetition.

The verified records were then added to the database and sorted by timestamp.



9) Keep only NEW rows: newer than max DB datetime and not duplicate datetimes

```
dtDbValid = dtDb(~isnat(dtDb));
if isempty(dtDbValid)
    maxDb = datetime(0,0,0); % very old origin
else
    maxDb = max(dtDbValid);
end

maskKeep = (dtNew > maxDb) & ~ismember(dtNew, dtDbValid);

dtAdd = dtNew(maskKeep);
cnAdd = cnNew(maskKeep);

if isempty(dtAdd)
    fprintf("No new rows to append to %s.\n", sheetTarget);
    return;
end
```

Figure 3-18 Preparing database for adding new data

3.7.5. Reproducibility and extensibility of the workflow

The developed MATLAB process ensures an automated and repeatable procedure for updating the database by utilizing the new software output files. The user is required to export the measurement results in an Excel format and process the script to incorporate the new information.

Since the detection of the datetime and measurement variables is generic and does not confine to a particular parameter (Cn), the approach can be used for various output variables and future data without any changes. This ensures scalability and future use of the database for further modeling.

Chapter 4. Results

4.1. The meaning of Cn in PDAnalyze

It was evaluated based on eight criteria that CN is an indicator of fine/ultrafine particle concentrations.

4.1.1. daily CN concentration

The overall CN concentrations remain relatively stable over the seven-year period, with recurrent peaks and no clear long-term upward or downward trend. Sharp clustered peaks appear repeatedly each year, suggesting episodic short-term emission increases.

Interannual variability is evident: some years show more pronounced CN peaks, indicating stronger particle formation episodes or meteorological conditions favoring accumulation.

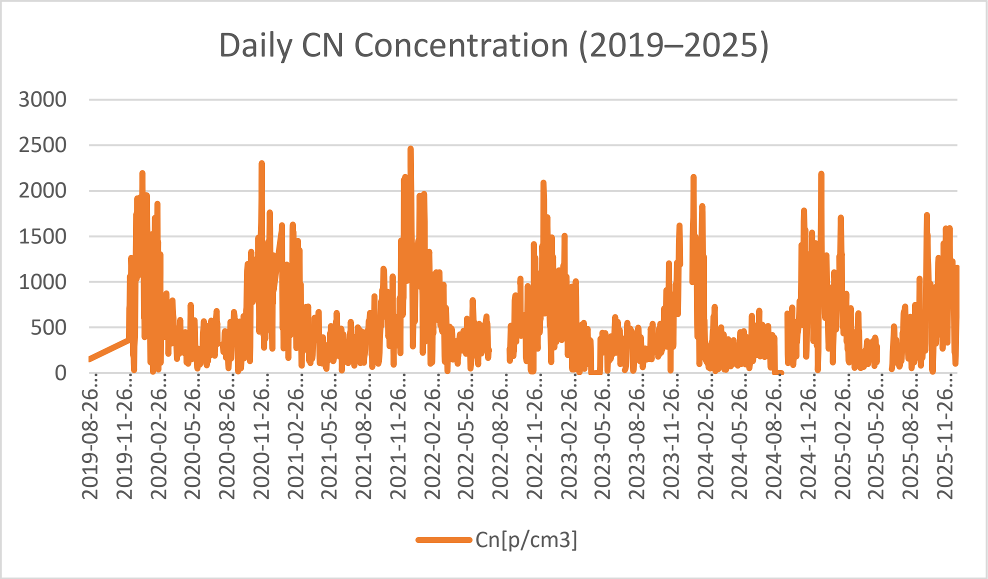


Figure 4-1 Daily CN Concentration (2019–2025)

Figures 4-2 to 4-5 illustrate the daily CN concentration for each individual year, ranging from 2019 to 2025.

From the figures, it can be seen that there is a seasonal pattern in each year. In each year, it can be noted that the CN values increase during the colder months and decrease during the warmer months. This further confirms the influence of seasonal meteorological conditions on particle number concentration.

From the figures, it can also be noted that the amplitude and frequency of the CN peaks differ in each year. For instance, some years, such as 2020 and 2022, show an increase in the frequency and amplitude of the peaks, indicating an increase in pollution and conditions favoring particle concentration.

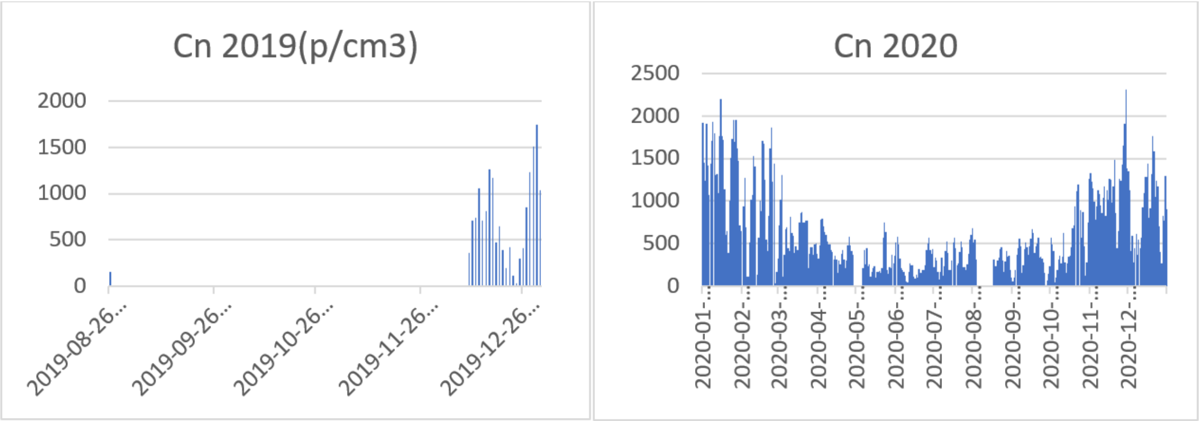


Figure 4-2 Daily average CN for 2019&2020

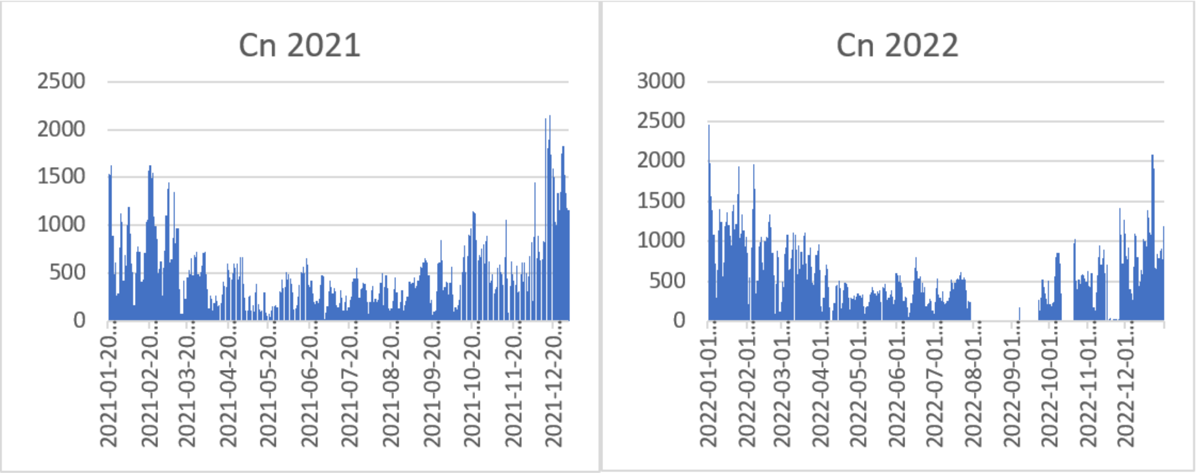


Figure 4-3 Daily average CN for 2020&2022

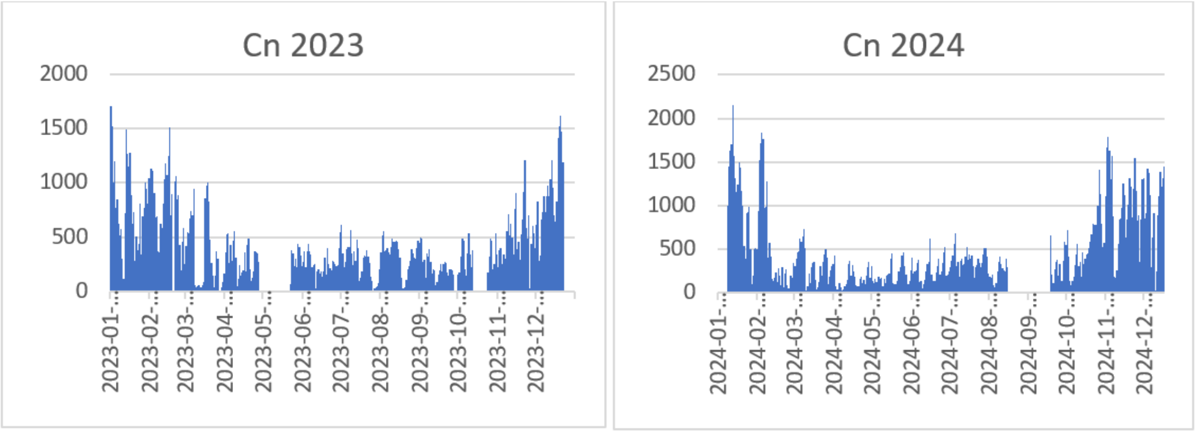


Figure 4-4 Daily average CN for 2023&2024

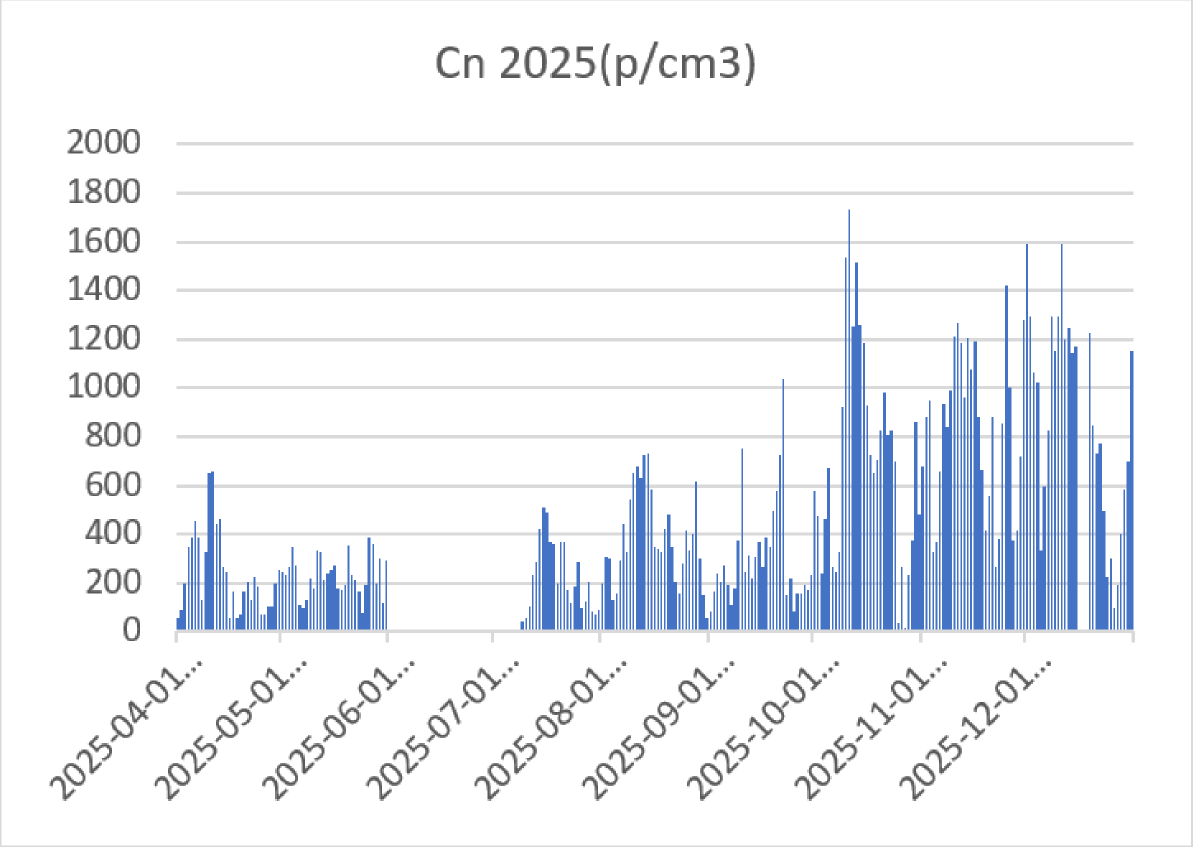


Figure 4-5 Daily average CN for 2025

This trend can, however, not be identified for the seven-year period. Instead, the values of CN exhibit interannual variability, with periods of increased and reduced concentrations depending on the emissions and atmospheric conditions.

The fact that the peaks are clustered suggests that the concentrations of CN are mainly controlled by episodic events rather than continuous emissions. These episodic events can be related to traffic intensity, combustion, and unfavorable atmospheric conditions such as temperature inversion. The similarity of the patterns for different years suggests the stability of the seasonal cycle, confirming that the concentrations of CN are strongly controlled by environmental conditions.

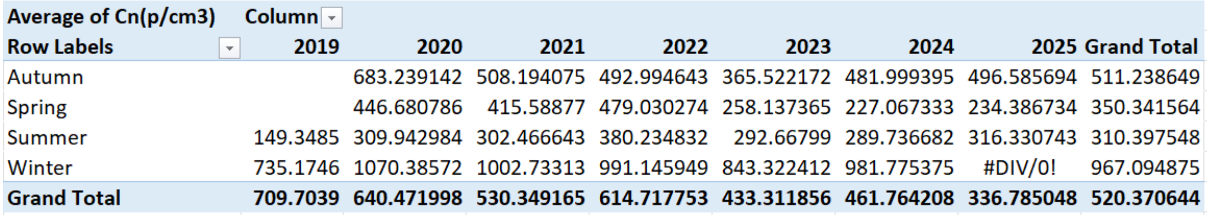
The yearly plots for CN offer further support that the number of particles is strongly variable on short timescales, has a strong seasonal cycle, and remains relatively stable on long timescales.

4.1.2. Seasonal CN pattern

Extended low-CN periods (close to zero value) likely coincide with favourable meteorological conditions such as precipitation or high wind speeds that remove ultrafine particles from the atmosphere.

The distribution of daily values suggests a seasonal cycle, with higher CN concentrations generally occurring during colder months and lower values during summer periods.

The amplitude of CN variability differs among years, indicating that meteorology and emission dynamics change significantly between periods. For example, 2020 will see a bigger drop, probably due to the lockdown.



Average of Cn(p/cm3)	Column							
Row Labels	2019	2020	2021	2022	2023	2024	2025	Grand Total
Autumn		683.239142	508.194075	492.994643	365.522172	481.999395	496.585694	511.238649
Spring		446.680786	415.58877	479.030274	258.137365	227.067333	234.386734	350.341564
Summer	149.3485	309.942984	302.466643	380.234832	292.66799	289.736682	316.330743	310.397548
Winter	735.1746	1070.38572	1002.73313	991.145949	843.322412	981.775375	#DIV/0!	967.094875
Grand Total	**709.7039**	**640.471998**	**530.349165**	**614.717753**	**433.311856**	**461.764208**	**336.785048**	**520.370644**

Figure 4-6 Table of seasonal average CN (2019-2025)

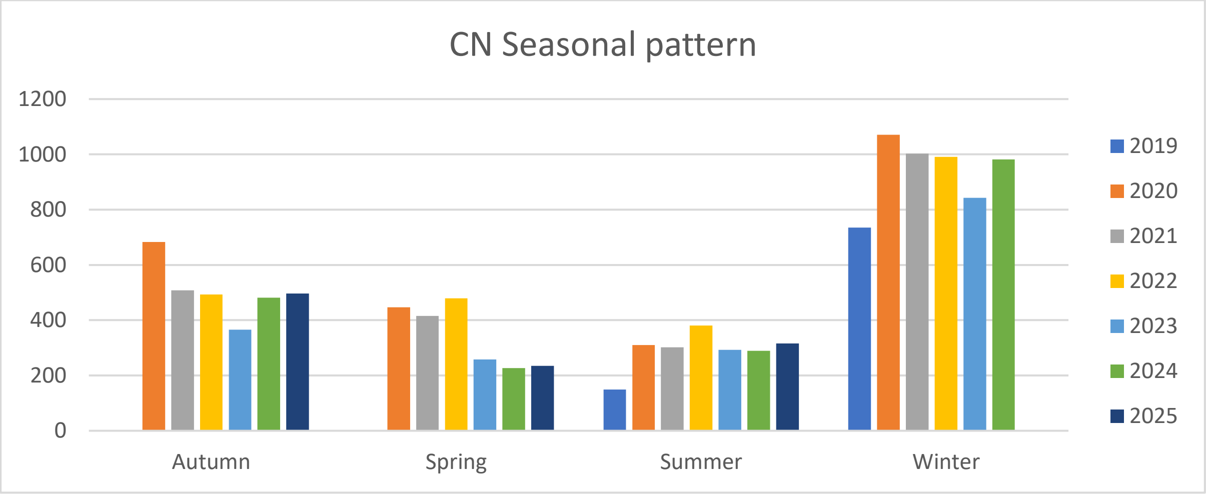


Figure 4-7 CN Seasonal pattern (2019-2025)

Winter has the highest CN in all years and more polluted years are 2020, 2021 and 2024 with very high Winter peaks (~1000–1100 p/cm³) while in all years, summer has the lowest value (approximately 250–350 p/cm³) that means cleanest air conditions for ultrafine particles and because of small interannual variations is the most stable season. In Autumn pollutants start to accumulate. Although the seasonal pattern remains consistent across years, the magnitude of CN concentrations varies significantly, indicating that interannual meteorological differences strongly influence ultrafine particle levels.

In conclusion, all years show highest Cn in winter and lowest Cn in summer that is special behavior of Ultrafine Particles.

4.1.3. Annual CN Trends Analysis

CN is on a downward trend from 2019 to 2025 (from about 700 to 330). This decline suggests that:

Either the sources of UFP production (e.g. traffic, combustion, industrial activity) have decreased, or meteorological conditions have created better ventilation.

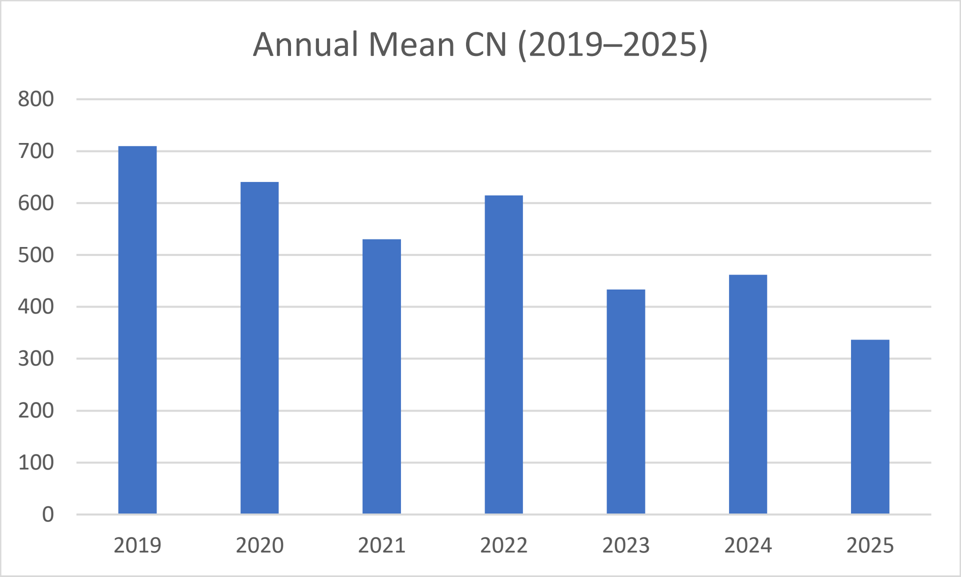


Figure 4-8 Annual Mean CN (2019–2025)

4.1.4. PM2.5 & PM10 Annual Trend

In PM, like CN, a clear annual decline is obvious in which the convergence of CN and PM means that CN is most likely an indicator of fine/ultrafine particle concentrations, and its annual behavior is consistent with PM, but its rate of change is larger. In overall CN tends to increase in conditions where PM2.5/PM10 do not increase very much. This means that CN, unlike total PM, is representative for the number of ultrafine particles (UFP), not the mass of the particles. What was important about 2019 was that the data was very small, and we only had data recorded for 08/26/2019, and the period of 12/10/209 to 12/31/2019. That is why the graphs have an odd shape for 2019.

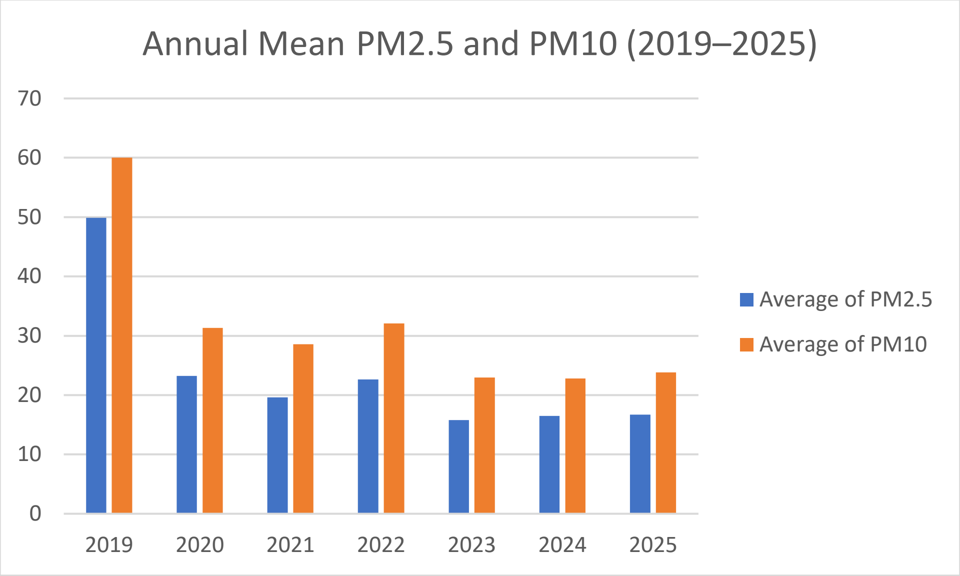
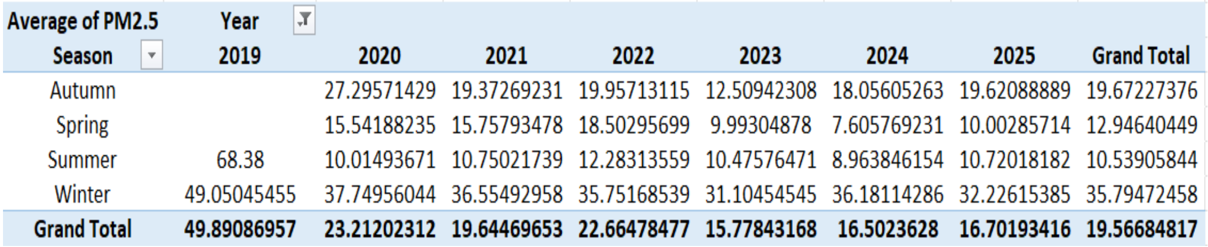


Figure 4-9 Annual Mean PM2.5 and PM10 (2019–2025)

Figures 4-10 and 4-13 present the average concentration values of PM2.5 and PM10 during each season in different years from 2019 to 2025.

The figures show that there is a clear and consistent pattern in the concentrations of PM2.5 and PM10 during different seasons in different years.

The concentration values of PM2.5 show that the maximum values occur during the winter season, and the minimum values occur during the summer season. Similarly, the concentration values of PM10 also show that the maximum values occur during the winter season and the minimum values occur during the summer season. However, the concentration values of PM10 show greater variation compared to PM2.5 during different seasons. For example, in some years, the values in the summer season are higher than those in PM2.5. This could be due to dust particles that are usually suspended in the air during the summer season.



Average of PM2.5	**Year**							
Season	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**	**2025**	**Grand Total**
Autumn		27.29571429	19.37269231	19.95713115	12.50942308	18.05605263	19.62088889	19.67227376
Spring		15.54188235	15.75793478	18.50295699	9.99304878	7.605769231	10.00285714	12.94640449
Summer	68.38	10.01493671	10.75021739	12.28313559	10.47576471	8.963846154	10.72018182	10.53905844
Winter	49.05045455	37.74956044	36.55492958	35.75168539	31.10454545	36.18114286	32.22615385	35.79472458
Grand Total	**49.89086957**	**23.21202312**	**19.64469653**	**22.66478477**	**15.77843168**	**16.5023628**	**16.70193416**	**19.56684817**

Figure 4-10 Table of seasonal average PM2.5(2019-2025)

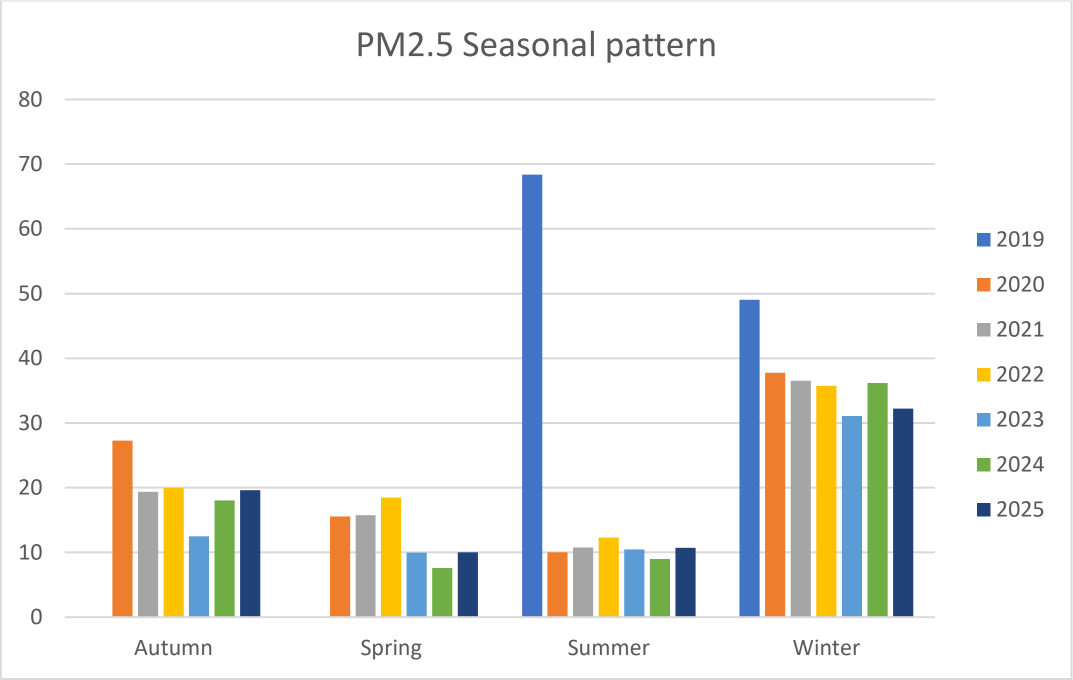
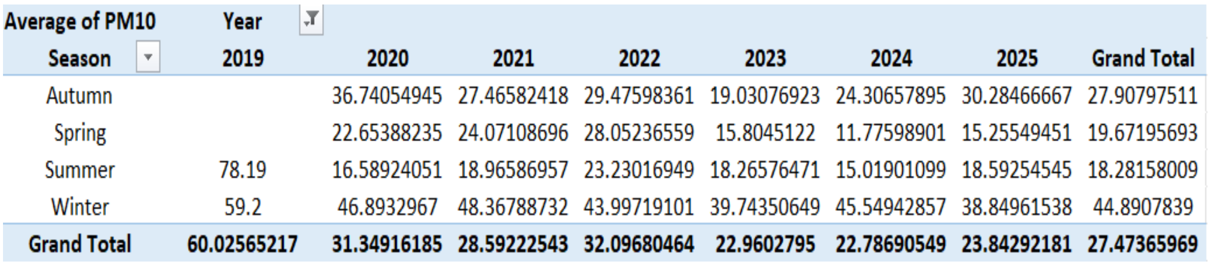


Figure 4-11 PM2.5 seasonal pattern (2019-2025)



Average of PM10	Year							
Season	2019	2020	2021	2022	2023	2024	2025	Grand Total
Autumn		36.74054945	27.46582418	29.47598361	19.03076923	24.30657895	30.28466667	27.90797511
Spring		22.65388235	24.07108696	28.05236559	15.8045122	11.77598901	15.25549451	19.67195693
Summer	78.19	16.58924051	18.96586957	23.23016949	18.26576471	15.01901099	18.59254545	18.28158009
Winter	59.2	46.8932967	48.36788732	43.99719101	39.74350649	45.54942857	38.84961538	44.8907839
Grand Total	**60.02565217**	**31.34916185**	**28.59222543**	**32.09680464**	**22.9602795**	**22.78690549**	**23.84292181**	**27.47365969**

Figure 4-12 Table of seasonal average PM10 (2019-2025)

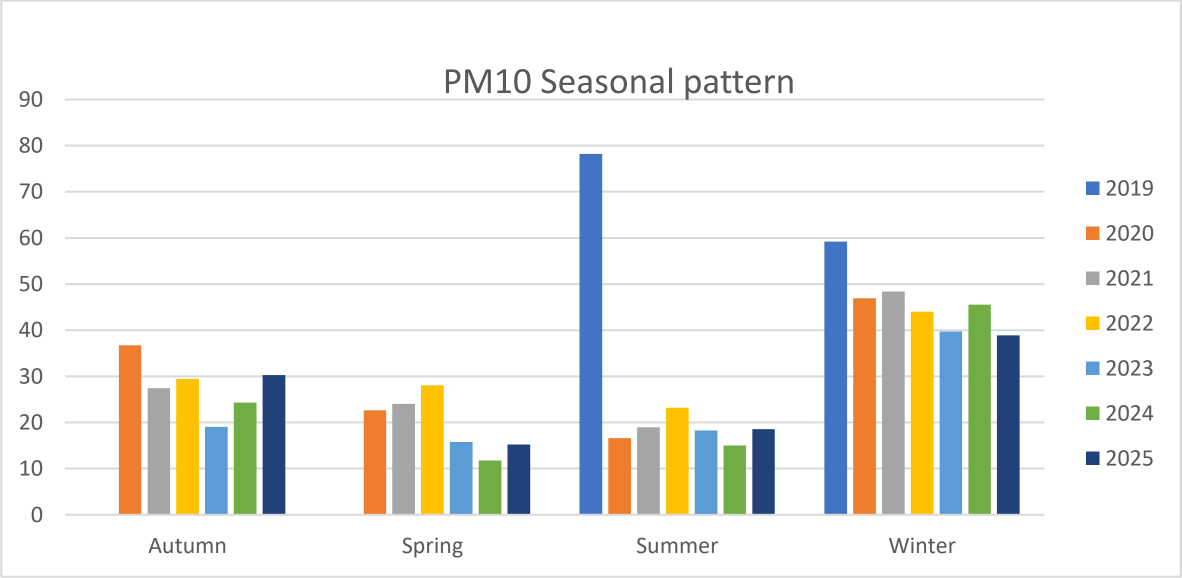


Figure 4-13 PM10 seasonal pattern (2019-2025)

Interannual changes can also be seen in PM2.5 and PM10. For example, previous years like 2020 have greater seasonal averages, and later years have generally lower values. This confirms the annual decreasing trend in PM values and supports the hypothesis of an overall improvement in air quality over time.

When comparing PM2.5 and PM10 with CN, similar long-term and seasonal changes can be seen. However, greater changes and greater sensitivity to changes in conditions can be noted for CN. This supports the hypothesis that CN represents the number of ultrafine particles, whereas PM2.5 and PM10 represent the mass of particles. Thus, CN can increase in conditions in which PM2.5 and PM10 do not change much.

In summary, seasonal plots of PM2.5 and PM10 support the hypothesis of strong control of PM values by seasonal factors. Winter pollution events dominate the annual changes, and summer conditions have generally lower and more constant values.

4.1.5. Hourly CN high-pollution week

A week with highest Cn peak in winter between 6 years was founded and the hourly CN profile during one selected winter high-pollution week (29 December 2021 to 4 January 2022) shows a pronounced peak exceeding 4500 p/cm³ at the beginning of the period, which is likely to be associated with a strong emission event or low boundary-layer conditions. After that peak, CN rapidly decreases and stabilizes between 1000-1500 p/cm³, showing typical wintertime accumulation under stagnant atmospheric conditions. Several small-scale hourly fluctuations occur during this whole week due to temperature, wind speed, and local emissions. At the end of the week, CN levels gradually decay below 1000 p/cm³, indicating improved ventilation in the atmosphere.

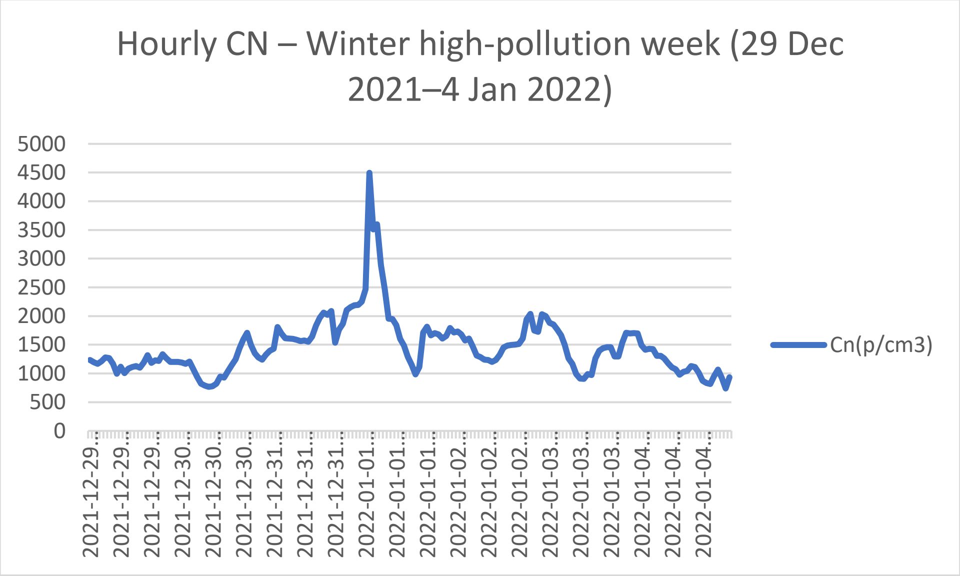


Figure 4-14 Hourly CN – Winter high-pollution week (29 Dec 2021–4 Jan 2022)

4.1.6. Hourly CN low-pollution week

A week with lowest Cn peak in winter between 6 years was clarified and the hourly CN concentrations during the selected summer week (21–27 June 2024) reflect a typical low-pollution summer pattern. CN levels are relatively high at the very beginning of the period but drop sharply on the second day of the period and remain mostly in the range between 50 and 150 p/cm³ for several days, indicating strong mixing in the atmosphere and efficient dispersion. From 24 June onward, CN increases gradually and on 26–27 June reaches several pronounced peaks, very likely associated with short-term local emissions. Overall, the week is characterized by low baseline CN values with periodic high-frequency peaks, reflecting dynamic but generally clean summer atmospheric conditions.

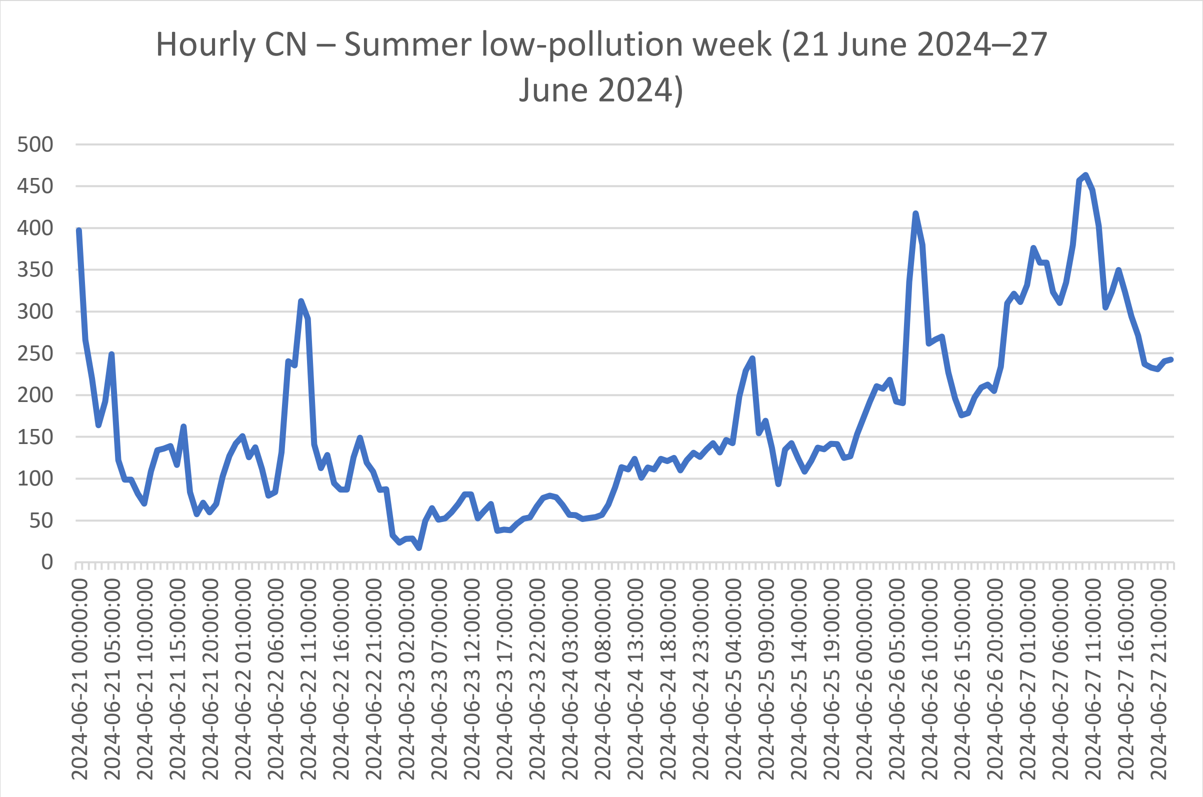


Figure 4-15 Hourly CN – Summer low-pollution week (21 June 2024–27 June 2024)

4.1.7. Two-day extreme peak

2 days with highest Cn peak between 6 years were selected. To better understand the behaviour of ultrafine particle surges, the two most polluted days of the six-year dataset were analysed at hourly resolution. It follows that the analysis of the two most polluted days in the six-year dataset, 31 December 2021–1 January 2022, delivers the characteristic behaviour of an extreme ultrafine particle event. CN concentrations rapidly rise to values above 4500 p/cm³, revealing a strong emission or high-intensity emission episode under very stagnant winter atmospheric conditions. The peak is followed by a steep decline, showing that once mixing increases,

ultrafine particles disperse quickly. Throughout the rest of the period, CN remains elevated between 1000 and 2000 p/cm³, reflecting the persistent accumulation typical of wintertime low boundary-layer heights. These plots of extreme events provide insight into the maximum intensity, temporal structure, and decay rate of severe CN pollution episodes-information that cannot be conveyed through daily or seasonal averages.

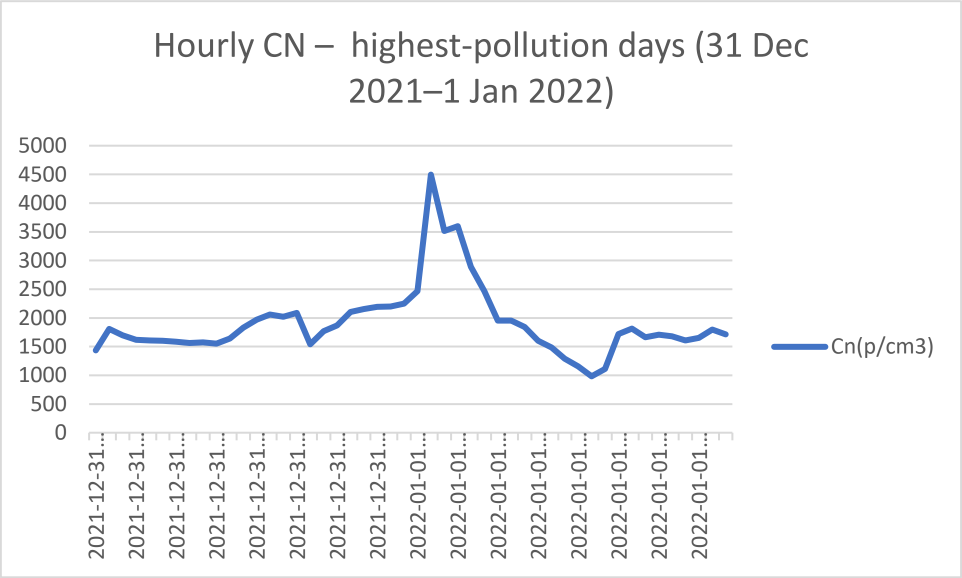


Figure 4-16 Hourly CN – highest-pollution days (31 Dec 2021–1 Jan 2022)

4.1.8. Comparison of Daily CN with PM2.5 and PM10

Figures 4-17 to 4-20 show the comparison between the daily variation of concentrations of PM2.5 and PM10 and the daily CN for different years.

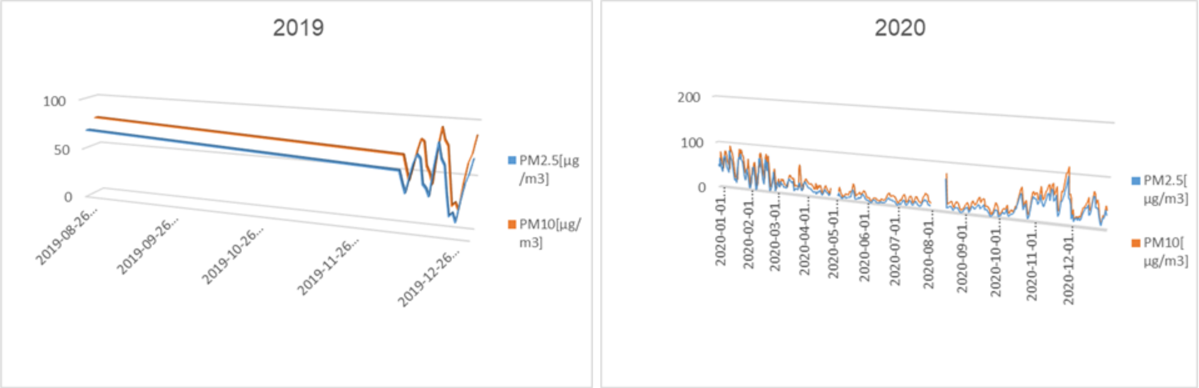


Figure 4-17 Daily average of PM2.5 & PM10 for 2019&2020

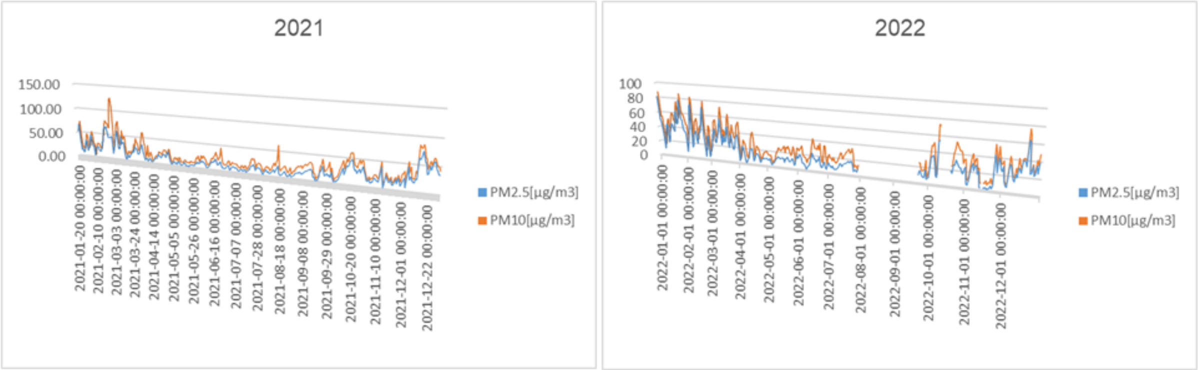


Figure 4-18 Daily average of PM2.5 & PM10 for 2021&2022

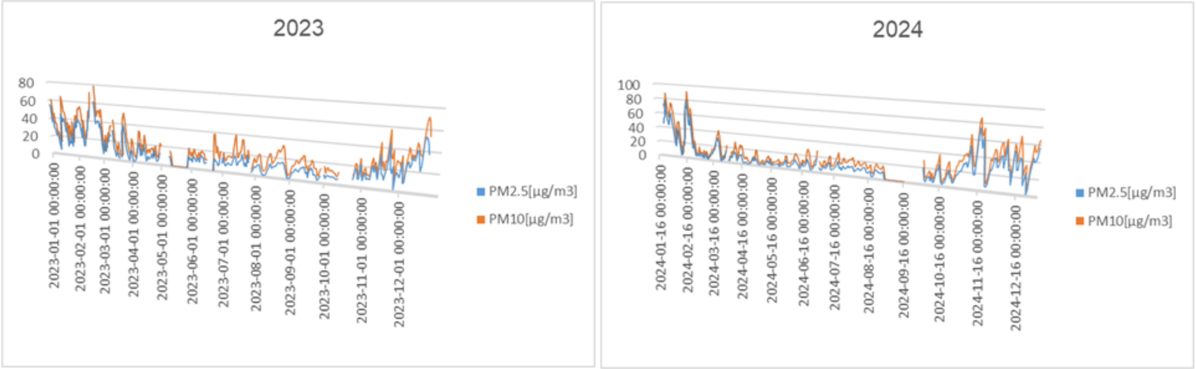


Figure 4-19 Daily average of PM2.5 & PM10 for 2023&2024

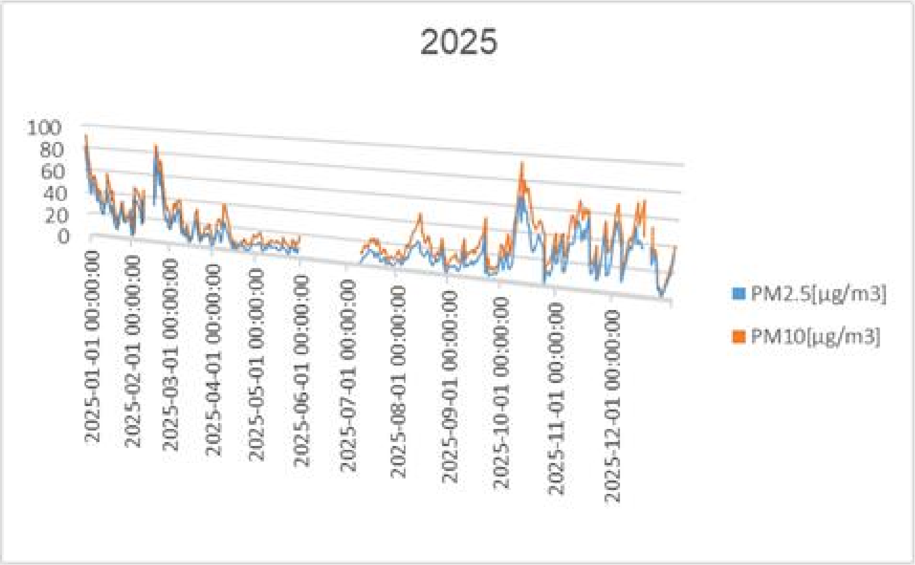


Figure 4-20 Daily average of PM2.5 & PM10 for 2025

The variation in PM2.5 and PM10 concentrations is smooth over time, and their peak patterns are similar to one another. This indicates that these two parameters are primarily associated with particle mass concentration.

The variation in the concentration of CN is larger compared to PM2.5 and PM10. There are certain periods where the concentration of particulate matter increases rapidly while the concentrations of PM2.5 and PM10 change slightly. This indicates that the concentration of particulate matter is associated with a large number of ultrafine particles.

Even though the concentrations of particulate matter and particulate matter number vary similarly over time, the concentration of particulate matter number is more responsive to short-term pollution events. These results suggest that particulate matter number concentration is associated with ultrafine particle number concentration rather than particle mass concentration.

4.2. Comparison between years and seasons

This section will display the results of the annual and seasonal comparisons of the air pollution data from 2019 to 2025. The trends and patterns that can be observed from the data will be presented through graphical analysis. The results will aid in understanding the variations of the particle concentrations over time and under different meteorological conditions.

4.2.1. Correlation between CN and Air Pressure

Figure 4-9 illustrates the correlation between CN and air pressure. The correlation coefficient is -0.0194963, which is almost zero, meaning there is no correlation between these two factors. This implies that air pressure is not a factor that affects CN. Pressure has the least effect on the variation of CN compared to other factors.

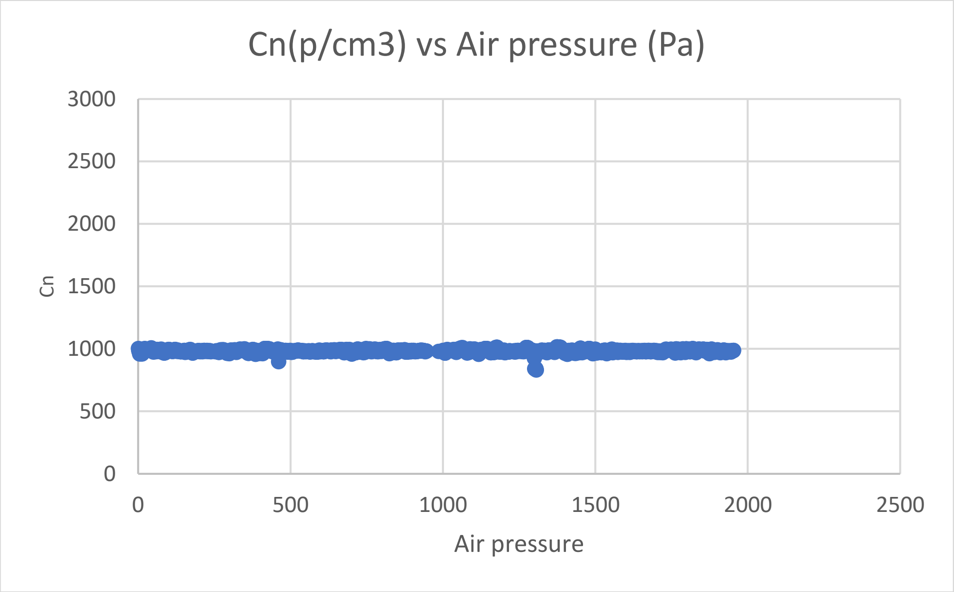


Figure 4-21 Influence of Air pressure (Pa) on Cn(p/cm3)

4.2.2. Correlation between CN and Dew Point / Humidity (Low Humidity Conditions)

Figure 4-10 shows the correlation between CN and dew point (or low humidity). There is a moderate negative correlation, with a correlation coefficient of about -0.40. This indicates that as CN increases, humidity and dew point decrease. Dry air is conducive to the formation of ultrafine particles, possibly because of combustion and a lack of particle growth. Such conditions are common during the colder and drier months of the year.

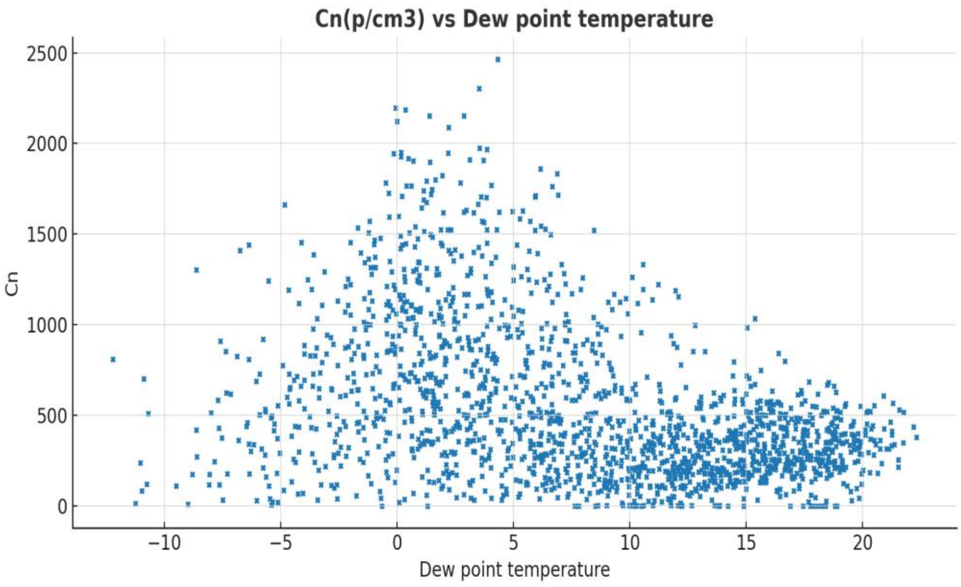


Figure 4-22 Influence of Dew point temperature (c) on Cn(p/cm3)

4.2.3. Correlation between CN and Relative Humidity

The correlation between CN and relative humidity is shown in Figure 4-11. A weak positive correlation is found, with a correlation coefficient of about 0.27. This indicates that a higher relative humidity leads to a slight increase in CN. This might be because higher humidity can facilitate the formation of clusters or growth of particles. However, this effect is not very significant compared to the effect of temperature and seasonality.

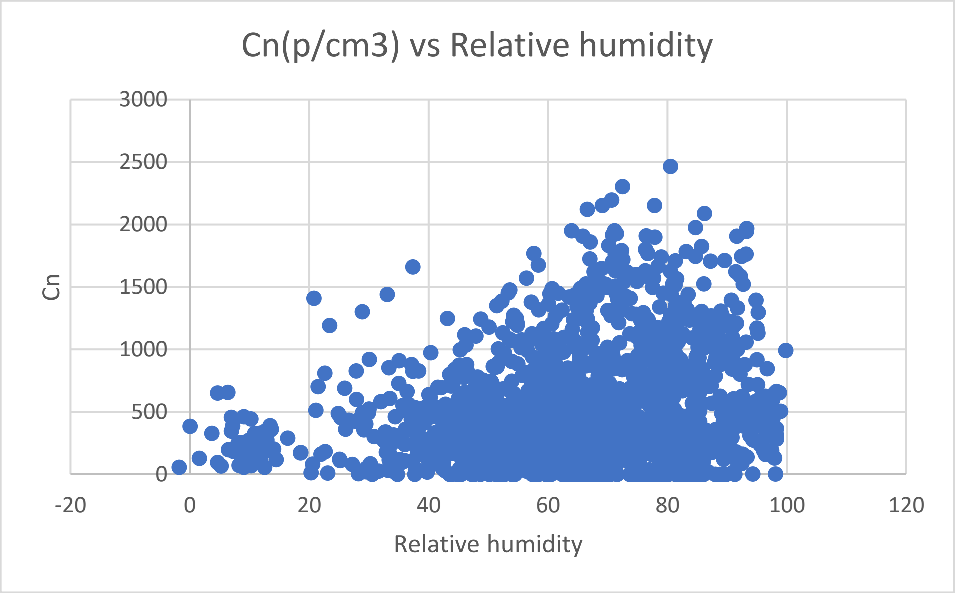


Figure 4-23 Influence of Relative humidity (%) on Cn(p/cm3)

4.2.4. Correlation between CN and Temperature

The relationship between CN and temperature is presented in Figure 4-12. A strong negative correlation is found, with a correlation coefficient of about -0.5274614. This is the strongest meteorological effect among all variables considered. Higher CN concentrations are found at lower temperatures. This is as expected during winter, when temperature inversion, reduced atmospheric mixing, and increased emissions from heating systems result in higher particle concentrations. The strong temperature dependence clearly indicates the importance of meteorological factors in ultrafine particle concentration regulation.

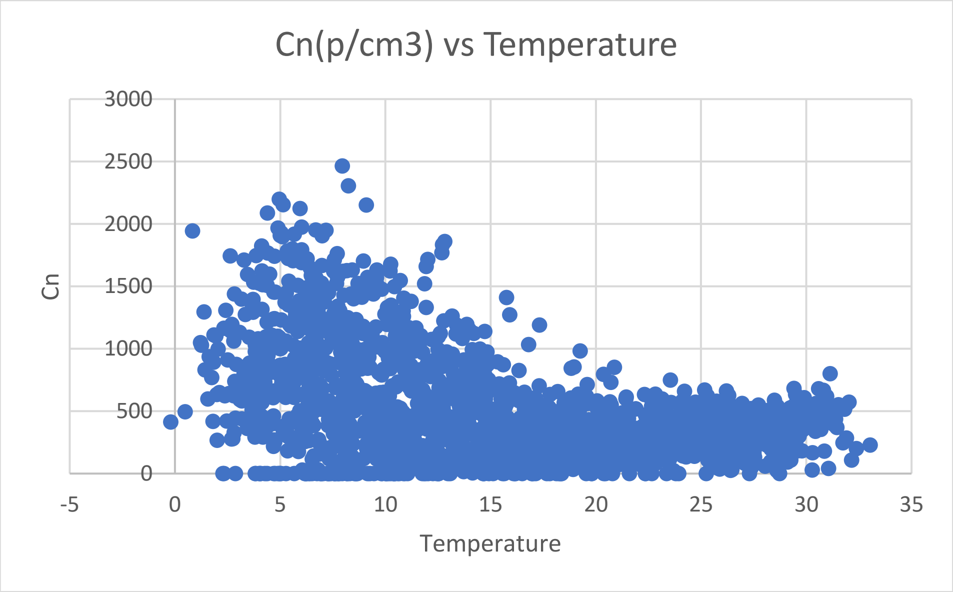


Figure 4-24 Influence of Temperature (c) on Cn(p/cm3)

4.2.5. Hourly PM2.5 & PM10 High-Pollution Week

Figure 4-13 illustrates the hourly changes in PM2.5 and PM10 concentrations during the winter pollution episode from 29 December 2021 to 4 January 2022. Both PM2.5 and PM10 display high concentration levels throughout the chosen week. Large daily variations can be noticed, with periodic peaks at specific times of the day. These peaks are probably linked to increased traffic and heating in residential areas, which are common sources of particulate matter during winter. Moreover, the absence of air mixing due to stable atmospheric conditions and temperature inversion during winter further contributes to the accumulation of pollutants close to the ground.

The continuous high levels of PM concentrations during this period indicate adverse meteorological conditions for the dispersion of pollutants. The high levels of PM2.5 are consistent and display a similar pattern to that of PM10, indicating that fine particulate matter is a major contributor to pollution episodes during winter.

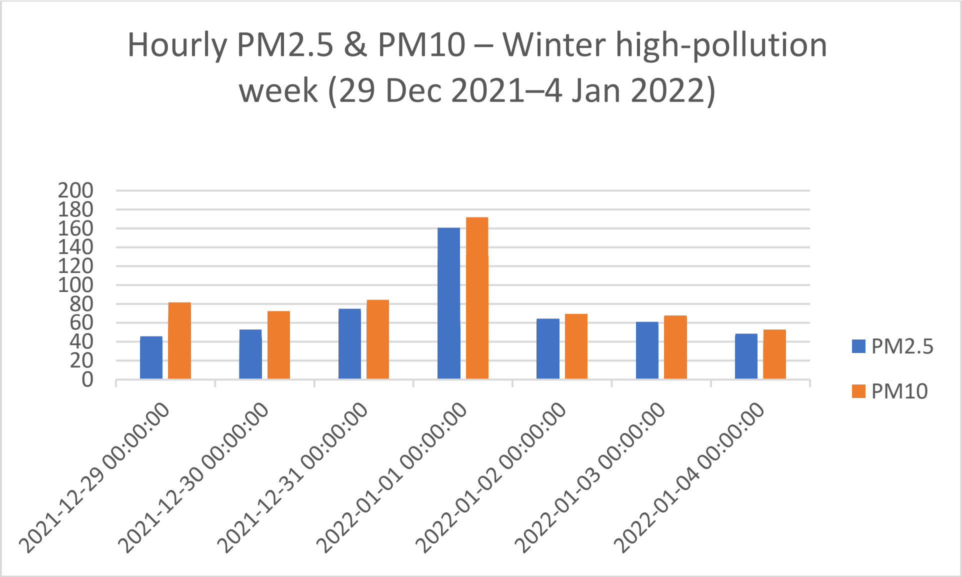


Figure 4-25 Hourly PM2.5 & PM10 – Winter high-pollution week (29 Dec 2021–4 Jan 2022)

4.2.6. Hourly PM2.5 & PM10 Low-Pollution Week

Figure 4-14 presents the hourly changes of PM2.5 and PM10 concentration levels during the summer clean-air period from 21 June 2024 to 27 June 2024.

Compared to the winter periods, the concentration levels of both PM2.5 and PM10 are significantly lower. The curves are smoother with less steep peaks and smaller variations. This is because of the favorable meteorological conditions, which include higher temperatures, increased mixing of the atmosphere, and better dispersion of pollutants. The concentration levels of PM2.5 and PM10 remain relatively constant throughout the day, indicating that the local sources of pollution have less effect during summer.

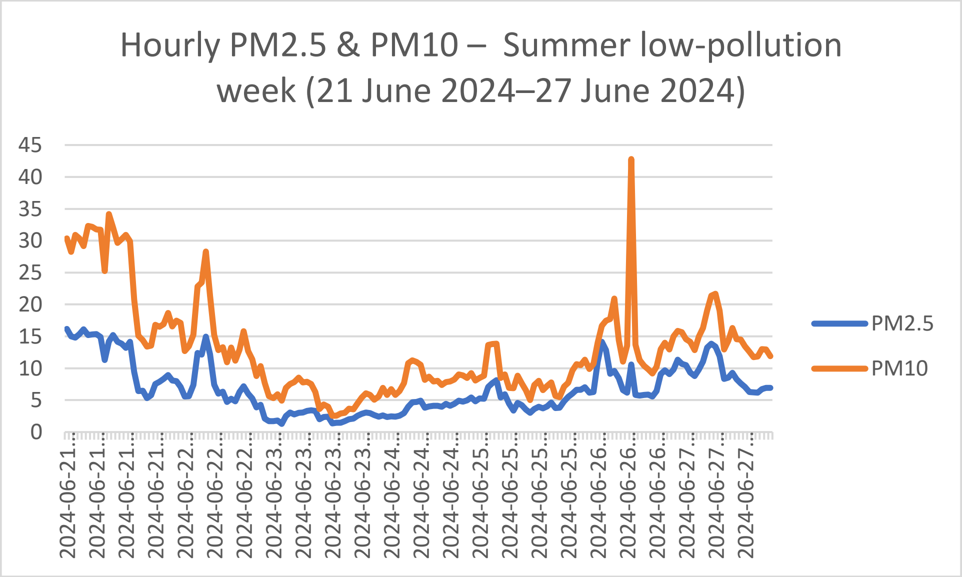


Figure 4-26 Hourly PM2.5 & PM10 – Summer low-pollution week (21 June 2024–27 June 2024)

4.2.7. Hourly PM2.5 & PM10 Highest-Pollution Days

Figure 4-15 introduces the hourly concentrations of PM2.5 and PM10 for the two days with the highest recorded pollution levels (31 December 2021 and 1 January 2022).

Very sharp peaks can be observed for these days, indicating extreme pollution events. The absolute values for PM10 are higher than those of PM2.5, but both variables show similar trends. The highest peaks occur in the evening and nighttime hours, which could be related to enhanced combustion from domestic heating and perhaps firework displays during the New Year celebrations. The large changes over short time periods indicate that pollution events are dynamic processes and that high-resolution temporal analysis is important. These findings show that extreme pollution events are not only associated with high average concentrations but also with large hourly changes.

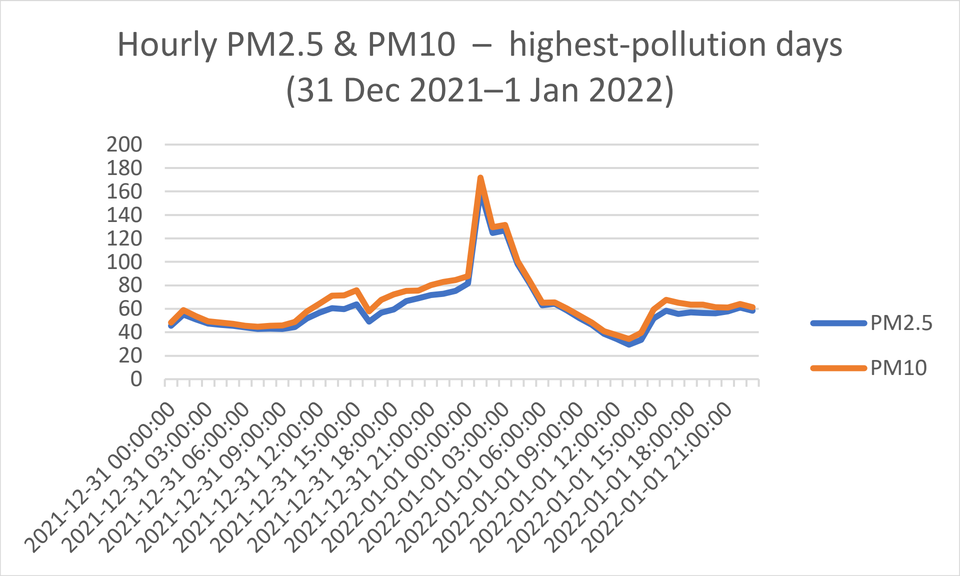


Figure 4-27 Hourly PM2.5 & PM10 – highest-pollution days (31 Dec 2021–1 Jan 2022)

4.2.8. Overall Interpretation of CN Behavior

Taking into consideration the annual, seasonal, and meteorological analyses as a whole, it can be concluded that CN behaves in a manner characteristic of ultrafine particle number concentration. CN is higher during winter and low temperatures, and lower during summer and higher temperatures. The less significant effect of air pressure and the moderately significant effect of humidity confirm this conclusion. The above results indicate that CN is not a measure of particle mass concentration but rather a measure of the number of ultrafine particles in the atmosphere. This is why CN varies more significantly than PM2.5 and PM10.

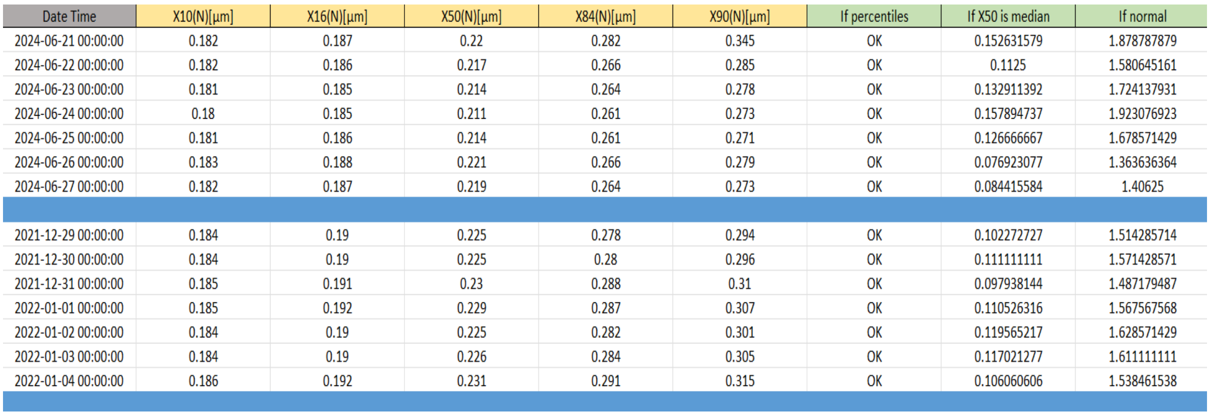
The comparison between winter high-pollution periods and summer low-pollution periods clearly demonstrates the strong seasonal influence on particulate matter concentrations. Winter episodes are characterized by high and highly variable PM2.5 and PM10 levels, while summer periods show lower and more stable concentrations. These results emphasize the role of meteorological conditions and emission patterns in controlling short-term air pollution events.

4.3. The meaning of the X10, X16, X50, X84 and X90 in PDAnalyze

The columns X10, X16, X50, X84, and X90 in the software output are percentiles of the particle size distribution. That is, X10 is the size below which 10% of the particles are smaller, X50 represents the median value at which 50% of the particles are smaller and 50% larger, and X90 is the size below which 90% of the particles are smaller. Examination of the data order showed that in all samples the logical relationship X10 < X16 < X50 < X84 < X90 holds, which is a characteristic behavior of percentiles of a cumulative distribution and confirms that these variables truly represent percentiles of the particle size distribution.

For further validation, the statistical relationship between X16, X50, and X84 were examined. The results showed that X50 was approximately in the center of the interval between X16 and X84, which is consistent with the statistical definition of the median. However, the ratio of the distance (X84 − X50) to (X50 − X16) was obtained to be greater than 1, indicating that the distribution was asymmetric and had a right-skewed distribution because coarse particles create a "tail". This behavior is quite normal in the analysis of particle size distribution and indicates that although the data are not statistically normal, they still reliably represent the true percentiles of the particle size distribution.

And also, the reported X10, X50 and X90 values represent the 10th, 50th and 90th percentiles of the particle size distribution, as defined by ISO 9276-2.



Date Time	X10(N)[µm]	X16(N)[µm]	X50(N)[µm]	X84(N)[µm]	X90(N)[µm]	If percentiles	If X50 is median	If normal
2024-06-21 00:00:00	0.182	0.187	0.22	0.282	0.345	OK	0.152631579	1.878787879
2024-06-22 00:00:00	0.182	0.186	0.217	0.266	0.285	OK	0.1125	1.580645161
2024-06-23 00:00:00	0.181	0.185	0.214	0.264	0.278	OK	0.132911392	1.724137931
2024-06-24 00:00:00	0.18	0.185	0.211	0.261	0.273	OK	0.157894737	1.923076923
2024-06-25 00:00:00	0.181	0.186	0.214	0.261	0.271	OK	0.126666667	1.678571429
2024-06-26 00:00:00	0.183	0.188	0.221	0.266	0.279	OK	0.076923077	1.363636364
2024-06-27 00:00:00	0.182	0.187	0.219	0.264	0.273	OK	0.084415584	1.40625
2021-12-29 00:00:00	0.184	0.19	0.225	0.278	0.294	OK	0.102272727	1.514285714
2021-12-30 00:00:00	0.184	0.19	0.225	0.28	0.296	OK	0.111111111	1.571428571
2021-12-31 00:00:00	0.185	0.191	0.23	0.288	0.31	OK	0.097938144	1.487179487
2022-01-01 00:00:00	0.185	0.192	0.229	0.287	0.307	OK	0.110526316	1.567567568
2022-01-02 00:00:00	0.184	0.19	0.225	0.282	0.301	OK	0.119565217	1.628571429
2022-01-03 00:00:00	0.184	0.19	0.226	0.284	0.305	OK	0.117021277	1.611111111
2022-01-04 00:00:00	0.186	0.192	0.231	0.291	0.315	OK	0.106060606	1.538461538

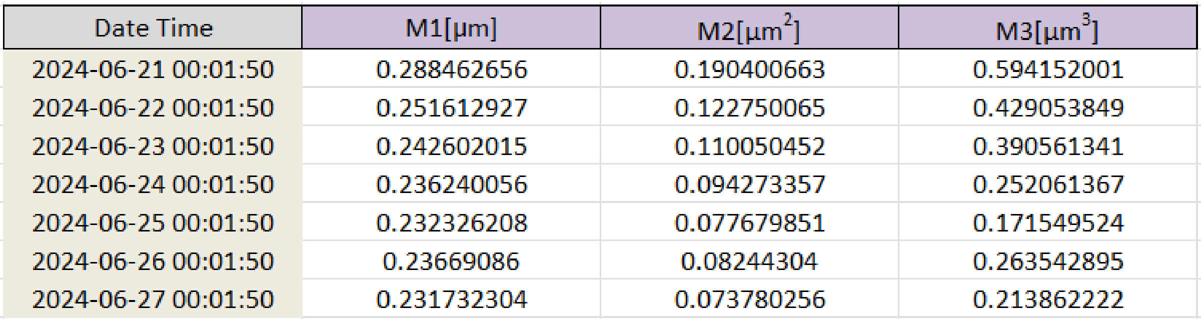
Figure 4-28 Two specific weeks of X-parameters

4.4. The meaning of M1, M2 and M3

M1, M2, and M3 are the first to third order statistical moments of the particle size distribution. They describe the physical properties of aerosols. The first order moment (M1) indicates the mean particle size in micrometers (µm). The calculated values of M1, M2, and M3, as obtained from the Excel database, were compared with values obtained from the PDAnalyze software. From the comparison, it can be observed that the results are very close for all days. For instance, on 21 June 2024, the calculated M1 value was 0.288463 µm, whereas the obtained value from the software was also 0.288463 µm. A similar comparison was observed for M2 and M3. Although some discrepancies are observed in the values, these are due to rounding. Therefore, from this comparison, it can be concluded that the results obtained from the implemented calculations are consistent with the results obtained from the PDAnalyze software.

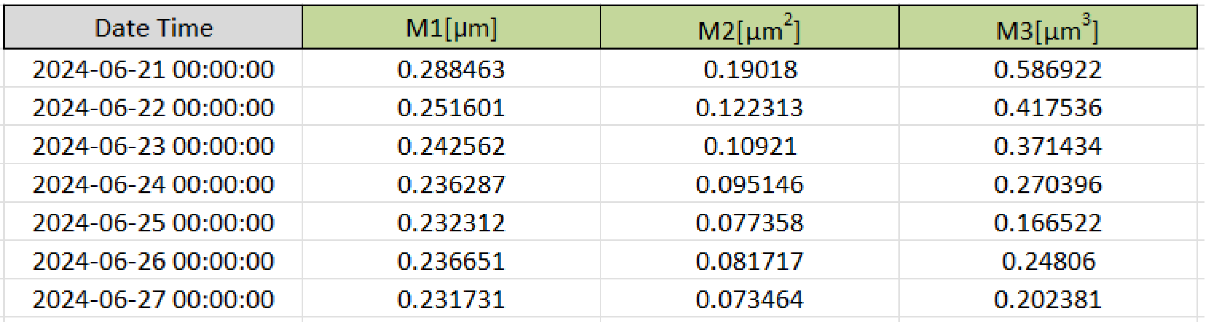
The results also indicate that the parameters are following a certain trend over the period of analysis. For example, between 21st June and 27th June, it is observed that the value of M1 gradually decreases from 0.288 µm to approximately 0.232 µm. A similar pattern can be observed for M2 and M3, where the values are observed to be decreasing over the same period. This is indicative of the fact that the size of the particles is decreasing over the period of analysis. As the values of M2 and M3 are dependent on the square and cube of the diameter of the particles, respectively, it is observed that the values of M2 and M3 are decreasing much faster compared to M1. This is consistent with the mathematical definitions of the moments of the size distribution of particles.

This agreement between the calculated values and the software output verifies that the parameters M1, M2, and M3 are indeed the statistical moments of the particle size distribution. In particular, the first moment, M1, is the average value corresponding to the average diameter of the particles. The second moment, M2, corresponds to the surface area of the particles, whereas the third moment, M3, corresponds to the volume of the particles. The validation of the interpretation of the parameters within the data set is achieved by the successful reproduction of the software results using the same mathematical formulation in Excel.



Date Time	M1[µm]	M2[$µm^2$]	M3[$µm^3$]
2024-06-21 00:01:50	0.288462656	0.190400663	0.594152001
2024-06-22 00:01:50	0.251612927	0.122750065	0.429053849
2024-06-23 00:01:50	0.242602015	0.110050452	0.390561341
2024-06-24 00:01:50	0.236240056	0.094273357	0.252061367
2024-06-25 00:01:50	0.232326208	0.077679851	0.171549524
2024-06-26 00:01:50	0.23669086	0.08244304	0.263542895
2024-06-27 00:01:50	0.231732304	0.073780256	0.213862222

Figure 4-29 The M1, M2, and M3 values calculated in Excel using the particle size distribution data



Date Time	M1[µm]	M2[µm^2]	M3[µm^3]
2024-06-21 00:00:00	0.288463	0.19018	0.586922
2024-06-22 00:00:00	0.251601	0.122313	0.417536
2024-06-23 00:00:00	0.242562	0.10921	0.371434
2024-06-24 00:00:00	0.236287	0.095146	0.270396
2024-06-25 00:00:00	0.232312	0.077358	0.166522
2024-06-26 00:00:00	0.236651	0.081717	0.24806
2024-06-27 00:00:00	0.231731	0.073464	0.202381

Figure 4-30 Output values from the original software

4.5. The fastest way to update database

The developed automated process was able to successfully merge new measurement files into the master database without any human intervention. The script was able to properly identify the datetime and measurement columns from various Excel output formats and categorize the data into either hourly or daily measurements depending on the time interval.

The database was updated by adding only new data points with timestamps beyond the existing maximum value, and any duplicates and invalid data were automatically eliminated. The procedure appended 2,450 new hourly records and 120 daily records to the database while removing duplicated timestamps. The application of this technique has greatly optimized the process of data handling by minimizing human error and processing time. The automated process was more reliable and reproducible compared to the manual process of merging data, as it offered a consistent means of maintaining an updated database structure optimal for subsequent statistical analysis.

The MATLAB code developed in this study successfully automated the process of updating the air pollution database with the newly exported data from the analysis software. The code was able to successfully append new data to the existing data structure, maintaining consistency in variables and data formats. This automated process significantly reduced the time taken for updating the data and also reduced human error possibilities involved in manually entering data. The database was successfully made more efficient for data storage and future use.

In addition, the process was flexible and scalable, as it could be used not only for concentration data (Cn) but also for other variables of measurement generated by the software. Processing time was reduced from several minutes of manual editing to a few seconds using the automated workflow.

Chapter 5. Conclusions

This thesis focuses on managing data and improving it in the air domain. It transforms raw environmental and meteorological measurements into a structured, reliable, and reusable database. As scientific decisions increasingly rely on environmental data, organizations need to collect and store this data to support research, policy development, and air quality monitoring. The main goal of this work was not just to collect and store data; it aimed to improve its interpretability, accessibility, and usability for future users.

The first stage of the research involved understanding the software created by the company PDAanalyze. There were clear issues with the documentation that affected the scientific readability of the data set. By interpreting and evaluating the principles of particle measurement, we identified the X-ranges (X10, X16, X50, X84, X90 [µm]) as particle size percentiles. These percentiles indicate the diameter below which a certain percentage of particles are found, thus displaying the particle size. Correctly interpreting these indices was vital since particle size significantly affects atmospheric behavior, transport dynamics, and potential health impacts.

The M parameters (M1[µm], M2[μm^2], M3 [μm^3]) were also defined as the statistical mean particle diameters from number distribution, unlike mass-derived metrics, and as such provide a means to better understand particle size from a numerical standpoint. Therefore, the clarification of these variables was critical in improving the usability of the database for scientific purposes and decreasing the potential for misinterpretation of future analyses. This contributes to the overall integrity of environmental datasets which are based on properly defined variables. The thesis also worked towards the goal of streamlining the database updating process. Data from environmental monitoring systems continues to be generated in large amounts and manually integrating this data into an Excel format is very time-consuming and susceptible to error. To resolve this problem, a MATLAB script was developed which would automatically append newly exported data from PDAnalyze to an existing Excel database. The automation of this process will significantly decrease the amount of time needed for database updates while providing for better accuracy and consistency among records. Therefore, the warehousing of new records will increase the operational efficiency of the database and promote its long-term sustainability. This method of automating database updates is essential for environmental databases that are anticipated to grow continuously into the future.

The behavior of CN is consistent with that of ultrafine particle number concentration. CN increases during winter and under low-temperature conditions, while it decreases during summer and at higher temperatures. The weak influence of air pressure and the moderate effect of humidity further confirm that CN mainly represents the number of small particles rather than their mass. These findings indicate that CN is more sensitive to meteorological conditions and emission sources compared to PM2.5 and PM10, making it a useful indicator for studying short-term and seasonal variations in air pollution.

Comparison between various years and seasons reveals a definite decreasing trend in particle concentration values from 2019 to 2025, with higher values in winter and lower values in summer. The trends are consistent for all years, thereby establishing the significant impact of meteorological factors on air pollution levels. The above findings establish the significance of temporal analysis in comprehending long-term and seasonal variations in air quality.

Using the database, we were able to conduct a temporal analysis of polluted and particle characteristics for different weather conditions, allowing us to compare differences in emissions from 2019 to 2025 on a season-by-season and year-by-year basis. Seasonal comparisons are relevant to air quality studies due to how the elements of atmospheric stability, temperature variations, and recurring emission patterns contribute to fluctuations in air quality. Long-term data sets also allow for more accurate identification of long-term trends, which are critical to understanding how successful environmental policies and mitigation efforts are working. Although this project did not employ a great deal of advanced statistical modelling, the database is available for future studies where that type of analysis may be applicable.

It's critical to improve the way we structure and maintain environmental data so that we can have a better understanding of pollution and protect public health. Even with the positive accomplishments so far, there are still some limitations that need to be addressed. For now, the database largely consists of data that were extracted from one software program. This may create interoperability issues when trying to share that data with other monitoring systems. The database is designed primarily to store data offline; therefore, it is necessary to export data from the monitoring system into the database for updating purposes, rather than having the ability to access it on an as-needed basis immediately once the automated sensors provide that information in real-time. Using automated sensors and cloud-based storage systems would allow for greater scalability and access to the database. Addressing these limitations can help foster the use of the database as an effective tool for long-term environmental monitoring.

The future of this work will be able to develop in a number of ways. For example, real-time data ingestion would facilitate continuous updates without going through an intermediate export step. Also, by moving to the cloud, the database can improve its data-sharing capability while providing secure storage and version control. Also, due to the dataset's structure, there are many different ways that it could benefit from advanced methods of analysis, such as using machine-learning techniques for predicting pollution, detecting anomalies, and creating scenario analyses. All of these capabilities would greatly enhance the practical use of this database for both researchers and decision-makers.

In conclusion, this thesis presents a framework for utilizing structured data management to create useful information from environmental measurements into usable information. By achieving clarity regarding both the major variables and the development of an automated mechanism to regularly update databases (and thus have temporal comparisons available), a user-friendly data structure has been established. These factors create a solid foundation from which future air quality projects can develop. With growing complexity associated with environmental problems, the importance of reliable and well managed data is greater now than ever before. Thus, the methods developed in this project provide a mechanism to support improved decision making relating to air quality management and highlight the importance of optimizing data in advancing knowledge within environmental science.

References

1. Smith, J., & Kumar, R. (2022). *Advances in data management systems*. Elsevier.
2. Miller, A., & Rossi, L. (2023). *Data methods and management innovations*. ResearchGate.
3. Chen, Y., & Dubois, P. (2021). *Environmental big data and optimization*. Springer.
4. Johnson, T., & Wang, L. (2020). *Optimization techniques in environmental data analysis*. Elsevier.
5. Brown, H., & Patel, S. (2019). *Data governance and environmental monitoring*. Springer.
6. Green, P., & Allen, R. (2018). *Machine learning applications in air pollution studies*. Elsevier.
7. White, M., & Becker, J. (2017). *Ontology-driven approaches for environmental data management*. Springer.
8. Thompson, D., & Rivera, F. (2016). *Provenance and traceability in environmental data systems*. Elsevier.
9. Garcia, E., & Chen, H. (2015). *Database technologies for large-scale environmental monitoring*. ScienceDirect.
10. Kumar, V., & Singh, A. (2014). *Pipelines for environmental data processing*. Elsevier.
11. Li, Z., & Zhao, Q. (2013). *Integration of heterogeneous environmental datasets*. Springer.
12. O'Connor, J., & Smith, L. (2012). *Data governance in large-scale projects*. Elsevier.
13. Fernandez, R., & Kim, S. (2011). *Improving data quality in environmental research*. Springer.
14. Roberts, K., & Evans, P. (2010). *Optimization in environmental policy modeling*. Elsevier.
15. Chang, H., & Lee, J. (2009). *Predictive models using environmental big data*. Springer.
16. Silva, M., & Costa, N. (2008). *Information systems for environmental monitoring*. ScienceDirect.
17. Huang, Y., & Lin, C. (2007). *Data-driven approaches to pollution management*. Elsevier.
18. Fischer, T., & Schmidt, G. (2006). *Managing uncertainty in environmental databases*. Springer.
19. Allen, B., & Cooper, D. (2005). *Workflow optimization in scientific data management*. Elsevier.
20. Ahmed, S., & Rashid, M. (2004). *Scalable systems for air quality data*. Springer.
21. Carter, J., & Holmes, R. (2003). *Policy-oriented data management frameworks*. Elsevier.
22. Walker, P., & Nelson, G. (2002). *Challenges in environmental data access*. Springer.
23. Jackson, M., & Li, F. (2001). *Data integration for policy support*. Elsevier.
24. Edwards, K., & Baker, J. (2000). *Historical perspectives on environmental data management*. Springer.
25. Clarke, S., & Howard, E. (1999). *Optimizing environmental databases*. Elsevier.

26. Roberts, D., & Taylor, M. (1998). *Governance models for scientific data*. Springer.
27. ITRC Environmental Data Management Best Practices Team. (2022). *Environmental data management systems*. ITRC.
28. Rosini, I., Rahmatunnisa, M., Sunardi, S., & Priyanto, I. F. (2025). *Research data management in environmental studies: Scoping review and bibliometrics analysis*. Data Science Journal.
29. U.S. Environmental Protection Agency (EPA). (2025). *Guide to managing air quality data*. EPA.
30. Temkov, S. (2025). *Air pollution data: A dataset gathered through a crowd sensing IoT platform*. Elsevier.
31. Zhang, W., Lin, Y., & Wang, P. (2024). *Analytics in data management methods*. MDPI.
32. Zhang, W., et al. (2024). *Methods and applications of data management and analytics*. MDPI.
33. Singh, R. K., & Pandey, A. (2023). *DMPFrame: A conceptual metadata framework for data management plans*. Taylor & Francis.
34. Mositsa, R. J. (2023). *Towards a conceptual framework for data management in business intelligence*. MDPI.
35. Marcucci, S., Gonzalez Alarcon, N., Verhulst, S. G., & Wullhorst, E. (2023). *Mapping and comparing data governance frameworks: A benchmarking exercise*. arXiv.
36. Rosales, C. M., et al. (2025). *Open air quality data platforms for environmental research*. Springer.
37. Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2016). *Towards air quality estimation using collected multimodal environmental data*. arXiv.
38. Berrisford, J., & Menezes, R. (2024). *Environmental insights: Democratizing access to ambient air pollution data and predictive analytics with an open-source Python package*. arXiv.
39. Za'al Alma'aitah, W., Al-Sharaeh, S., Al-Sharaeh, F., & Almomani, O. (2024). *Integration approaches for heterogeneous big data: A survey*. Sciendo.
40. Ranatunga, S., Ødegård, R. S., Jetlund, K., & Onstein, E. (2025). *Use of semantic web technologies to enhance the integration and interoperability of environmental geospatial data*. MDPI.

41. Geng, G., Xiao, Q., Liu, S., & Zhang, Q. (2021). *Tracking air pollution in China: Near real-time PM2.5 retrievals from multiple data sources*. arXiv preprint arXiv:2103.06520.

42. Gianquintieri, A., Ricciardelli, I., & Ferrero, L. (2025). *State-of-art in modelling particulate matter concentration: A scoping review*. Environmental Science and Pollution Research, 32(1), 1–20.

43. Katzenstein, A., & Etcheverry, S. (2025). *FAIR and quality-aware air quality data management*. CEUR Workshop Proceedings, 4002, 1–8.

44. Rosales, R., Martínez, A., & López, J. (2025). *Open air quality data platforms for environmental health*. Environmental Health Perspectives, 133(2), 1–12.

45. Wang, Y., Zhang, H., Liu, X., & Chen, J. (2024). *Seasonal particle size distribution and its influencing factors in a typical polluted city in North China*. Aerosol and Air Quality Research, 24(3), 1–15. https://doi.org/10.4209/aaqr.230127

46. Wójcik-Gront, E., & Gozdowski, D. (2025). *Air pollution monitoring and modeling: A comparative study*. Atmosphere, 16(10), 1199. https://doi.org/10.3390/atmos16101199

47. International Organization for Standardization. (2015). *ISO 9276-2:2015 Statistical methods — Principles for the presentation of data — Part 2: Graphical representation (Principles)*. ISO.

Appendix

```
%% append_cn_to_db_FIXED.m
% Robust append NEW Cn Excel to main DB (Daily/Hourly auto-detect)
% Works on older MATLAB versions (uses xlsfinfo instead of sheetnames)
% Auto-detects datetime & Cn columns safely and converts types
clear; clc;
% --------- USER SETTINGS ----------
dbFile  = "D:\DB_cleaned(test).xlsx";   % main database path
newFile = "D:\CN(7-9-2025).xlsx";        % new file path
% ----------------------------------
%% 1) Read DB sheets
Tdaily  = readtable(dbFile, "Sheet","Daily",  "VariableNamingRule","preserve");
Thourly = readtable(dbFile, "Sheet","Hourly", "VariableNamingRule","preserve");
%% 2) Read NEW file (first sheet) - compatible way
[~, sheetList] = xlsfinfo(newFile);
if isempty(sheetList)
    error("No sheets found in new file: %s", newFile);
end
Tnew = readtable(newFile, "Sheet", sheetList{1}, "VariableNamingRule","preserve");
newVars = string(Tnew.Properties.VariableNames);
lvNew   = lower(newVars);
%% 3) Robustly detect NEW datetime column by trying conversions
bestDtIdx = [];
bestValid = -1;
bestDt = [];
for i = 1:width(Tnew)
    col = Tnew{:,i};
    dt = [];
    try
        if isdatetime(col)
            dt = col;
        elseif isnumeric(col)
            % could be Excel serial datetime
            dt = datetime(col, "ConvertFrom","excel");
        else
            % try parse text/cell/string
            if iscell(col), col = string(col); end
            dt = datetime(string(col), "InputFormat",""); %#ok<DTCH>
            % if it fails, catch will handle
        end
        % count valid
        v = sum(~isnat(dt));
        if v > bestValid
            bestValid = v;
            bestDtIdx = i;
            bestDt = dt;
        end
    catch
        % ignore this column
    end
end

if isempty(bestDtIdx) || bestValid < max(5, round(height(Tnew)*0.3))
    % fallback by header keywords if conversion approach failed
    dtIdx = find(contains(lvNew,"time") | contains(lvNew,"date"), 1, "first");
    if isempty(dtIdx)
        error("Datetime column not found in new file.");
    end
    bestDtIdx = dtIdx;
    col = Tnew{:,bestDtIdx};
    if isdatetime(col)
        bestDt = col;
    elseif isnumeric(col)
        bestDt = datetime(col,"ConvertFrom","excel");
    else
        if iscell(col), col = string(col); end
        bestDt = datetime(string(col));
    end
end
dtNewAll = bestDt;
dtNewAll = dtNewAll(:);
% remove NaT rows
validNew = ~isnat(dtNewAll);
dtNew = dtNewAll(validNew);

%% 4) Detect NEW Cn column robustly
% Strategy:
%  - Prefer columns whose name contains 'cn' but not 'time/date'
%  - Then test convertibility to numeric and pick the one with most finite values
cnCandidates = find(contains(lvNew,"cn") & ~contains(lvNew,"time") &
~contains(lvNew,"date"));
if isempty(cnCandidates)
    cnCandidates = find(contains(lvNew,"cn")); % fallback
end
if isempty(cnCandidates)
    error("No column containing 'Cn' found in new file.");
end
bestCnIdx = [];
bestCnScore = -1;
for j = cnCandidates(:)'
    raw = Tnew{:,j};
    raw = raw(validNew,:);
    % reject if datetime
    if isdatetime(raw)
        continue;
    end
    % convert to numeric safely
    try
        x = raw;
        if iscell(x), x = string(x); end
        if isstring(x) || ischar(x)
            x = str2double(string(x));
        end
        x = double(x);
        score = sum(isfinite(x));  % count numeric usable rows
        if score > bestCnScore
            bestCnScore = score;
            bestCnIdx = j;
        end
    catch
        % ignore
    end
end
if isempty(bestCnIdx) || bestCnScore < max(5, round(numel(dtNew)*0.3))
    error("Cn column detection failed (could not find a numeric-like Cn column).");
end
cnRaw = Tnew{:, bestCnIdx};
cnRaw = cnRaw(validNew,:);

% convert to numeric
if iscell(cnRaw), cnRaw = string(cnRaw); end
if isstring(cnRaw) || ischar(cnRaw)
    cnNew = str2double(string(cnRaw));
else
    cnNew = double(cnRaw);
end
cnNew = cnNew(:);
%% 5) Detect Daily vs Hourly (median time step)
dtSorted = sort(dtNew);
if numel(dtSorted) < 2
    error("Not enough datetime rows in new file.");
end
stepMed = median(diff(dtSorted));
isDaily = stepMed >= hours(20);
%% 6) Pick target DB table/sheet
if isDaily
    Tdb = Tdaily;
    sheetTarget = "Daily";
else
    Tdb = Thourly;
    sheetTarget = "Hourly";
end

dbVars = string(Tdb.Properties.VariableNames);
lvDb   = lower(dbVars);
%% 7) Find DB datetime column (prefer 'Date Time' exact-ish, else best datetime-like)
dbDtIdx = find(lvDb=="date time" | lvDb=="date_time", 1, "first");
if isempty(dbDtIdx)
    % keyword fallback
    dbDtIdx = find(contains(lvDb,"date") | contains(lvDb,"time"), 1, "first");
end
if isempty(dbDtIdx)
    % conversion fallback: choose column that converts to datetime with most valid rows
    best = [];
    bestV = -1;
    for i = 1:width(Tdb)
        col = Tdb{:,i};
        try
            if isdatetime(col)
                dt = col;
            elseif isnumeric(col)
                dt = datetime(col, "ConvertFrom","excel");
            else
                if iscell(col), col = string(col); end
                dt = datetime(string(col));
            end
            v = sum(~isnat(dt));
            if v > bestV
                bestV = v;
                best = i;
            end
        catch
        end
    end
    dbDtIdx = best;
end
if isempty(dbDtIdx)
    error("Datetime column not found in DB sheet %s.", sheetTarget);
end
dtDb = Tdb{:, dbDtIdx};
if ~isdatetime(dtDb)
    try
        if isnumeric(dtDb)
            dtDb = datetime(dtDb, "ConvertFrom","excel");
        else
            if iscell(dtDb), dtDb = string(dtDb); end
            dtDb = datetime(string(dtDb));
        end
    catch
        error("Could not convert DB datetime column to datetime.");
    end
end
dtDb = dtDb(:);
%% 8) Find DB Cn column (avoid date/time)
dbCnIdx = find(contains(lvDb,"cn") & ~contains(lvDb,"time") & ~contains(lvDb,"date"),
1, "first");
if isempty(dbCnIdx)
    % fallback: any cn
    dbCnIdx = find(contains(lvDb,"cn"), 1, "first");
```

```
end
if isempty(dbCnIdx)
    error("Cn column not found in DB sheet %s.", sheetTarget);
end
%% 9) Keep only NEW rows: newer than max DB datetime and not duplicate datetimes
dtDbValid = dtDb(~isnat(dtDb));
if isempty(dtDbValid)
    maxDb = datetime(0,0,0); % very old origin
else
    maxDb = max(dtDbValid);
end
maskKeep = (dtNew > maxDb) & ~ismember(dtNew, dtDbValid);

dtAdd = dtNew(maskKeep);
cnAdd = cnNew(maskKeep);
if isempty(dtAdd)
    fprintf("No new rows to append to %s.\n", sheetTarget);
    return;
end
%% 10) Build Tadd safely with correct height and types
n = numel(dtAdd);
% Create table with same vars and n rows
Tadd = Tdb(1:n,:);
% Fill all columns with missing values, preserving type
for k = 1:width(Tadd)
    x = Tadd{:,k};
    if isdatetime(x)
        Tadd{:,k} = NaT(n,1);
    elseif isnumeric(x)
        Tadd{:,k} = NaN(n,1);
    elseif isstring(x)
        Tadd{:,k} = strings(n,1);
    elseif iscell(x)
        Tadd{:,k} = cell(n,1);
    elseif islogical(x)
        Tadd{:,k} = false(n,1);
    else
        % for categorical or other types
        try
            Tadd{:,k} = repmat(missing, n, 1);
        catch
        end
    end
end
% Assign datetime
Tadd{:, dbDtIdx} = dtAdd(:);
% Assign Cn respecting DB column type
targetCol = Tdb{:, dbCnIdx};
if isdatetime(targetCol)
    error("DB Cn column seems datetime (dbCnIdx wrong). Please check DB headers.");
```

```
end
if isnumeric(targetCol)
    Tadd{:, dbCnIdx} = double(cnAdd(:));
elseif iscell(targetCol)
    Tadd{:, dbCnIdx} = cellstr(string(cnAdd(:)));
elseif isstring(targetCol)
    Tadd{:, dbCnIdx} = string(cnAdd(:));
else
    % fallback
    Tadd{:, dbCnIdx} = double(cnAdd(:));
end
%% 11) Append + sort
Tout = [Tdb; Tadd];
dtAll = Tout{:, dbDtIdx};
if ~isdatetime(dtAll)
    dtAll = datetime(dtAll);
end
[~,ord] = sort(dtAll);
Tout = Tout(ord,:);
%% 12) Write back only target sheet
writetable(Tout, dbFile, "Sheet", sheetTarget, "WriteMode","overwritesheet");
fprintf("✅ Appended %d rows to %s sheet.\n", n, sheetTarget);
fprintf("   NEW file datetime col: %s\n", Tnew.Properties.VariableNames{bestDtIdx});
fprintf("   NEW file Cn col:       %s\n", Tnew.Properties.VariableNames{bestCnIdx});
fprintf("   DB sheet datetime col: %s\n", Tdb.Properties.VariableNames{dbDtIdx});
fprintf("   DB sheet Cn col:       %s\n", Tdb.Properties.VariableNames{dbCnIdx});
fprintf("   Last appended datetime: %s\n", string(max(dtAdd)));
```