NOVA IMS Information Management School

# MGI

Master Degree Program in

**Information Management**

# Business Intelligence Framework for Air Quality Management

## A Case Study in the Municipality of Lisbon

Carlota Valadão Pimenta

Project Work

presented as partial requirement for obtaining a Master's Degree in Information Management

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# A Business Intelligence Framework for Air Quality Management

A Case Study in the Municipality of Lisbon

by

Carlota Valadão Pimenta

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence.

## Supervised by

Bruno Jardim, PhD, NOVA Information Management School

July, 2025

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

*Lisbon, July 2025*

*Carlota Valdão Pimenta*

# ACKNOWLEDGEMENTS

I want to thank Bruno Jardim for his guidance and valuable feedback, as well as all my fellow Master's colleagues and teachers.

To my parents, for their unconditional support and for always encouraging me to aim higher. To Leonor, for helping me overcome the difficult moments of this journey. And to my family and friends, thank you for the support, kindness and friendship throughout this entire process.

# ABSTRACT

Air quality is a constant public health issue in urban areas, where traffic density, meteorological conditions, and population growth contribute to high concentration levels. In this research we propose to develop a Business Intelligence solution to monitor and analyze air quality sensor data in the Municipality of Lisbon by integrating environmental and traffic congestion data. The solution followed the Kimball lifecycle methodology was implemented through Microsoft Fabric tools, incorporating a full architecture. We propose a tool with both current air quality data and historical data analysis, to discover hidden patterns in the time of the day, day of the week and months, and identify hotspot areas in Lisbon. Additionally, the incorporation of Key Performance Indicators allows to develop an alert system to inform authorities of health risk implications. A forecasting component using both SARIMA and Prophet models supports short-term monitoring of key pollutants. The results reveal that $PM_{10}$, $PM_{2.5}$ and $NO_2$ are the most critical pollutants in Lisbon, with frequent high concentration values in different locations, as well as Prophet outperforming SARIMA across most pollutants. This study contributes to a fully integrated solution by combining Business Intelligence tools and forecasting techniques in an interactive tool, allowing proactive decision-making.



# KEYWORDS

Air Quality; Business Intelligence; Kimball Lifecycle; Data Visualization; Key Performance Indicators; Forecasting

# TABLE OF CONTENTS


Statement of Integrity ..... 1

Acknowledgements ..... 2

Abstract ..... 3

List of Figures ..... 6

List of Tables ..... 7

List of Abbreviations and Acronyms ..... 9

1. Introduction ..... 10

2. Literature review ..... 12

2.1. Air Quality Background ..... 12

2.1.1. Air Quality Monitoring ..... 13

2.1.2. Air Quality Prediction ..... 14

2.1.3. Indexes and Indicators for Air Quality ..... 16

2.1.4. Research Limitations ..... 20

3. Methodology ..... 21

3.1. Business Understanding & Requirements ..... 22

3.2. Data Collection ..... 23

3.2.1. Study Area ..... 23

3.2.2. Air Quality Data ..... 24

3.2.3. Complementary Data ..... 25

3.3. Technical Architecture Design ..... 28

3.4. Dimensional Modelling ..... 29

3.4.1. Business Process ..... 29

3.4.2. Granularity ..... 29

3.4.3. Dimension Tables ..... 30

3.4.4. Fact Tables ..... 30

3.4.5. Schema Approach ..... 31

3.5. Data Transformation (ETL process) ..... 32

3.5.1. Extract ..... 32

3.5.2. Transform ..... 33

3.5.2.1. Air Quality dataset ..... 33

3.5.2.2. Meteorological dataset ..... 34

3.5.2.3. Alerts Type dataset ..... 35

3.5.2.4. Date and Time Dimensions ..... 36
3.5.3. Load ..... 37
3.6. Forecasting ..... 40
3.6.1. ARIMA ..... 41
3.6.2. Prophet ..... 41
3.6.3. Evaluation Metrics ..... 42
3.6.4. Pipeline integration ..... 42
3.7. Semantic Model ..... 44
3.7.1. Relationships ..... 45
3.7.2. Hierarchies ..... 46
3.7.3. Key Performance Indicators and Calculated Measures ..... 47
4. Results and Discussion ..... 49
4.1. Air Quality Current data ..... 49
4.1. Air Quality Overview ..... 51
4.2. Hotspots and High Risk Areas ..... 52
4.3. Anomalies ..... 54
4.4. Meteorological conditions Impact ..... 55
4.5. Traffic Congestion Impact ..... 56
4.6. Forecasting ..... 58
4.6.1. SARIMA model results ..... 58
4.6.2. Prophet model results ..... 59
4.6.3. Comparative Analysis ..... 60
4.7. Discussion ..... 61
5. Conclusions and Future Research ..... 62
Bibliographical References ..... 64
Appendix A ..... 70
Appendix B ..... 72

# LIST OF FIGURES

Figure 3.1 - Business dimensional lifecycle diagram, retrieved from Kimball & Ross (2013).. 21
Figure 3.2 - Spatial distribution of stations in Lisbon by parish. ..... 23
Figure 3.3 - Dimensional Model, Star Schema approach. ..... 31
Figure 3.4 - Data flow during ETL to achieve DW. ..... 32
Figure 3.5 - Load Stating Area pipeline. ..... 38
Figure 3.6 - Staging Area Validation pipeline. ..... 39
Figure 3.7 - Load Data Warehouse pipeline. ..... 39
Figure 3.8 - Master pipeline to automate the running process. ..... 40
Figure 3.9 - Machine learning models' development flow chart in MF. ..... 41
Figure 3.10 - Load Data Warehouse pipeline with the addition of the forecast notebook and tables. ..... 43
Figure 3.11 – Forecast tables validation pipeline. ..... 44
Figure 3.12 – Semantic model design. ..... 45
Figure 3.13 – Date dimension hierarchy definition. ..... 46
Figure 3.14 – Time dimension hierarchy definition. ..... 46
Figure 3.15 – Location dimension hierarchy definition. ..... 46
Figure 3.16 – Alert Type dimension hierarchy definition. ..... 46
Figure 4.1 – Dashboard page 1 'Current Air Quality'. ..... 49
Figure 4.2 - Dashboard page 2 'Air Quality Overview'. ..... 51
Figure 4.3 - Dashboard page 3 'Hotspots'. ..... 53
Figure 4.4 - Dashboard page 4 'Air Quality Anomalies'. ..... 54
Figure 4.5 - Dashboard page 5 'Air Quality vs Weather'. ..... 56
Figure 4.6 - Dashboard page 6 'Air Quality vs Alert Type'. ..... 57
Figure 4.7 - Real vs Forecast values with SARIMA Model for $NO_2$. ..... 59
Figure 4.8 – Real vs Forecast values with Prophet Model for $NO_2$. ..... 60

# LIST OF TABLES

Table 2.1 - Summary of research on Air Quality. ..... 14
Table 2.2 - Air Quality Index and Indicators. ..... 16
Table 2.3 - WHO recommendations for Air Quality Guideline levels, in μg/m$^3$, adapted from World Health Organization (2021). ..... 17
Table 2.4 - Atmospheric Pollution Index classification values, adapted from Seibert et al. (2022). ..... 18
Table 2.5 - Air Quality Influencing Factors. ..... 19
Table 3.1 – Business Needs and Questions definition. ..... 22
Table 3.2 - Name and description of air quality data attributes. ..... 24
Table 3.3 - Statistics for air quality data discriminated by pollutant. ..... 25
Table 3.4 - Category, features and source for complementary data. ..... 25
Table 3.5 - Name and description of weather data attributes. ..... 26
Table 3.6 - Name and description of traffic alert type data attributes ..... 27
Table 3.7 - Applied Technologies and their Roles. ..... 29
Table 3.8 - Granularity definition per fact. ..... 30
Table 3.9 - Name and description of dimension tables in the BI model. ..... 30
Table 3.10 - Name and description of fact tables in the BI model. ..... 31
Table 3.11 - Data transformations for Air Quality dataset. ..... 33
Table 3.12 - Feature Engineering Air Quality dataset. ..... 34
Table 3.13 - Data transformations for Meteorological dataset. ..... 35
Table 3.14 - Data transformations for Traffic Alert dataset. ..... 35
Table 3.15 - Holidays and special days definition for Date dimension. ..... 36
Table 3.16 - TIME_FRAME field definition for Time dimension. ..... 37
Table 3.17 - Description of the rules applied for data validation. ..... 38
Table 3.18 - *FACT_FORECAST_VALUE* fields and description. ..... 43
Table 3.19 - *DIM_FORECAST* fields and description. ..... 43
Table 3.20 - Sematic model relationships between fact and dimension tables. ..... 45
Table 3.21 - Lisbon pollutant thresholds for Air Quality in μg/m$^3$. Adapted from Câmara Municipal de Lisboa (2021). ..... 47
Table 4.1 - Current Air Quality (BN 1) insights. ..... 50
Table 4.2 - Air Quality Overview (BN 1) insights. ..... 52
Table 4.3 - Hotspots and High Risk Areas (BN 2) insights. ..... 53
Table 4.4 - Anomalies (BN 3) insights. ..... 55
Table 4.5 – Weather impact (BN 4) insights. ..... 56
Table 4.6 – Traffic congestion (jam alerts) impact (BN 5) insights. ..... 57

Table 4.7 – SARIMA model parameters definition for each parameter. ..... 58
Table 4.8 – Prophet models regressors selection per parameter..... 59
Table 4.9 – SARIMA vs Prophet Evaluation metrics. ..... 60

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **API** | Application Programming Interface |
| **AQG** | Air Quality Guidelines |
| **AQI** | Air Quality Index |
| **BI** | Business Intelligence |
| **BN** | Business Need |
| **BQ** | Business Question |
| **CO** | Carbon Monoxide |
| **DW** | Data Warehouse |
| **ETL** | Extract, Load and Transform |
| **KPI** | Key Performance Indicator |
| **MAE** | Mean Absolute Error |
| **MF** | Microsoft Fabric |
| **NO** | Nitrogen Oxide |
| **$NO_2$** | Nitrogen Dioxide |
| **$O_3$** | Ground-level Ozone |
| **PM** | Particulate Matter |
| **$PM_{2.5}$** | Particulate Matter with diameter less than 2.5 micrometers |
| **$PM_{10}$** | Particulate Matter with diameter less than 10 micrometers |
| **RMSE** | Root Mean Absolute Error |
| **$SO_2$** | Sulfur Dioxide |
| **VOC** | Volatile Organic Compounds |
| **WHO** | World Health Organization |
| **REZ** | Reduced Emission Zones |

# 1. INTRODUCTION

Urbanization and the continuous growth of population in major cities like Lisbon have significant impact on air pollutant concentrations, contributing to health issues due to prolonged exposure (Zhan et al., 2023). Thus, urban centres are exposed to elevated emissions from vehicle, industrial activities, and climate conditions, leading to elevated levels of pollutants, such as PM2.5 and O3 (Sicard et al., 2021). Despite the effective results on low emission zones in Lisbon, some areas continue to struggle to meet air quality standards (Ferreira et al., 2015), needing further research and assessment to continue improvement. Addressing this issue is crucial for public health, environmental protection, and sustainability, to help policymakers make informed decisions on urban planning.

While there has been significant progress in raising public awareness of air quality issues, efforts to engage citizens through interactive tools remain underdeveloped (Dashkevych & Portnov, 2023). Moreover, previous studies have often relied on statistical approaches to analyse air pollutants (Athira et al., 2018); however, there is limited research in incorporating machine learning models within a Business Intelligence (BI) framework as well as development of Key Performance Indicators (KPIs) to monitor air quality (Alrashed, 2020). Developing a BI solution with interactive dashboards, powered by machine learning models, could not only enhance predictive accuracy but also improve public awareness by providing citizens with accessible insights into air quality data. Regarding the topic at hand, the research will explore an implementation of a BI pipeline to store and analyse data, the definition of KPIs to monitor air quality and identify patterns, and the integration of machine learning models.

This study follows the Kimball lifecycle methodology, to develop a BI framework using Microsoft Fabric tools. The solution integrates key components such as a centralized Data Warehouse, structured ETL processes (extraction, transformation and load) and a design of a semantic model to define KPIs tailored to air quality monitoring in Lisbon. Furthermore, these components enable a visualization tool to incorporate air quality trend analysis, external factors with impact in air quality, such as meteorological conditions and traffic congestion, and short-term pollutant levels prediction. The developed framework is designed to support informed decision-making and continuous monitoring.

Several key insights were revealed, on the main hotspots in Lisbon, the pollutants that exceed more frequently the air quality health limits, peak hours, days of the week and months analysis, and identify how traffic and weather conditions impact air quality. This solution contributes both methodologically and practically in air quality research. Firstly, it demonstrates how a BI approach can combine structured environmental data with forecasting and spatial analysis in user-oriented interface. Practically, the dashboard enables city planners, environmental agencies, and the population to monitor air quality trends and identify pollution hotspots.

The research is organized into five chapters. The next chapter represents a detailed literature review covering pollutant sources, public health impacts, air quality monitoring approaches and predictive methods. Followed by the methodology description, including business understanding, data sources, processing tools and the dimensional modelling process. The results chapter presents the dashboard and the answers to the business needs defined previously. Finally, the last chapter concludes with a summary of findings, discusses limitations and suggests future research directions to enhance the current solution.

# 2. LITERATURE REVIEW

Urbanization and the continuous growth of population in urban cities has significantly changed economies and societies, but it has led to decreased air quality, affecting not only the environment, but also the populations (Sokhi et al., 2021). This literature review assesses the existing research on air quality impact both on the environment and public health, identifies the main pollutants in the outdoor air composition, explores the advancements on air quality monitoring and identifies existing methodologies to monitor through prediction and visualization methods.

## 2.1. AIR QUALITY BACKGROUND

Recent studies suggest that the main air pollution sources are emissions from vehicle and industrial activities, and climate conditions (Sicard et al., 2021). Thus, urban areas are the most affected by poor air quality due to higher and constant exposure to pollutant concentrations when compared to rural areas (Jena et al., 2023; Wajeetongratana, 2023). As the levels of pollutants such as nitrogen dioxide (NO2), particulate matter (PM2.5 and PM10), and volatile organic compounds (VOCs) increase, the concern for public health also increases, as it is often linked to health issues (Zhan et al., 2023).

The health implications of poor air quality include respiratory and cardiovascular systems malfunction, affecting primarily the most vulnerable population, such as elderly, children and those with previous health conditions (Abelsohn & Stieb, 2011), and in extreme cases, long-term exposure can lead to increased morbidity and mortality rates (Zheng et al., 2021). For instance, according to Lelievel et al. (2015), it was estimated to be the reason for approximately 60% of global premature deaths in 2010. However, during COVID-19 lockdown, these values decreased due to lower human activity, mainly vehicle activities, which lead to a positive correlation between human activity and air quality (Lelievel et al., 2015).

Additionally, the environment is also affected by poor air quality, leading to climate changes overtime, as well as modification in ecosystems with biodiversity loss and material damage Sokhi et al. (2021). Processes such as acidification, eutrophication and ground-level ozone formation are associated with environmental decay.

Outdoor air composition includes pollutants such as particulate matter (PM), nitrogen oxides (NO), sulfur dioxide ($SO_2$), carbon monoxide (CO), and ground-level ozone ($O_3$) (Jena et al., 2023; Lelieveld et al., 2015). These pollutants, with different characteristics and impact on both public health and the environment, are described below.

Among these pollutants, particulate matter ($PM_{2.5}$ and $PM_{10}$) is one of the most critical in outdoor air composition, composed of solid and liquid particulates. Due to its capability to penetrate into the lungs, this pollutant composes high risk health impacts, which influences respiratory and cardiovascular systems (Jena et al., 2023), and can shorten life expectancy

(Koolen & Rothenberg, 2019). Several sources from human activity can lead to particulate matter emissions, such as vehicle and industrial activities, construction dust and residential heating (Todorov et al., 2023). Furthermore, its interaction with other pollutants can form harmful compounds (Zhu et al., 2011), such as pollution complex and grey haze.

Beyond PM, vehicle emissions are also the main contributor to Nitrogen Oxide ($NO_2$) emissions, mainly produced from combustion processes (Wajeetongratana, 2023; Zhou et al., 2024). When in contact with Volatile Organic Compounds (VOCs), it contributes to the formation of ground-level ozone ($O_3$), which causes serious health risks, including respiratory issues, lung malfunction and worsen asthma effects; and is known for increased hospital admissions for respiratory diseases, as well as respiratory and cardiovascular deaths (Manisalidis et al., 2020). Additionally, for an environmental point of view, its presence can lead to soil degradation, decreased water quality, and vegetation damage.

Sulfur Dioxide ($SO_2$), emitted mainly from fossil fuels burning and industrial activities, can form fine particulate matter and acid rains through reactions in the atmosphere (Manisalidis et al., 2020). Humans and animals are affected by its capability of penetrating into the lungs, causing respiratory irritation (Anurogo et al., 2023); in the other hand, acidification of soil and acid rain often affect plants and the ecosystems (Manisalidis et al., 2020).

Like sulfur dioxide, carbon monoxide (CO) is release from combustion processes (Todorov et al., 2023), when combustion of carbon-containing fuels is incomplete (Todorov et al., 2023), through processes such as vehicle and industrial emissions, and household appliances. Exposure to this colourless, odourless gas can interfere with oxygen quantity in the blood, leading to headaches, dizziness, and, in the more critical cases, to cardiovascular disease (Manisalidis et al., 2020), particularly in vulnerable populations such as children and elderly. Additionally, CO can impact global temperature and cause climate change, as it influences greenhouse gases.

Regarding all health and environmental impacts, several organizations have had various initiatives to mitigate outdoor air pollution, in an attempt to improve public health and the environment. For example, the World Health Organization (WHO) updated its Global Air Quality Guidelines in 2021, with stricter recommendations for pollutant's levels (World Health Organization, 2021), so countries can adopt them and improve their air quality levels. Moreover, recent studies associated reductions of $PM_{2.5}$ with air quality improvements (Sicard et al., 2021; Zhou et al., 2024), which emphasizes the importance and impact of reducing their concentration on the atmosphere.

### 2.1.1. Air Quality Monitoring

Air quality monitoring has evolved significantly over the years, alongside the development of Internet of Things (IoT). It has transitioned from basic manual sampling methods to sophisticated real-time monitoring technologies.

In the beginning of the 20th century, the first methods used deposit gauges that systematically collected and filtered soot in order to monitor the existence of pollutants in the air (Des Voeuz & Owens, 1912). Despite its effectiveness in identifying high level concentrations, traditional methods using fixed monitoring stations are often seen as failing to capture individual exposure due to their limited spatial coverage and temporal resolution (Steinle et al., 2013).

Advancements in sensor technologies and data analytics have enabled the development of mobile monitoring systems and low-cost sensors that can capture high-resolution air quality data across urban landscapes (Tarazona Alvarado et al., 2024). Current monitoring technologies include remote sensing, satellite observations, and ground-based sensors (Steinle et al., 2013; De Vito et al., 2021). Using these methods collectively can improve capture high-resolution air quality data across urban landscapes and better understand air pollution dynamics (Alvarado et al., 2019).

Moreover, the integration of IoT in air quality monitoring has enabled the development of innovative solutions for specific challenges in the field. For instance, Sunarno et al. (2020) explored a web-based wireless sensor system for monitoring sulfur dioxide levels, as well as for other indicators, such as temperature and humidity. Furthermore, a new system that operates on wireless sensors was proposed (Oyo-Ita et al., 2023), which creates real-time alerts through an application when air quality exceeds recommended thresholds.

### 2.1.2. AIR QUALITY PREDICTION

In highlight of the growing concerns about the negative impacts on public health and the environment linked to poor air quality, various authors worldwide started to develop air quality research. Their studies have focused on developing and optimizing effective methods to monitor air pollution and designing predictive models in order to mitigate its effects. Table 2.1 reviews studies on air quality topic based on different approaches.

Table 2.1 - Summary of research on Air Quality.

| Research Motivation | Methodology | Author(s) |
|---|---|---|
| Assess and simulate the impact of traffic road emissions on air quality in Lisbon | Combination of models (VISUM, TREM, VADIS) to simulate the effects of vehicle emissions | Borrego et al. (2004) |
| Explore and compare the performance of deep learning models in forecasting air quality on time-series data | Perform data training on air pollution data using deep learning models (RNN, LSTM, GRN) | Athira et al. (2018) |
| Development and implementation of a deep learning model to predict air pollutant concentrations | Temporal Sliding Long Short-Term Memory Extended, that combines | Mao et al. (2021) |

| | | |
|---|---|---|
| | multiple layers to use historical data | |
| Approach to improve air quality prediction accuracy | Implementation of combined chi-square (CT) and long short-term memory (LSTM) methods | Wang et al. (2021) |
| Focus on short-term air quality forecasting by evaluating models with multiple performance criteria on time series data | Comparative study using statistical, Machine Learning and Deep Learning models | Espinosa et al. (2021) |
| Framework that links vehicle emissions to air pollutants data and develops predictive analysis and visualizations to assist decision making | Perform data processing, visual analysis and machine learning models | Datia et al. (2022) |
| Implementation of a full process that provide real-time air pollution data collecting and monitoring. Data processing and display allows users to understand better air quality information | Development of sensors to detect air pollutants and meteorological conditions; data transition wirelessly; and data visualization interface | Oyo-Ita et al. (2023) |

In Lisbon, Borrego et al. (2004) study assesses the impact of road traffic emissions on urban air quality, using three different models for different purposes. First, VISUM model is used to characterize traffic demand and assess its impact on users, the data about vehicles and their routes is then input for the next model. TREM receives data from VISUM model and quantifies vehicles emissions based on their speed and category. Finally, VADIS model will evaluate and simulate the dispersion and concentration of traffic-related pollutants. The combination of these three models leads to a cohesive framework that effectively estimates pollutant's emissions on the atmosphere.

Moreover, the incorporation of deep learning models to improve air quality predictions has increased due to their ability to analyse large and complex data. Athira et al. (2018) used deep learning models to improve air quality forecasting, opposing to traditional prediction methods. The authors selected three deep learning models, specifically Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Network (GRN) architectures, expected to lead to good performance in predicting future $PM_{10}$ values with pollutants values data and factors influencing air quality, such as temperature and humidity. The research determines that all three models perform similarly well when performing prediction in the data. Additionally, Mao et al. (2021) used sensors, meteorological and temporal data over the course of six years to improve air quality forecasting. By incorporating a neural network with a Temporal Sliding Long Short-Term Memory Extended (TS-LSTME), a

deep learning model combines multiple layers to use historical data as inputs and the output of each iteration to predict the concentration of $PM_{2.5}$ for the next 24 hours. The study concludes that the proposed model outperforms traditional models such as Support Vector Regression (SVR).

Wang et al. (2021) proposed a new approach to develop air quality prediction, CT-LSTM, which combines chi-square (CT) and long short-term memory (LSTM). With the evaluation of the correlation between environmental variables and air quality, the chi-square identifies which meteorological factors can influence AQI. In this research, the CT method determined that, for the tested factors, wind scale influences air quality but humidity does not. The AQI is then predicted using LSTM, a type of Recurrent Neural Network (RNN), that receives as input historical data of pollutant levels and the meteorological factors selected from CT. When in comparison to different neural networks approaches, such as simple RNN and SVR, the CT-LSTM has improved accuracy and decreased error metrics for RMSE and MAE.

Furthermore, Espinosa et al. (2021) introduced a multi-criteria decision-making framework with focus on forecasting accuracy and stability. By testing different model approaches, with both machine and deep learning models, such as LSTM, Random Forest and Lass Regression, the results are reliable 24-hour forecasts.

The study from Datia et al. (2022) aims to assist decision-makers in Lisbon in monitoring and prediction of urban air quality. With data collected from low-cost portable sensors and the incorporation of traffic and vehicle emissions data, the authors were able to apply machine learning regression models to predict pollutant concentrations. Finally, an interactive dashboard facilitates the interpretation of the predictive model, as well as enables users to explore air quality data leading to informed decision making.

In conclusion, the diverse approaches explored in air quality predictions share a common purpose: improve the accuracy of air quality forecasting. With these efforts, countries can take timely actions to help mitigate adverse public health effects and reduce environmental harm trough informed decision making.

### 2.1.3. Indexes and Indicators for Air Quality

Through Key Performance Indicators the task of monitoring and predict air quality data is facilitated, as only by tracking certain values it is possible to assess air quality levels and derive relevant insights to improve decision-making (Seibert et al., 2022). Thus, they are essential for assessing the effectiveness of air quality management strategies and ensuring public health safety, especially concerning vulnerable population (Shikwambana et al., 2024). Regarding the existing research, Table 1.1 summarizes current KPIs for air quality monitoring.

Table 2.2 - Air Quality Index and Indicators.

| KPI | Description | Reference |
|---|---|---|

| | | |
|---|---|---|
| Air Pollutant Concentration Levels | Single value per pollutant | World Health Organization, (2021)<br>Wang et al. (2021) |
| Air Quality Index (AQI) | Traffic Light index for air quality indication | Shikwambana et al. (2024)<br>Datia et al. (2022) |
| - AQI for Health (AQIH)<br>- AtmoFrance Index<br>- Belgium AQI<br>- Comprehensive Air Quality Index (CAI)<br>- Common AQI (CAQI)<br>- Daily AQI (DAQx)<br>- Daily AQI (DAQI)<br>- European AQI<br>- Índice de Qualidade do Ar (IQAr) | Air quality measure variations for each country/ region:<br>- Ireland<br>- France<br>- Belgium<br>- South Korea<br>- European Union<br>- Freiburg Germany<br>- United Kingdom<br>- Europe<br>- Portugal | Karavas et al. (2021)<br>Tan et al. (2021) |
| Atmospheric Pollution Index (API) | Aggregation of air pollution data, providing a single value | Seibert et al. (2022) |
| Air Quality Health Index (AQHI) | Numerical scale that measures air quality based on pollutant's health risks | Abelsohn & Stieb (2011) |
| General Air Quality Health Index (GAQHI) | Improved local health-based AQI, based on pollutant's health risks | Tan et al. (2021) |

In 2021, with the arising concerns on poor air quality impacts, WHO updated their air quality guidelines (AQG) with stricter recommendations for major pollutants with higher risk to public health (World Health Organization, 2021). These guidelines can be used as an inspiration for countries to incorporate them as a legally enforceable standard.

Table 2.3 - WHO recommendations for Air Quality Guideline levels, in $\mu g/m^3$, adapted from World Health Organization (2021).

| **Indicator** | **Averaging Period** | **AQG Level** |
|---|---|---|
| $PM_{2.5}$ | 24-hour | 15 |
| $NO_2$ | 24-hour | 25 |
| $PM_{10}$ | 24-hour | 45 |
| $SO_2$ | 24-hour | 40 |

| | | |
|---|---|---|
| CO | 24-hour | 4000 |
| $O_3$ | 8-hour | 100 |

Air Quality Index (AQI) provides an understandable way to interpret more complex data, as it is calculated based on different major pollutants' concentrations on the atmosphere, including particulate matter, nitrogen dioxide, sulfur dioxide, carbon monoxide, and ozone (Shikwambana et al., 2024). This index is communicated through colour coding, indicating how polluted outdoor air currently is; usually, predictions on how the air quality will be are also available. Studies have shown that the AQI is predictive of health outcomes, such as for hospital admissions and respiratory issues, making it an important tool to monitor air quality and mitigate possible known outcomes (Heintz et al., 2022; Zacharko et al., 2021). This index transforms complex data into a more understandable way, aiding decision-making.

Although many countries and regions remain using their own index, some research try and adapt a common air quality index for the European Union. Karavas et al. (2021) selected and compared existing indexes in the EU, namely DAQI, CAQI, AQIH, from United Kingdom, European Union and Ireland, respectively. The study concludes that Air Quality Index for Health from Ireland is a suitable option to be adapted as a common air quality index in European Union due to the generality of its calculation and its central normalized values. Additionally, Tan et al. (2021) also criticized the existence of many different indexes to monitor and communicate air quality, due to its differences in calculation methods which makes it difficult to compare indexes for different countries and regions. Therefore, the General Air Quality Health Index (GAQHI) was proposed as an alternative index which incorporates pollutants and recognizes its health impacts, by evaluating the relationships between air pollutants and health outcomes.

The Atmospheric Pollution Index (API) is another index available to evaluate air quality by integrating secondary data into their predictions (Seibert et al., 2022). The calculation of API occurs in two moments. A) Using Analytic Hierarchy Process (AHP) method, data from various sources is integrated into the decision-making tool to select the factors that influence the most air quality. Such data is usually pollutant's concentration levels and meteorological conditions. B) The second moment consists of defining different criteria based on the selected data and weighting them according to its impact on air quality. Finally, IPA is calculated by assigning scores to these weighted factors, which are then normalized to produce a single representative value; the higher the API value, the worst the air quality.

Table 2.4 - Atmospheric Pollution Index classification values, adapted from Seibert et al. (2022).

| **API values** | **Classification** | **Colour Code** |
|---|---|---|
| API <0,20 | Good | Green |

| | | |
|---|---|---|
| 0,20 < API < 0,40 | Moderate | Yellow |
| 0,40 < API < 0,60 | Poor | Orange |
| 0,60 < API < 0,80 | Very Poor | Red |
| API <0,80 | Extremely Poor | Purple |

Additionally, concerning on the high risks of poor air quality in population's health, Abelsohn & Stieb (2011) used the Air Quality Health Index to serve as a health risk communication tool to help inform the population on the level of risk to outdoor air exposure. This index is calculated based on the concentration levels of three pollutants considered to have higher impact on health, ozone, particulate matter and nitrogen oxides, and their associated weights. As the AQHI uses a colour-coded scale ranging from 1 to 10 where higher values indicates greater risks, it is very easy for the population to interpret and make informed decisions.

Moreover, other than specifically defined indexes to monitor air quality, influencing factors on outdoor air composition are systematically incorporated to improve forecasting and conduct analysis. This data includes source specific emissions, such as vehicle and industrial activity emissions (Datia et al., 2022), meteorological factors, as temperature and humidity (Oyo-Ita et al., 2023), and demographical data.

Table 2.5 - Air Quality Influencing Factors.

| **Indicator** | **Reference** |
|---|---|
| Traffic Data | Borrego et al. (2004) |
| Vehicle Speed | |
| Traffic Congestion Levels | |
| Emission Quantification | |
| Vehicle Characteristics | Datia et al. (2022) |
| Fuel type | |
| Usage | |
| Age | |
| Engine capacity | |
| Technology | |
| Meteorological Data | Mao et al. (2021) |
| Precipitation | Wang et al. (2021) |
| Air pressure | |
| Relative humidity | |
| Temperature | |
| Wind direction | |
| Temporal Data | Wang et al. (2021) |

### 2.1.4. RESEARCH LIMITATIONS

As mentioned before, research on air quality has had a significant impact over recent years, with valuable insights both in health risks and environment management. However, despite the advancements, some limitations persist across various studies, which can affect the accuracy, scalability, and usability of air quality models and dashboards, highlighting improvement areas for further research.

Several studies have identified common challenges such as the capability to deal with high volume data generated continuously by sensors (Athira et al., 2018; Mao et al., 2021; Wang et al., 2021), in addition to difficulties in dealing with model complexity. For instance, Borrego et al. (2004) emphasize the complexity of the incorporation of three different models to improve air quality prediction, as well as how models rely on the availability and quality of the data collected. Finally, Mao et al. (2021) and Datia et al. (2022) highlight the insufficient incorporation of influencing factors on poor air quality, which can limit the model's ability to generalize effectively and incorporate models in bigger areas (Borrego et al., 2004; Mao et al., 2021).

In terms of dashboard creation, it has some user interactions limitations related to data representation; users can't see the usability of the estimators in the dashboard (Datia et al., 2022).

In conclusion, air quality is a very common topic on the scientific community, offering different studies and approaches. Some of them are more directed into analysing the impacts on public health, society and the environment; however, others offer more technical approaches while trying to improve air quality predictions applying different big data processing and transformation techniques. Nevertheless, these limitations emphasize the importance of continuous model enhancement and data integration strategies, which will contribute to more air quality monitoring systems and improve decision-making. A Business Intelligence framework is able to work through big data limitations and pre-processing automation, while incorporating the development of KPIs to improve air quality analysis, and predictive models to progress air quality monitoring systems.

# 3. METHODOLOGY

Business Intelligence has become an essential tool for decision-making across industries, providing organizations and policymakers with actionable insights. Due to high amounts of data from companies, government agencies, and research institutions, BI tool are leveraged to transform raw data into strategic knowledge, optimizing performance and public policies (Kelly et al., 2023). For air quality management, BI enables the integration of large volumes of environmental data, leading to trend analysis, and predictive modelling. This makes it crucial for urban planning, public health strategies, and environmental sustainability initiatives (Tan et al., 2021).

This chapter outlines the methodology used in developing a BI model for air quality analysis in Lisbon, focusing on the Kimball 'Business Dimensional Lifecycle' (Kimball & Ross, 2013), a widely adopted framework for BI projects. The development of an effective business intelligence model laid the ground for the development and implementation of a machine learning model, enabling a more effective assessment of air quality in Lisbon. Figure 3.1 illustrates the approach employed.

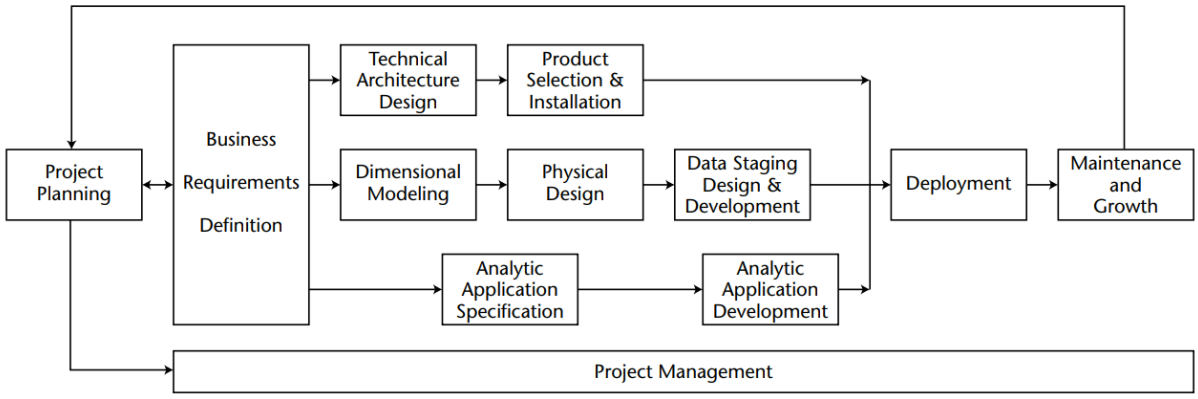


Figure 3.1 - Business dimensional lifecycle diagram, retrieved from Kimball & Ross (2013).

Therefore, Microsoft tools such as Microsoft Fabric and Power BI had a crucial role in the development approach. In the early stages, Microsoft Fabric allowed for the ETL processes and storage of the transformed data into the data warehouse, whereas Microsoft Power BI proved to be useful in the construction of new metrics, and implementation of dashboards and reports. Finally, with the data properly processed and structured, a machine learning model was implemented in Microsoft Fabric.

The process starts by an early identification of the business requirements, crucial to understand the needs and develop a model accordingly. The dimensional model design will follow the four-step dimensional design process, aiding as the foundation for building data warehouses and BI solutions. Approaching the deployment step, the ETL process is a crucial component, which cleans and transforms data into a desired structure, and loads it into a warehouse, ensuring consistency and usability. This structured approach enables

organizations to build scalable and efficient BI models that support informed decision-making and data-driven strategies. The following sections will describe the process of designing the dimensional model using Kimball's lifecycle approach.

This approach has been successfully implemented across various sectors, such as health, banking and management (Aziz et al., 2021; Cavalheiro & Carreira, 2016; Delgado et al., 2019) showing significant improvements in query development and dashboard usability (Landütama & Chowanda, 2023).

## 3.1. Business Understanding & Requirements

According to Kimball (Kimball & Ross, 2013), understanding the business needs and gathering the requirements is one of the most crucial steps in developing a successful model. This ensures the solution aligns with the goal, by improving efficiency and decision-making processes.

Thus, the following business needs and questions were strategically formulated to develop a visualization tool that, based on multiple indicators, can provide useful insights to leverage decision-making regarding air quality management in Lisbon (Table 3.1).

Table 3.1 – Business Needs and Questions definition.

| Business Need | Business Question |
|---|---|
| 1 - Air Quality Overview | • What are the current pollution levels at different monitoring stations in Lisbon?<br>• How do pollutant concentrations fluctuate over time (hourly, monthly, and seasonally)?<br>• Which locations experience the highest levels of pollution, and what pollutants are most prevalent?<br>• How often do pollutant levels exceed the maximum expected values?<br>• What is the AQI trend over time, and how does it compare across different locations? |
| 2 - Identifying Pollution Hotspots and High-Risk Areas | • Which locations exceed air quality thresholds most frequently?<br>• Are certain areas more affected by specific pollutants? e.g., high $NO_2$ in traffic-heavy zones.<br>• Which locations consistently report high levels of air pollution? |
| 3 - Detecting Anomalies in Air Quality Data | • Are there unusual spikes in pollution that do not align with historical trends? |
| 4 - Understanding the Impact of Meteorological | • How does different meteorological conditions affect pollution levels? |

| | |
|---|---|
| Conditions on Air Quality | • What meteorological conditions are most associated with poor air quality events? |
| 5 - Understanding the Impact of Traffic on Air Quality | • How do air pollution levels change during peak traffic hours compared to non-peak hours?<br>• Is there a strong correlation between vehicle flow and $NO_2$ or CO concentrations?<br>• How does reduced traffic during weekends or holidays affect air quality?<br>• Can traffic control measures (e.g., low-emission zones) lead to measurable improvements in air quality? |

## 3.2. DATA COLLECTION

The air quality data collected by Lisbon's council enables the analysis of parameters in the city and the development of thoughtful insights. Leveraging data from 80 different sensors spread across Lisbon, this chapter recognizes and describes relevant data for better understanding air quality in the city.

### 3.2.1. STUDY AREA

This research focuses its study in the municipality of Lisbon, the capital and largest city of Portugal, integrating twenty-four parishes. Situated in the western part of the Iberian Peninsula and north to the Tejo River, the city has an estimated population of over 540,000 (Câmara Municipal de Lisboa, 2021), covering an area of approximately 100 square kilometres. Figure 3.2 illustrates the distribution of the sensors in the city.

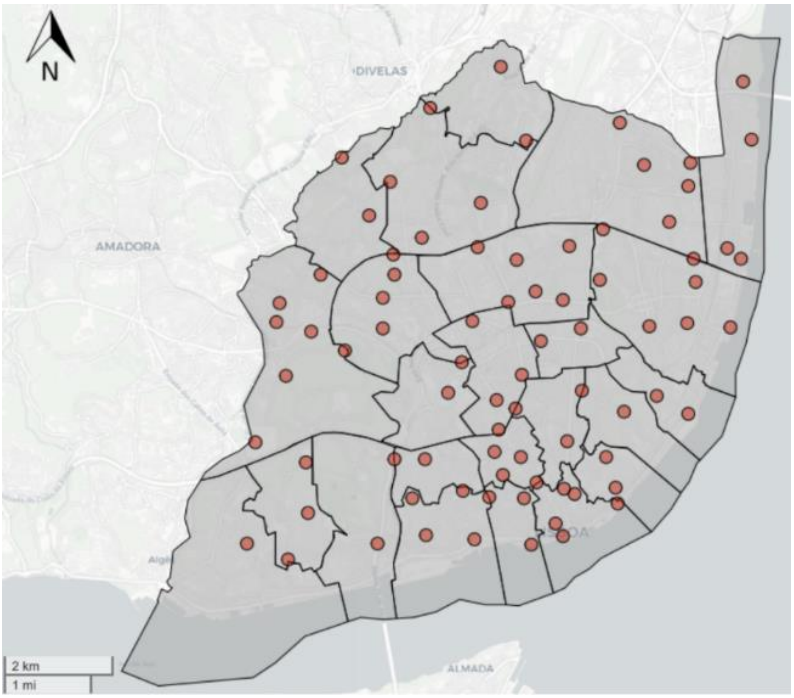


Figure 3.2 - Spatial distribution of stations in Lisbon by parish.

The air quality data is collected and managed by Câmara Municipal de Lisboa, through a network of sensors. This collected data is then made available to the public through Lisboa Aberta[1] website, an open data platform that provides access to various datasets related to the city.

The data in this study ranges from July 15, 2021, to January 15, 2024. With specific locations aiming to monitor diversity of regions and spaces with different characteristics, the sensors are distributed homogenously to be representative of the municipality. Thus, their placement is in accordance with different criteria, being a grid with 2x2 km spacing each one of them.

### 3.2.2. AIR QUALITY DATA

The data collected by a network of 80 sensors spread across the city enables for the analysis of the evolution of major pollutants values as well as some meteorological factors over time. Thus, each of this sensors monitors parameters such as particulate matter (PM), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$), carbon monoxide (CO), and ground-level ozone ($O_3$), as well as sound level (LAeq), atmospheric pressure, relative humidity, wind, precipitation, temperature, global radiation, ultraviolet radiation, and hourly traffic volume (HTV). Furthermore, it is important to highlight that not all parameters are monitored in every sensor location.

The dataset consists of 9 *csv* files and a total of 5,188,188 records. Table 3.2 provides the features in air quality dataset, as well as a brief description and data types.

Table 3.2 - Name and description of air quality data attributes.

| Attribute | Data Type | Description |
|---|---|---|
| DTM_UTC | DateTime | Date and time of report in Coordinated Universal Time (UTC). |
| DTM_LOCAL | DateTime | Date and time the measurement occurred (each measurement occurs every 15 minutes, the averaged value is reported hourly). |
| TEMATICA | String | Theme of the report (QA for Air Quality). |
| COD_PARAMETRO | String | Code of the parameter being measured (4-digit). |
| PARAMETRO | String | Name of the parameter being measured (2-digit chemical formula). |
| NR_ESTACAO | Number (INT) | Station number (1-80). |
| COD_SENSOR | String | Identifier of the measurement: composed of the topic Air Quality (2-digit), parameter (4-digit) and location (4-digit). |
| LOCAL | String | Name of the street where the sensor is deployed. |

[1] https://lisboaaberta.cm-lisboa.pt

| | | |
|---|---|---|
| LATITUDE | Decimal | Latitude of the sensor. |
| LONGITUDE | Decimal | Longitude of the sensor. |
| UNIDADE | String | Measurement unit applied for the parameter ($\mu g/m^3$; $mg/m^3$). |
| ETIQUETA_NIVEL | String | Level label of the parameter state according to predefined measures ('*NORMAL*', '*MODERATE*', '*HIGH*', '*VERY HIGH*'). |
| COR_NIVEL | String | Colour label of the parameter state according to predefined measures ('*NORMAL*', '*MODERATE*', '*HIGH*', '*VERY HIGH*'). |
| VALOR | Decimal | Sensors reported value of parameter. |

Table 3.3 describes the six pollutants in the air quality dataset, and some statistics on them. The values for each pollutant are in $\mu g/m^3$.

Table 3.3 - Statistics for air quality data discriminated by pollutant.

| | **$NO_2$** | **NO** | **CO** | **$O_3$** | **$PM_{10}$** | **$PM_{2.5}$** | **$SO_2$** |
|---|---|---|---|---|---|---|---|
| % of the total of rows | 20.76% | 8.71% | 15.52% | 9.13% | 22.77% | 22.65% | 0.46% |
| Minimum | -99.0 | -99.0 | -99.0 | -99.0 | -66.0 | -66.0 | -99.0 |
| Maximum | 14539.0 | 2000.0 | 13270.3 | 897.0 | 285152.0 | 52852.0 | 5441.0 |
| Mean | 59.17 | 38.09 | 193.07 | 73.93 | 35.97 | 10.27 | 49.23 |

### 3.2.3. COMPLEMENTARY DATA

Beyond direct air quality measurements, other data can provide valuable insights in air quality analysis. These datasets, while not strictly air quality metrics, support context to air quality trends and patterns. Regarding the additional data complementing the air quality dataset, table 3.4 provides a summary along with relevant features and data sources.

Table 3.4 - Category, features and source for complementary data.

| **Data category** | **Feature(s)** | **Source** | **Nr records** |
|---|---|---|---|
| Meteorological | Date-time, parameter, value, location. | Lisboa Aberta | 2,909,537 |
| Traffic | Date-time, location, address, type, subtype, confidence, nthumbsup, reliability. | Waze Partner Support | 1,185,386 |

| | | | |
|---|---|---|---|
| Reduced Emission Zones | Zone identifier, latitude, longitude. | Urban Access Regulations[2] | 2 |

As previously mentioned, meteorological data is widely used in air quality analysis since atmospheric conditions significantly influence the dispersion of pollutants (Mao et al., 2021; Wang et al., 2021). The data provided by Câmara Municipal de Lisboa through, once again, Lisboa Aberta's website, includes factors for meteorological conditions such as atmospheric pressure, relative humidity, wind speed and direction, precipitation, temperature, and global and ultraviolet radiation, spanning from January 2022 to January 2024. Similar to the air quality data, the weather data is recorded by the same 80 sensors spread across Lisbon and have a similar structure (Table 3.5).

Table 3.5 - Name and description of weather data attributes.

| **Attribute** | **Data Type** | **Description** |
|---|---|---|
| DTM_UTC | DateTime | Date and time of report in Coordinated Universal Time (UTC). |
| DTM_LOCAL | DateTime | Date and time the measurement occurred (each measurement occurs every 15 minutes, the averaged value is reported hourly). |
| TEMATICA | String | Theme of the report (ME for Meteorology). |
| COD_PARAMETRO | String | Code of the parameter being measured (4-digit). |
| PARAMETRO | String | Name of the parameter being measured (2-digit chemical formula). |
| NR_ESTACAO | Number (INT) | Station number (1-80). |
| COD_SENSOR | String | Identifier of the measurement: composed of the topic Air Quality (2-digit), parameter (4-digit) and location (4-digit). |
| LOCAL | String | Name of the street where the sensor is deployed. |
| LATITUDE | Decimal | Latitude of the sensor. |
| LONGITUDE | Decimal | Longitude of the sensor. |
| UNIDADE | String | Measurement unit applied for the parameter ($\mu g/m^3$; $mg/m^3$). |
| ETIQUETA_NIVEL | String | Level label of the parameter state according to predefined measures ('*NORMAL*', '*MODERATE*', '*HIGH*', '*VERY HIGH*'). |
| COR_NIVEL | String | Colour label of the parameter state according to predefined measures ('*NORMAL*', '*MODERATE*', '*HIGH*', '*VERY HIGH*'). |

[2] https://geodados-cml.hub.arcgis.com/datasets/4933d8f832474ad2bff558cae59c5207_5/explore?location=38.713264%2C-9.155328%2C13.76

| | | |
|---|---|---|
| VALOR | Decimal | Sensors reported value of parameter. |

In addition to weather data, traffic data was retrieved from the Waze Partner Support website[3], with data from January 1, 2022, to February 26, 2024. This dataset has data related to traffic alerts reported by users at a specific time and location; for each incident reported, information is provided on the type and subtype of the event and the level of confidence on it (Table 3.6).

Table 3.6 - Name and description of traffic alert type data attributes[4]

| **Attribute** | **Data Type** | **Description** |
|---|---|---|
| City | String | City and state name. [City, State] in case both are available; [State] if there is no city available. |
| Confidence | Number (INT) | Precision score based on user reactions (0-10). |
| nThumbsUp | Number (INT) | User 'Like' count. |
| Street | String | Street name. |
| uuid | String | Unique system ID. |
| Country | String | Country code. |
| Type | String | Event type |
| Subtype | String | Event subtype. |
| roadType | Number (INT) | Type of road. |
| Reliability | Number (INT) | Confidence score based on user reactions and level (0-10). |
| Magvar | Number (INT) | Event direction (the direction in which the driver is moving at the time of the alert, 0 degrees pointing north, based on their device). |
| reportRating | Number (INT) | User classification (1-6; 6 = best classification). |
| Geometry | String | Geocode of the alert report. |
| Start_time | DateTime | Start date and time of the alert report. |
| Lat | Float | Latitude of the alert report. |
| Lon | Float | Longitude of the alert report. |
| Geom | Coordinates | Latitude (X) and Longitude (Y) coordinates per alert. |

[3] https://support.google.com/waze/partners/

[4] https://support.google.com/waze/partners/answer/13458165?hl=en#zippy=%2Ctraffic-alerts%2Cdata-elements%2Cjson

| | | |
|---|---|---|
| Timestamp | DateTime | Date and time of the report. |
| Date-only | Date | Date of the report. |
| reportByMunicipalityUser | Boolean | Indication if the alert was sent by a municipality user. |
| reportDescription | String | Alert description (when available). |
| pubMillis | Timestamp | Publication Date. |

Moreover, to ensure a more complete dataset with relevant information, two datasets were incorporated. Starting with the parish relative to each sensor, as Lisbon is divided into twenty-four parishes, understanding the air quality patterns in each one might unveil relevant insights. The external dataset named '*Freguesias 2012*', also available in the portal of the Lisbon city council, set the limits of each parish in Lisbon. Additionally, aiming to improve air quality, the Lisbon's council limited specific areas in the city where the access is restricted to certain vehicles. This measure defines that only vehicles from year 2000 onwards are allowed in zone one, and only vehicles from year 1996 onwards are allowed in zone two, with no restrictions in the remaining areas[5]. Thus, *'REZ_zones*' is a dataset with the borders of the two zones, by having the zone where each sensor lies can enable the research to better understand the impact and result these have on air quality.

Overall, by combining major air pollutants data in Lisbon and relevant supporting data with possible impact in air quality, ensures a robust foundation for assessing air pollution trends and their influencing factors.

### 3.3. TECHNICAL ARCHITECTURE DESIGN

The BI solution will be supported by the chosen technological components for each step of the project, by developing a well-structured technical foundation. Thus, Microsoft Fabric (MF) will be the core platform for this solution, an analytics cloud-based platform which incorporates different components that allow for data ingestion, storage, transformation and visualization (Borra, 2024).

This tool will play a crucial role in the development of the dimensional model, by leveraging data flow and allowing for the creation of the desired structure for the data. Therefore, the raw data collected from different sources, as described in chapter 3.1.2., is stored in a Lakehouse, that stores both structured and unstructured data. The data is then transformed through Data Flows and Python notebooks, acting as tools to clean, transform and integrate the data through the Staging Area, serving as an intermediate stage to data transformation processes before transferring into the Data Warehouse. All these processes will be leveraged through Data Pipelines which help automate the data flow. Regarding visualization, Power BI

[5] https://www.lisboa.pt/temas/mobilidade/modos-de-transportes/veiculo-privado

(Borra, 2024) will leverage dashboard creation. Table 3.7 describes each technology used and its purpose in the project.

Table 3.7 - Applied Technologies and their Roles.

| Component | Technology | Purpose |
|---|---|---|
| Data Storage | Lakehouse | Stores raw data. |
| ETL processes | Dataflow | Data preparation processes from multiple sources. |
| ETL processes | Python Notebook | Data preparation processes from multiple sources using Python libraries. |
| Staging Area | Data Warehouse | Intermediate Data Warehouse for data preparation and validation processes. |
| Data Warehouse | Data Warehouse | Stores structured transformed data for analytical queries. |
| Managing the Data Flow | Data Pipeline | Automate data flow: load Staging, validate Staging, load Data Warehouse. |
| Visualization | Power BI | Business Intelligence and reporting. |

## 3.4. Dimensional Modelling

As stated before, according to Kimball & Ross (2013), the key four steps for an effective process are as follows: (1) Select the business process that generate data; (2) Declare the grain to specify the level of detail; (3) Identify the dimensions to capture the descriptive context; (4) Identify the facts with the numeric performance metrics. This structured approach enables to build scalable and efficient BI model that support informed decision-making and data-driven strategies.

### 3.4.1. Business Process

The study focuses on monitoring and analysing air pollutant levels and to classify them according to predefined thresholds. Therefore, the business process is Air Quality Tracking. This includes collecting air quality data from 80 sensors distributed across Lisbon, monitoring key pollutants such as $PM_{2.5}$, $PM_{10}$, $NO_2$, CO, and $O_3$, and conducting seasonal, and temporal analysis. Consequently, the designed model ensures the data is structured to support critical queries, such as pollution fluctuations over time, and correlations with weather conditions, enabling decision-making, and the improvement of current strategies.

### 3.4.2. Granularity

Additionally, defining the level of detail for each fact row is crucial, since it ensures consistency in the model, avoiding mismatches, inconsistent metrics and redundancy (Kimball & Ross, 2013). Concerning the air quality data, each transaction corresponds to the value for each parameter at a specific date, station and category level, allowing analysis in detail for each

collection. This granularity ensures that the business intelligence model supports a range of analytical queries. Table 3.8 refers to each fact and the granularity specified.

Table 3.8 - Granularity definition per fact.

| Table | Granularity Definition |
|---|---|
| Air Pollution | Parameter; Unit of Measurement; Sensor; Date; Hour. |
| Weather | Parameter; Sensor; Date; Hour. |
| Traffic Alert | Alert Type; Alert Subtype; Street; Date; Hour. |

### 3.4.3. DIMENSION TABLES

As the dimension tables define the descriptive attributes, they are crucial to provide context to the transactions stored in the fact table. Thus, the development of six different dimension tables helped improve the model, by providing extra information without overcrowd the fact table: parameter, location, level, date, time and traffic, a summarized description of each dimension is in table 3.9. Organizing context data into dimension tables makes it easier to generate reports that communicate pollution severity in an understandable manner.

Table 3.9 - Name and description of dimension tables in the BI model.

| Name | Description |
|---|---|
| Date | Date-related attributes, such as day, month, year, and holiday indicator, to support time-based analysis. |
| Time | Time-related details, as hour, peak identifier and time frames. |
| Location | Stores information about stations, including location, parish and REZ identifier. |
| Parameter | Context to pollutants parameters, including name and unit of measurement. |
| Level | Categorizes pollutant values into air quality index levels based on threshold values. |
| Alert Type | Support traffic data, with information about alerts' type and subtype. |

### 3.4.4. FACT TABLES

The final step is to define the fact tables, the central component of the model (Kimball & Ross, 2013), containing values of air quality data. These tables link to the dimension tables through foreign keys, this way connecting all information about the transactions. Other than the date, time, location, pollutant parameter and level category keys that link to the dimension tables, the fact AirQuality has a numeric field that represents the value collected by each station per pollutant. The structured design of the fact table enables efficient aggregation and filtering of

data, supporting various analysis. Other than the air quality data fact table, the weather and traffic data are incorporated into the model to help identify trends with impact in air quality. Table 3.10 summarizes the tables created and their purpose in the developed Business Intelligence model.

Table 3.10 - Name and description of fact tables in the BI model.

| Name | Description |
|---|---|
| Air Quality | Stores pollutant concentration readings from sensors, linked to time, station, parameter, and pollution level dimensions. |
| Weather | Contains meteorological data such as temperature, wind speed, humidity, atmospheric pressure, and radiation. |
| Traffic | Stores traffic alert data and event types. |

### 3.4.5. SCHEMA APPROACH

The star schema approach introduced by Kimball & Ross (2013) consists of a central fact table that stores all transactional data, and multiple dimension tables that provide context to each transaction. Among several advantages, this approach is simple and easily understandable. In this project, a constellation schema model was implemented, which consists of multiple interconnected star schemas sharing common dimension tables. This structure ensures each fact table can focus on a specific subject area (Figure 3.3), where the location, date and time dimensions are shared between all three fact tables.

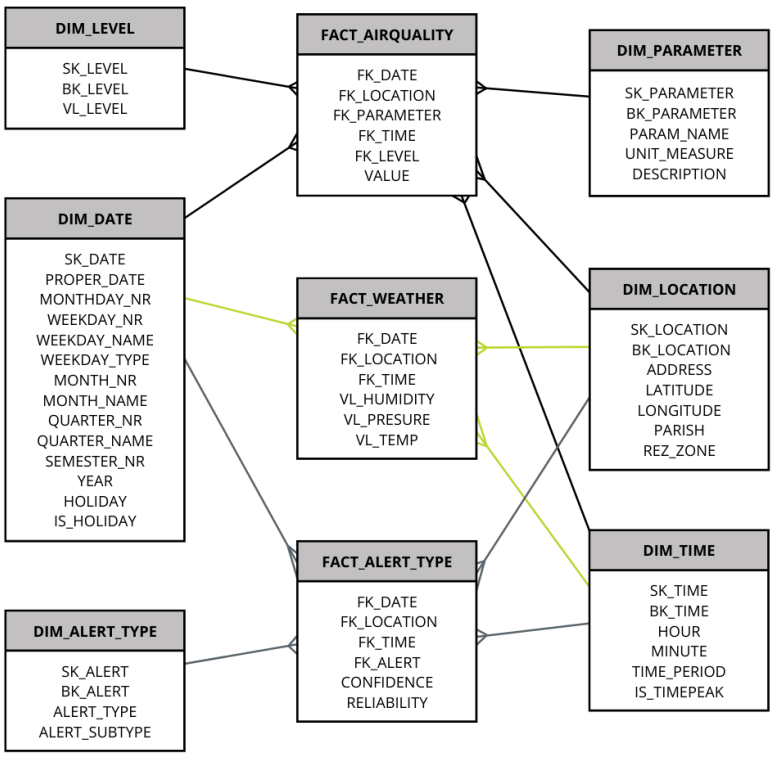


Figure 3.3 - Dimensional Model, Star Schema approach.

## 3.5. DATA TRANSFORMATION (ETL PROCESS)

After defining the model's prerequisites, it is important to conduct the ETL (Extract, Load & Transform) processes, which ensure the data from monitoring stations is properly ingested, validated and stored in the data warehouse (Kimball & Ross, 2013). By conducting these processes, the data would be refined before future developments dashboard wise, facilitating visual analysis and enabling decision-making.

Therefore, by developing a structured workspace environment in Microsoft Fabric, and leveraging its available tools, enables the creation of an efficient and scalable ETL process. Thus, the majority of ETL is handled in a staging area, created with the purpose of developing, refining and validating the data, the final stage is to populate the final data warehouse tables with the corresponding staging ones. The data flow is outlined in Figure 3.4.

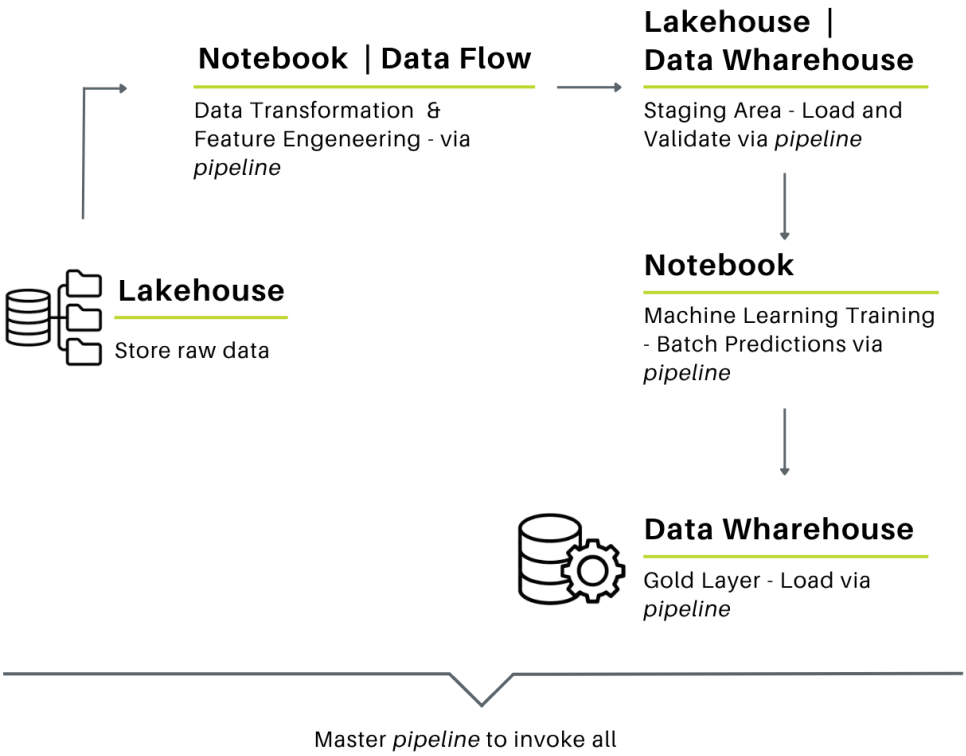


Figure 3.4 - Data flow during ETL to achieve DW.

### 3.5.1. EXTRACT

The original data was extracted and stored into the environment within the Microsoft Fabric Lakehouse ('*LH_AIRQUALITY*'), which serves as a data architecture platform to store data. Thus, all datasets described in section 3.3 were stored in a single location, which simplifies their later use.

For both Air Quality and Meteorological data, there are 9 files for air quality data and 5 for meteorological data, each containing different quarters and semesters, respectively. To merge

these fragmented datasets, the corresponding data frames for each were concatenated into one, easing further manipulation.

### 3.5.2. TRANSFORM

The transformation phase focused on cleaning, enriching, and restructuring the extracted datasets to ensure consistency and usability. Although each dataset required specific manipulations, several transformations were applied across all datasets to maintain a standardized structure. The changes applied in the raw data are aligned to the constellation schema previously designed.

Therefore, the majority of the pre-processing was developed in a Python notebook for each dataset, used to extract the data into data frames. By leveraging notebooks to transform the data allowed for greater flexibility and control.

#### 3.5.2.1. AIR QUALITY DATASET

To ensure consistency across the dataset, the air quality data was under transformations (Table 3.11). Additionally, column selection was carefully reviewed to eliminate redundancy and retain only relevant attributes for analysis.

Table 3.11 - Data transformations for Air Quality dataset.

| Field | Transformation |
|---|---|
| *LABEL* | Values translated to English:<br>▪ 'NORMAL' -> 'NORMAL'<br>▪ 'MODERADO' -> 'MODERATE'<br>▪ 'ELEVADO' -> 'HIGH'<br>▪ 'MUITO ELEVADO' -> 'VERY HIGH'<br>'@NA' values associated with the correct classification. |
| *LEVEL_COLOUR* | Values translated to English:<br>▪ 'VERDE' -> 'GREEN'<br>▪ 'AMARELO' -> 'YELLOW'<br>▪ 'LARANJA' -> 'ORANGE'<br>▪ 'ENCARNADO' -> 'RED'<br>'GRAY' values associated with the correct classification. |
| *UNIT_MEASUREMENT* | Clean blank spaces. |
| COD_PARAMETER | NO values dropped since this pollutant is not currently monitored by Lisbon's Council (8.71%);<br>$SO_2$ values dropped since it is not monitored in 50% of the parishes (0.46%). |
| *VALUE* | Dropped rows when negative values (0.61%). |

| | |
|---|---|
| | Dropped rows when $PM_{10}$ > 150 μg/$m^3$<br>Dropped rows when $PM_{2.5}$ > 100 μg/$m^3$<br>Dropped rows when $NO_2$ > 200 μg/$m^3$<br>Dropped rows when $O_3$ > 300 μg/$m^3$<br>These rows represented 0.26% of the data. |

Some values were considered outliers due to sensor errors. Thus, to identify these high values, rather than applying a fixed threshold to each parameter, and after further analysis, cases when a single station reported exceptionally high values while all other stations during the same timestamp recorded significantly lower levels where flagged. These discrepancies suggest that the high values reported are due to faulty sensor behaviour, leading to isolating and removing these values.

In addition, several new columns were created to enrich the dataset and the BI model, derived from existing ones (Table 3.12).

Table 3.12 - Feature Engineering Air Quality dataset.

| **Field** | **Purpose** |
|---|---|
| DATE, TIME, HOUR | Derived from *LOCAL_DATETIME*. |
| PARAMETER | Full chemical names to make the information accessible and understandable to the general population. |
| PARISH, REZ_ZONES | Derived from LATITUDE, LONGITUDE and json boundaries files, to provide more context and clear information. |

To determine the parish each station is located on, the features *LATITUDE* and *LONGITUDE* were used, as well as an external dataset named '*Freguesias 2012*', which set the limits of each parish. After reading this new data, the GeoJSON file was extracted and the polygon of each boundary was defined, these were converted into a GeoDataFrame used to check which sensors fell within each parish. A similar process was conducted to determine if each combination was in a Reduced Emission Zone or not, using *'REZ_zones*' dataset containing limits for both zone 1 and 2.

#### 3.5.2.2. METEOROLOGICAL DATASET

Similarly to the air quality dataset pre-processing, the meteorological data was also standardized by removing blank spaces as well as ensuring only relevant, non-redundant attributes were selected. Refer to Table 3.13 for data cleaning and transformations in the weather dataset.

Table 3.13 - Data transformations for Meteorological dataset.

| Field | Transformation |
|---|---|
| - | Drop rows with null values (17.93%). |
| COD_PARAMETER | Filtered to maintain 'TEMPERATURE', 'ATMOSPHERIC PRESSURE' and 'RELATIVE HUMIDITY', the most representative categories (35.5%, 31.2% and 29.9%, respectively), representing over 96% of the data. |
| *VALUE* | Filter maximum and minimum values according to history data in Lisbon[6]:<br>▪ 'HUMIDITY': 0-100 %<br>▪ 'PRESSURE': 970-1050 mbar<br>▪ 'TEMPERATURE': -2-40 $^{o}$C<br>Total of 1.7% rows dropped. |

The defined thresholds for each value were fixed according to expected minimum and maximum values in Lisbon for relative humidity, atmospheric pressure and temperature.

Finally, the dataset structure was transposed, where instead of having one row per parameter, timestamp and station, a new format was adopted where each row represents a unique combination of date, time, and station, and the values for each parameter were stored as individual columns, one per parameter. This new layout ensures more intuitive analysis and simplifies the design of the model.

#### 3.5.2.3. ALERTS TYPE DATASET

For the alerts type dataset an initial validation of null count per field was conducted, which lead to dropping the fields *nThumbsUp*, *reportByMunicipalityUser*, *reportDescription* and *pubMillis*, each with over 80% null values. The transformations applied in the remaining fields are described in table 3.14.

Table 3.14 - Data transformations for Traffic Alert dataset.

| Field | Transformation |
|---|---|
| - | Drop duplicated rows (6.1%). |
| *STREET* | Names standardized to match *ADDRESS* field in air quality and weather datasets.<br>Use coordinates with known street to fill rows with null street field (6.6% filled rows of total null count).<br>Drop rows with remaining null values (3.5%). |

[6] https://www.ipma.pt/pt/oclima/monitoriza.mensal/mmm-clima-PT100-cm.jsp

| | |
|---|---|
| *SUBTYPE* | Null values (8.5%) replaced with 'NO_SUBTYPE', which is a valid subtype. |
| *START_TIME* | Portugal local time, selected to match *LOCAL_DATETIME* field in air quality and weather datasets. |

Regarding feature engineering, a similar approach to air quality dataset (Table 3.12) was conducted.

After all transformation processes, *csv* files for parameter and level dimensions were exported to a common location into the Lake House. For the remaining tables, the data was exported into the same location as parquet files, which maintain the original data type making them compatible with integer and decimal fields in the staging tables.

#### 3.5.2.4. DATE AND TIME DIMENSIONS

Since the original data covers values from July 2021 to February 2024, the date dimension was created using a Data Flow to generate it and related fields. Thus, by generating dates between the expected ones in the original data, leveraging Data Flow tools, the Day, Day of the week name, Month, Month name and Year were easily generated. Additionally, Weekday Type, Semester and Holiday fields were generated based on the previous ones. This approach facilitates the generation of further dates when applied to new data for more recent years. Table 3.15 describes the holidays and respective dates for each year.

Table 3.15 - Holidays and special days definition for Date dimension.

| Year | Date | Holiday |
|---|---|---|
| All | January 1st | New Year's Day |
| 2021 | February 16[16] | Carnival |
| 2022 | March 1st | |
| 2023 | February 21st | |
| 2024 | February 13th | |
| 2021 | April 4th | Easter Day |
| 2022 | April 17th | |
| 2023 | April 9th | |
| 2024 | March 31st | |
| 2021 | April 2nd | Holy Friday |
| 2022 | April 15th | |
| 2023 | April 7th | |
| 2024 | March 29th | |
| All | April 25th | Liberty Day |

| | | |
|---|---|---|
| All | May 1st | Labour Day |
| All | June 10th | Portugal Day |
| All | June 13th | Saint Anthony Day |
| 2021 | June 3rd | Corpus Christi |
| 2022 | June 16th | |
| 2023 | June 8th | |
| 2024 | May 30th | |
| All | August 15th | Assumption of Our Lady Mary |
| All | October 5th | Proclamation of the Republic |
| All | November 1st | All Saints' Day |
| All | December 1st | Restoration of Independence |
| All | December 8th | Imaculada Conceição Day |
| All | December 24th | Christmas Eve |
| All | December 25th | Christmas Day |
| All | December 31st | New Year's Eve |

Additionally, the time dimension was also created using a Data Flow, generating times by the defined granularity, hour. The hour field allows for the creation of the *IS_TIMEPEAK* field, which indicates peak hours, assigning a value of 1 when the time falls between 7–9 a.m. or 5–7 p.m., inclusive, and 0 otherwise. Similarly, the *TIME_FRAME* field categorizes each record into broader time periods, to flag records occurring during high-traffic or activity peak times (Table 3.16).

Table 3.16 - TIME_FRAME field definition for Time dimension.

| **HOUR** | **TIME FRAME** |
|---|---|
| 0-5 | Late Night |
| 6-8 | Early Morning |
| 9-11 | Morning |
| 12-13 | Early Afternoon |
| 14-17 | Afternoon |
| 18-20 | Evening |
| 21-23 | Night |

### 3.5.3. LOAD

After the pre-processing stage, to achieve the end goal to load the Data Warehouse, the design of four pipelines for different purposes was crucial, to automate the processes of transformation, validation and load of the data.

Therefore, the first pipeline '*PL_AQ_Load_STG*' automates the staging area process (Figure 3.5), through running pre-processing notebooks individually for Air Quality, Weather and Traffic Alert Types datasets. Additionally, individual notebooks for each table in the model were created, to ease the process of debugging if needed, and the output files were stored in a Lakehouse folder 'Silver Layer – Productive Data', representing the silver layer. The model tables are then populated sequentially, in a first moment the dimension tables, and then the fact tables.

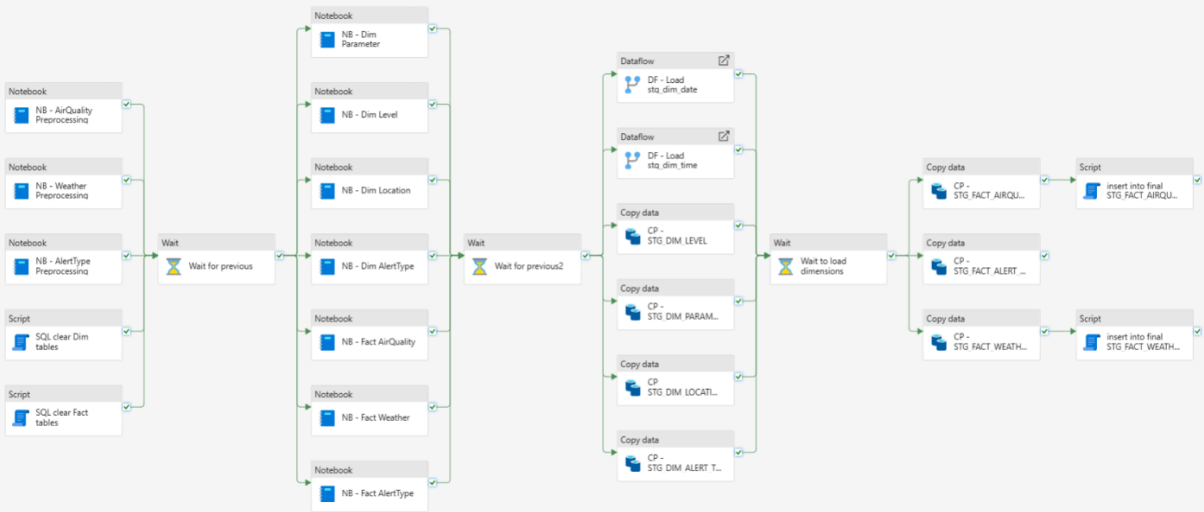


Figure 3.5 - Load Stating Area pipeline.

Once the data was loaded into the staging area tables, it must be validated to ensure it is structured according to best practices in a star schema approach. Thus, the pipeline '*PL_AQ_Validate_STG*' (Figure 3.6) ensures these validations are conducted both on fact and dimension tables based on predefined rules (Table 3.17).

Table 3.17 - Description of the rules applied for data validation.

| Rule | Target tables | Description |
|---|---|---|
| 1 | Dimension tables | Check integrity of business key, by ensuring there are no duplicated BK. |
| 2 | Dimension tables | Check uniqueness of dimension attributes, ensuring the non-business key fields are unique. |
| 3 | Fact tables | Check integrity of primary keys, by guaranteeing the combination of all foreign keys is unique. |
| 4 | Fact tables | Check relationship of dimension and fact tables, by assuring each foreign on the fact table is an existing business key in the respective dimension table. |

Each validation inserts a new row to a LOG_QUALITY_CHECKS table, where a field ETL_RESULT contains either '*FAIL*' or '*OK*' values, according to the validation status. Therefore, if there is

any '*FAIL*' value the system automatically generates and sends an email to a predefined user, allowing for quick action in the proper correction of the data.

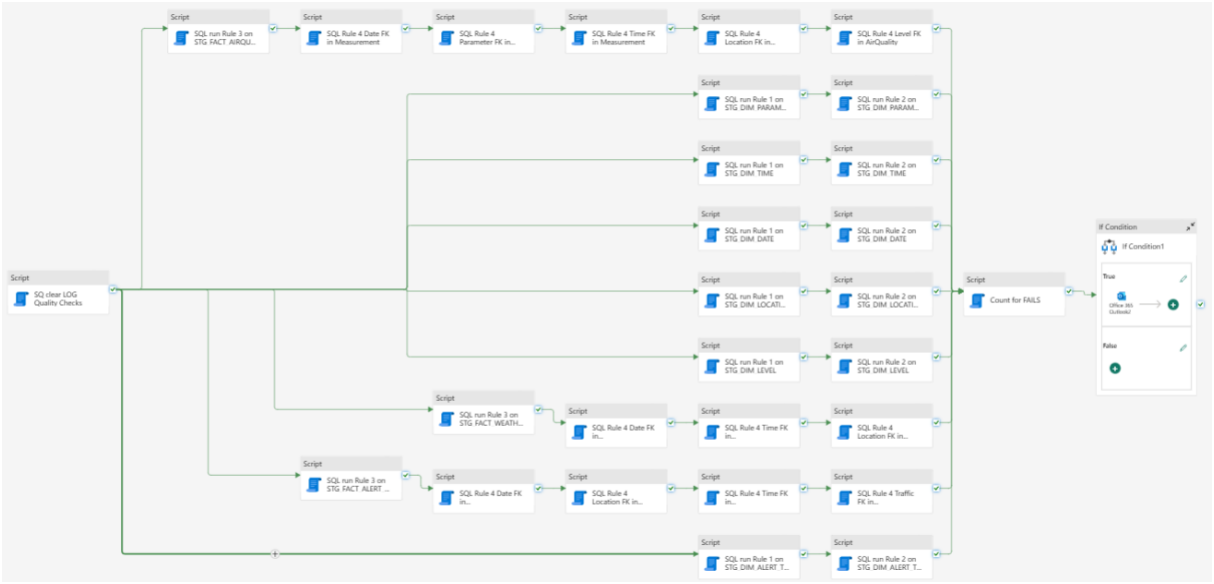

Figure 3.6 - Staging Area Validation pipeline.

Finally, after the staging tables are populated and the data validations are completed, the data is loaded into the final central repository, the data warehouse, where the data is cleaned, organized and structured, and ready to be used for BI reporting. To achieve this, pipeline '*PL_AQ_Load_DW*' will directly populate the final tables in the Data Warehouse, making the process more efficient and automated (Figure 3.7).

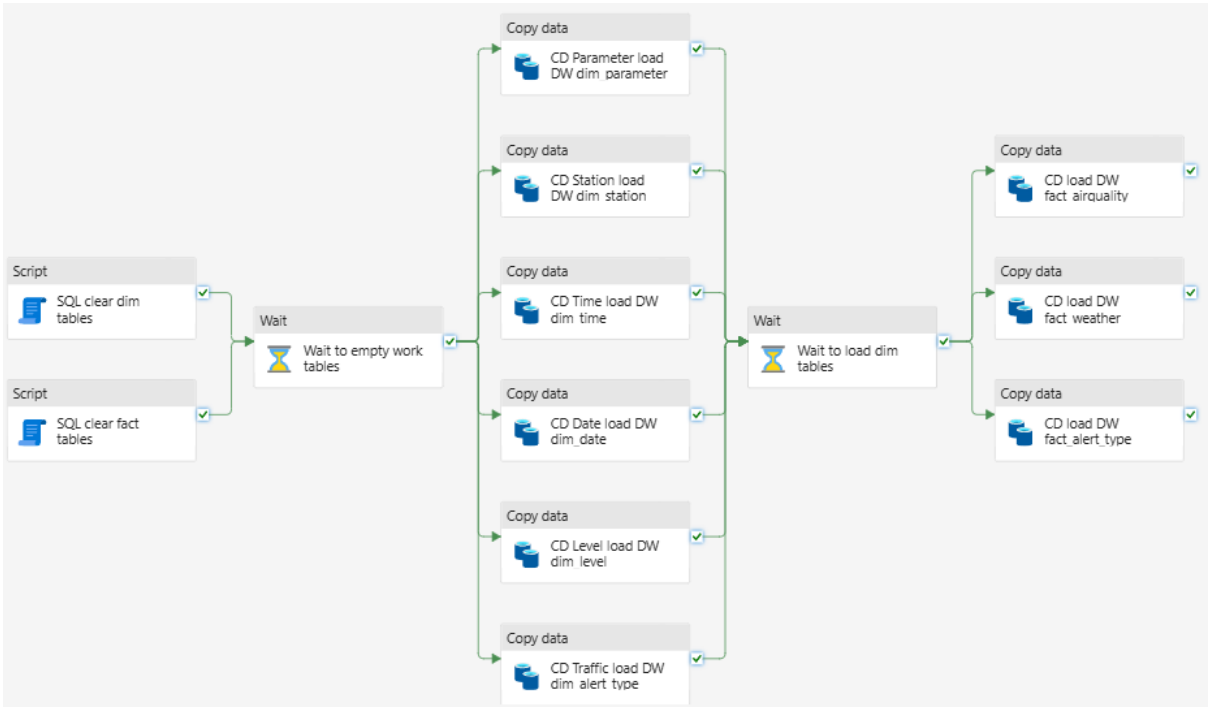


Figure 3.7 - Load Data Warehouse pipeline.

A final master pipeline '*PL_Master*' was developed to execute all previously created pipelines, automating the data loading and validation processes (Figure 3.8). This way simplifying execution, reducing intervention and ensuring consistency, eliminating the need to run each pipeline individually. This final pipeline will generate an email if one of the three main pipelines fail indicating which one failed and the time, allowing for quick action in solving the error. Additionally, this pipeline is scheduled to run every hour, aligning with the hourly generation of new data, this ensures the model remains as up to date as possible and provides timely, relevant information.

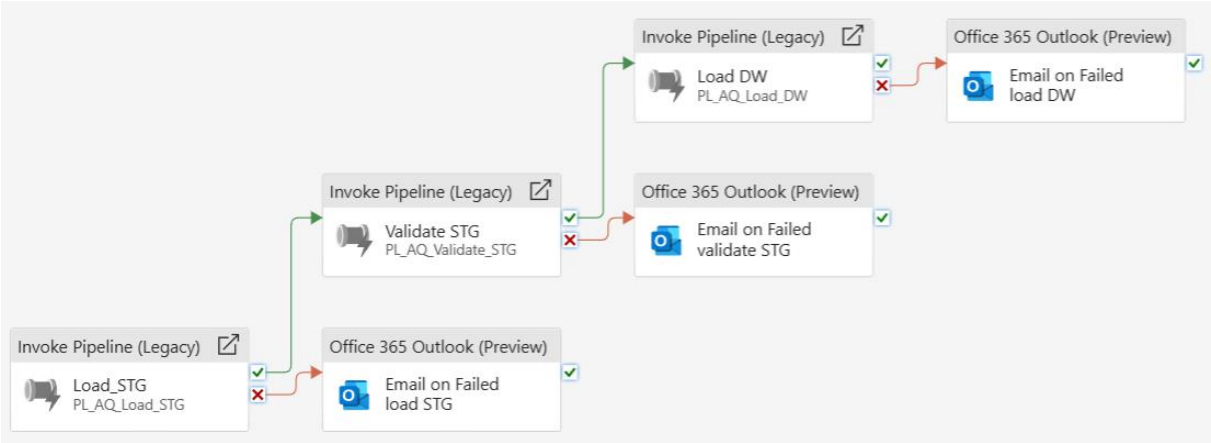


Figure 3.8 - Master pipeline to automate the running process.

With the development of a robust dimensional model with clean and structured data, advanced analysis is possible, retrieving valuable insights on air quality. Thus, this builds a foundation for a development and implementation of prediction models, to improve and enhance decision-making.

## 3.6. FORECASTING

In this section, two models are proposed using time-series analysis to predict air quality data, ARIMA and Facebook Prophet. Thus, a notebook '*NB_TEST_FORECAST*' was used to train the models and achieve desired results. With this approach, is then possible to select the model that better predicts air quality and save them as machine learning models in MF. Additionally, a notebook to run the forecasts is created and integrated into a pipeline to automate the process (Figure 3.9).

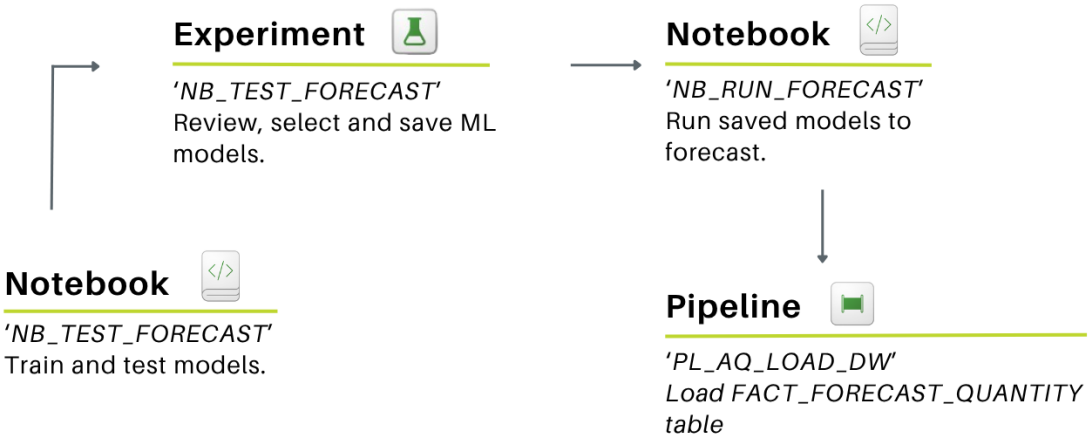


Figure 3.9 - Machine learning models' development flow chart in MF.

Thus, to forecast values for each parameter, a model was developed and implemented for each parameter, resulting in a total of 10 models developed, 2 for each parameter (both ARIMA and Prophet).

### 3.6.1. ARIMA

ARIMA (Auto Regressive Integrated Moving Average) models are statistical models used for time series forecasting, introduced by Box and Tiao (1975). These models capture the structure of time series by combining three components:

- Autoregressive (AR): modelling the dependence between an observation and several lagged observations.
- Integrated (I): differencing to make a time series stationary.
- Moving Average (MA): modelling the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Additionally, SARIMAX extends this model by including one or more external regressors expected to influence the target variable (X). Whereas, a seasonal component (S) handles temporal patterns within the data, such as monthly or yearly.

### 3.6.2. PROPHET

The Prophet model, proposed and developed by Facebook (Taylor & Letham, 2018), offers an interpretable model that balances the strengths of automated statistical forecasting with human domain knowledge. It is a decomposable time series model, with three main components (trend, seasonality and holidays), see equation 3.1.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (3.1)$$

Where:

- $g(t)$ is the trend component
- $s(t)$ is the seasonality component

- $h(t)$ captures of holidays or irregular events days
- $\epsilon_t$ is the residual component (presumed as with a normal distribution)

With this approach and components, it is possible to capture behaviour influenced by hourly, daily, weekly or seasonal cycles, and holiday patterns. By applying the holidays defined in table 3.14, a baseline of days with an expected different flow and influence in air quality is established.

### 3.6.3. EVALUATION METRICS

To assess the forecasting models, two evaluation metrics were selected, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics are widely used in various research approaches (Athira et al., 2018; Espinosa et al., 2021; Wang et al., 2021).

Therefore, MAE measures the average difference between the actual and forecasted values (Equation 3.2). While easy to interpret, this metric does not distinguish if the forecasted values are higher or lower than the actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.2}$$

Additionally, RMSE also reflects the average of prediction errors but penalizes lager errors more heavily, making it a more conservative metric (Equation 3.3).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

By applying both metrics to evaluate the models it was possible to not only determine which model produced more accurate results, but also which model was more robust against large errors.

### 3.6.4. PIPELINE INTEGRATION

The forecast integration into the scheduled pipelines is essential to ensure the forecasted values are systematically generated, stored and kept up to date in the data pipeline. Thus, after training the models and accomplishing the desired results, two notebooks '*NB_RUN_FORECAST*' and '*NB_RUN_FORECAST2*' were useful to apply the five prophet models into recent data and create new forecasts. These notebooks were then implemented into the '*PL_Load_DW*' pipeline (Figure 3.10) previously described, automating forecast generation and ensuring the system produces timely forecast as new data becomes available.

The necessity to develop two notebooks instead of one is due to an error on maximum capacity reached for that session, which lead to populate the fact table in two steps with the copy data feature. Additionally, two auxiliary tables were created to mitigate errors in the ingestion of the notebooks' output parquet files.

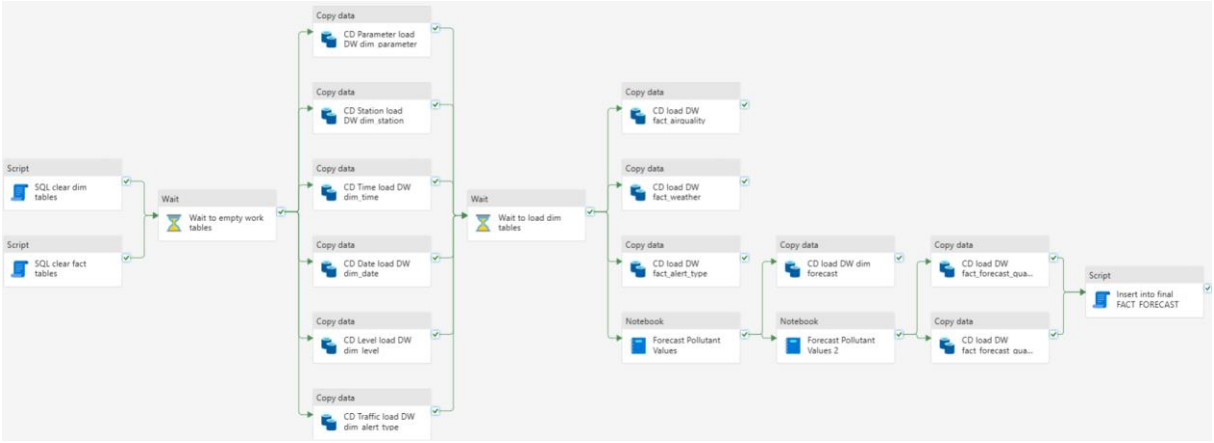


Figure 3.10 - Load Data Warehouse pipeline with the addition of the forecast notebook and tables.

The forecast values populate, through the pipeline, a table *FACT_FORECAST_VALUE*. Refer to table 3.18 for a description of the table fields.

Table 3.18 - *FACT_FORECAST_VALUE* fields and description.

| Field | Data type | Description |
|---|---|---|
| FK_DATE | INTEGER | Foreign key to date dimension. |
| FK_TIME | INTEGER | Foreign key to time dimension. |
| FK_PARAMETER | INTEGER | Foreign key to parameter dimension. |
| FK_MODEL | INTEGER | Foreign key to model dimension. |
| FORECAST_VALUE | FLOAT | Forecasted value for the date, time, parameter and model. |

Additionally, a new dimension table to identify which model the forecasted values refer to, was developed (Table 3.19).

Table 3.19 - *DIM_FORECAST* fields and description.

| Field | Data type | Description |
|---|---|---|
| SK_MODEL | INTEGER | Surrogate key of model dimension. |
| BK_MODEL_NAME | VARCHAR | Business key, the values refer to the name of the model with the forecasted values:<br>▪ 'SARIMA'<br>▪ 'PROPHET' |

Furthermore, the new tables were also subject to validation, through the rules previously established in Table 3.16, applied in '*PL_AQ_Validate_Forecast*' pipeline (Figure 3.11) and integrated into the master pipeline to automate the process. Similarly to the staging tables validation process, if there are any '*FAIL*' values in LOG_QUALITY_CHECKS table, an email is generated.

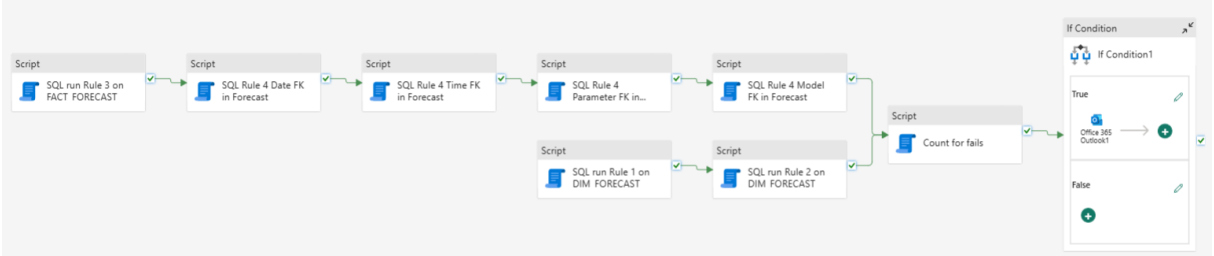


Figure 3.11 – Forecast tables validation pipeline.

These tables were also integrated into the semantic model, described in the further chapter.

## 3.7. SEMANTIC MODEL

The development of a semantic model denotes a crucial layer that ensures a bridge between raw structured data and the end user (Kimball & Ross, 2013). In this step business logic, relationships between tables, hierarchies and calculated measures are defined, and fields are selected meaningfully. Therefore, by organizing the data into a user-friendly model, it ensures consistency in calculations and simplifies report building. The following chapters focus on the development of the semantic model, as well as key steps to its construction.

The semantic model has a similar structure to the star schema model previously established (Figure 3.3) with three fact tables and six dimension tables, where location, date and time dimensions are shared between all three fact tables. Nevertheless, the fact and dimension forecast tables were added, with the fact table linked to the forecast, date and time dimensions (Figure 3.12).

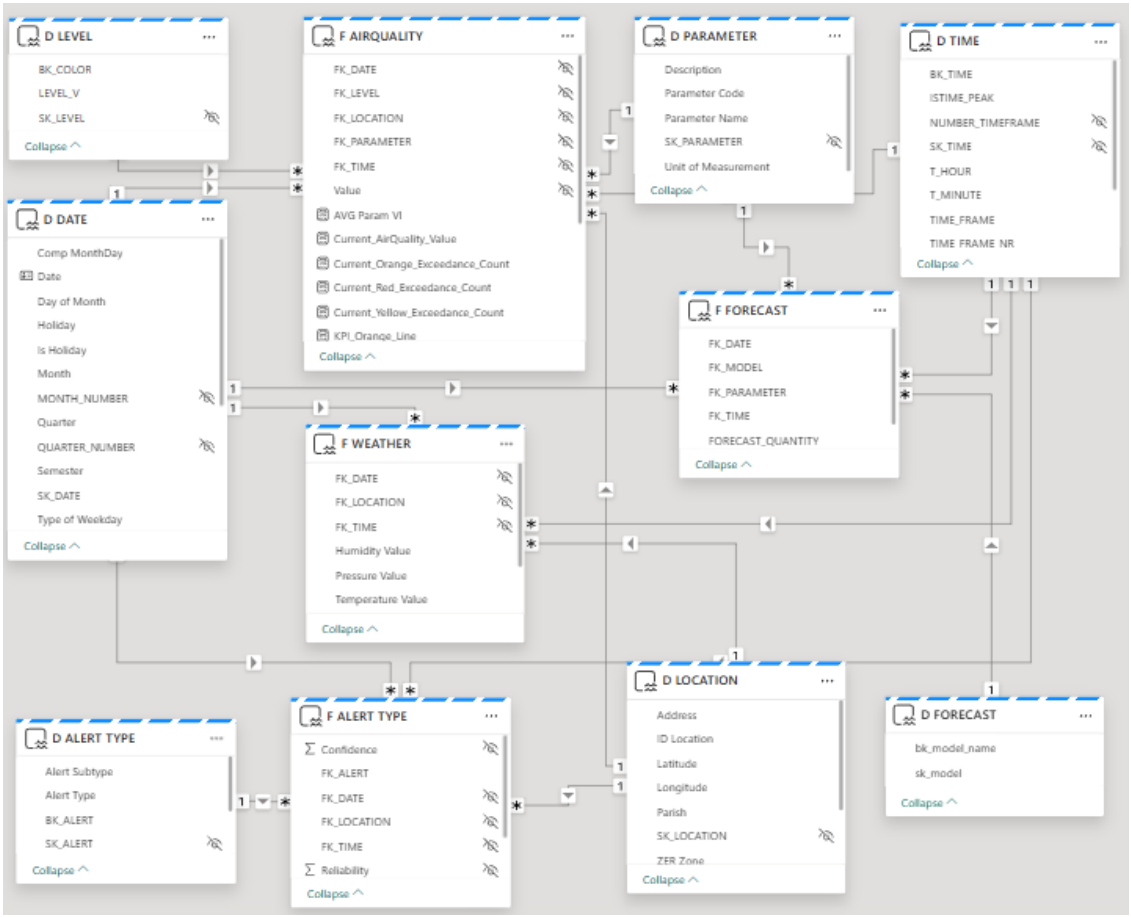


Figure 3.12 – Semantic model design.

### 3.7.1. RELATIONSHIPS

Ensuring the correct relationship between fact and dimension tables is crucial for accurate analysis (Table 3.20). These relationships are established based on the foreign keys in the fact tables and surrogate keys in the dimension tables, both integer fields.

Table 3.20 - Sematic model relationships between fact and dimension tables.

| Fact Table (column) | Relationship | Dimension Table (column) |
|---|---|---|
| Air Quality (FK_DATE) | Many to One | Date (SK_DATE) |
| Air Quality (FK_TIME) | Many to One | Time (SK_TIME) |
| Air Quality (FK_LOCATION) | Many to One | Location (SK_LOCATION) |
| Air Quality (FK_LEVEL) | Many to One | Level (SK_LEVEL) |
| Air Quality (FK_PARAMETER) | Many to One | Parameter (SK_PARAMETER) |
| Alert Type (FK_DATE) | Many to One | Date (SK_DATE) |
| Alert Type (FK_TIME) | Many to One | Time (SK_TIME) |
| Alert Type (FK_LOCATION) | Many to One | Location (SK_LOCATION) |
| Alert Type (FK_ALERT) | Many to One | Alert Type (SK_ALERT) |

| | | |
|---|---|---|
| Weather (FK_DATE) | Many to One | Date (SK_DATE) |
| Weather (FK_TIME) | Many to One | Time (SK_TIME) |
| Weather (FK_LOCATION) | Many to One | Location (SK_LOCATION) |
| Forecast (FK_DATE) | Many to One | Date (SK_DATE) |
| Forecast (FK_PARAMETER) | Many to One | Parameter (SK_PARAMETER) |
| Forecast (FK_TIME) | Many to One | Time (SK_TIME) |
| Forecast (FK_MODEL) | Many to One | Forecast (SK_MODEL) |

### 3.7.2. HIERARCHIES

From the data provided, there are four main hierarchies identified, regarding date, time location and alert types. These hierarchies are directly related to the granularity previously defined in chapter 3.4.2.

Regarding temporal granularity (Figure 3.13 and 3.14), the smallest granularity is the date, which provides the day of the transaction, followed by month, quarter, semester and year, enabling to capture useful trends in the data. Moreover, time granularity is described by hour and time frame.

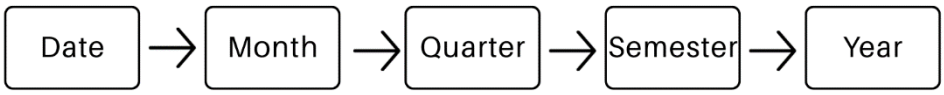


Figure 3.13 – Date dimension hierarchy definition.

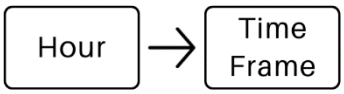


Figure 3.14 – Time dimension hierarchy definition.

Concerning geographical granularity (Figure 3.15), it is at street, or address level, followed by the parish, allowing for a deeper analysis on the location of each pollutant level.

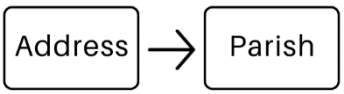


Figure 3.15 – Location dimension hierarchy definition.

Finally, the hierarchy for the alert type is defined by the alert subtype and type, allowing to analyse, if needed, the traffic alerts in different perspectives (Figure 3.16).

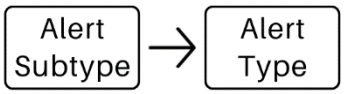


Figure 3.16 – Alert Type dimension hierarchy definition.

### 3.7.3. KEY PERFORMANCE INDICATORS AND CALCULATED MEASURES

Defining KPIs is an important step of the development of a BI model, as they help quantify relevant data points to achieve specific goals. According to Lisbon's council (Câmara Municipal de Lisboa, 2021), there are defined criteria for the parameters, in order to report accordingly when the values exceed the expected limits (Table 3.21). These standards will be used in this project as KPIs to help evaluate air quality over time.

Whereas for CO there is no specified threshold for the mentioned classification, values for the green, yellow, orange and red categories were established. As WHO (World Health Organization, 2021) updated their thresholds, for carbon monoxide, the hourly average for this pollutant remains valid since last update in 2005, set at 35 $mg/m^3$. Having into consideration that even short-term exposure can lead to several negative health effects (Manisalidis et al., 2020), 25 and 30 $mg/m^3$ were settled as yellow and orange threshold, respectively, and the WHO guideline of 1-hour average would be settled as the red level threshold.

Table 3.21 - Lisbon pollutant thresholds for Air Quality in $\mu g/m^3$. Adapted from Câmara Municipal de Lisboa (2021).

| | Air Quality Levels | | | |
|---|---|---|---|---|
| **Indicator** | **Green** | **Yellow** | **Orange** | **Red** |
| $NO_2$ | 0-100 | 101-200 | 201-400 | >401 |
| $O_3$ | 0-100 | 101-180 | 181-240 | >241 |
| $SO_2$ | 0-200 | 201-350 | 351-500 | >501 |
| $PM_{10}$ | 0-35 | 36-50 | 51-100 | >101 |
| $PM_{2.5}$ | 0-20 | 21-25 | 26-50 | >51 |
| CO | 0-2500 | 2501-3000 | 3001-3500 | >3500 |

Defining these thresholds is an important step for the monitoring of air quality and decision makers. By evaluating its results and identifying early existing fluctuations improves timely warnings to the population, helping to protect population health. Refer to Table A1 to the *dax* code leading to the development of the KPIs described previously. Each threshold is defined in individual variables, which assumes a different value based on the parameter selected.

Besides KPIs, some metrics were developed to help monitor pollutants' values over time and location (Table A2). Calculating the average of a pollutant allows for comparison across different regions or time periods. Moreover, the yellow thresholds exceeds count, as well as for orange and red thresholds, allows for a quantitative assessment on how often a certain pollutant concentrations reach concerning levels. Additionally, a current value for each parameter allows for a current air quality analysis, which enables the early detection of harmful pollution levels. Finally, the measure pollutant level label helps identify the level

(‘*NORMAL*’, ‘*Moderate*’, ‘*HIGH*‘ and ‘*VERY HIGH*’) each pollutant is categorized allowing to aggregate data, whether for temporal or region analysis.

The measures developed and KPIs acknowledged are applied within the semantic model.

# 4. RESULTS AND DISCUSSION

By applying various tools available in Power BI, along with the KPIs and calculated measures developed in the semantic model, it is possible to effectively address the business questions developed in the previous chapter 3.1. The follow-on dashboard[7] represents an interactive summary of the data, offering actionable insights of the current and historical air quality data in Lisbon, enabling decision makers to act in time according to air quality fluctuations.

## 4.1. AIR QUALITY CURRENT DATA

The first page of the dashboard simulates current air quality data (Figure 4.1). Although the data in this project is historical data, this page was included to serve as a reference for how air quality data can be visualized in a timely and accessible format, supporting immediate visualization across Lisbon. The goal of developing this page is also forward-looking, in a scenario where the dashboard is ever adapted for regular use with continuously updated data.

By selecting individually each pollutant, it is possible visualize the current air quality value for that pollutant, as well as the air quality status and warning counts for each threshold (yellow, orange and red). The bar chart allows for an analysis on the locations with higher values of air quality, and the bubble map of the dispersion of air quality across Lisbon. Finally, the bottom line chart provides the prediction for the next 24 hours with forecast data, supporting short-term air quality prediction and proactive decision-making.

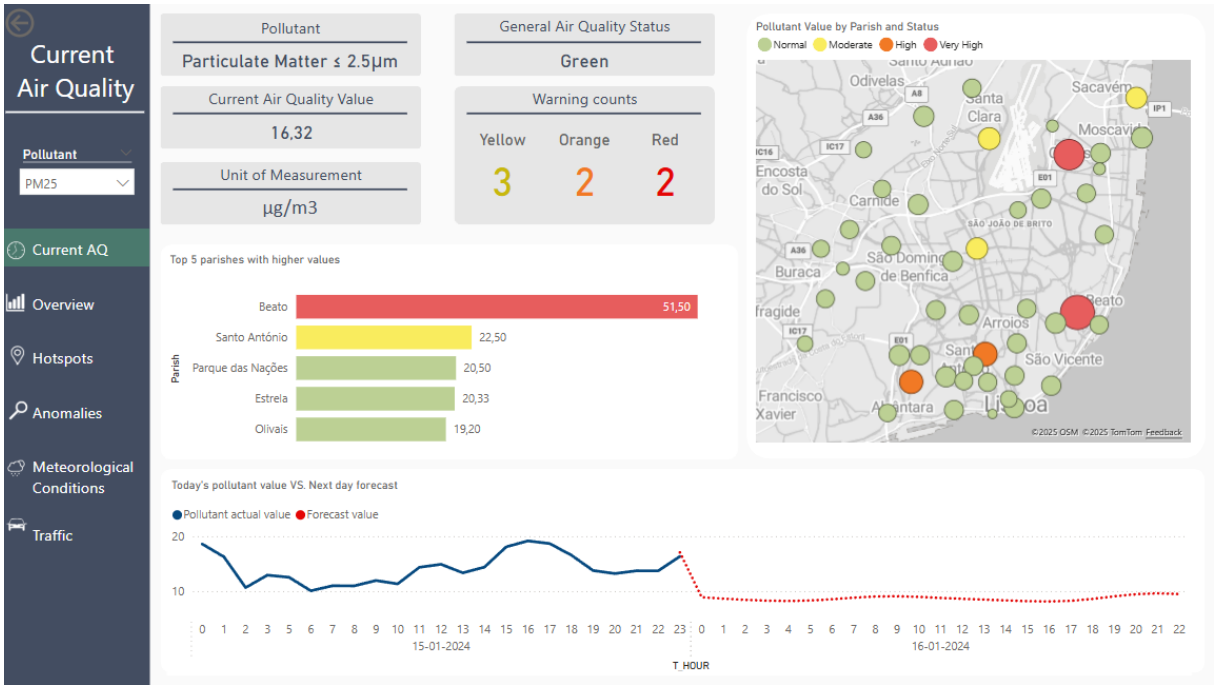


Figure 4.1 – Dashboard page 1 'Current Air Quality'.

[7] https://app.powerbi.com/links/VvyjRpczrg?ctid=e4bd69ff-e6f7-4c2e-b247-41b54ba2490e&pbi_source=linkShare&bookmarkGuid=79f4aed1-f709-485a-9f1e-2b818f154e19

Regarding air quality status, all five pollutants averages show normal air quality values. However this value represents an average for the current date and time values across Lisbon, making it crucial to analyse each stations individually to identify hotspots. The warning counts show that $PM_{10}$ and $PM_{2.5}$ are more propitious to surpass the god air quality limits.

Therefore, the bar plot identifies the five parishes experiencing higher values for each pollutant. The most affected parishes vary depending on the pollutant, however, Lumiar, Beato and Santo António parishes stand out, where high pollutant levels are more frequently observed (CO, $PM_{2.5}$, $O_3$). Additionally, $NO_2$ is highest in Arroios and Lumiar, while Estrela and Lumiar tops the list for $PM_{10}$.

Moreover, the map visualization helps detect pollution clusters or geographic hotspots, allowing for comparing across different locations. Thus, Carbon Oxide, Nitrogen Dioxide and Ozone show normal values across the Lisbon area. However, Particulate Matter ($PM_{10}$ and $PM_{2.5}$) are shown to be more prevalent in terms of geographic spread.

Finally, the line chart provides air quality trend over the past 24 hours, as well as forecast for the next 24 hours, allowing early detection of deviations.

Table 4.1 depicts the first business need and the answers to each business question.

Table 4.1 - Current Air Quality (BN 1) insights.

| **Business Question** | **Answer** |
|---|---|
| What are the current pollution levels? | CO -> 0.21 $mg/m^3$ (normal)<br>$NO_2$ -> 33.86 $\mu g/m^3$ (normal)<br>$O_3$ -> 58.60 $\mu g/m^3$ (normal)<br>$PM_{10}$ -> 28.49 $\mu g/m^3$ (normal)<br>$PM_{2.5}$ -> 16.32 $\mu g/m^3$ (normal)<br>These values concern the average current value for Lisbon. |
| Which locations experience the highest levels of pollution? | CO -> Beato; Olivais.<br>$NO_2$ -> Arroios.<br>$O_3$ -> Beato.<br>$PM_{10}$ -> Estrela; Lumiar.<br>$PM_{2.5}$ -> Beato. |
| What pollutants are more prevalent? | $PM_{10}$ and $PM_{2.5}$. |

With the insights on what pollutants have higher levels allows for a better understanding of the specific precautions needed based on the most frequent pollutant. Moreover, knowing which locations have those high values enables to alert the population about regions where greater caution is advised or even avoid outdoor exposure.

## 4.1. Air Quality Overview

The second page of the dashboard (Figure 4.2) provides an overview of air quality evolution over time. Its main purpose is to highlight temporal patterns in pollution levels and to identify periods of increased health risks. The dropdown allows for a selection of the parameter and a deeper individual analysis.

The bar chart shows how many times the selected pollutant exceeded the yellow, orange and red thresholds each month, enabling seasonal risk analysis. The three line chart allow for temporal analysis of each pollutant by hour, weekday and month.

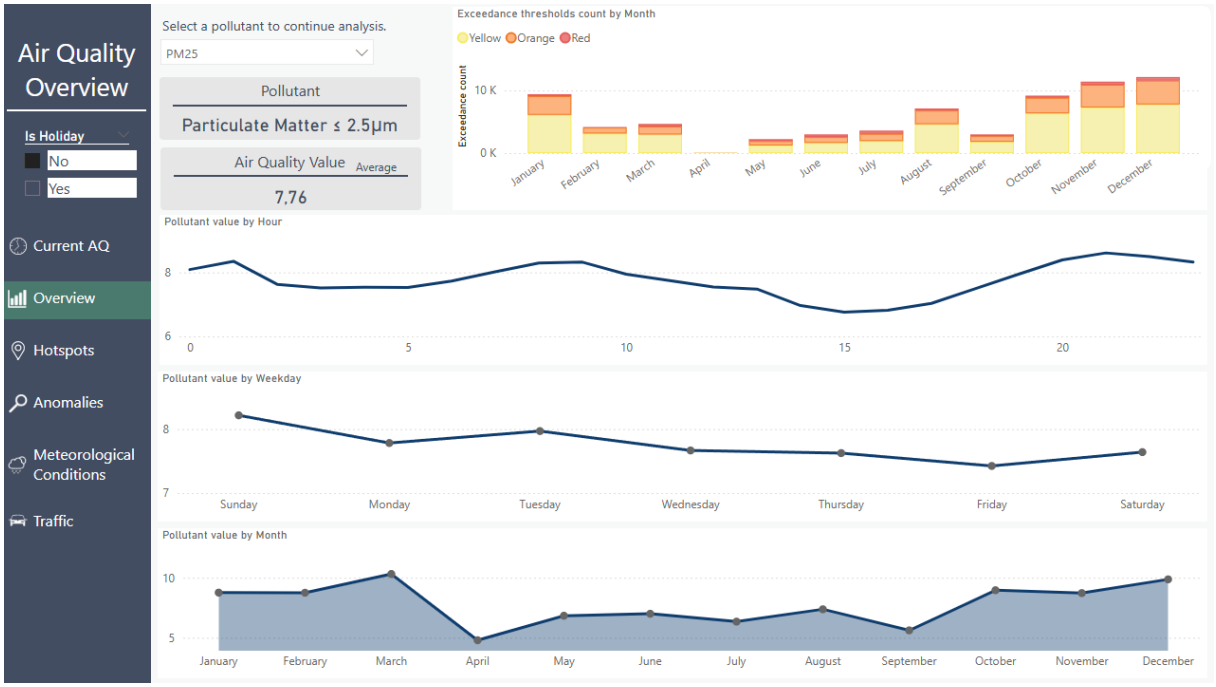


Figure 4.2 - Dashboard page 2 'Air Quality Overview'.

Through the analysis of the bar plot, with the count of cases exceeding air quality limits, $NO_2$ exhibits the highest frequencies exceeding the yellow limit over 10 thousand times almost every month from 2021 to 2023, followed by $O_3$. Although CO rarely exceeds threshold limits, the majority of the surpassing occurs in the heat months (June, July and August). Finally, both Particulate Matter pollutants show spikes during winter months.

Regarding hourly evolution, $NO_2$, CO, $PM_{10}$ and $PM_{2.5}$ values peak during morning and evening rush hours, which aligns to traffic patterns and hours with more movement in the city. However, $O_3$ peaks in the early afternoon, consistent with the sunlight.

Concerning weekdays, $NO_2$ and CO show higher values on weekdays, reinforcing high traffic intensity influence. $O_3$ on the other hand, show higher values during the weekends, and $PM_{2.5}$ and $PM_{10}$ indicate decreasing values along the week.

Finally, the last time evolution plots concern monthly analysis. Most pollutants indicate higher values during winter months (October to February) and decreases during spring and summer months. However, $O_3$ show higher values in the high temperature months (April to July).

This page of the dashboard also allows for a holiday selection, 'Yes' for analysing the data where it is a special day, such as holidays celebrated in Lisbon, and 'No' otherwise. When selecting 'Yes' there is mainly a difference in CO and $NO_2$ which can be associated with less vehicle emissions and industrial activities.

Refer to Table 4.2 for the remaining business questions and respective answers for Business Need 1.

Table 4.2 - Air Quality Overview (BN 1) insights.

| **Business Question** | **Answer** |
|---|---|
| How do pollutant concentrations fluctuate over time? | Hourly, daily and monthly analysis. |
| How often do pollutant levels exceed the maximum expected values? | CO -> Summer months. Rare events (<20 cases per month).<br>$NO_2$ -> Similar throughout the year (~10K per month).<br>$O_3$ -> Similar throughout the year (~5K per month).<br>$PM_{10}$ -> Winter months (>10K per month).<br>$PM_{2.5}$ -> Winter months (>5K per month). |
| Do holidays influence air quality? | Slight positive impact in $NO_2$ an CO. |

Understanding how and when pollution spikes occur is crucial for urban planning, public awareness and to apply thoughtful temporal regulations. Therefore, peak hours usually mean higher pollutant values as well as winter months.

## 4.2. HOTSPOTS AND HIGH RISK AREAS

The page 'Hotspots' of the dashboard focuses on highlighting the areas in Lisbon with the most critical air quality concerns (Figure 4.3), designed to answer business need 2 defined in chapter 3.1. By selecting a specific pollutant, users can visualize the parishes with higher values, a distribution of the averages per parish in a heat map, and the parishes with the higher threshold exceeds for that pollutant. Additionally, the location hierarchy allows for a drop-down approach in the top bar plot, enabling both parish and street analysis.

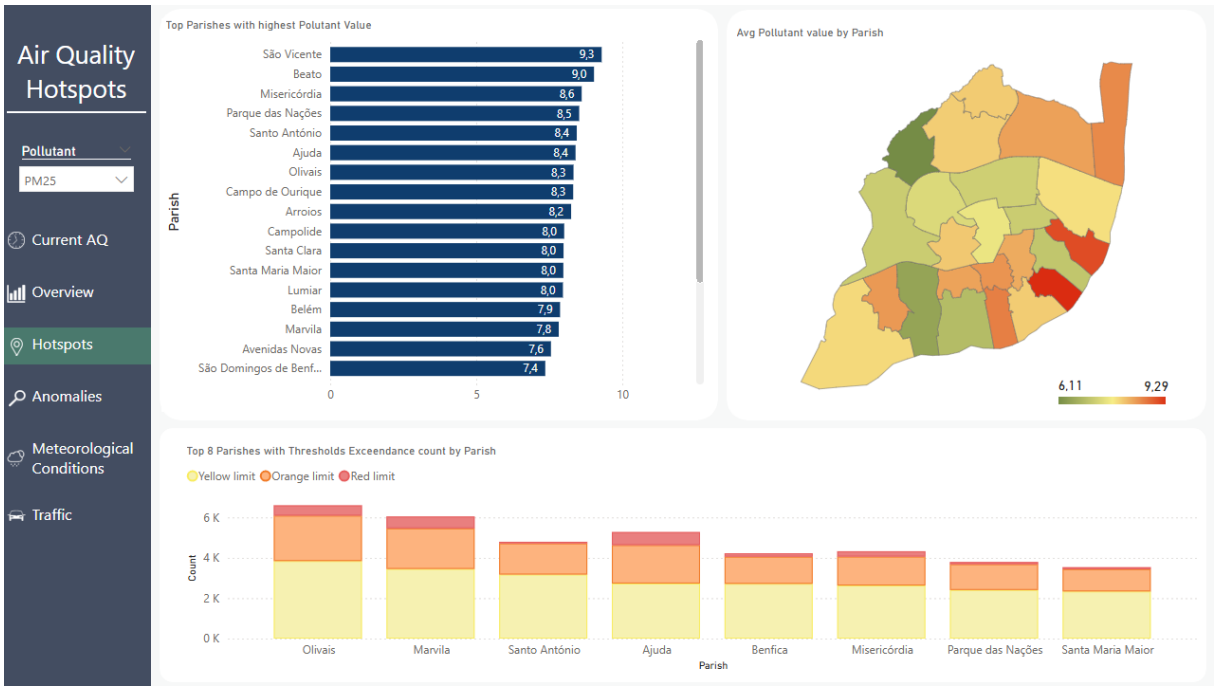


Figure 4.3 - Dashboard page 3 ‘Hotspots’.

The top two bar plots identify the parishes and locations with higher values per pollutant, therefore with consistent high values. By analysing the concentration values, Beato frequently has higher values (CO, $PM_{10}$ and $O_3$) as well as São Vicente (CO, $PM_{10}$ and $NO_2$) and Campo de Ourique ($PM_{2.5}$). For $NO_2$ the presence of high values in parishes located near major roads and highways, such as Lumiar, aligns with the known traffic-related pollution behaviour (Borrego et al., 2004).

Additionally, shape maps visually identify hotspots and high risk areas, allowing for a deeper analysis on each pollutant dispersion across Lisbon. For instance, while $O_3$ values are more widespread affecting both central and peripheral areas, $NO_2$ highest values are more located in areas with dense traffic. $PM_{2.5}$, on the other hand, has higher values more centrally located.

Finally, the bottom bar chart reveals that for pollutants like $PM_{10}$ and $PM_{2.5}$, parishes such as Olivais and Marvila dominate the frequency exceeds, which is consistent with these parishes being near main traffic and industrial zones.

Table 4.3 answers the business questions related to air quality hotspots in a summarized table.

Table 4.3 - Hotspots and High Risk Areas (BN 2) insights.

| **Business Question** | **Answer** |
|---|---|
| Which locations exceed air quality thresholds most frequently? | CO -> Rare cases.<br>$NO_2$ -> Lumiar and Alvalade.<br>$O_3$ -> Marvila and Alvalade.<br>$PM_{10}$ -> Olivais and Marvila. |

| | |
|---|---|
| | $PM_{2.5}$ -> Olivais and Marvila. |
| Are certain areas more affected by specific pollutants? | CO -> São Vicente and Avenidas Novas.<br>$NO_2$ -> Lumiar, Alvalade and Arroios.<br>$O_3$ -> Beato, Benfica and Parque das Naçoes.<br>$PM_{10}$ -> Arroios and Beato.<br>$PM_{2.5}$ -> São Vicente and Beato. |
| Which locations consistently report high levels of air pollution? | Beato, São Vicente and Lumiar. |

The visual patterns align with known urban dynamics validating the relevance of the insights. In Lisbon, different pollutants affect different areas, however identifying pollution hotspots with consistent high levels reported can support targeted environmental strategies, health advisories, and long-term planning.

### 4.3. ANOMALIES

This page of the dashboard aims to identify unusual spikes and deviation in pollutant concentrations over time, and to flag them as extreme events or sensor errors. Although most outlier values were already identified and dropped from the analysis in the previous ETL phase, this page can help target some cases that were not identified yet. Including the yellow and orange limits it helps identify these peak values easier by putting them into the context of the limits they surpass. Additionally, location and time filters, through the selection of parish and month, was added to allow for a more granular analysis.

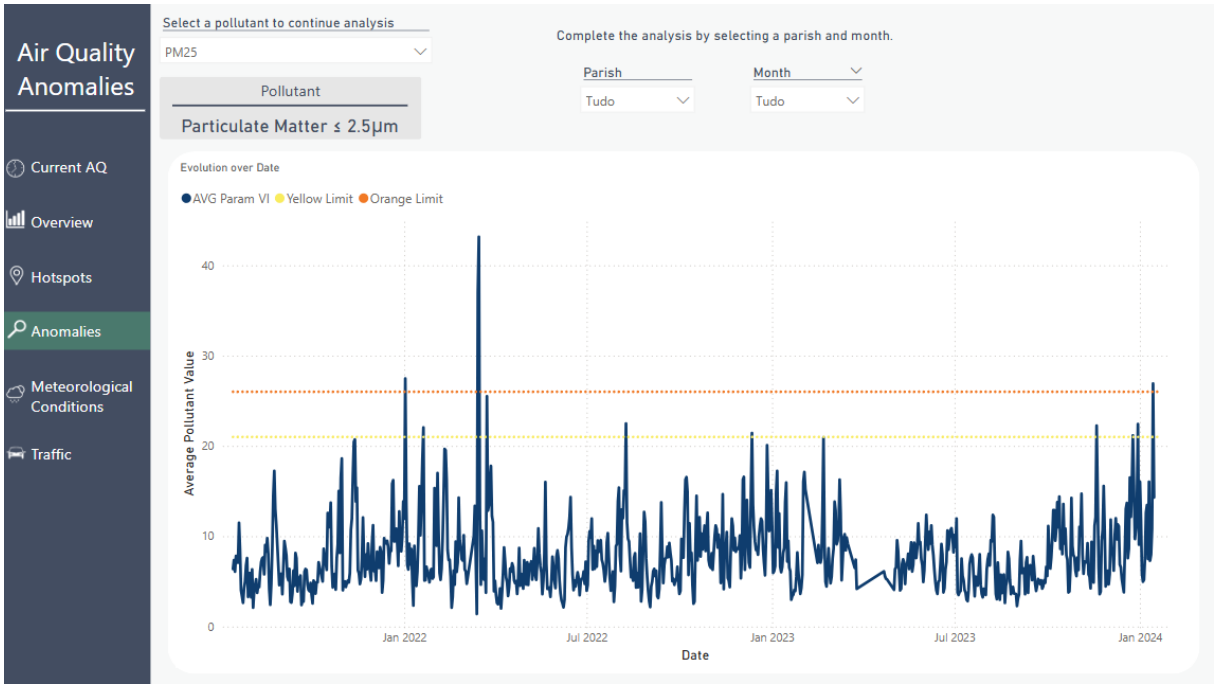


Figure 4.4 - Dashboard page 4 ‘Air Quality Anomalies’.

For $PM_{2.5}$ and $PM_{10}$ a spike is visible between March 15 and 17, 2022, where values exceeded 40 µm/$m^3$ in Lisbon, which is above the orange threshold line and typical background levels. This anomaly is likely linked to a known Saharan dust intrusion[8] event that affected Spain and large parts of Portugal, including Lisbon. $NO_2$ also has a peak during this time period, which might have been caused by the dust intrusion, preventing the dispersion of local pollutants. The remaining pollutants, CO and $O_3$, have periodic peaks however with low impact.

Regarding sensor errors, there is no extremely high or low values that suggest possible monitoring faults. Table 4.4 answers the business question previously established for anomaly analysis.

Table 4.4 - Anomalies (BN 3) insights.

| **Business Question** | **Answer** |
|---|---|
| Are there unusual spikes in pollution? | Peak values mid-March, 2022, due to dust intrusions. |

## 4.4. METEOROLOGICAL CONDITIONS IMPACT

The fifth page of the dashboard entitled 'Air Quality vs Weather' aims to explore the relationship between specific meteorological conditions variables, temperature and humidity, and air quality across different pollutants (Figure 4.5). By analysing weather monthly trends and compare them to air quality trends, enables to uncover how environmental conditions influence pollutant behaviour.

Regarding temperature, CO and $NO_2$ values tend to increase in cooler months, which aligns with high traffic activity in the city. Moreover, $O_3$ shows a possible strong positive relation with temperature, where months with high temperature values match with high Ozone values, consistent with its photochemical nature. Particulate Matter tend to increase during cooler months, but do not show a strong temperature dependency.

Humidity wise, CO and $O_3$ show negative relation with relative humidity values, where when humidity is high these pollutant values are low and the other way around. Furthermore, $NO_2$, $PM_{10}$ and $PM_{2.5}$ have a positive correlation with humidity, showing higher impact in $NO_2$.

[8] https://observador.pt/2022/03/15/os-ceus-de-portugal-comecam-a-ficar-pintados-de-laranja-com-a-chegada-da-chuva-de-barro/


Figure 4.5 - Dashboard page 5 'Air Quality vs Weather'.

Refer to table 4.5 for a summarized answer to the business questions for business need 4.

Table 4.5 – Weather impact (BN 4) insights.

| Business Question | Answer |
|---|---|
| How does different meteorological conditions affect pollution levels? | High temperatures are linked to increased $O_3$ and lower CO and $NO_2$ values.<br>High humidity is associated with high $NO_2$ as well as particulate matter values; whereas low humidity is linked to higher $O_3$ values. |
| What meteorological conditions are most associated with poor air quality events? | Vary by pollutant.<br>Hot and dry conditions -> high $O_3$ values.<br>Cold and dry conditions -> high CO, $PM_{10}$ and $PM_{2.5}$. |

This page of the dashboard show that meteorological factors have impact in air quality, namely temperature and relative humidity.

## 4.5. Traffic Congestion Impact

The final page of the dashboard is dedicated to understand the impact of vehicle activity on air quality, focusing on traffic congestion (Figure 4.6). Before exploring this page, the user has to select a pollutant to analyse. Once selected, it aims to help understand if Reduced Emission Zones have a significant impact in air quality by comparing them with zones with no vehicle restrictions. Moreover, the bar charts compare pollutant concentrations and traffic alert volumes between weekdays and weekends, and across different times of the day. Additional

filters such as to identify if is peak hour (Yes/No) or holiday (Yes/No) allow for further analysis on traffic congestion impact during rush hours or non-labour days.


Figure 4.6 - Dashboard page 6 'Air Quality vs Alert Type'.

Regarding the restricted traffic zones (zone 1 and 2), zone 2 generally shows higher pollutant values, although the differences are not significant. These results reveals that this measure does not have a positive impact in Lisbon's air quality, as the restricted access zones show equal or even higher pollutant values compared to unrestricted zones (Zone 0).

The bar chart comparing weekdays to weekends reveals that pollution levels, especially for traffic-related pollutants like $NO_2$ and CO, are significantly higher during weekdays, in line with commuter traffic intensity. Similarly, time-of-day analysis further reinforces these findings, although jam alerts peak in the afternoon, pollutant concentrations often rise in the evening and night. This pattern is especially visible for $PM_{10}$, $PM_{2.5}$, and $NO_2$.

Moreover, regarding holidays, the jam alerts count increases during the night and late-night periods, between 9 PM and 5 AM, compared to other times of the day. Similarly, there is a noticeable improvement in air quality during the late-night period for CO, $PM_{10}$, and $PM_{2.5}$ when compared to non-holiday periods, which can be due to less traffic jam. The remaining periods, especially for $NO_2$, reveal a slight decrease in air quality values during holidays.

Table 4.6 summarizes the business questions and answers for business need 5.

Table 4.6 – Traffic congestion (jam alerts) impact (BN 5) insights.

| Business Question | Answer |
|---|---|

| | |
|---|---|
| How do air pollution levels change during peak traffic hours compared to non-peak hours? | Main impact in $NO_2$, CO and $O_3$, where air pollution levels increase during peak traffic hours. |
| Is there a strong correlation between vehicle flow and $NO_2$ or CO concentrations? | Yes, CO and $NO_2$ have significant higher values during weekdays rather than weekends, similar behaviour for $NO_2$ during holidays. |
| How does reduced traffic during weekends or holidays affect air quality? | Weekends and holidays show a positive impact in $NO_2$ and CO. |
| Can traffic control measures (e.g., low-emission zones) lead to measurable improvements in air quality? | No significant impact in air quality in the reduced emission zones in Lisbon. |

Together, these visuals show that urban air quality is closely tied to traffic volume and timing, highlighting the need for emissions control strategies that target rush hours and high-density zones.

## 4.6. FORECASTING

As already mentioned, two forecasting models were implemented to simulate short-term air pollution concentrations, SARIMAX and Prophet, both applied for the five selected pollutants. Therefore, an individual data frame was created for each pollutant, using as input timestamps in the 'YYYY-MM-DD HH:MM:SS' format, and the data was grouped in advanced by date and time. The location factor was not considered, instead the predictions were generated as a single aggregated value for the entire city, rather than producing separate forecasts for each individual location. This approach simplifies the forecasting process by reflecting citywide trends.

To evaluate model performance, the dataset was split into training and testing sets, thus the last 200 observations were reserved for testing and the remaining were used to train the models, representing just over 8 days.

### 4.6.1. SARIMA MODEL RESULTS

To define the parameter for each model, a loop of combinations of p, d and q was developed in order to select the better one. Table 4.7 describe the parameters selection for each pollutant for the SARIMA model. It is worth noting that for this approach the external variables component with influence in the pollutants was not included.

Table 4.7 – SARIMA model parameters definition for each parameter.

| Parameter | Order | Seasonal Order |
|---|---|---|

| | | |
|---|---|---|
| $NO_2$ | (0,1,4) | (0,1,1,24) |
| $O_3$ | (1,1,1) | (1,1,1,24) |
| $PM_{10}$ | (1,1,1) | (1,1,1,24) |
| $PM_{2.5}$ | (2,1,1) | (1,1,1,24) |
| CO | (0,1,4) | (1,1,1,24) |

SARIMA models demonstrated fair predictive performance in capturing general seasonal trends and smoother fluctuations. However, in cases with more abrupt variations, such as for $PM_{10}$ and $PM_{2.5}$, the model under predicts extreme events (Figure B.1). Figure 4.7 includes the comparison between real and forecasted values for the last 8 days for $NO_2$.

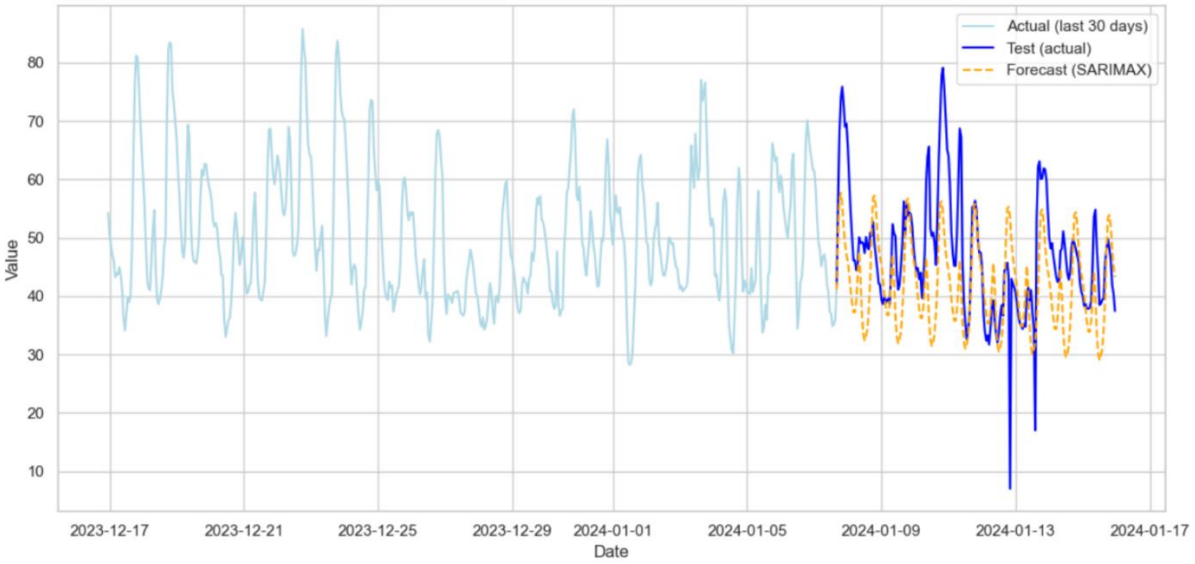


Figure 4.7 - Real vs Forecast values with SARIMA Model for $NO_2$.

### 4.6.2. PROPHET MODEL RESULTS

The Prophet model, developed by Facebook was applied to hourly pollutant averages, with different regressors for each pollutant. After correlation analysis and different tests (Table B.1), table 4.8 summarizes the selected variables for each parameter prophet model.

Table 4.8 – Prophet models regressors selection per parameter.

| **Parameter** | **Regressor** |
|---|---|
| $NO_2$ | IS_HOLIDAY; HUMIDITY; PRESURE, TEMPERATURE |
| $O_3$ | IS_HOLIDAY; HUMIDITY; TEMPERATURE |
| $PM_{10}$ | IS_HOLIDAY; HUMIDITY; TEMPERATURE |
| $PM_{2.5}$ | IS_HOLIDAY; HUMIDITY |
| CO | IS_HOLIDAY; HUMIDITY |

The Prophet model showed varying levels of performance across different pollutants, with a general tendency to underestimate short-term spikes, similar to SARIMA model results. For $NO_2$ (Figure 4.8) and $O_3$, the models maintained an accurate seasonal pattern, aligning with daily fluctuations, lagging behind only on higher peaks. However, for CO, $PM_{10}$ and $PM_{2.5}$, the forecasts were smooth compared to actual values (Figure B.2), with low responsiveness in identifying sharp anomalies.

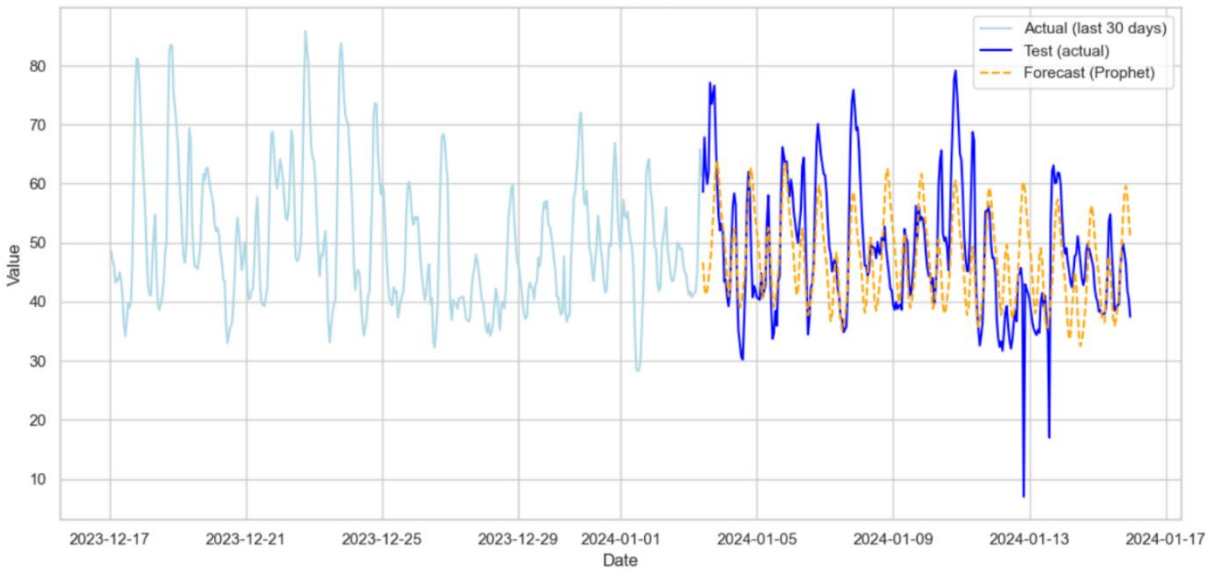


Figure 4.8 – Real vs Forecast values with Prophet Model for $NO_2$.

### 4.6.3. Comparative Analysis

Based on the evaluation metrics, MAE and RMSE, Prophet model outperforms SARIMA across most pollutants. Particularly for $NO_2$, $O_3$ and $PM_{10}$, this model notably shows lower errors; for instance, the MAE values for $O_3$ is 13.72 in SARIMA while 9.58 in Prophet, and the RMSE is 15.81 against 12.24 in Prophet (Table 4.9).

Table 4.9 – SARIMA vs Prophet Evaluation metrics.

| | **MAE** | | **RMSE** | |
|---|---|---|---|---|
| | SARIMA | Prophet | SARIMA | Prophet |
| $NO_2$ | 8.25 | 7.54 | 10.80 | 9.68 |
| $O_3$ | 13.72 | 9.58 | 15.81 | 12.24 |
| $PM_{10}$ | 12.46 | 11.64 | 19.21 | 16.57 |
| $PM_{2.5}$ | 5.58 | 5.54 | 8.21 | 7.85 |
| CO | 0.12 | 0.12 | 0.15 | 0.14 |

The previous visual evaluation analysis, also show that while Prophet sometimes underestimates extreme spikes, the SARIMA models often have smoothed forecast line and struggles to capture higher peaks.

Based both on the quantitative measures and visual evaluation previously conducted, the Prophet model was incorporated into the final dashboard solution.

## 4.7. DISCUSSION

The findings from this study seek to highlight the locations in Lisbon with high impact in air quality, identify critical time periods of the day and the year, and evaluate how meteorological conditions and traffic congestion in air pollution. Thus, through deeper analysis it is possible to obtain valuable insights to take timely action.

Therefore, the results aligns with several studies previously mentioned. The recognition of $NO_2$ as one of the most frequently exceeded pollutants corresponds with the fact that vehicular and combustion activities are major sources of air pollution, as highlighted by Sicard et al. (2021) and Wajeetongratana (2023). Additionally, the results revealed rush hour peaks for $NO_2$ and CO, as well as stronger pollution levels in traffic-dense parishes like Olivais, Lumiar and Marvila, and seasonal increases in winter for particulates and in summer for $O_3$. These patterns reinforce the idea that air quality is highly influenced by urbanization (Jena et al., 2023) and by both traffic patterns and meteorological conditions (Mao et al., 2021; Wang et al., 2021).

However, upon analysis on policy measures employment, namely Reduced Emission Zones in Lisbon, new insights can be derived. Despite the implementation of these zones, the results suggest no significant improvements in air quality when comparing restricted areas to unrestricted ones. This contrast with Ferreira et al. (2015), who shows how the traffic measures implemented in Lisbon have had a positive impact in Lisbon air quality over the years. As Datia et al. (2022) highlighted that spatial policy measures require better enforcement and real-time monitoring to truly make an impact, aligning with the results.

Finally, while many studies utilize deep learning (Athira et al., 2018; Mao et al., 2021), the use of time series forecasting with Prophet model and integration into the BI framework offers an accessible tool for decision-makers.

The interactive nature of the dashboard contributes to existing work by providing a scalable and replicable solution, focusing on multiple business questions, along with geographic and temporal analysis, and short-term predictions. Moreover, this study demonstrates how BI tools can successfully translate complex air quality data into actionable insights.

# 5. CONCLUSIONS AND FUTURE RESEARCH

The implemented BI framework aims to provide a data-driven, interpretable and actionable solution for air quality data in Lisbon. It integrates multiple datasets, ranging from pollutant concentrations to meteorological conditions and traffic patterns, into a unified BI pipeline, incorporating a full data flow.

The main final contribution of this study is a complete framework leading to a Power BI visualization tool that not only provides monitoring of pollutant levels but also incorporates spatial and temporal breakdowns, and comparative analyses across multiple dimensions, such as REZ zones, weekdays, and time periods. Moreover, the development and integration of KPIs leveraged the analysis on AQI trends, exceed counts and hotspot detection. This solution also incorporates short-term forecasting, merging both visualization and prediction components. Thus, this study aims to support informed decision-making regarding urban management and public health, as the versatile nature of this solution allows for it to be extended to other cities or scenarios.

Despite the results, this study is subject to a number of limitations. One of the main limitations is in regard of data availability and coverage. The air quality dataset covers only a short historical period, which limited analysis over the years specially to assess traffic control measures. Additionally, missing data for the sulfur dioxide parameter prevented its inclusion in the analysis. Similarly, the meteorological data was restricted to the availability of certain factors, as most of the meteorological variables had predominantly null values. This lack of completeness reduced the robustness of the analysis and limited deeper insight into the interactions between pollution and weather conditions. Regarding both air quality and meteorological datasets, although they had the same structure and originate from the same sensor network, the volume of available meteorological data was noticeably lower. In many cases, air quality records did not have corresponding weather data for the same timestamp and location. This mismatch limited the scope of analysis between pollutants and meteorological conditions, reducing the number of usable observations and potentially affecting the representativeness of the results in certain external variable evaluations. Also, the semantic model is not automatically refresh, as the sensor data is not received in a hurly or daily basis, restraining the possibility to maintain real-time analysis. Regarding Machine Learning models development, due to large volume of data and computational demands, memory limitations in the notebook session often led to the kernel crashing. This made it necessary to split the model execution across separate notebooks reducing efficiency. Lastly, due to time constraints, it was not possible to deeply optimize or fine-tune parameters and test alternative approaches. As a result, while the forecast adds value, its performance could be improved with further testing and validation.

While this study successfully developed a BI solution for air quality in Lisbon, future research can incorporate several improvements to the current solution. Firstly, it would be interesting

to explore additional environmental factors, as well as other factors expected to affect air quality, such as industrial or agricultural activities. Additionally, the solution would benefit from expanding the historical data window. The current analysis is limited to a few years, by incorporating data across a broader historical period, it would enable year evolution analysis to better assess the impact of environmental regulations and detect shifts in pollutant behaviour. Despite the efforts to automate the dashboard, some improvements could simplify even more the process. For instance, creating a dynamically updated table with public holidays for each year and linking it to the date dimension would allow for further years to be analysed without manual intervention. Moreover, establishing an Application Programming Interface (API) would enable the integration of updated air quality data directly into Microsoft Fabric and the entire BI process to be run continuously with minimal manual input. This would make the solution suitable for daily monitoring and rapid response scenarios. Finally, it would be of interest to test different prediction approaches that would fit better the available data and amplify the predictions per location, rather than at the city level.

## BIBLIOGRAPHICAL REFERENCES


Abelsohn, A., & Stieb, D. M. (2011). Health effects of outdoor air pollution: Approach to counseling patients using the Air Quality Health Index. *Canadian Family Physician*, *57*(8), 881–887.

Alrashed, S. (2020). Key performance indicators for Smart Campus and Microgrid. *Sustainable cities and society*, *60*, 102264.

Alvarado, M. J., McVey, A. E., Hegarty, J. D., Cross, E. S., Hasenkopf, C. A., Lynch, R., Kennelly, E. J., Onasch, T. B., Awe, Y., & Sanchez-Triana, E. (2019). Evaluating the use of satellite observations to supplement ground-level air quality data in selected cities in low-and middle-income countries. *Atmospheric Environment*, *218*, 117016.

Anurogo, D., Sulaeman, S., Yamtana, Y., & Andarmoyo, S. (2023). Assessing the Impact of Air Quality on Respiratory Health in Urban Environments: A Case Study of Tangerang. *West Science Interdisciplinary Studies*, *1*(10), 940–951.

Athira, V., Geetha, P., Vinayakumar, R., & Soman, K. (2018). Deepairnet: Applying recurrent networks for air quality prediction. *Procedia computer science*, *132*, 1394–1403.

Aziz, A., Saha, S., & Arifuzzaman, M. (2021). Analyzing Banking Data Using Business Intelligence: A Data Mining Approach. *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, 245–256.

Borra, P. (2024). Microsoft Fabric Review: Exploring Microsoft's New Data Analytics Platform. *International Journal of Computer Science and Information Technology Research*, *12*(2), 34–39.

Borrego, C., Tchepel, O., Salmim, L., Amorim, J. H., Costa, A. M., & Janko, J. (2004). Integrated modeling of road traffic emissions: Application to Lisbon air quality management. *Cybernetics and Systems: An International Journal*, *35*(5–6), 535–548.

Box, G. E., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, *70*(349), 70–79.

Câmara Municipal de Lisboa. (2021). *Metainformação relativa à Monitorização de Parâmetros Ambientais da Cidade de Lisboa*. https://www.lisboa.pt/temas/ambiente/qualidade-ambiental/ar

Cavalheiro, J., & Carreira, P. (2016). A multidimensional data model design for building energy management. *Advanced Engineering Informatics*, *30*(4), 619–632.

Dashkevych, O., & Portnov, B. A. (2023). Does city smartness improve urban environment and reduce income disparity? Evidence from an empirical analysis of major cities worldwide. *Sustainable Cities and Society*, *96*, 104711.

Datia, N., Pato, M., Taborda, R., & Pires, J. M. (2022). ML Approach to Predict Air Quality Using Sensor and Road Traffic Data. Em *Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery* (pp. 379–401). Springer.

De Vito, S., Esposito, E., Massera, E., Formisano, F., Fattoruso, G., Ferlito, S., Del Giudice, A., D'Elia, G., Salvato, M., & Polichetti, T. (2021). Crowdsensing IoT architecture for pervasive air quality and exposome monitoring: Design, development, calibration, and long-term validation. *Sensors*, *21*(15), 5219.

Delgado, A., Rosas, F., & Carbajal, C. (2019). System of business intelligence in a health organization using the kimball methodology. *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, 1–5.

Des Voeuz, H., & Owens, J. (1912). The Sootfall of London: Its Amount, Quality, and Effects. *The Lancet*, *179*, 47–50. https://doi.org/10.1016/S0140-6736(00)51732-2

Espinosa, R., Palma, J., Jiménez, F., Kamińska, J., Sciavicco, G., & Lucena-Sánchez, E. (2021). A time series forecasting based multi-criteria methodology for air quality prediction. *Applied Soft Computing*, *113*, 107850.

Ferreira, F., Gomes, P., Tente, H., Carvalho, A., Pereira, P., & Monjardino, J. (2015). Air quality improvements following implementation of Lisbon's Low Emission Zone. *Atmospheric Environment*, *122*, 373–381.

Jena, M. C., Mishra, S. K., & Moharana, H. S. (2023). Challenges and way forward to maintain air quality standard in urban areas. *The Global Environmental Engineers*, *10*, 33–43.

Karavas, Z., Karayannis, V., & Moustakas, K. (2021). Comparative study of air quality indices in the European Union towards adopting a common air quality index. *Energy & Environment*, *32*(6), 959–980.

Kelly, C., Fawkes, J., Habermehl, R., de Ferreyro Monticelli, D., & Zimmerman, N. (2023). PLUME Dashboard: A free and open-source mobile air quality monitoring dashboard. *Environmental Modelling & Software*, *160*, 105600.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling*. John Wiley & Sons.

Koolen, C. D., & Rothenberg, G. (2019). Air pollution in Europe. *ChemSusChem*, *12*(1), 164–172.

Landütama, J. F., & Chowanda, A. (2023). Applied design thinking for kimball lifecycle to improve business intelligence dashboard usability. *International Journal of Innovative Computing, Information and Control*, *19*(4), 1139–1152.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, *525*(7569), 367–371.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in public health*, *8*, 14.

Mao, W., Wang, W., Jiao, L., Zhao, S., & Liu, A. (2021). Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustainable Cities and Society*, *65*, 102567.

Oyo-Ita, E., Ekah, U., Ana, P., & Ewona, I. (2023). Development of a Smart Air Quality Monitoring System Using Wireless Sensors. *Advances in Research*, *24*(6), 50–59.

Seibert, O. G., Pinto, W. de P., & Monte, E. Z. (2022). Índice de poluição atmosférica: Uma proposta baseada em dados secundários para avaliação da qualidade do ar. *Engenharia Sanitaria e Ambiental*, *27*(6), 1209–1219.

Shikwambana, L., Kganyago, M., Mbatha, N., & Mhangara, P. (2024). First time calculation of the spatial distribution of concentration and air quality index over South Africa using TROPOMI data. *Journal of the Air & Waste Management Association*, *just-accepted*.

Sicard, P., Agathokleous, E., De Marco, A., Paoletti, E., & Calatayud, V. (2021). Urban population exposure to air pollution in Europe over the last decades. *Environmental Sciences Europe*, *33*, 1–12.

Sokhi, R. S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., & Halenka, T. (2021). Advances in air quality research–current and emerging challenges. *Atmospheric Chemistry and Physics Discussions*, *2021*, 1–133.

Steinle, S., Reis, S., & Sabel, C. E. (2013). Quantifying human exposure to air pollution—Moving from static monitoring to spatio-temporally resolved personal exposure assessment. *Science of the Total Environment*, *443*, 184–193.

Sunarno, S., Purwanto, P., & Suryono, S. (2020). *The Monitoring System of Sulfur Dioxide Gas Using a Web-based Wireless Sensor*. Proceedings of the 13th International Interdisciplinary Studies Seminar, IISS 2019, 30-31 October 2019, Malang, Indonesia.

Tan, X., Han, L., Zhang, X., Zhou, W., Li, W., & Qian, Y. (2021). A review of current air quality indexes and improvements under the multi-contaminant air pollution exposure. *Journal of environmental management*, *279*, 111681.

Tarazona Alvarado, M., Salamanca-Coy, J., Forero-Gutièrrez, K., Núñez, L., Pisco-Guabave, J., Escobar-Diaz, F., & Sierra-Porta, D. (2024). Assessing and monitoring air quality in cities and urban areas with a portable, modular and low-cost sensor station: Calibration challenges. *International Journal of Remote Sensing*, *45*(17), 5713–5736.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45.

Todorov, A., Gicheva, P., Stoykova, V., Karapetkov, S., Uzunov, H., Dechkova, S., & Zlatev, Z. (2023). Environmental Monitoring in Bus Transportation Using a Developed Measurement System. *Urban Science*, *7*(3), 90.

Wajeetongratana, S. (2023). *The role of ecological management and green infrastructure in improving air quality*. *452*, 05001.

Wang, J., Li, J., Wang, X., Wang, J., & Huang, M. (2021). Air quality prediction using CT-LSTM. *Neural Computing and Applications*, *33*, 4779–4792.

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.

Zhan, C., Xie, M., Lu, H., Liu, B., Wu, Z., Wang, T., Zhuang, B., Li, M., & Li, S. (2023). Impacts of urbanization on air quality and the related health risks in a city with complex terrain. *Atmospheric Chemistry and Physics*, *23*(1), 771–788.

Zheng, P., Chen, Z., Liu, Y., Song, H., Wu, C.-H., Li, B., Kraemer, M. U., Tian, H., Yan, X., & Zheng, Y. (2021). Association between coronavirus disease 2019 (COVID-19) and long-term exposure to air pollution: Evidence from the first epidemic wave in China. *Environmental Pollution*, *276*, 116682.

Zhou, J., Liu, J., Zhou, Y., Xu, J., Song, Q., Peng, L., Ye, X., & Yang, D. (2024). The impact of fine particulate matter on chronic obstructive pulmonary disease deaths in Pudong New Area, Shanghai, during a long period of air quality improvement. *Environmental Pollution*, *340*, 122813.

Zhu, T., Shang, J., & Zhao, D. (2011). The roles of heterogeneous chemical processes in the formation of an air pollution complex and gray haze. *Science China Chemistry*, *54*, 145–153.

# APPENDIX A

Table A.1 - KPIs *dax* code definition.

| Metric | *Dax* code |
|---|---|
| YELLOW_LINE | `VAR SelectedParameter = SELECTEDVALUE('D PARAMETER'[Parameter Code])`<br>`RETURN`<br>`SWITCH(`<br>`SelectedParameter,`<br>`"NO2", 101,`<br>`"O3", 101,`<br>`"PM10", 36,`<br>`"PM25", 21,`<br>`"CO", 2.501, -- Adjust if needed`<br>`BLANK())` |
| ORANGE_LINE | `VAR SelectedParameter = SELECTEDVALUE('D PARAMETER'[Parameter Code])`<br>`RETURN`<br>`SWITCH(`<br>`SelectedParameter,`<br>`"NO2", 401,`<br>`"O3", 241,`<br>`"PM10", 101,`<br>`"PM25", 51,`<br>`"CO", 3.501,`<br>`BLANK())` |
| RED_LINE | `VAR SelectedParameter = SELECTEDVALUE('D PARAMETER'[Parameter Code])`<br>`RETURN`<br>`SWITCH(`<br>`SelectedParameter,`<br>`"NO2", 401,`<br>`"O3", 241,`<br>`"PM10", 101,`<br>`"PM25", 51,`<br>`"CO", 3.501,`<br>`BLANK())` |

Table A.2 - Calculated measures development and DAX code for BI solution.

| Calculated Measure | Dax code |
|---|---|
| Average Pollutant Value | `AVERAGE('F AIRQUALITY'[Value])` |
| Yellow exceeds count<br>(Applied similarly to Orange and Red thresholds.) | `VAR SelectedParameter = SELECTEDVALUE('D PARAMETER'[Parameter Code])`<br>`VAR YellowThreshold = [KPI_Yellow_Line]`<br><br>`RETURN`<br>`CALCULATE(` |

```
        COUNTROWS('F AIRQUALITY'),
        'F AIRQUALITY'[Value] > YellowThreshold,
        'D PARAMETER'[Parameter Code] = SelectedParameter
        )
```

Parameter Current Value

```
VAR MaxDateInFact =
    CALCULATE (
        MAX ( 'D DATE'[Date] ),
        FILTER (
            'D DATE',
            CALCULATE ( COUNTROWS ( 'F AIRQUALITY' ) ) > 0
        ))

VAR MaxHourOnDate =
    CALCULATE (
        MAX ( 'D TIME'[T_HOUR] ),
        FILTER (
            ALL ( 'D TIME' ),
            CALCULATE (
                COUNTROWS ( 'F AIRQUALITY' ),
                'F AIRQUALITY'[FK_DATE] = MAXX ( FILTER ( 'D
DATE', 'D DATE'[Date] = MaxDateInFact ), 'D DATE'[SK_DATE] )
            ) > 0
        ))

RETURN
CALCULATE (
    AVERAGE ( 'F AIRQUALITY'[Value] ),
    'D DATE'[Date] = MaxDateInFact,
    'D TIME'[T_HOUR] = MaxHourOnDate )
```

Current Pollutant Level Label

```
VAR Value_vl = [Current_AirQuality_Value]
VAR Yellow = [KPI_Yellow_Line]
VAR Orange = [KPI_Orange_Line]
VAR Red = [KPI_Red_Line]

RETURN
    SWITCH(
        TRUE(),
        ISBLANK(Value_vl), BLANK(),
        Value_vl > Red, "Red",
        Value_vl > Orange, "Orange",
        Value_vl > Yellow, "Yellow",
        "Green"
    )
```

# APPENDIX B

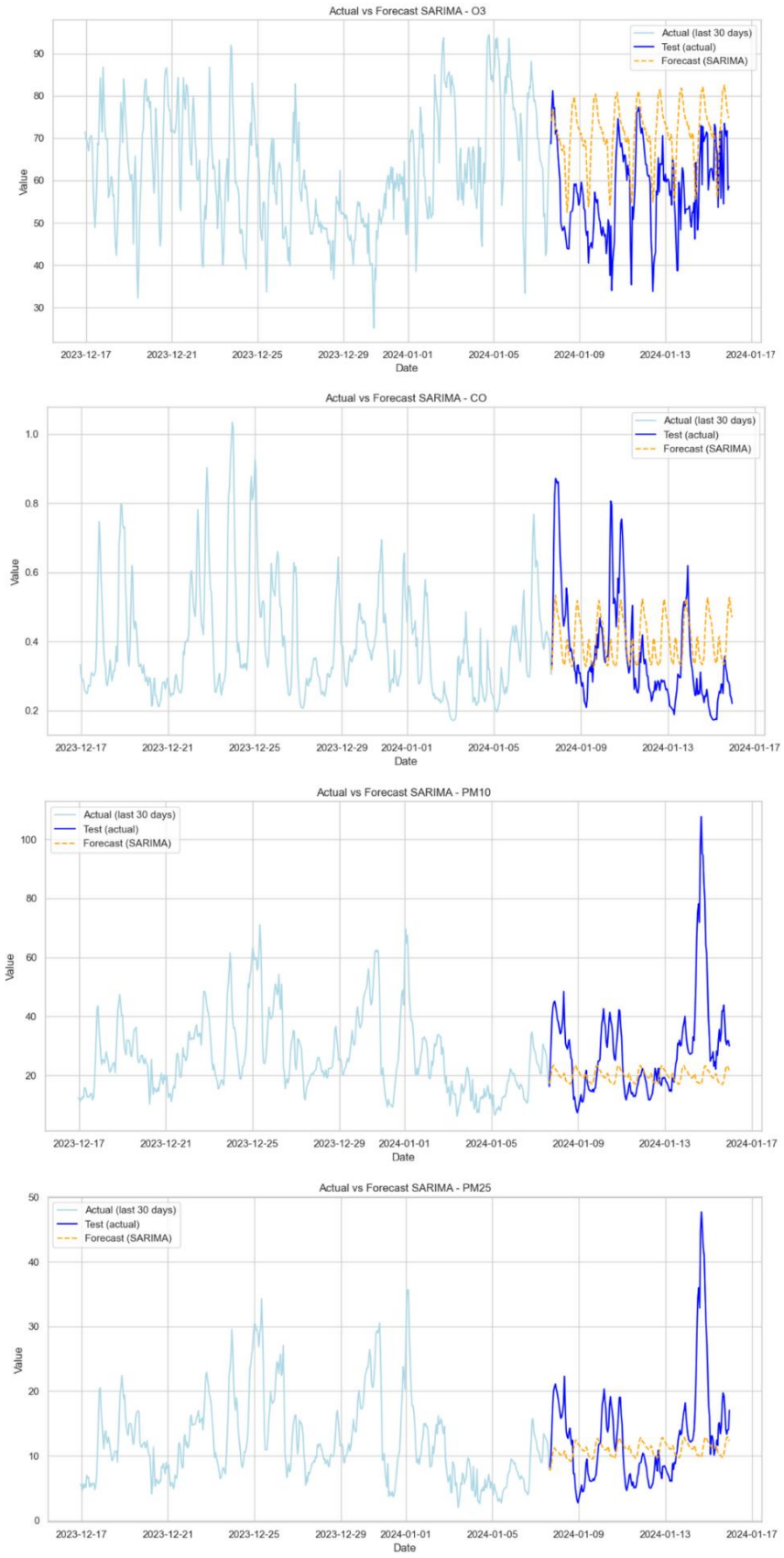


Figure B.1 - Real vs Forecast values with SARIMA Model for $O_3$, CO, $PM_{10}$ and $PM_{2.5}$.

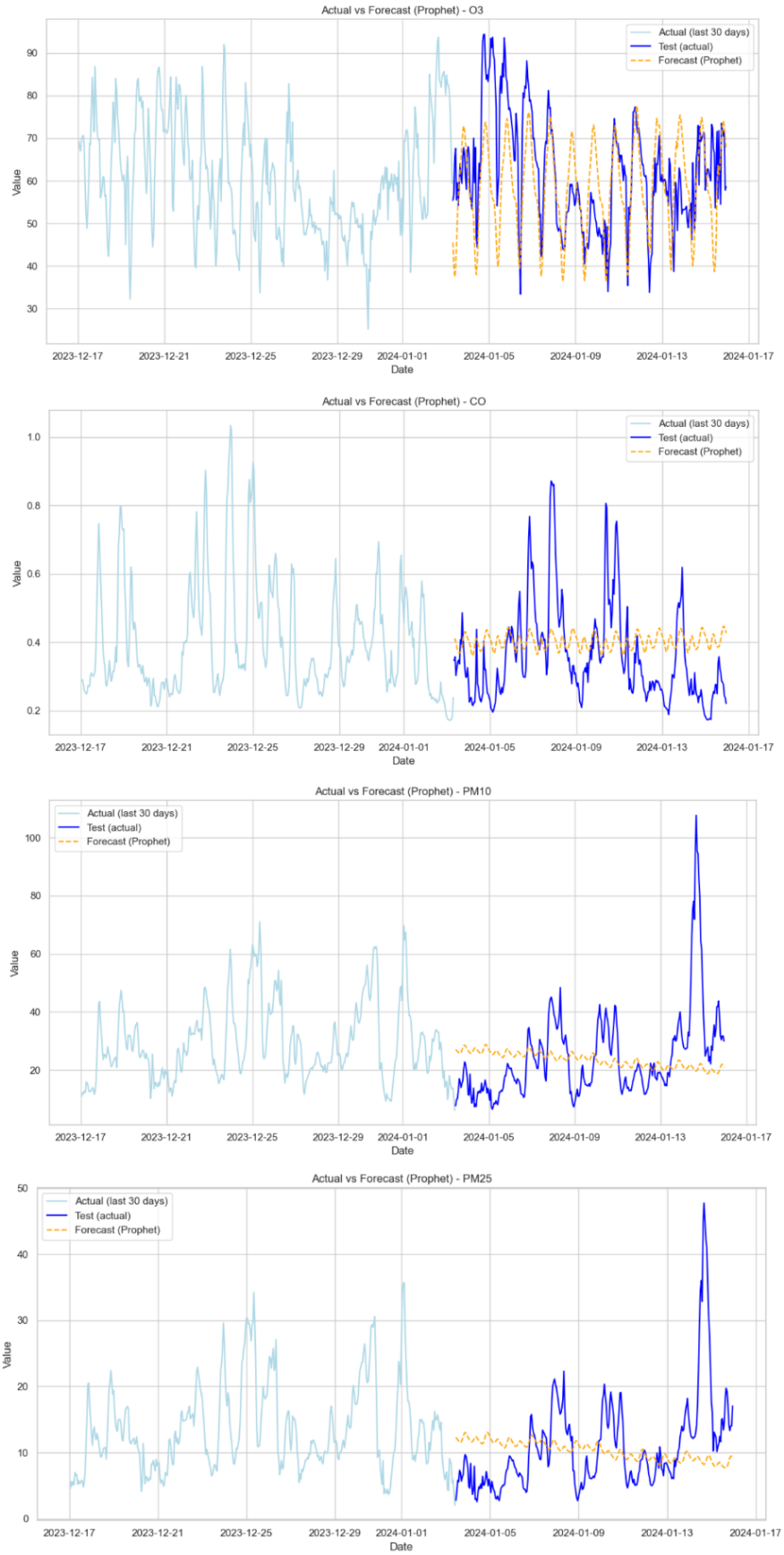


Figure B.2 - Real vs Forecast values with PROPHET Model for $O_3$, CO, $PM_{10}$ and $PM_{2.5}$.

Table B.1 - Correlation analysis between pollutants and external variables.

| | $NO_2$ | $O_3$ | $PM_{10}$ | $PM_{2.5}$ | CO |
|---|---|---|---|---|---|
| **Weekday** | -0.05 | 0.03 | 0.01 | 0.0001 | -0.002 |
| **Is Holiday** | -0.02 | 0.007 | 0.06 | 0.06 | 0.04 |
| **Is Time Peak** | 0.16 | 0.01 | 0.01 | 0.003 | 0.09 |
| **Humidity** | 0.31 | -0.21 | 0.18 | 0.14 | -0.1 |
| **Pressure** | 0.4 | 0.04 | 0.01 | -0.04 | -0.08 |
| **Temperature** | 0.19 | 0.14 | -0.06 | -0.07 | -016 |
| **Type Hazard** | -0.04 | -0.04 | 0.04 | 0.03 | 0.13 |
| **Type Jam** | 0.001 | 0.03 | -0.08 | -0.08 | -0.004 |
| **Type No Alert** | 0.001 | -0.00006 | 0.06 | 0.05 | -0.003 |
| **Type Road Closed** | -0.03 | 0.02 | -0.02 | -0.01 | 0.03 |
| **Type Weather Hazard** | -0.02 | -0.06 | 0.03 | 0.03 | -0.05 |


**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa