# Data enrichment toolchain: A use-case for correlation analysis of air quality, traffic, and meteorological metrics in Madrid's smart city

Amir Reza Jafari, Víctor González, Laura Martín, Luis Sánchez, Jorge Lanza, Syed Mohsan Raza, Maira Alvi, Kanawut Kaewnoparat, Roberto Minerva, Noel Crespi

# Data Enrichment Toolchain: A Use-Case for Correlation Analysis of Air Quality, Traffic, and Meteorological Metrics in Madrid's Smart City

Amir Reza Jafari [a], Víctor González [b], Laura Martín [b], Luis Sánchez [b], Jorge Lanza [b], Syed Mohsan Raza [a], Maira Alvi,[a], Kanawut Kaewnoparat[a], Roberto Minerva [a], Noel Crespi,[a]

[a] *Samovar, Telecom SudParis, Institut Polytechnique de Paris, Palaiseau, 91120, France*

[b] *Network Planning and Mobile Communications Lab., Universidad de Cantabria, Santander, 39005, Spain*

**Abstract**

In the era of burgeoning data diversity in heterogeneous sources, unlocking valuable insights becomes pivotal. Raw data often lack context and meaning, necessitating the deployment of services that link and enhance data, thereby extracting meaningful patterns and information. For example, exploring the significance of IoT sensors in measuring air quality across cities emphasizes the potential to establish connections between air quality and associated metrics like traffic intensity and meteorological conditions.

Introducing the Data Enrichment Toolchain (DET), this study underscores its role in harmonizing and curating diverse datasets. DET operates on linked-data principles and adheres to the NGSI-LD standard, enabling seamless integration and correlation analysis across disparate data domains. The research delves into the intricate relationship between traffic patterns and prevalent air pollutants, utilizing enriched datasets from European cities focusing on the smart city of Madrid as a use-case.

Considering the COVID-19 pandemic's impact on traffic flow and meteorological influences on air quality, the study examines pre-pandemic, pandemic, and post-pandemic traffic scenarios in Madrid. By leveraging DET-enhanced datasets, the investigation aims to unravel nuanced insights into the interplay between traffic, meteorological factors, and air quality, offering valuable

implications for urban planning and pollution mitigation strategies.

## 1. Introduction

Many major cities around the world have been facing an increased concentration of air pollution as a result of daily human activities. As it has been stated in recent World Health Organization (WHO) reports, air pollution is responsible of an estimated seven million deaths globally in 2016 and 91% of the global population are living in areas below the WHO air quality criteria [1, 2]. Additionally, in pure economic terms, air pollution has an associated cost, globally, that exceeds 3.7 billion euros per year [3]. For example, in Spain it represents between 1.7% and 4.7% of its Gross Domestic Product (GDP) [4]. This has led to Administrations establishing policies and legally enforcing the reduction of air pollutants, mainly at urban areas. For example, in July 2019 the European Commission announced its decision to refer Spain to the European Court of Justice (CJEU) for exceeding legal air pollutant emission limits – NOx in particular, in the urban areas of Madrid and Barcelona[1].

In this regard, several studies have described how urban areas have higher concentrations of air pollutants than any other ecosystems. These studies have concluded that the main reason for this is the large number of vehicles and reduced road capacity [5, 6]. In European cities, about 70% of environmental pollution is caused by motorized transport [7]. In order to address this issue, some large cities have introduced policies to reduce traffic-related pollution by encouraging the use of public transport.

This paper aims to assess what are the relationships between traffic and some of the most relevant air pollutants by analyzing fine-grained datasets coming from Internet of Things (IoT) deployment at the city of Madrid (Spain). Concretely, we are focusing on the information on specific time periods before, during and after the COVID-19 pandemic.

The rationale for selecting this time periods is that during the pandemic in several countries, traffic has been severely reduced to limit the circulation of people and the spread of the virus. This has created a unique situation

[1] http://europa.eu/rapid/press-release_MEMO-18-4486_en.html

to study the effect of reduced traffic on pollution. The traffic and pollution data in March, April and May 2019 are very different from traffic data in the same months of 2020. As an effect of some relaxation of lockdown, data in 2021 in the same period are also different. This gives us the opportunity to understand the effect of reduced traffic in 2020 and to compare data from 2019 and 2021 in order to understand how the traffic was different. At the same time, pollution should have changed in the same period. So a comparison can bring in additional elements. However, the pandemic forced people at home so it is likely that heating could have had a different impact on the pollution. We also consider the data in August 2020 in order to compare the level of (reduced) traffic, and the pollution (without heating).

Mostly we consider the traffic-related emission as a major cause of air pollution such as Nitrogen dioxide ($NO_2$), which is the reason of asthma in four million humans annually, Carbon monoxide (CO), Particulate matter 2.5 ($PM_{2.5}$), Particulate matter 10 ($PM_{10}$), and Sulphur dioxide ($SO_2$).

As it has already been said, for the study that is presented in this paper, we are leveraging large datasets that have been obtained from existing IoT deployments in the city of Madrid. In this sense, the proliferation of data sources associated with IoT deployment is creating an abundance of information that is called to bring benefits for both the private and public sectors. However, data itself is worthless (as gas would be if the combustion engine did not exist), data is only valuable as it can be swiftly consumed so that actual knowledge can be extracted from it. Thus, before actually analyzing the available datasets, it is necessary to consider a number of aspects that must be taken into account to guarantee swiftness in the consumption of IoT data. In this work, we have leveraged a so-called Data Enrichment Toolchain (DET) developed within the EU-funded research project SALTED[2] (Situation-Aware Linked heTerogeneous Enriched Data) to deal with two of them, namely data harmonization and data curation.

This paper endeavours to address the following pivotal research inquiries:
**1)** Can general-purpose data collection platforms adeptly accommodate specialized services?
**2)** How can collected data be harmonised and curated in order to build a data injection chain?

[2]https://salted-project.eu/

**3)** In what manner can harmonised and curated data be correlated and interlinked through the implementation of a DET?

Paper key contributions are as follows:
**1)** In this study, we have employed DET to effectively model and unify datasets sourced from diverse providers, spanning heterogeneous domains and formats. Through the application of linked-data principles and adherence to the NGSI-LD standard [3], DET facilitated the seamless harmonization of these datasets. This critical harmonization, achieved via pre-processing, established a standardized format essential for enabling correlation analysis. Our work successfully unified datasets not only across distinct domains, such as traffic and air quality data, allowing for the computation of the mutual impact between road traffic and air pollution but also various cities, allowing for extending the study carried out in this paper to other cities where similar data is available.

**2)** Furthermore, DET played a pivotal role in curating the dataset by systematically removing faulty entries and rectifying duplicates or missing data items. This rigorous pre-processing step was imperative in ensuring the integrity of our analyses, safeguarding against the potential bias introduced by low-quality data in our correlation analyses, and it was possible to address it in a systematic and homogeneous manner only after the data modelling and harmonization enabled by the DET.

**3)** Finally, utilizing the harmonized and curated datasets integrated into our system's data broker, we conducted an extensive correlation analysis. Our investigation focused on unraveling discernible patterns associated with air quality issues by correlating air quality metrics with various relevant parameters across diverse locations and temporal spans. This comprehensive analysis aimed to illuminate intricate relationships and uncover meaningful insights into the dynamics of air quality across different contexts.

The remaining of the paper is structured as follows. Section 2 provides an overview of background knowledge and related works. Section 3 details the architecture of the DET. In Section 4, we introduce the various data sources collected and integrated into our system while Section 5 outlines the method used for data curation and treatment. Moving forward, Section 6 focuses

[3]https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

on a practical use case illustrating correlation analysis. Finally, Section 7 concludes the paper by summarizing key findings and presenting avenues for future research directions.

## 2. Related work

In this section, we aim to present an overview encompassing prior research and ongoing initiatives that explore the multifaceted realms of DET functionalities. Additionally, we offer an in-depth analysis of previous works concerning data correlation analysis and related use cases, providing valuable insights into the methodologies, advancements, and applications within this critical domain of study.

### *2.1. General purpose Data Enrichment Toolchain*

The concept of DET embodies a fusion of diverse microservices, synergistically enhancing the quality and intrinsic value of initial information extracted from data. Conceptually, envision a DET as a dynamic pipeline, comprising distinct sets of components, each meticulously addressing a pivotal stage within the data source enhancement cycle. This iterative cycle iterates systematically, affording adaptability by enabling certain components to be dynamically parameterized, thus allowing for a flexible and responsive operational framework.

The realm of DET for harmonizing diverse datasets and an experimental evaluation of DET's implementation is described in detail in [8] to show the potential of enriching data into a semantic knowledge graph, creating new data via linking, aggregation, reasoning and leveraging linked-data modelling and semantics extends metadata for enhanced utility. However, the interoperability of different components is an important aspect of this toolchain. Sharing data within each system necessitates a complex implementation process and robust standardization initiatives [9]. In IoT ecosystems, platforms' interoperability is an important problem; for example, [10] utilized five interoperability patterns crucial for cross-platform interoperability, aiding in the establishment of successful IoT ecosystems. Moreover, semantic interoperability is another important aspect of this subject, especially with the help of machine learning; this can be applicable across standardization domains [11], [12].

In our work, the DET solution facilitates interoperability and incorporates processing steps for enriching datasets sourced from heterogeneous data sources, extending beyond IoT platforms.

### *2.2. Data correlation analysis*

Many prior studies have consistently highlighted the connection between traffic patterns and air pollution. It has been well-documented that pollution stemming from vehicular traffic exerts deleterious impacts, not only on the natural environment but also on the health of individuals residing in areas influenced by high pollution levels, such as cardiovascular disease, leading to premature death. Considering these consequences, governments at local and national levels across the globe have proactively implemented a range of policies designed to curtail vehicular emissions. These policies encompass measures such as low emission zones, designated car-free days, and innovative schemes like the odd-even license plate rotation system [13, 14, 15].

In 2020, human activities experienced a significant reduction due to the implementation of stringent government measures, including city lockdowns and restrictions on outdoor activities. A direct outcome of these policies was the notable decrease in transportation and subsequent reduction in air pollution [16]. Numerous studies have been conducted to investigate and assess the impact of lockdown-induced reductions in traffic on air pollution concentrations. In China, a study by Chen et al. [17] utilized regression analysis to examine the correlation between changes in vehicle restriction policies and air pollution levels before and after the onset of COVID-19 in 49 cities. The key finding suggests that cities with slower economic growth experienced more substantial improvements in air quality due to reduced traffic, compared to cities characterized by rapid development. In the [18], the NO level around schools in the United Kingdom is found to reduce 35.1% and 40.8% in Urban-South and Urban-North traffic during the stay-at-home order in March 2020. Another study [14] analyzed the relationships between traffic flow and air quality from 2017, 2018 and 2020 in Padova, Italy. It found that NO and NOx are significantly associated with the vehicle flows but no clear evidence for $PM_{10}$. In the context of particulate matter such as PM, humidity stands out as a crucial factor influencing its concentration. An adaptive correction framework introduced by authors in [19] effectively addresses this challenge by dynamically modeling hygroscopicity, thereby mitigating humidity's influence on particle measurements.

Understanding and managing air quality is a major focus for researchers investigating monitoring, modeling, and forecasting in urban areas. In this regard, machine learning and deep learning methods are utilized to predict the quality based on the observed patterns [20]. For instance, Iskandaryan et al. utilized Graph Neural Networks to analyze data from Madrid, evaluating metrics like Root Mean Square Error and Mean Absolute Error to assess prediction accuracy [21, 22]. Additionally, they explored the effectiveness of LSTM, which showed promising results in predicting metrics such as nitrogen dioxide concentrations [23, 24]. Also, other methods such as long short-term memory (LSTM) recurrent neural network (RNN) were used in traffic forecasting based on various air pollutant and meteorological metrics [25].

In this study, we aim to analyze the correlation between various pollutants and traffic, as well as meteorological metrics in the city of Madrid during the pre, during, and post-COVID eras.

## 3. Data Enrichment Toolchain Architecture

In this section, the functional architecture of the DET is described, along with a brief explanation of its role as the key enabler for data enrichment and linking. Further insights on the DET can be found at [8].

The main objective of the DET is to enable the enhancement of datasets and data-streams by way of enrichment mechanisms based on the application of linked-data, semantics, and AI technologies.

The DET architecture is divided into two different planes based on the relation to data or control. The data plane allows utilizing data, and the control plane enables configuring the components.

### *3.1. Data plane*

Figure 1 depicts the DET functional architecture and illustrates the flow of data through different modules. The DET is composed of microservices that progressively transform and enhance the data. In general terms, the DET can be seen as a pipeline with a set of modules that each target an atomic step within the overall process. Particularly, the aim of the data enrichment phase is to improve the quality and value of the original information.

The core components of the architecture, as seen in Figure 1, are the injection chain, the context broker (one or multiple in a federation) and the

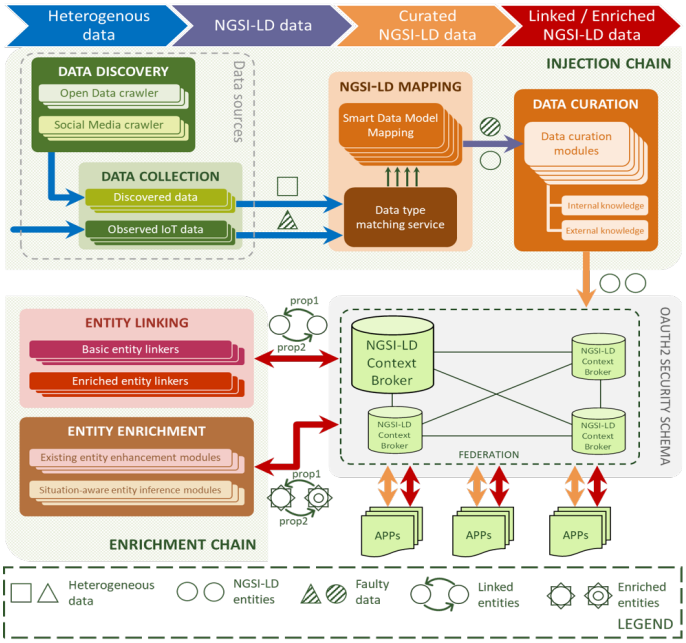


Figure 1: DET Data Plane architecture [8].

enrichment chain. The injection chain is responsible for transforming raw data into curated NGSI-LD data. The processed data can be accessed by external applications through the context broker, which facilitates communication, storage, and historic data management. Finally, the enrichment chain handles the linking and enrichment of NGSI-LD data obtained through the broker. A more detailed explanation of each step is provided below:

- **Data Discovery and Collection** modules acquire raw data from heterogeneous sources. These may include, but are not limited to, IoT based deployments, social media, web-stored, statistical catalogues or meteorological agencies. The output of this phase consists of the raw data collected from various data sources, which are, by definition, heterogeneous in both type and format.
- **NGSI-LD Mapping** modules transform the raw data into the NGSI-LD information model, and more specifically, the resulting data is compliant with FIWARE's Smart Data Models initiative [26]. The transformed data is then forwarded to the next phase.

- **Data Curation** modules ensure that the data injected in the NGSI-LD Context Broker is adequate to be processed by data processing modules. As an example, curation may include data quality mechanisms such as outlier detection, deduplication, loss management or taggers for data quality metrics (accuracy, timeliness and so on). These modules inject their resulting clean data into the NGSI-LD Context Broker.
- **Entity Linking** modules create NGSI-LD Relationships between two or more NGSI-LD Entities, regardless of their data source. This is done by finding and establishing common aspects among data, whether semantic, spatial, temporal, or otherwise. These relationships facilitate any further processing by simplifying navigation through connected data. Once the linking process is finished, the newly linked data is injected back into the NGSI-LD Context Broker.
- **Entity Enrichment** modules generate new NGSI-LD Entities or new NGSI-LD Properties in existing Entities. This is usually done by leveraging information from external knowledge sources. These modules are typically specific to a particular domain and are designed with a specific application or use case in mind; however, this is not always the case, as domain-agnostic enrichment is also a possibility. After the enrichment process is complete, the enhanced data is injected back into the NGSI-LD Context Broker.

### *3.2. Control plane*

The control plane enables external or internal applications to configure the DET components providing additional functionalities through parametrization. The control plane is decoupled from the data plane to avoid sharing the same component for control-related functionalities and the control interfaces are kept simple to avoid implementing different interfaces depending on the component. Compared to the data plane, the NGSI-LD context broker is replaced by the control broker in the control plane. The control broker provides the communication between the DET component and the application configures the DET component.

The Inversion of Control (IoC) pattern is used to facilitate communication among DET components and prevent them from learning the specifics of other components or their endpoints. Applications or DET components can call the components only through the control broker to reduce the security

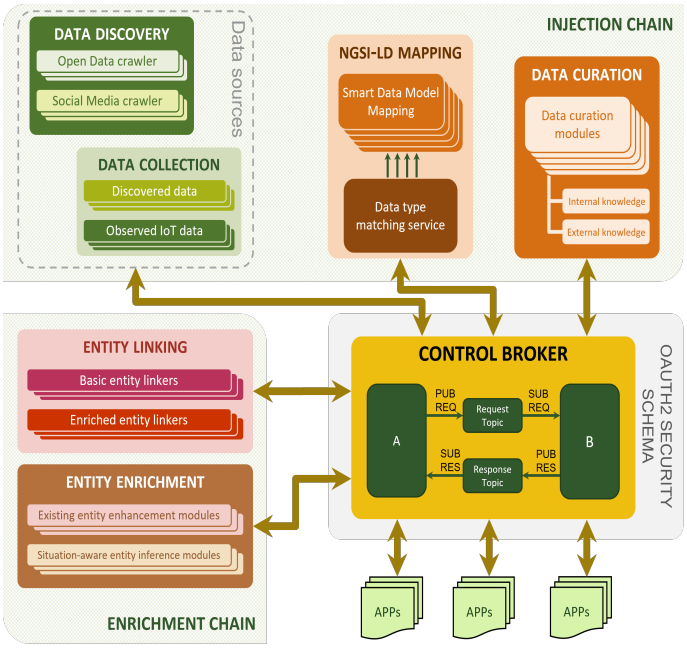


Figure 2: DET Control Plane architecture.

risks by not exposing the components to external applications directly. The control broker uses a pub/sub-event-based mechanism for the communication. Components are subscribed to specific topics and users or applications can publish to these topics to change the configuration of the DET components. A predefined format is used based on the configurable parameters available in each component.

### *3.3. DET deployment details*

The deployment architecture of the DET is based on a federated setup that includes Scorpio Context Broker Federator that connects the different "satellites" of partners in the data plane. In addition, there is a control broker that handles the control plane functions of the DET.

Access to the Federator Scorpio Broker and the EMQX Control Broker is restricted using OAuth 2.0. The technology used for the Identity and Access Management is Keycloak. This enables authorisation restrictions based on JSON Web Tokens (JWT), and the communication is encripted with Transport Layer Security (TLS).

We have deployed the Scorpio NGSI-LD Brokers from their latest docker images (i.e. scorpiobroker/all-in-one-runner:java-kafka-latest). On the other hand, we have used Python 3 as the programming language for the DET components. These are deployed as separate Python scripts acting independently, which enhances their modularity and reusability. Communications between components are achieved though HTTP, with most components implementing their own lightweight HTTP server with the flask and waitress Python libraries.

The DET used in this article is publicly available and can be found at [27].

Furthermore, several Injection Chains are implemented: IoT Data Injection Chain, Web Data Injection Chain and Social Media Data Injection Chain. This list can be extended by application developers by introducing new Injection Chains. Technical descriptions and deployment details of the listed injection chains can be found in [28]. As a result of this implementation, a bunch of enriched datasets have been created and are available at [29].

## 4. Data Sources

One of the critical stages in the DET platform is the initial phase of data discovery and collection. This entails extracting raw data from diverse sources, followed by a curation and pre-processing stage before the data is fed into the brokers. The data collection spans various sources, including IoT-based data, national and international meteorological data, and information from social media platforms. However, the primary focus of this paper revolves around the utilization of IoT-based data, specifically data obtained from a wide array of IoT sensors, such as traffic, pollution, and weather sensors.

Furthermore, the analysis also incorporates meteorological data, as elaborated in Section 6. As part of the work carried out in the framework of the SALTED project, which constitutes the context for this paper, the selection of cities for data acquisition is based on a set of rigorous criteria, including the factors of public accessibility and data update frequency, with an emphasis on data being updated at intervals of one hour or less. In this sense, data on traffic and pollution from several European cities have been gathered through the corresponding DET injection chain. For a comprehensive overview, Tables 1 and 2 summarize the essential characteristics of the traffic and pollution datasets that have been collected so far. It is important to

Table 1: Characteristics of the traffic datasets.

| City | Format | Frequency | Size | #Sensors |
|---|---|---|---|---|
| Santander | JSON | 1 minute | ∼900 MB | ∼300 |
| Barcelona | CSV | 5 minutes | ∼100 MB | ∼525 |
| Oslo | JSON | 1 hour | ∼15 MB | ∼130 |
| Madrid | CSV | 15 minutes | 74-650MB | ∼5000 |

Table 2: Characteristics of the pollution datasets.

| City | Format | Frequency | Pollutants | #Stations |
|---|---|---|---|---|
| Santander | CSV | 1 hour | $PM_{10}$,$SO_2$,$NO_2$,CO | 2 |
| Barcelona | JSON | 1 hour | $PM_{10}$,$SO_2$,NO,$NO_2$,$O_3$,CO | 7 |
| Oslo | JSON | 1 hour | $PM_{10}$,$SO_2$,NO,$NO_2$,$O_3$,CO | 13 |
| Madrid | CSV | 1 hour | $PM_{10}$,$SO_2$,$NO_2$,$O_3$,CO, $PM_{2.5}$, NO | 25 |

mention that data from more cities can be made available through the DET just by integrating its corresponding Data Discovery and Data Collection modules. Moreover, Tables 1 and 2 are only showing the details about the data that has been used for establishing the correlations and relationships that are within the scope of this paper. but since the DET is intended for general purpose goals, its sources are not only related to these cities and these data, but they are spanning over other application domains such as social media, agriculture, socioeconomic statistical data, and, in perspective, more.

Focusing on the scope of the paper, as shown in Tables 1 and 2, the most common formats are CSV and JSON. The update frequency of the measurements tends to be faster in the traffic data, resulting in additional processing to align both datasets. This means either generating synthetic measurements for the less frequent dataset or aggregating the measurements of the more frequent one. The size of the datasets in terms of disk usage is quite relevant for the traffic data since it can easily grow up to gigabytes. The sizes displayed in Table 1 are indicative of one month of data. Naturally, the more frequent the measurements, the more sizable the dataset will be. This characteristic is not as relevant in the case of the pollution data, given the fact

that the number of pollution stations is significantly lower than the number of traffic sensors. Finally, the pollutants covered by the different cities may not totally overlap, but the most relevant ones for this study (as mentioned in Section 1) are present in all of the cities surveyed.

There are data available from some well-monitored cities like Santander, Madrid, Oslo and Barcelona. They share the accuracy in determining the traffic intensity (per hour, the worst case of Oslo) and relevant measurement in terms of pollutants ($PM_{10}$, $SO_2$, $NO_2$ and CO). These cities are also interesting because they have different traffic patterns depending on the touristic period and the way the cities "behave" because of differences in weather conditions. Other cities like Dublin and Aarhus have been considered, but the datasets on traffic have not comparable accuracy.

The analysis presented in the following sections has focused on the largest city of the ones from which we had data available, Madrid. However, as it will be presented in Section 7 the same analysis will be performed over the data available from the other three cities so that, on the one hand, the DET platform capacities for allowing the development of interoperable application and services over harmonized data sources are further validated (as it will not only show the correlation of heterogeneous types of data, but also enable multi-site correlation), and, on the other hand, multi-site results' evaluation (including comparison and transfer learning) might allow for the extraction of further conclusions that are not evident on the single-site case.

## 5. Curation and treatment of the data

DET's injection chains (cf. Figure 1) analyze raw data sources and generate the correspondingly normalized data elements using the NGSI-LD information model. By formatting heterogeneous data into a single, standardized format, the mapping function enables uniform processing for the following components in the pipeline. Due to its heterogeneous nature, the input can be represented using several different data formatting standards such as Comma Separated Values (CSV) or JavaScript Object Notation (JSON). Moreover, both the names of the properties and the values can be highly different from one another, as a result of the heterogeneous data sources and their internal policies, language, units, and several other factors.

Moreover, in the process of transforming the raw data to NGSI-LD-formatted high quality data, rigorous checks were conducted to ensure data consistency

and reliability. This involved refining the data to eliminate sequences generated by malfunctioning sensors and incorporating missing sequences. Notably, specific procedures and checks were implemented for the considered use case to ensure high data usability and quality. It is important to highlight that thanks to the initial harmonization of data, subsequent processing and curation of data can be performed homogeneously independently of which data source it has been collected from. Detailed descriptions of these essential processes are presented in the subsequent sections.

### *5.1. Pollution data*

In the context of air pollution data collection and analysis, our approach involves careful curation and handling of the employed datasets. As defined in the previous section, pollution data was obtained from open data portals, that comprised various pollutants, including CO, $SO_2$, $PM_{10}$, $NO_2$, and NO, recorded on an hourly basis. To ensure the accuracy and applicability of the data, we initiated the process with extensive data cleaning and pre-processing to ensure that it is suitable for analysis.

This involved identifying the essential components of our study, as well as checking for missing or inconsistent data, and transforming the data into a format that can be easily analyzed. Some of our datasets showed various inconsistencies over the course of a year, with periodic gaps in the recorded values related to pollution levels. These gaps can occur for several reasons, including equipment failure, delays in the data collection process, or other unanticipated events. To keep the dataset coherent and manageable, we used an interpolation method to approximate the missing values by looking for patterns in the available data.

#### *5.1.1. Interpolation technique*

A linear interpolation technique was used to fill the missing values in the data. This approach is predicated on the idea that pollution levels fluctuate gradually over time and that trends in the existing data can be utilized to forecast values for the missing intervals. Linear interpolation is the simplest method to estimate the missing value based on the known values. Equation 1 below shows the mathematical expression for the linear interpolation where $a$ represents the timestamp where we want to estimate the value.

- $(a - a_1)$ shows the variation between the earliest known timestamp and the target timestamp.

- $(b_2 - b_1)$ shows the variation in pollutant levels between the second and first known timestamps.
- $(a_1 - a_2)$ shows the time interval between the two known values.

$$b(interpolated) = b_1 + \frac{(a - a_1) \cdot (b_2 - b_1)}{a_2 - a_1} \tag{1}$$

*5.2. Traffic data*

The traffic data utilized in this study is sourced from the Madrid City Council Open Data portal[4]. This dataset serves as a comprehensive reference for traffic-related attributes, and its observation frequency is deemed sufficient for our analysis. The city is equipped with approximately 4,746 traffic sensors, each capable of providing real-time traffic intensity data with a commendable level of precision. Each sensor in the dataset records various attributes at 15-minute intervals, contributing to a substantial volume of data. For instance, a single day's traffic observations encompass data from all 4,746 sensors over 24 hours, with each sensor generating four observations per hour. This results in a large dataset, reflecting the detailed and continuous nature of the traffic observations.

Initially, our focus was on a select set of attributes from the sensors capturing the collected data, specifically, Occupancy and sensor status. Occupancy represents the percentage of time a vehicle occupies the detector within a 15-minute timeframe. The details regarding traffic attributes, such as the Occupancy time interval, are elaborated in the traffic data structure and content document specification.

The curation process applied to the raw data aimed to identify and address missing data. During this process, we discovered 145 instances of missing values for a single sensor in April 2019, specifically related to the Occupancy attribute. While these individual instances are negligible in the context of the overall data volume for a month, the cumulative total of missing values across all sensors amounted to 22,992 in a single month. This cumulative figure could potentially introduce inconsistencies and hinder a coherent analysis of the data.

[4]https://datos.madrid.es/portal/site/egob

To handle the missing values, we employed interpolation techniques, utilizing functions supported by the Panda library, as detailed in Subsection 5.1.1. This approach ensures a more complete dataset and facilitates a more accurate and comprehensive exploration of the data.

## 6. Data correlation use case

The well-structured data collected in our system offers many opportunities for both predictive and descriptive analyses. In the realm of smart cities, diverse datasets make it apparent that many of these datasets are intuitively related to each other. An example of this correlation is the interaction between meteorological data and traffic data with air pollution data. Meteorological factors such as wind, temperature, and humidity significantly influence the behaviour of various pollutants and their dispersion within the environment.

Furthermore, the intensity of traffic plays a pivotal role as a major contributor to pollution levels. It stands to reason that higher traffic intensity would correspond to a heightened concentration of pollutants. However, it is important to note that city datasets encapsulate intricate phenomena, making it challenging to discern straightforward and easily identifiable relationships among the various datasets. Therefore, a critical need arises for tools and processes that can facilitate the assessment of correlations between these phenomena and the datasets that encapsulate them. The DET architecture has been developed aiming at easing the linking and enrichment of different types of data.

In order to evaluate the potential of the DET, a specific use case requiring correlation and integration of different datasets is presented. The selected Use Case for challenging the linking capabilities serves as the analysis, in well-monitored cities, of the relation between traffic intensity, meteorological conditions and pollution values before, during and after the COVID period. The aim is to evaluate how the lockdown period impacted on these values considering that it is reasonable to assume an almost negligible contribution of the road traffic during the strictest periods of lockdown (i.e. April 2020).

### *6.1. Use case introduction*

As previously mentioned, the main objective of the data fusion enabled through the DET is to delve into the correlation between traffic patterns and meteorological conditions with the air pollution factors in various city

scenarios. In this paper, we specifically focus on the city of Madrid. We will achieve this by conducting a comprehensive analysis of historical data pre-processed through DET injection chains from three distinct periods: pre-COVID (2019), during the COVID pandemic (2020), and post-COVID (2022). Our initial focus on Madrid is motivated by three key factors obtained after thorough examination:

- Madrid's highly developed transport network infrastructure makes it one of the biggest European cities with exceeding air pollution from WHO standards and thus requires demanding traffic pollution policies [30].
- Madrid, with almost 6 million inhabitants, has been one of the most strongly impacted cities by COVID-19, accounting for the greatest share of cases and deaths in the country [31]. The lockdown policy in such a densely populated city could give us a clearer impact on the factors influencing air pollution changes.
- Madrid prides itself on being a smart city equipped with over hundreds of thousands of IoT sensors and open data platforms. The study takes advantage of open data to understand the nature and interplay between meteorological factors, traffic volume and, ultimately, air quality.

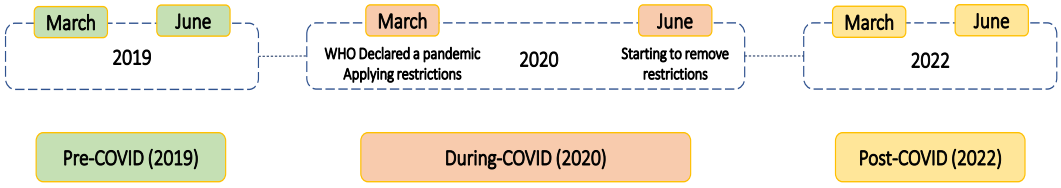


Figure 3: Selected data of Madrid city in 3 time periods.

In any case, as it will be highlighted and further discussed in Section 7, considering the fact that similar data, already harmonised and curated through the DET pre-processing, is available from three other cities (i.e. Barcelona, Oslo and Santander), the same correlation analysis will be carried out on those datasets. In this sense, the motivation for limiting the use case to the analysis in the city of Madrid is, besides article extension limitation, the fact that it already allows fulfilling the research objectives addressed in

the paper in terms of the harmonisation and curation of heterogeneous data sources, and the interlinking of such harmonised datasets in look for discovering higher-level knowledge (in our case, impact of traffic reduction policies in the city air pollution levels).

### *6.2. Madrid historical data*

In the use-case, the data was extracted from the NGSI-LD Broker within the DET to conduct the analysis. The COVID era data serves as an invaluable context for our research, given the stringent lockdown measures implemented to mitigate the pandemic's spread and the consequential reduction in traffic along with meteorological metrics, which provides a pivotal point of interest for assessing pollution metrics. Madrid, in particular, experienced several stages of restrictions during this period, from the pre-state of alarm on March 8th to the lifting of restrictions on June 20th, 2020.

To carry out our analysis, as shown in Figure 3, we focused on a four-month duration during the COVID pandemic from the start of March to the end of June. We also studied equivalent timeframes in 2019 (pre-COVID) and 2022 (post-COVID) to facilitate correlation pattern analysis. The selected data was subsequently subjected to pre-processing to establish a daily unit within the defined interval, encompassing all stages of the COVID-19 pandemic lockdown in Madrid.

To further investigate whether there have been alterations in pollutant concentrations and traffic flows in selected time periods and to what extent these

Table 3: Summary of selected stations.

| **Type** | **No. of Stations** | **Features** | **Unit** |
|---|---|---|---|
| Air Station | 3 | NO<br>$NO_x$<br>$PM_{2.5}$<br>$PM_{10}$ | µg/$m^3$ |
| Meteorological Station | 3 | Wind Speed<br>Temperature<br>Solar Radiation<br>Humidity | m/s<br>C<br>$W/m^2$<br>% |
| Traffic Station | 33 | Passing Cars | Num/hour |

changes have occurred, we have employed various features in the selected stations as indicated in Table 3.

*6.2.1. Air quality data*

From the pool of 24 air quality monitoring stations scattered throughout Madrid, we picked three stations that are strategically located in distinct areas, each offering unique characteristics for our comprehensive analysis:

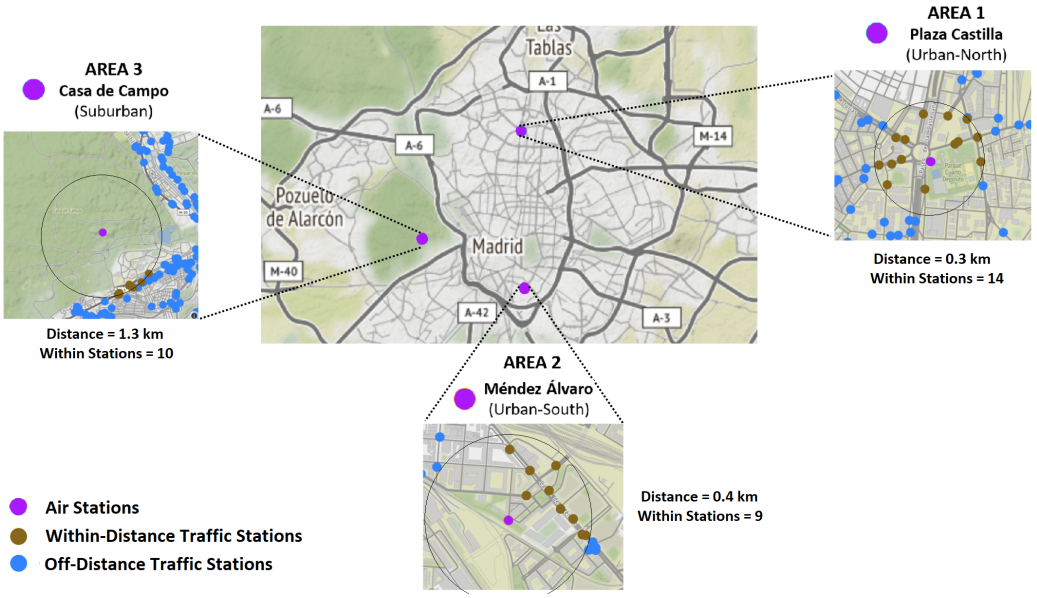


Figure 4: Illustration of Madrid selected air and traffic stations. For three selected air stations in each area, the distance difference threshold and the number of within-distance traffic stations are shown.

- **Area 1 (A1: Urban-North):** Located in the northern part of Madrid, called "Plaza Castilla" which serves as a pivotal transport hub, facilitating the convergence of traffic from multiple routes. This area is renowned for its ease of commuting and is a sought-after residential neighborhood, making it an ideal candidate for analyzing traffic flow and pollution. Numerous offices and residential communities surround it.
- **Area 2 (A2: Urban-South):** Situated in the southern region of Madrid at "Méndez Álvaro", this station is strategically positioned near one of the city's major bus and train stations. Given its proximity

to these transport hubs, this area experiences heavy traffic flow, serving as a primary gateway to the city center.

- **Area 3 (A3: Suburban):** Nestled in the west of Madrid, "Casa de Campo" offers a distinct setting away from the city center. Notably, this area features a sprawling public park, attracting residents, especially during weekends and holidays, as it is famous for its pleasant and healthy climate. Furthermore, the nearby "Avenue of Portugal" serves as a primary entry point into Madrid center, making it an intriguing case study for analyzing traffic flow from outside the city.

Among all the stations in these areas, we have applied the first filter to look at only stations with sensors that measure our interested air quality indexes: NO, NOx, $PM_{2.5}$ and $PM_{10}$.

Only seven stations pass the first filter, with only one from Urban-South and 1 from Suburban. To select the remaining Urban-North, we look at the furthest station. The finalists' station name, type and location are illustrated in Figure 4.

Pertaining to the sensors with which these stations are equipped, for the $PM_{2.5}$ and $PM_{10}$, they use a continuous dichotomous ambient air monitor composed of two Filter Dynamics Measurement Systems (FDMS) and two mass sensors housed in a single cabinet, with a measurement range of 0 to 1,000,000 µg/m$^3$ (1 g/m$^3$), a resolution of 0.1 µg/m$^3$, a precision: ±2.0 µg/m$^3$ (1-hour avg) and an accuracy for mass measurement: ±0.75%. Regarding the Nitrogen oxides, the sensor used is based on cross-flow modulation with reduced pressure chemiluminescence (CLD). Its measurement range is of 0-0.1/0.2/0.5/1.0 ppm, its repeatability ±1.0% of F.S. and a linearity of ±1.0% of F.S.

Furthermore, the chosen stations record air quality metrics on an hourly basis and periodic measurement campaigns are also conducted by means of a mobile unit. Additional miscellaneous information about the placement and pollutants measured by each station is readily available on their air quality web portal[5]. We collect this information through a REST interface provided by Madrid's Open Data portal. After the retrieval, the data is processed as described in Section 5.

[5]Madrid air quality portal (in Spanish)

### *6.2.2. Meteorological data*

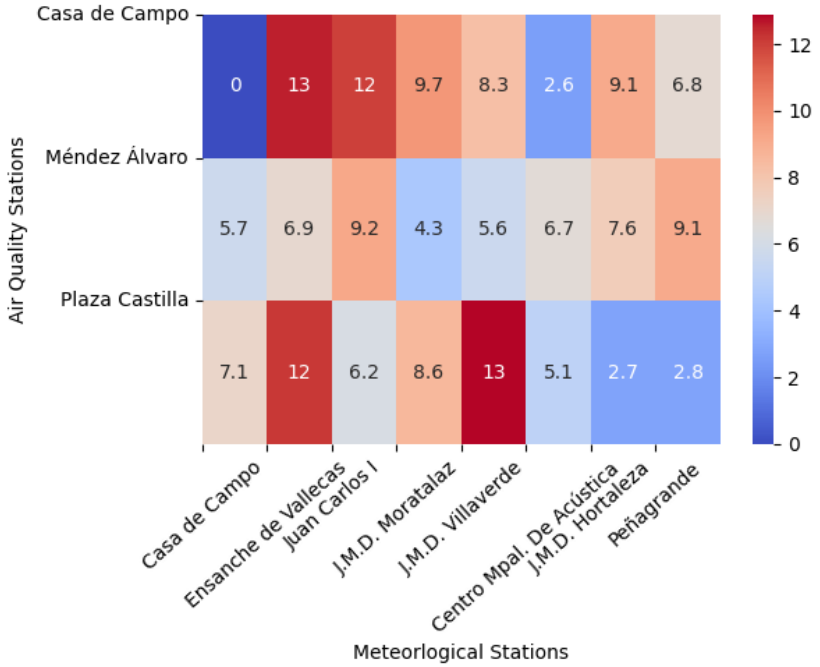


Figure 5: Pair-wise distance matrix between meteorological and air stations in KM.

Meteorological data plays a crucial role in understanding the atmospheric conditions that impact air quality. Our study focuses on four key meteorological factors: wind speed, temperature, solar radiation and humidity, as they are pivotal in assessing air quality in Madrid. However, it is essential to note that not all meteorological stations in the region provide data on all four of the mentioned indicators. In fact, out of the 36 meteorological stations scattered across Madrid, only 8 are equipped to measure wind speed, temperature, solar radiation and humidity simultaneously. To ensure the highest level of precision and accuracy in our analysis, we create a pair-wise distance matrix based on the latitude and longitude of each station (as depicted in Figure 5). By doing so, we identify the meteorological stations that are in closest proximity to the selected air stations, as illustrated in Figure 6.

These stations are equipped with thermo-hygrometer sensors in addition to, in some cases, other devices that match the meteorological parameter being measured (e.g., if the station measures wind speed, then it is equipped with anemometers). In the case of these sensors, their specifications are

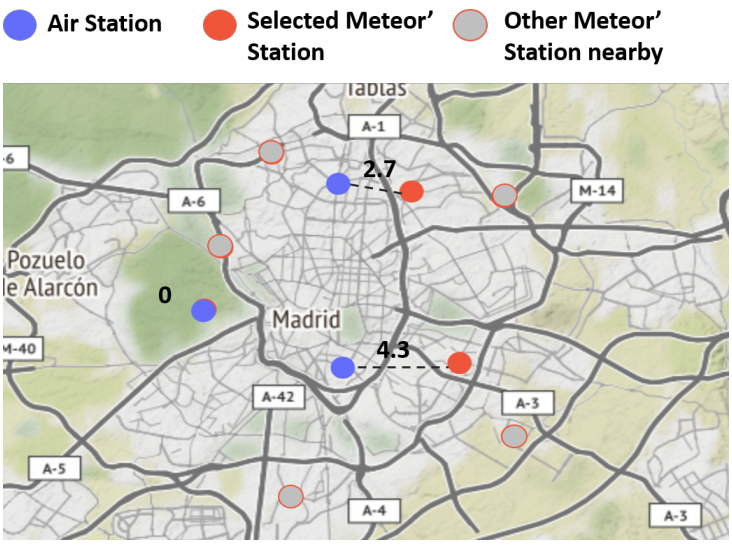


Figure 6: Madrid selected meteorological stations and their pair-wise distance to the selected air station.

not provided by the city; nevertheless, information about the placement and parameters measured by each station is readily available on their meteorological web portal[6]. We collect this information hourly, using the same method pointed out in the air quality data subsection.

It is essential to clarify that our primary objective is not solely to establish the direct correlation between meteorological factors and air quality at individual air stations. Instead, we opt for a more comprehensive approach. We aggregate the meteorological data by calculating the daily average values from the three selected meteorological stations. As a result, each of the three air stations will have a singular value for each meteorological factor per day, creating a harmonized dataset that simplifies further analysis. This approach ensures consistency and comparability in our study, enhancing the accuracy and reliability of our findings.

*6.2.3. Traffic data*

In our study, we aim to comprehensively assess the air quality in Madrid, considering not only meteorological factors but also the influence of traf-

[6]Madrid meteorological portal (in Spanish)

fic conditions. The number of traffic stations continuously rises, with new installations each year. To ensure a fair and consistent comparison over a 3-year period, we specifically selected traffic stations from 2019, resulting in a total of 4,153 stations for our analysis. These traffic stations record data at a 15-minute interval.

These traffic measurement points are based on vehicle detection. A small number of them are equipped with cutting-edge modules: around 1% of the sensors include number plate readers, and almost 2% incorporate optical sensors that enable artificial vision processing at the city's Mobility Management Center. Their basic functionality, on the other hand, is based on semaphore control to measure the number of vehicles passing through a specific lane. We process this information as pointed out in the previous subsections.

To align the data for meaningful comparisons, we aggregate all data points into daily units, representing each sensor's data on a daily basis. The key feature we are interested in from this traffic dataset is the number of passing cars, which can significantly impact air quality in urban areas.

Similar to our approach with meteorological stations, we create a pair-wise distance matrix based on the latitude and longitude of each of the 4,153 traffic stations concerning the 3 air stations in our study. Determining an optimal proximity threshold for traffic stations in proximity to air stations differed in selected areas as traffic sensors are more likely to be located in two selected urban rather than suburban ones. Finally, as shown in Figure 4, we calculated the optimal threshold for our analysis to cover a couple of sensors near the selected air stations.

To derive meaningful insights, we average the number of daily passing cars from the closest traffic stations to determine a single traffic density value for each air station's area. This approach enables us to assess the impact of traffic conditions on air quality in various regions of Madrid, contributing to a more comprehensive and accurate analysis of air quality factors.

### *6.3. Analysis of air quality, traffic and meteorological metrics changes*

To gain a comprehensive understanding of the data collected from these three types of monitoring stations, we have conducted a detailed analysis that focuses on both the statistical means and the percentage change from the previous year. This analysis encompasses air quality and traffic flow,

taking into account not only the temporal dimension but also the spatial aspect of station locations.

*6.3.1. Statistical metrics*

To assess alterations in selected air quality, traffic and meteorological metrics, we employed the Mann-Whitney U test, a non-parametric statistical method designed to ascertain significant differences between distributions of two independent groups. This test is particularly suitable when the data fails to meet the assumptions necessary for parametric tests such as the t-test. By evaluating whether one group's values consistently rank higher or lower than the other across the entirety of their distributions, the Mann-Whitney U test provides insights into potential disparities between the groups. The test statistic $U$ is calculated as follows:

For two independent samples, with $n_1$ observations in group 1 and $n_2$ observations in group 2:

- Ranking all values from both groups together in ascending order, assigning ranks from 1 to $N = n_1 + n_2$ to the combined dataset
- Calculating the sum of ranks ($R_1$) for the observations in group 1.
- The test statistic $U_1$ can be computed as:

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - R_1 \tag{2}$$

We also employed effect size or "probability of superiority" ($\phi$) measures to help quantify the magnitude of differences or the strength of association between groups. This metric is calculated as the probability that a randomly selected observation from one group will be greater than a randomly selected observation from the other group. This effect size can be obtained using the formula:

$$\phi = \frac{U}{n_1 \times n_2} \tag{3}$$

Where $\phi$ is the probability of superiority (effect size), $U$ is the Mann-Whitney U test statistic and $n_1$ and $n_2$ are the number of observations in group 1 and 2, respectively.

The $\phi$ value ranges from 0 to 1, with higher values indicating a greater probability that an observation from one group is higher than an observation from the other group.

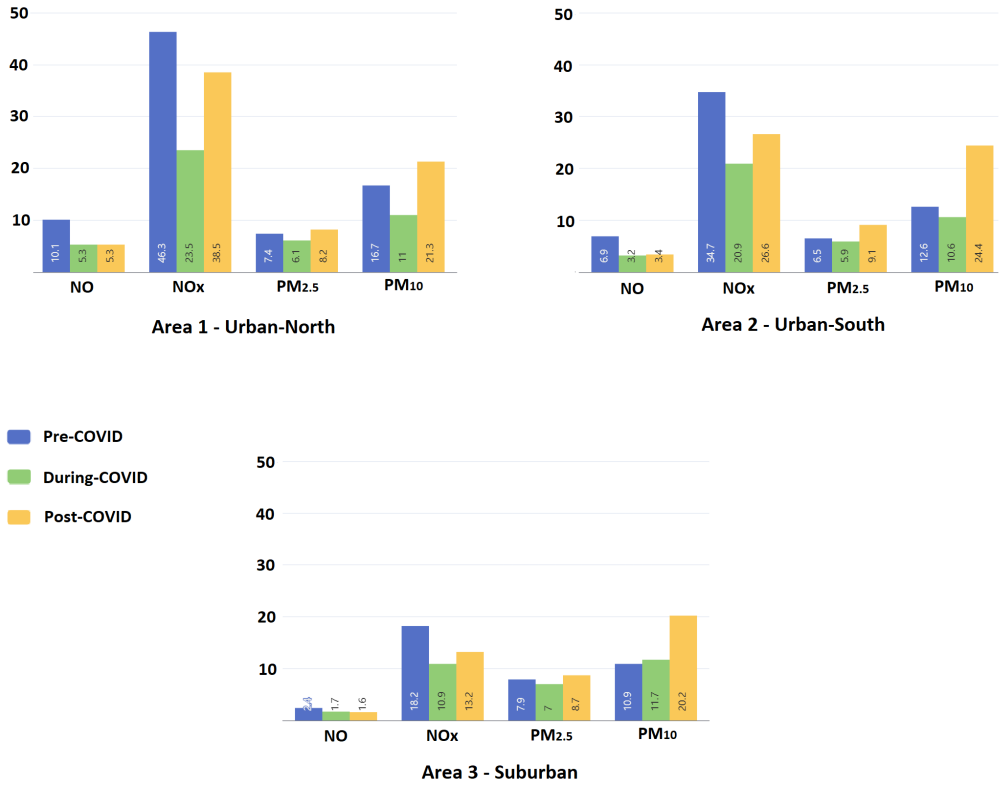


Figure 7: Descriptive statistics from air quality stations.

*6.3.2. Air quality changes*

The analysis of air quality changes over three different time spans in each area is depicted in Figure 7. A detailed examination of each pollutant reveals a consistent decrease in all pollutants during the COVID period (averaging a 46% reduction for NO and NOx, and a 15% reduction for $PM_{2.5}$ and $PM_{10}$), except for $PM_{10}$ in suburban areas, which exhibited a marginal increase.

Upon comparing the Post-COVID period to the During-COVID phase, on average, NO levels remained relatively stable across all areas (with a 2% change), whereas $NO_x$ experienced a substantial 41.6% increase, notably

Table 4: Mann Whitney U-test p-value and effect size on the change of air pollution concentration.

| | | p-value | | | effect size | | |
|---|---|---|---|---|---|---|---|
| **Pollutant** | **Area** | Pre&During | During&Post | Pre&Post | Pre&During | During&Post | Pre&Post |
| NO | A1 | 0.000 | 0.201 | 0.000 | 0.727 | 0.001 | 0.8 |
| | A2 | 0.000 | 0.000 | 0.37 | 0.261 | -0.043 | 0.26 |
| | A3 | 0.001 | 0.852 | 0.001 | 0.403 | 0.032 | 0.446 |
| NOx | A1 | 0.000 | 0.000 | 0.002 | 1.146 | -0.851 | 0.426 |
| | A2 | 0.000 | 0.000 | 0.025 | 0.571 | -0.324 | 0.384 |
| | A3 | 0.000 | 0.000 | 0.000 | 0.741 | -0.289 | 0.532 |
| $PM_{2.5}$ | A1 | 0.000 | 0.184 | 0.461 | 0.467 | -0.349 | -0.131 |
| | A2 | 0.041 | 0.000 | 0.35 | 0.174 | -0.452 | -0.372 |
| | A3 | 0.294 | 0.015 | 0.901 | 0.217 | -0.262 | -0.118 |
| $PM_{10}$ | A1 | 0.000 | 0.000 | 0.054 | 0.838 | -0.359 | -0.158 |
| | A2 | 0.062 | 0.000 | 0.000 | 0.282 | -0.43 | -0.365 |
| | A3 | 0.251 | 0.014 | 0.001 | -0.122 | -0.294 | -0.319 |

soaring by 64% in the Urban-North region. Conversely, both $PM_{2.5}$ and $PM_{10}$ showed higher average concentrations post-COVID, with increases of 20.3% and 68.8%, respectively. Particularly noteworthy is the surge in $PM_{10}$ levels in the Urban-South area, rising from 12.6 $\mu g/m^3$ during the post-COVID phase to 24.4 $\mu g/m^3$.

These variations underscore the intricate nuances in pollutant levels across different temporal periods and geographical areas, elucidating a notable divergence in the trends observed for various pollutants post-COVID, especially evident in $PM_{10}$ concentrations within the Urban-South locality.

To statistically examine the alterations, Table 4 presents the outcomes of the Mann-Whitney U-test alongside effect size measurements. The comparison of changes in each pollutant across different areas is categorized into three groups: Pre&During, During&Post, and Pre&Post COVID periods. Significant changes are indicated in the table by p-values of 0.

For a more comprehensive understanding of these changes, effect size values offer a broader perspective. A value greater than 0.5 signifies a substantial effect. However, in cases where the effect size is negative, it implies that the average of the second group has increased compared to the first group. These effect size measurements provide additional insights into the magnitude and

directionality of the observed changes in pollutant levels across distinct time periods and geographical regions.

*6.3.3. Meteorological changes*

We conducted an analysis of the average values for four specific meteorological metrics—temperature, wind speed, solar radiation and humidity as depicted in Figure 8. Temperature and wind speed exhibit notable stability across all the compared time spans. For instance, the temperature in Madrid remains relatively consistent, registering at 14.7°C from March to June during pre-COVID years and at 14.6°C during the COVID period, with a slight increase to 15.1°C post-COVID. Concurrently, the wind speed maintains a steady trend throughout the selected timeframe, averaging at 1.88 $m/s$.

However, the solar radiation factor displays some variation. It experienced a reduction from 270.7 $W/m^2$ in the pre-COVID year to 238.5 $W/m^2$ during the COVID period, before rebounding by 5% to 250.7 $W/m^2$ in the post-COVID year. This shift in solar radiation levels suggests fluctuations in environmental conditions during these time periods, potentially influencing various aspects of the observed air quality changes. The same trend has been observed in reverse for humidity. For this metric, it was 45% in the pre-COVID year while it increased to 57.73% during the COVID period, before decreasing by 5% to 52.32% in the post-COVID year.

Referring to Table 5, it's evident that humidity and solar radiation stand out with the most substantial effect size and the smallest p-values among all meteorological metrics. This prominence can be attributed to significant changes in averages observed across all three selected periods. Conversely, the trends for temperature and wind speed exhibit less pronounced effects, largely remaining steady throughout these periods.

*6.3.4. Traffic changes*

Traffic plays a significant role in air quality, and it stands as a primary factor in our analysis. As demonstrated in Figure 9, there was a substantial 61.5% decrease in traffic flow across selected areas during the COVID-19 period. While post-COVID restrictions were lifted, daily average traffic levels have not fully rebounded to pre-pandemic norms, signifying enduring alterations in traffic behavior. Analysis of traffic averages across regions indicates a marginally higher traffic volume in Urban-North, attributed to the presence of numerous corporate entities. Notably, the suburban region consistently

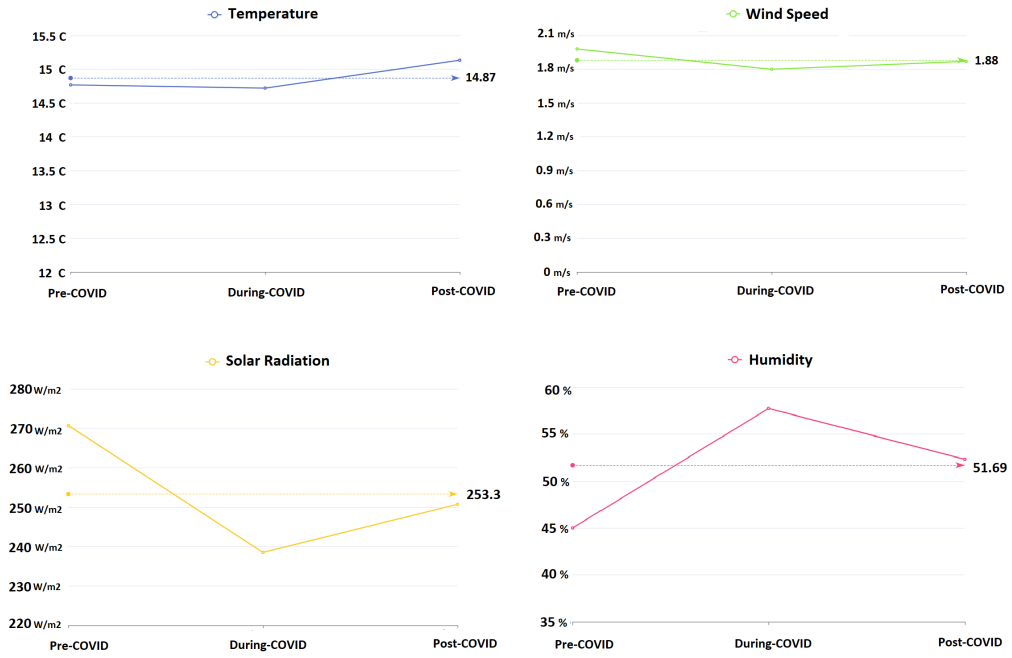


Figure 8: Descriptive statistics from meteorological stations.

records double the traffic volumes of urban areas, mainly due to the strategic importance of the A-5 motorway connecting Madrid to the Spanish-Portugal border and central Madrid.

As presented in Table 6, the analysis of p-values for traffic flow across all three time periods highlights substantial changes across different areas with average of less than 0.01 for all parts. In examining these changes more comprehensively, the effect size indicates significant differences, particularly when comparing the pre and during-COVID eras, with a notable effect size of 1.057 observed in the suburban area. Comparatively, when evaluating the pre and post-COVID periods, a smaller yet positive effect size is evident, signifying a slight reduction in the average traffic flow.

### *6.4. Results of correlation analysis*

For correlation analysis, we used Spearman's rank correlation coefficient, which is a statistical method used to assess the strength and direction of association between two variables. It measures the monotonic relationship between variables, which means it determines whether the variables tend to

Table 5: Mann Whitney U-test p-value and effect size on the change of meteorological concentration.

| | p-value | | | effect size | | |
|---|---|---|---|---|---|---|
| | Pre&During | During&Post | Pre&Post | Pre&During | During&Post | Pre&Post |
| **Temperature** | 0.924 | 0.903 | 0.759 | 0.006 | -0.054 | 0.048 |
| **Wind Speed** | 0.071 | 0.327 | 0.351 | 0.141 | -0.016 | 0.134 |
| **Solar Radiation** | 0.000 | 0.028 | 0.056 | 0.383 | -0.126 | 0.222 |
| **Humidity** | 0.000 | 0.000 | 0.000 | -0.791 | 0.303 | -0.463 |

Table 6: Mann Whitney U-test p-value and effect size on the traffic flow change.

| | p-value | | | effect size | | |
|---|---|---|---|---|---|---|
| **Area** | Pre&During | During&Post | Pre&Post | Pre&During | During&Post | Pre&Post |
| **A1** | 0.000 | 0.000 | 0.004 | 0.734 | -1.072 | 0.135 |
| **A2** | 0.000 | 0.000 | 0.005 | 0.969 | -0.883 | 0.142 |
| **A3** | 0.000 | 0.000 | 0.009 | 1.057 | -0.83 | 0.243 |
| **Avg** | 0.000 | 0.000 | 0.006 | 0.778 | -0.671 | 0.221 |

increase or decrease together, but not necessarily at a constant rate. Spearman's rank correlation coefficient ($\rho$) is calculated using the following formula:

$$\rho = 1 - \frac{6 \sum {d_{\mathrm{i}}}^2}{n(n^2 - 1)} \tag{4}$$

Where $\rho$ represents Spearman's rank correlation coefficient, $d_{\mathrm{i}}$ denotes the difference between the ranks of corresponding variables, and $n$ is the number of data points.

We conducted a correlation analysis to assess the interplay between meteorological features, traffic flow, and pollutant concentration levels. This involved three distinct correlation analyses: time-domain, area-domain, and cross time-and-area domain. These analyses aimed to deepen our understanding of how each measured feature impacts air quality across different domains.

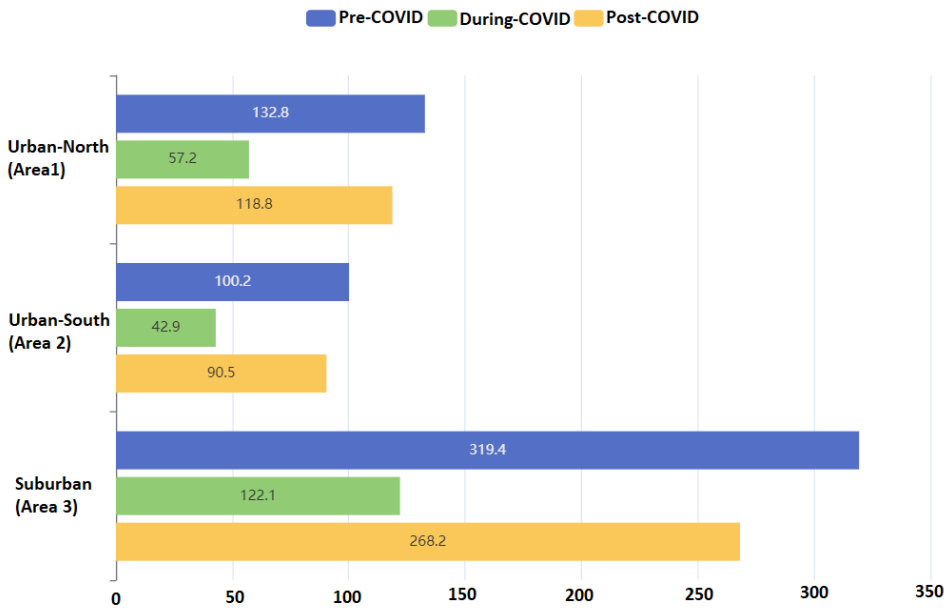


Figure 9: Descriptive statistics from traffic stations.

*6.4.1. From time domain*

The presented Table 7 showcases correlation coefficients between humidity, solar radiation, temperature, wind speed, and traffic flow, concerning NO, $NO_x$, $PM_{2.5}$, and $PM_{10}$ within distinct temporal phases: pre-COVID (2019), during COVID (2020), and post-COVID (2022). A significant discovery emerges from the data indicating a robust negative correlation between wind speed and all air pollutants across the examined years, signifying a considerable dispersion of pollutants facilitated by higher wind speeds. This negative relationship underscores the role of wind in reducing pollutant concentrations. The same negative correlation is observed in humidity as well. However, for humidity, a greater negative correlation exists for metrics $PM_{2.5}$ and $PM_{10}$ specially in pre and during the COVID era.

Another noteworthy finding is the consistent and statistically significant negative correlation observed between traffic flow and NO as well as NOx throughout the three-year timeframe. However, the correlation coefficients between traffic flow and $PM_{2.5}$, as well as $PM_{10}$, were notably low and lacked statistical significance. Conversely, temperature emerged as a dominant fac-

Table 7: Correlation coefficient between meteorological features and traffic flow with air pollutant concentration in time domain.

| | **NO** | | | **NOx** | | |
|---|---|---|---|---|---|---|
| | **Pre** | **During** | **Post** | **Pre** | **During** | **Post** |
| **Temperature** | -0.07 | 0.11 | 0.04 | -0.07 | 0.08 | 0 |
| **Wind Speed** | -0.29 | -0.22 | -0.24 | -0.38 | -0.3 | -0.28 |
| **Solar Radiation** | -0.08 | 0.12 | 0.02 | -0.14 | 0.02 | -0.13 |
| **Humidity** | -0.05 | -0.23 | -0.02 | -0.12 | -0.26 | 0.01 |
| **Traffic Flow** | -0.12 | -0.27 | -0.24 | -0.26 | -0.18 | -0.33 |
| | **$PM_{2.5}$** | | | **$PM_{10}$** | | |
| | **Pre** | **During** | **Post** | **Pre** | **During** | **Post** |
| **Temperature** | 0.41 | 0.3 | 0.36 | 0.39 | 0.55 | 0.37 |
| **Wind Speed** | -0.26 | -0.22 | -0.24 | -0.1 | -0.09 | -0.16 |
| **Solar Radiation** | 0.1 | 0.1 | -0.05 | 0.14 | 0.31 | -0.08 |
| **Humidity** | -0.36 | -0.21 | -0.15 | -0.43 | -0.47 | -0.2 |
| **Traffic Flow** | 0.05 | -0.03 | 0.03 | 0.11 | 0.09 | -0.08 |

tor strongly correlating with $PM_{2.5}$ and $PM_{10}$, exhibiting a substantial average coefficient of 0.4 with a p-value less than 0.01.

This detailed analysis highlights the influential role of wind speed and humidity in dispersing pollutants and the varying degrees of correlation between traffic flow, temperature, and different air pollutant levels across the studied time periods, providing valuable insights into the complex dynamics affecting air quality under different environmental conditions.

*6.4.2. From area domain*

In the area-domain correlation analysis as presented in Table 8, traffic flow exhibits notably high correlation coefficients with NO, $NO_x$, $PM_{2.5}$, and $PM_{10}$, particularly evident in the Urban-North traffic area, displaying statistical significance with p-values less than 0.01. However, contrasting results arise in the Urban-South and Suburban regions, where traffic flow shows statistical correlation solely with NO and $NO_x$, lacking significant associations with $PM_{2.5}$ and $PM_{10}$.

Analyzing meteorological features reveals distinct patterns. Solar radiation, for instance, exhibits correlation exclusively with the Suburban region across

Table 8: Correlation coefficient between meteorological features and traffic flow with air pollutant concentration in area domain.

| | **NO** | | | **NOx** | | |
|---|---|---|---|---|---|---|
| | Urban-North | Urban-South | Suburban | Urban-North | Urban-South | Suburban |
| **Temperature** | -0.04 | -0.01 | 0.1 | 0 | -0.03 | 0.04 |
| **Wind Speed** | 0.03 | -0.32 | -0.47 | -0.1 | -0.38 | -0.49 |
| **Solar Radiation** | 0.05 | 0.02 | 0.11 | 0.05 | -0.09 | -0.09 |
| **Humidity** | -0.13 | -0.22 | -0.18 | -0.23 | -0.23 | -0.18 |
| **Traffic Flow** | 0.59 | 0.46 | 0.27 | 0.74 | 0.48 | 0.34 |
| | **$PM_{2.5}$** | | | **$PM_{10}$** | | |
| | Urban-North | Urban-South | Suburban | Urban-North | Urban-South | Suburban |
| **Temperature** | 0.34 | 0.22 | 0.51 | 0.36 | 0.42 | 0.49 |
| **Wind Speed** | -0.22 | -0.21 | -0.25 | -0.09 | -0.06 | -0.17 |
| **Solar Radiation** | 0.06 | -0.07 | 0.21 | 0.1 | 0.13 | 0.18 |
| **Humidity** | -0.26 | -0.15 | -0.38 | -0.34 | -0.36 | -0.37 |
| **Traffic Flow** | 0.17 | 0.04 | -0.04 | 0.33 | 0.19 | -0.08 |

all air pollutants, displaying a relatively higher coefficient, especially notable for $PM_{2.5}$ and $PM_{10}$ at approximately 0.20. On the other hand, temperature demonstrates statistically significant positive associations solely with $PM_{2.5}$ and $PM_{10}$ across the three locations, implying a trend where higher temperatures coincide with increased concentrations of these pollutants.

The influence of wind speed on air pollutants varies concerning pollutant type and location. However, a consistent trend emerges in the suburban area, showcasing a robust and negative correlation between wind speed and all air pollutants. This pattern is explained by the suburban environment characterized by fewer obstructive buildings and construction sites, allowing unimpeded wind flow. This unobstructed airflow significantly aids in dispersing and diluting air pollutant concentrations.

Regarding humidity, the results indicate a negative correlation between humidity and all pollutants. Specifically, for $PM_{10}$, the negative correlation exceeds 0.3 in all three areas.

These observations highlight the nuanced relationships between traffic flow, meteorological features, and air pollutant concentrations across distinct geo-

Table 9: Correlation coefficient between meteorological features and traffic flow with air pollutant concentration in cross-domain (time and area).

| **A1:Urban-North** | **NO** | | | **NOx** | | |
|---|---|---|---|---|---|---|
| | Pre | During | Post | Pre | During | Post |
| **Temperature** | -0.29 | 0.1 | 0.03 | -0.25 | 0.14 | 0.11 |
| **Wind Speed** | 0.01 | 0.1 | -0.08 | -0.24 | -0.09 | -0.25 |
| **Solar Radiation** | -0.21 | 0.13 | 0.05 | -0.24 | 0.15 | 0.06 |
| **Humidity** | 0.25 | -0.26 | 0.05 | 0.09 | -0.38 | -0.03 |
| **Traffic Flow** | 0.45 | 0.75 | 0.37 | 0.37 | 0.77 | 0.44 |
| | **$PM_{2.5}$** | | | **$PM_{10}$** | | |
| | Pre | During | Post | Pre | During | Post |
| **Temperature** | 0.24 | 0.45 | 0.35 | 0.42 | 0.65 | 0.22 |
| **Wind Speed** | -0.31 | -0.27 | -0.15 | -0.08 | -0.11 | -0.17 |
| **Solar Radiation** | -0.07 | 0.24 | -0.07 | 0.14 | 0.43 | -0.24 |
| **Humidity** | -0.24 | -0.40 | -0.09 | -0.41 | -0.57 | 0.01 |
| **Traffic Flow** | -0.05 | 0.22 | 0.04 | 0.09 | 0.35 | 0.08 |
| **A2:Urban-South** | **NO** | | | **NOx** | | |
| | Pre | During | Post | Pre | During | Post |
| **Temperature** | -0.08 | 0.09 | 0.02 | -0.07 | 0.03 | -0.02 |
| **Wind Speed** | -0.43 | -0.36 | -0.38 | -0.49 | -0.46 | -0.49 |
| **Solar Radiation** | -0.11 | 0.1 | -0.06 | -0.19 | -0.03 | -0.25 |
| **Humidity** | -0.17 | -0.23 | 0.02 | -0.25 | -0.22 | 0.12 |
| **Traffic Flow** | 0.25 | 0.3 | 0.44 | 0.13 | 0.32 | 0.38 |
| | **$PM_{2.5}$** | | | **$PM_{10}$** | | |
| | Pre | During | Post | Pre | During | Post |
| **Temperature** | 0.33 | -0.01 | 0.33 | 0.42 | 0.4 | 0.47 |
| **Wind Speed** | -0.23 | -0.14 | -0.33 | -0.04 | -0.04 | -0.17 |
| **Solar Radiation** | 0 | -0.19 | -0.1 | 0.17 | 0.14 | 0.01 |
| **Humidity** | -0.33 | 0.07 | -0.13 | -0.48 | -0.32 | -0.27 |
| **Traffic Flow** | -0.11 | -0.14 | 0.05 | 0.06 | 0.23 | 0.04 |
| **A3:Suburban** | **NO** | | | **NOx** | | |
| | Pre | During | Post | Pre | During | Post |
| **Temperature** | 0.01 | 0.21 | 0.13 | 0.05 | 0.13 | -0.05 |
| **Wind Speed** | -0.58 | -0.44 | -0.47 | -0.69 | -0.53 | -0.46 |
| **Solar Radiation** | -0.06 | 0.21 | 0.12 | -0.1 | 0.01 | -0.35 |
| **Humidity** | -0.05 | -0.25 | -0.04 | -0.18 | -0.18 | 0.19 |
| **Traffic Flow** | 0.24 | 0.19 | 0.31 | 0.02 | 0.2 | 0.3 |
| | **$PM_{2.5}$** | | | **$PM_{10}$** | | |
| | Pre | During | Post | Pre | During | Post |
| **Temperature** | 0.63 | 0.52 | 0.4 | 0.44 | 0.63 | 0.44 |
| **Wind Speed** | -0.24 | -0.25 | -0.28 | -0.23 | -0.14 | -0.16 |
| **Solar Radiation** | 0.34 | 0.27 | 0 | 0.18 | 0.38 | 0 |
| **Humidity** | -0.53 | -0.43 | -0.22 | -0.47 | -0.57 | -0.28 |
| **Traffic Flow** | -0.11 | -0.31 | 0.07 | -0.08 | -0.32 | 0.06 |

graphical locations, emphasizing the impact of local environmental conditions on the dispersion and concentration levels of pollutants in specific regions.

*6.4.3. From time and area domain*

In Table 9, encompassing both time and area domains, distinct correlations between traffic flow and air pollutant concentrations are evident, particularly during the COVID period in 2020. Statistically significant positive correlations are observed solely between traffic flow and NO as well as $NO_x$ across all station types. Notably, the Urban-North traffic station displays the highest correlation, followed by the Urban-South and Suburban stations, with coefficients of 0.75, 0.3, and 0.19 for NO, exemplifying this trend. Another noteworthy metric during this period is humidity, which exhibits the highest negative correlation compared to all meteorological metrics, particularly concerning $PM_{2.5}$ and $PM_{10}$ in Urban-north and Suburban areas.

Across non-COVID years (2019 and 2022), a similar sequence in the degree of correlation persists, albeit with generally smaller coefficients. This trend underscores that restricting vehicular movement has a more pronounced impact on reducing NO and $NO_x$ concentrations in the specified order of Urban-North, Urban-South, and Suburban areas.

Contrarily, for $PM_{2.5}$ and $PM_{10}$, non-COVID years exhibit an absence of significant correlation between traffic flow and pollution concentrations across all stations. However, in 2020 during the city lockdown, a notable shift occurs: $PM_{2.5}$ and $PM_{10}$ demonstrate statistical correlations with traffic flow, exclusively observed at the Urban-North area, with coefficients of 0.22 and 0.35 for $PM_{2.5}$ and $PM_{10}$, respectively. In the Urban-South, $PM_{10}$ displays statistical correlation with traffic flow, but $PM_{2.5}$ does not exhibit such a relationship.

One significant factor driving this behavior could be particle size. $PM_{10}$ particles, being larger than $PM_{2.5}$ particles, settle more rapidly from the atmosphere. When emitted from vehicles, $PM_{10}$ particles tend to concentrate around areas with dense traffic. Additionally, local conditions play a role. $PM_{10}$, being larger and heavier, is less affected by atmospheric conditions and tends to stay closer to its sources, including traffic. In contrast, $PM_{2.5}$ particles can disperse over longer distances and may not exhibit as strong a correlation as $PM_{10}$. The results suggest that local environmental conditions, such as wind speed and humidity patterns, influence the dispersion and trans-

port of pollutants. For instance, humidity demonstrated a strong negative correlation with $PM_{10}$ particles in Urban-South and Suburban areas.

An anomalous phenomenon arises in the Suburban area: both $PM_{2.5}$ and $PM_{10}$ exhibit statistically negative correlations with traffic flow during the reduced vehicular period, suggesting that fewer vehicles result in higher air pollutant concentrations. Overall, the implementation of vehicle reduction measures results in decreased $PM_{2.5}$ and $PM_{10}$ concentrations solely in the Urban-North area, with $PM_{10}$ affected in the Urban-South area, while no improvements, and potentially worsening conditions, are noted in the Suburban area.

### *6.5. DET linking implementation details*

Statistical analysis and correlation functions were executed using Python libraries, constituting integral components developed within the SALTED project and housed in the project's GitHub repository[7]. Although they are currently tailored for specific discussed use cases, assessing correlation through statistical analysis and measuring correlation values is a common necessity across various applications integrating diverse datasets. The DET platform can potentially provide a suite of general-purpose functionalities for immediate application. Moreover, programmers using the SALTED platform will have the flexibility to customize these functionalities to suit specific platform requirements.

### *6.6. Discussion on results bias and limitation*

The study has considered traffic and pollution data before, during, and after the pandemic events in Madrid. An "Event bias" may have had an influence on the results. The pandemic has been an exception in the lives of people. Hence, the resulting behavior could be considered "special" and biased by the peculiarity of the situation and different regulations and laws imposed to contrast the diffusion of COVID. The data during the pandemic's peak may be characterized by people's impulsive and immediate responses and urgencies of people to the events. However, the important aspect of the study, i.e., a very reduced vehicle circulation during the considered period of the year is evidence. To avoid a Time interval bias, the study has considered comparable intervals of time from available data sets. Considering other

[7]https://github.com/SALTED-Project

years (after COVID) will help consolidate the results minimizing the interval time bias.

The City of Madrid has planned and deployed traffic sensors and pollution base stations according to their goals and purposes. These do not necessarily coincide with the needs in terms of the location of sensors to the best placement for the current study. However, to avoid or at least limit a "Location-based bias", the areas and the sensors taken into consideration are those that do show a good closeness between the pollution stations and the traffic intensity sensors. In addition, these areas have been chosen to exemplify different aspects of the traffic and pollution in a big city. Still, they may have particular local traffic trends or patterns, however, they were chosen to be general, diversified, and close enough to pollution stations to be considered as a core set of data that could be further elaborated and consolidated. We believe they are representative of the City of Madrid. However, a larger study may consolidate the results by considering more sensors in additional areas.

Another possible bent may be related to a "vehicle profiles" bias circulating in major cities. In recent years, national and European policies have supported a shift from thermal engines to electrical or hybrid ones. This can be seen as a susceptibility bias: the data could be influenced by a change in the type of vehicles with a different pollution impact. Even if the shift towards more effective vehicles progresses, its effects (as well as the normalization of usage of vehicles within the city) will take a while to substantially affect the pollution patterns in a large city (for instance, [8] the average age of the Spanish vehicle fleet is 13.5 years). Also in this case adding additional data sets for a longer interval (when available) will provide more consolidated data.

Weather conditions of the considered range of years could influence the results, a "Weather bias". Two considerations hold here: a trend towards an increase of temperatures (that may have effects on the pollution) and different specific climate events in the period considered (from March to May of each year). Also for this case, a longer period of study can reconduct the analysis to the "average" situation. However, the relevant climatic measure-

[8] according to https://www.acea.auto/figure/average-age-of-eu-vehicle-fleet-by-country/

ments in the analyzed periods are compatible and consistent. The results are related to a highly monitored city like Madrid and they can be influenced by how citizens "use" the city, a Behavioral bias. Some specific behaviors Madrid people put in place in daily life (e.g., peak traffic hours) can differ from other cities in other European Countries or the world. However, the general relationship between traffic and pollution still holds despite different citizens' habits and attitudes. Enlarging the study to other Spanish or international cities will consolidate this research.

The study has also considered some limitations. The study started with Madrid, but it will be extended to consider different types of cities (smaller, such as Santander in Spain, or similar in size, like Barcelona) or in different countries (like Dublin in Ireland). The attempt is to generalize and compare the results under several conditions and environments. Not all the available pollutants and traffic data have been considered, it is envisaged to operate on a double line of research: to progressively consider in the study all the available pollutants for Madrid, to focus on those stronger related to thermal engines for analysis of cities with a smaller range of data looking for similarities and differences. In line with this, the study could progressively introduce other weather conditions that could affect pollution. The current research has focused on the most common parameters for vehicle-generated pollutants and weather features. Enlarging the range of analyzed values can consolidate the work. However, the "transfer" and comparison with other cities could be more difficult due to the richness of the data set of Madrid compared to other cities. All these aspects are considered for further developments and improvements of the study in the near future.

## 7. Conclusion and Future Work

The escalating concerns regarding the impact of air pollution on public health have spurred the design and implementation of traffic reduction policies aimed at mitigating air pollutant concentrations. This study investigates the reduction in organic traffic resulting from COVID lockdowns as a parameter to assess changes in air pollutant levels. Through a comprehensive three-tiered statistical analysis across both place and time domains, notable reductions in NO and $NO_x$ are observed in descending order: most prominently at Urban-North area, followed by Urban-South locations, and finally, the suburban area. Conversely, statistically significant reductions in $PM_{10}$

concentrations are solely evident at Urban-North and Urban-South sites, while $PM_{2.5}$ demonstrates a significant decrease exclusively at Urban-North area.

The findings align with previous studies, affirming that vehicular flows notably impact NO and $NO_x$ levels, thus reinforcing this initial conclusion. Furthermore, our analysis provides a more nuanced insight by elucidating that $PM_{2.5}$ and $PM_{10}$ experience substantial reductions specifically in the Urban-North area. These observations suggest that policy interventions aimed at reducing traffic flows can effectively enhance air quality, particularly for NO and $NO_x$, in the Urban-North regions. However, addressing $PM_{2.5}$ and $PM_{10}$ concentrations in the Urban-South and suburban areas demands alternative strategies from the Madrid municipal administration to supplement traffic reduction policies for comprehensive air quality improvements.

In this sense, considering broader policy implications, the results obtained from the analysis carried out, and, even more importantly, the data linking and enrichment that the DET enables, highlight the benefits of integrating IoT technologies and advanced data analytics into decision-making processes in general, and urban planning in particular. The harmonization of reusable data and its integration within data processing pipelines should lead to the creation of accurate models that could feed precise Smart City Digital Twins where alternatives for urban planning and policy-making, especially in the context of sustainable mobility, can be evaluated and assessed.

Future work will be twofold. On the one hand, as has been previously indicated, similar analyses will be carried out using the data from Barcelona, Oslo and Santander. This way, not only the potentiality of the DET platform for enabling interoperable data fusion over harmonized data sources will be further validated by demonstrating multi-site correlation, but also, such multi-site results' comparison and cross-site discussion should allow for the extraction of further conclusions that are not evident on the single-site case.

On the other hand, the focus will be put on refining the automated evaluation of data correlation by incorporating more complex statistical measures that can handle diverse types of data relationships. Expanding beyond the traditional correlation coefficients, exploring nonlinear correlations and dynamic relationships within datasets could be a promising avenue. Additionally, delving deeper into the integration of machine learning models with interpretability features could enhance the applicability of advanced predictive

analysis. Research could further investigate techniques to make these sophisticated models more transparent and explainable, ensuring that correlations and predictions derived from these models are not only accurate but also comprehensible to domain experts and stakeholders. Furthermore, future studies might concentrate on devising adaptable frameworks that efficiently identify the most suitable predictive models for specific datasets, streamlining the process of fitting data into these advanced analytical functions. Such frameworks could facilitate automated model selection and parameter tuning, optimizing the utilization of available data and enhancing the overall predictive performance.

## 8. Acknowledgments

This work has been partially supported by the project SALTED (Situation-Aware Linked heterogeneous Enriched Data) from the European Union's Connecting Europe Facility program under the Action Number 2020-EU-IA-0274, and by means of the project THROTTLE (TrustwortHy uRban mObility daTa markeTpLacE) under Grant Agreement No. TED2021-131988B-I00 funded by MCIN/AEI/10.13039/501100011033 and the European Union Next GenerationEU/PRTR.

## References

[1] World Health Organization & Others Ambient air pollution: A global assessment of exposure and burden of disease. (World Health Organization,2016)

[2] World Health Organization World health statistics 2016: monitoring health for the SDGs sustainable development goals. (World Health Organization,2016)

[3] Álvarez, C., Galera, S., Campos-Celador, Á., Díaz, J., Linares, C., Barqueros, I. & Casadevante, J. INFORME SOBRE SOSTENIBILIDAD EN ESPAÑA 2019 - Por qué las ciudades son clave en la transición ecológica. (Fundación Alternativas,2019)

[4] Ceballos, M., Segura, P., Gutiérrez, E., Gracia, J., Ramos, P., Reaño, M. & Garcıéa, B. La calidad del aire en el Estado español durante 2018. *Ecologistas En Acción: Madrid, Spain.* (2018)

[5] Silva, C., Saldiva, P., Amato-Lourenço, L., Rodrigues-Silva, F. & Miraglia, S. Evaluation of the air quality benefits of the subway system in São Paulo, Brazil. *Journal Of Environmental Management*. **101** pp. 191-196 (2012)

[6] Rahman, A., Luo, C., Khan, M., Ke, J., Thilakanayaka, V. & Kumar, S. Influence of atmospheric PM2. 5, PM10, O3, CO, NO2, SO2, and meteorological factors on the concentration of airborne pollen in Guangzhou, China. *Atmospheric Environment*. **212** pp. 290-304 (2019)

[7] Rojas-Rueda, D., Nazelle, A., Teixidó, O. & Nieuwenhuijsen, M. Replacing car trips by increasing bike and public transport in the greater Barcelona metropolitan area: a health impact assessment study. *Environment International*. **49** pp. 100-109 (2012)

[8] Sánchez, L., Lanza, J., Santana, J., Sotres, P., González, V., Martín, L., Solmaz, G., Kovacs, E., Dietzel, M., Summa, A., Jafari, A., Minerva, R. & Crespi, N. Data Enrichment Toolchain: A Data Linking and Enrichment Platform for Heterogeneous Data. *IEEE Access*. **11** pp. 103079-103091 (2023)

[9] Borgogno, O. & Colangelo, Data sharing and interoperability: Fostering innovation and competition through APIs. *Computer Law & Security Review*. **35**, 105314 (2019)

[10] Bröring, A., Schmid, S., Schindhelm, C., Khelil, A., Käbisch, S., Kramer, D., Le Phuoc, D., Mitic, J., Anicic, D. & Teniente, E. Enabling IoT ecosystems through platform interoperability. *IEEE Software*. **34**, 54-61 (2017)

[11] Nilsson, J. & Sandin, F. Semantic interoperability in industry 4.0: Survey of recent developments and outlook. *2018 IEEE 16th International Conference On Industrial Informatics (INDIN)*. pp. 127-132 (2018)

[12] Mazayev, A., Martins, J. & Correia, N. Interoperability in IoT Through the Semantic Profiling of Objects. *IEEE Access*. **6** pp. 19379-19385 (2018)

[13] Laña, I., Del Ser, J., Padró, A., Vélez, M. & Casanova-Mateo, C. The role of local urban traffic and meteorological conditions in air pollution:

A data-based case study in Madrid, Spain. *Atmospheric Environment*. **145** pp. 424-438 (2016)

[14] Rossi, R., Ceccato, R. & Gastaldi, M. Effect of road traffic on air pollution. Experimental evidence from COVID-19 lockdown. *Sustainability*. **12**, 8984 (2020)

[15] Salas, R., Perez-Villadoniga, M., Prieto-Rodriguez, J. & Russo, A. Were traffic restrictions in Madrid effective at reducing NO2 levels?. *Transportation Research Part D: Transport And Environment*. **91** pp. 102689 (2021)

[16] Hwang, H. & Lee, J. Impacts of COVID-19 on air quality through traffic reduction. *International Journal Of Environmental Research And Public Health*. **19**, 1718 (2022)

[17] Chen, Z., Hao, X., Zhang, X. & Chen, F. Have traffic restrictions improved air quality? A shock from COVID-19. *Journal Of Cleaner Production*. **279** pp. 123622 (2021)

[18] Brown, L., Barnes, J. & Hayes, E. Traffic-related air pollution reduction at UK schools during the Covid-19 lockdown. *Science Of The Total Environment*. **780** pp. 146651 (2021)

[19] Casari, M., Po, L. MitH: A framework for Mitigating Hygroscopicity in low-cost PM sensors. *Environmental Modelling and Software*. **173** pp. 105955 (2024)

[20] Iskandaryan, D., Ramos, F. & Trilles, S. Application of deep learning and machine learning in air quality modeling. *Current Trends And Advances In Computer-Aided Intelligent Environmental Data Engineering*. pp. 11-23 (2022)

[21] Iskandaryan, D., Ramos, F. & Trilles, S. Graph neural network for air quality prediction: A case study in madrid. *IEEE Access*. **11** pp. 2729-2742 (2023)

[22] Iskandaryan, D., Ramos, F. & Trilles, S. Spatiotemporal prediction of nitrogen dioxide based on graph neural networks. *Environmental Informatics*. pp. 111-128 (2022)

[23] Iskandaryan, D., Ramos, F. & Trilles, S. Bidirectional convolutional LSTM for the prediction of nitrogen dioxide in the city of Madrid. *PloS One*. **17**, e0269295 (2022)

[24] Iskandaryan, D., Ramos, F. & Trilles, S. Comparison of nitrogen dioxide predictions during a pandemic and non-pandemic scenario in the city of Madrid using a convolutional LSTM network. *International Journal Of Computational Intelligence And Applications*. **21**, 2250014 (2022)

[25] Awan, F., Minerva, R. & Crespi, N. Improving road traffic forecasting using air pollution and atmospheric data: Experiments based on LSTM recurrent neural networks. *Sensors*. **20**, 3749 (2020)

[26] FIWARE Foundation. Smart Data Models. Available online: https://smartdatamodels.org/ (accessed on 2024-04-02).

[27] Universidad de Cantabria. DET on GitHub. Available online: https://github.com/tlmat-unican/salted-det-uc (accessed on 2024-04-02)

[28] Situation-Aware Linked heterogeneous Enriched Data (SALTED), "D2.2: Report on data modelling and linking, Project Deliverable, 2023.

[29] SALTED Project. Enhanced datasets (European Data Portal). Available online: https://data.europa.eu/data/catalogues/salted?locale=en (accessed on 2024-04-02)

[30] Sánchez, J., Ortega, E., Lopez-Lambas, M. & Martıén, B. Evaluation of emissions in traffic reduction and pedestrianization scenarios in Madrid. *Transportation Research Part D: Transport And Environment*. **100** pp. 103064 (2021)

[31] Pozo, R., Wilby, M., Dıéaz, J. & González, A. Data-driven analysis of the impact of COVID-19 on Madrid's public transport during each phase of the pandemic. *Cities*. **127** pp. 103723 (2022)