BULGARIAN ACADEMY OF SCIENCES
INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Petar Rumenov Zhivkov

MODELING THE STATE OF AIR QUALITY BASED ON HEALTH AND ECONOMIC ASPECTS

ABSTRACT

for obtaining
the educational and scientific degree of "Doctor"
in the professional field 4.6
"Informatics and Computer Science"

supervisor:
Prof. D.Sc. Stefka Fidanova

2024

Chapter 1

Introduction

Air pollution, particularly fine particulate matter (PM), is a leading environmental cause of morbidity and premature mortality, having risen from fifth position in 1990 [46]. Deaths attributed to ambient PM have increased over the past 25 years [14], and exposure to higher levels of atmospheric pollution is associated with increased mortality [8]. The latest report on the

PM is classified based on its size, with particles smaller than 10 micrometers (PM10) and 2.5 micrometers (PM2.5) causing the greatest concern due to their ability to penetrate deep into the lungs and cause adverse health effects. PM constitutes a complex mixture of solid particles and liquid droplets suspended in the air, with varying sizes and chemical compositions. These particles can originate from natural sources such as dust and pollen, as well as human activities including industrial emissions, vehicle exhaust, and combustion of biomass and solid fuels.

Studies have shown that exposure to PM2.5 and PM10 is associated with premature mortality from several diseases, including cardiovascular, respiratory, lung cancer, and upper and lower respiratory tract infections [2], [27], [31]. Urban air pollution increases the risk of pulmonary and systemic oxidative stress, hypoxemia, immunological modifications, atherosclerosis, and accelerated progression of cardiovascular diseases and chronic obstructive pulmonary disease (COPD) [41], while epidemiological studies indicate that PM exposure increases the risk of diabetes [73].

The severity of health consequences can vary depending on the concentration and duration of exposure, as well as the individual's health status and sensitivity to pollutants. Some population groups, such as children, the elderly, and those with pre-existing respiratory or cardiovascular diseases, are particularly vulnerable to the health effects of air pollution [56].

Studies show a strong association between air pollution and hospitalizations for respiratory diseases [65, 20, 42]. A cohort review of 175 articles summarizes that exposure to air pollution increases the likelihood of developing various diseases including respiratory diseases (asthma, chronic obstructive pulmonary disease (COPD), and lung cancer), cardiovascular diseases (heart attacks, strokes, and hypertension), kidney, liver, and others [1].

1.1 Relevance and Motivation of the Topic

The growing concern regarding the impact of air pollution on public health necessitates the development of predictive models to forecast acute morbidity levels based on air quality data [43]. The latest report by the European Union (EU) on air quality assesses the EU's efforts to improve air quality as only "partially effective" and specifically highlights that actions to reduce PM2.5 are insufficient. Figure 1.1 presents the results of a study involving 432 cities in the EU, which analyzes the economic consequences of the three main pollutants in European cities: particulate matter (PM), nitrogen dioxide, and ozone [17]. The results indicate that PM is the primary cause (82.5%) of economic damage resulting from illnesses and mortality due to air pollution. The study utilizes the AIRQ+ modeling program, developed by the World Health Organization (WHO), incorporating pollution levels and economic data for each city.

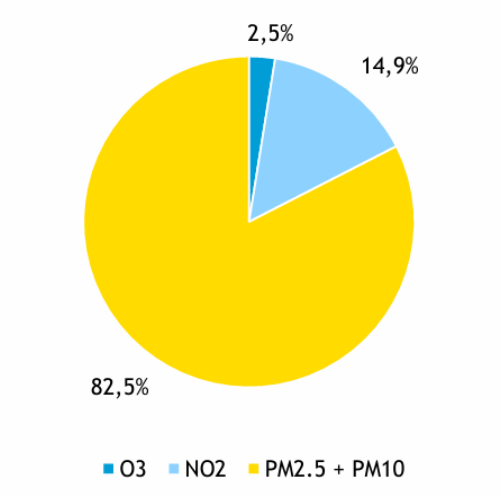

Figure 1.1: Contribution of the three main urban pollutants to the total morbidity and mortality damage in 432 European cities

In light of these findings, there is a growing need for models that can examine acute morbidity based on air quality data. Despite increasing awareness of the health risks associated with air pollution, there is a need for an in-depth analysis of specific pollutants and their effects on various health outcomes. Existing studies show links between air pollution and respiratory diseases, cardiovascular diseases, adverse birth outcomes, and even mental health issues. However, a comprehensive investigation of the extent and mechanisms of these connections is still required to inform evidence-based interventions and policies at the national and local levels.

The levels of PM exceed EU air quality standards in Bulgaria. The European Commission won a case against Bulgaria in April 2017 due to excessive air pollution in several cities. Bulgaria also leads in lost healthy life years due to dirty air. According to an international report, our citizens lose 2.5 years of their lives due to air pollution, whereas the EU average is 0.7 years. The main reasons for air quality surpassing the standards are household heating, car traffic, and geographical factors that trap pollutants in the atmosphere. The lack of solid research and statistical data on the health effects of PM

pollution in Sofia undermines the city's ability to develop evidence-based strategies to mitigate its impact. Without accurate and detailed information, policymakers and public health officials face significant challenges in implementing targeted interventions and raising public awareness of the severity of the problem. Hence, there is an urgent need for comprehensive research efforts and improved data collection methods to overcome existing knowledge gaps and pave the way for informed decision-making in combating PM pollution and improving health quality in Sofia.

1.2 Aim and objectives of the thesis

The aim of the dissertation is to study the impact of fine particulate matter (PM) on acute illnesses in Sofia and to identify methods for prevention.

To achieve this aim, the following four tasks have been formulated:

- Investigate the relationship between fine particulate matter and health indicators for acute morbidity in Sofia;
- Improve data from citizen air quality monitoring stations through calibration using a machine learning-based two-step method;
- Develop a software tool for optimizing and evaluating cycling routes by characterizing cyclists' exposure to air pollution;
- Develop an IoT platform for aggregating and modeling air quality sensor data.

1.3 Research Methodology

The present dissertation employs a multidisciplinary approach, combining literature review, quantitative analysis, and advanced statistical modeling. A systematic review of existing studies has been conducted to summarize and synthesize the current evidence on the relationship between air quality and health outcomes. Additionally, epidemiological studies and data from health surveillance agencies have been analyzed to quantify the relationship between specific pollutants and health consequences. Modern statistical techniques, such as regression modeling and spatial analysis, have been utilized to investigate complex interactions and spatial patterns of air pollution and health outcomes.

The methodology used in this dissertation involves a systematic approach to studying the impact of fine particulate matter (PM) on acute diseases in Sofia and developing strategies to mitigate their consequences. The methodology includes data collection from hospitals and monitoring stations, data analysis to establish correlations, calibration of citizen laser stations, development of an IoT platform for data visualization, and optimization of cycling routes.

Data collection includes admissions and hospitalizations from two major hospitals in Sofia, emergency medical service registry data, and PM measurements from both official monitoring stations of the Executive Environmental Agency (EEA) and citizen laser stations. Meteorological data, including humidity, atmospheric pressure, and temperature, as well as traffic data and other relevant factors, are also collected.

Data analysis techniques are applied to investigate the relationship between PM levels and acute diseases. Statistical analysis and comparisons of PM levels below and above the World Health Organization (WHO) health standards are performed. The calibration

process focuses on calibrating data from citizen laser stations using machine learning models with controlled and uncontrolled methods. Reference data from EEA monitoring stations are used for precise calibration, considering factors such as humidity, atmospheric pressure, and temperature.

A model for optimizing cycling routes to minimize PM inhalation is developed. This involves using a modified algorithm to find the shortest path and conducting real-field tests to validate the methodology. A software system is developed to visualize the location and indicators of monitoring stations on a geographical map. A new approach is applied for aggregating, organizing, processing, modeling, and exchanging data in the IoT system.

During the research process, a wide range of literature was utilized to investigate the problem. The newly developed methods and algorithms were published in [[**?**], [75], [76], [77]]. The presented methodology allows for a comprehensive analysis of the impact of fine particulate matter on acute illnesses, calibration of citizen laser stations through machine learning and neural networks, and optimization of cycling routes considering PM pollution. In line with the set objectives, algorithms and methodologies have been developed to address specific problems. Software applications for each of the developed algorithms have been created, and data visualization has been implemented through an IoT platform. The platform and software are written in Python and utilize the Django framework.

The results of this thesis will contribute to scientific knowledge, policy development and public awareness, ultimately leading to improved air quality management and better protection of human health.

Chapter 2

Link between PM Pollution and Acute Health Indicators

The links between urban air pollution (UAP) and human health have been consistently and clearly established by a number of researchers [66], [59], [39], with prominent groups of diseases being cardiovascular [68], neurovascular [64], [10], and pulmonary [67]. According to recent refined modeling, it is estimated that there are nearly 9 million deaths annually from UAP [53]. About 25% of premature deaths associated with UAP are respiratory by presumption [5].

Significant literature from epidemiological studies suggests a link between acute morbidity and exposure to PM2.5 air pollution [19]. Most of these data come from time-series analyses [23], comparing variations in hospitalizations with average variations in particulate matter [28]. UAP is responsible for respiratory tract inflammations [13], and considering that the respiratory system is a common gateway of entry, minimizing respiratory tract exposure also minimizes cardiovascular problems [62]. Additionally, there are studies that model pollution across multiple cities, such as the European Air Pollution and Health: A European Approach (APHEA) [35] and the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) [60], both providing consistent evidence of the health-air pollution link for multiple cities by covering large geographic areas.

The consequences of air pollution can be viewed as either an increased risk of illness or injury to an individual or as an additional general risk to the well-being of the population [22]. The goal of air quality management is to control or avoid the adverse effects of air pollution on public health. Therefore, it is important to define such impacts that are considered "adverse" and distinguish them from those impacts that are not considered adverse, thus concentrating protection efforts on pollutants that cause the most extreme health impacts.

2.1 PM Standards: Analysis of EU Legislation and WHO Guidelines

The EU legislative framework for PM levels in the air is based on the Ambient Air Quality Directive (2008/50/EC) and is complemented by the National Emission Ceilings Directive (2016/2284/EU). These regulations require Member States to meet air quality standards

Table 2.1: Air Pollution: WHO Recommendations and EU Legislation (in $\mu g/m^3$)

Air pollution standards	WHO		EU	
	PM10	PM2.5	PM10	PM2.5
daily average	50	25	50	–
yearly average	20	10	40	25

according to specific threshold values derived from scientific assessments. Compliance is monitored by the European Commission, and non-compliance can lead to legal actions and fines. In comparison, WHO provides general air quality guidelines without legislative obligations, leaving implementation to individual countries.

In the comparative analysis between EU legislation and WHO guidelines regarding PM2.5 levels in the air, several key differences emerge (Table 2.1). While both the EU and WHO recommend a daily average concentration threshold for PM10 of 50 $\mu g/m^3$, the EU and WHO significantly differ in their annual average recommendations, with WHO standards being twice as stringent (20 $\mu g/m^3$ compared to 40 $\mu g/m^3$ from the EU). For PM2.5, the EU does not have a daily average concentration threshold, and there are significant differences in annual recommendations, with WHO advocating for stricter limits. Furthermore, EU standards are region-specific, while WHO guidelines are non-binding and universal, providing guidance to countries worldwide.

For the purposes of this dissertation, we will use WHO air quality guidelines as they are recommended globally, have a daily limit for PM2.5, and are based solely on health aspects, compared to the EU directive, which focuses on regional and national needs and includes policy-economic and social aspects in the directives.

In its latest report, WHO pays special attention to PM pollution and considers exposure to PM2.5 as the greatest air-related killer globally. About 80% of global deaths attributable to PM2.5 exposure could be prevented if countries adhere to an annual PM2.5 threshold value of 5 $\mu g/m^3$. Achieving interim targets will also have significant health benefits. In the case of PM2.5, achieving interim target 4 (10 $\mu g/m^3$) would lead to approximately a 48% reduction in the total number of deaths attributable to PM2.5 exposure.

2.2 Geographical Study Area

Sofia is the only European capital located in a basin and is characterized by high levels of anthropogenic emissions and frequent occurrences of stagnant meteorological conditions. The city has a population of 1.2 million people [47] and is situated in the Sofia Basin. The area is recognized as problematic, particularly during winter when numerous exceedances of the norms in European legislation for air pollution control occur.

For years, Sofia has been struggling with significant levels of air pollution, with PM concentrations consistently exceeding the recommended limits set by WHO and the EU. These elevated levels of PM pollution in the city pose a serious threat to the health and well-being of its residents. Figure 2.2 shows the average concentration of PM2.5 in the most polluted capitals in Europe in 2022. It can be seen that Sofia ranks among the top positions in this ranking. Additionally, it is important to note that air quality measurement data are insufficient for Sofia. Data for PM2.5 levels in the capital are provided by only one station - AIS Hipodruma, which is insufficient for quality statistical

analyses for a city the size and population of Sofia.

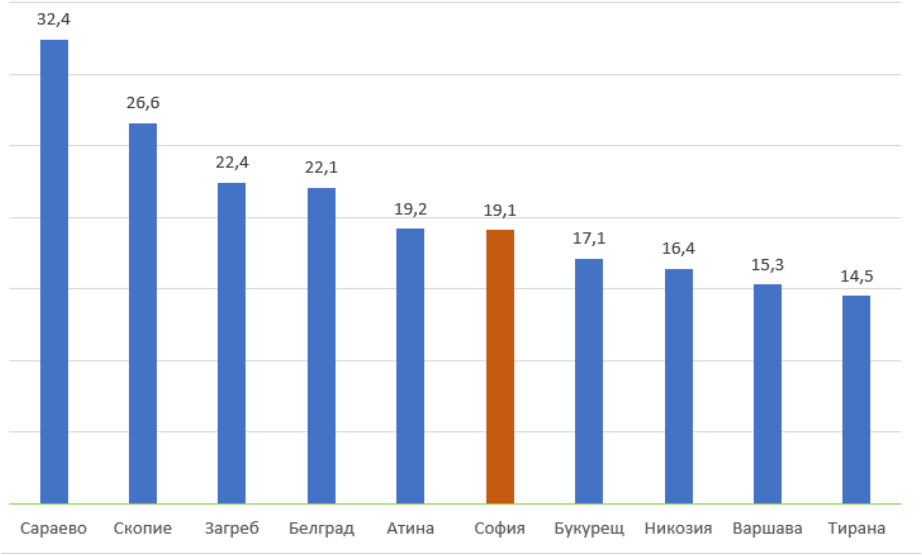

Figure 2.1: The average concentration of PM2.5 in the most polluted capitals in Europe in 2022 (in $\mu g/m^3$)

The combination of cold winters with subzero temperatures and the city's location predisposes it to temperature inversions that can last for several days to a week. Many urban areas situated in valleys with poor air exchange encounter significant air pollution problems, which are associated with local atmospheric conditions [57]. Strong inversions and a lack of precipitation and/or wind are the main reasons for retaining pollutants in the air, primarily during winter.

Historically, air quality indicators in Sofia during winter have been significantly worse than those during summer. In addition to temperature inversions typical of colder conditions, during the heating season, the burning of fossil fuels for domestic heating such as wood and coal, and sometimes waste, significantly increase PM levels.

Therefore, this difference in air quality between summer and winter conditions will serve as a good model for testing our hypothesis regarding a significant link between pollution and health outcomes in Sofia. The literature abounds with models that can be used to assess the hypothesis of pollution/health relationship. Specifically for Sofia, no such studies have been conducted. Correlation between temperature and mortality around Sofia is made using linear and nonlinear models [52]. Other studies examining the genotoxicity of atmospheric air in 3 European cities, including Sofia, show that air pollution in winter is 6 to 10 times higher compared to summer air [24].

2.3 Methodology of the Study

This is the first such study with real data from official and citizen sources of information on particulate matter air pollution, comprehensive data from the emergency medical service center (emergency aid), and data from two of the largest hospitals associated

with access to emergency care in Sofia - UMBSM Pirogov and UMBSM Tokuda. The results can be compared and contrasted with other international studies using local data.

The aim of this study is to determine which acute morbidity increases and by how much on days when PM levels do not meet WHO guidelines, using time-series analysis.

The following four subtasks have been addressed: comparing air quality data with health data obtained from hospitals and emergency care; comparative analysis and highlighting of key problem areas; testing the hypothesis that data from inexpensive sensors can be useful in such epidemiological studies; identifying future prospects and summarizing key areas where additional research is needed to improve the model's effectiveness.

Air Pollution Data

First, we will identify which days, within the observed period, exceed the daily average limits set by the WHO for PM concentrations. Then, we will compare the health data for the days when air quality is within WHO guidelines to those that are not.

The air quality data in the form of hourly average concentrations of PM2.5 and PM10 are for the years 2018 and 2019 and represent independent series for each element from official sensors provided by (a) the Executive Environment Agency (EEA) and (b) laser stations from the luftdaten network (a global sensor network run by volunteers that generates open environmental data).

The data is collected from 5 official monitoring sites (Druzhba, Nadezhda, Hipodruma, Pavlovo, and Mladost). The method of beta attenuation is used for measuring PM10 and PM2.5 according to European Directive 2008/50/EC (Directive, 2008).

According to previous studies, the use of data from inexpensive laser stations aims to enhance the potential benefits of traditional monitoring networks with additional geographic and temporal resolution of measurements [9]. Additionally, inexpensive stations can complement official measuring devices and particularly address the limited data quantity and station distribution [78]. It is believed that inexpensive sensors are suitable for many specific purposes, including expanding conversations with communities and enhancing public awareness of air pollution issues [45]. In Sofia, at the time of the study, there are several citizen networks of laser stations that complement, with additional geographic and temporal resolution of measurements, the official stations of the EEA.

Health Data

Due to the voluntary nature of data provision, there are various temporal characteristics and formats from different sources:

- Aggregated data on the activity of the Sofia Emergency Medical Center by diagnoses from 01.01.2017 to 14.03.2019;
- Information on diagnostic activity from Tokuda Primary Care Center - from 02.01.2018 to 31.12.2018;
- Information on diagnostic activity from University Hospital "St. Marina" - from 01.01.2018 to 31.05.2019;
- Information on hospitalized diagnoses from St. Marina Hospital - from 01.01.2018 to 31.12.2018.

The study uses the International Classification of Diseases (ICD), specifically ICD-10, to segment diseases and identify morbidity. ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems published by the World Health Organization (WHO), which is used up to the time of this study. It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases [48], [49].

Methods

This study employs time series analysis with correlation methods to analyze air quality and health data. The statistical methods in the study fall into two categories: parametric and non-parametric. Parametric methods, such as Pearson's correlation coefficient, are preferred when the data are normally distributed or there is a sufficiently large volume of measurements, allowing for the assumption that the data follow a normal distribution. These methods are suitable for assessing the linear relationship between two continuous variables.

The parametric Pearson correlation test (2.1) is used to compare the two sources of air quality data. It provides a measure of the linear relationship between the two continuous variables (commonly referred to as just the correlation coefficient). Correlation coefficients for each pair (x, y) are calculated for assessment, and the values of x and y are replaced with their ranks, respectively. The results of the test are applied to the correlation coefficient in the range of -1 to 1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

Non-parametric comparative methods, such as Spearman's rho and the Mann–Whitney U test, are used when the data are not normally distributed or there is uncertainty about whether they follow a normal distribution. These methods are suitable for assessing the rank correlation between variables or for comparing differences between two groups without requiring specific assessment of the linear relationship. The choice in this dissertation is as follows: descriptive analysis for individual variation series; correlation analysis (non-parametric - Spearman's rho) between pairs of variables; between-group comparison (non-parametric Mann–Whitney U test) of mean values.

The model developed in this study aims to explore the relationship between air quality indicators and health outcomes based on data on fine particulate matter and hospital admissions and Emergency Medical Service (EMS) visits. At the advice of medical professionals and health data analysts, we incorporate lagged effects of 1, 2, and 3 days. This is because health effects typically require time to manifest after exposure to air pollutants [34]. By doing so, our analysis of hospital admissions and EMS records will not only reveal which disease cases undergo changes due to variations in air quality but also demonstrate how these effects evolve over time. Figure 3.1 illustrates the exposure-response method, where the dose is the exposure multiplied by time.

2.4 Results

In a recent study, it is suggested that low-cost sensors can provide rough details about air quality but may not be suitable for applications requiring high accuracy [32]. In this chapter, we test the hypothesis that low-cost air quality sensors from Luftdaten

can be a valuable addition to the official monitoring stations operated by the EEA. By cleaning and summarizing the data, we aim to identify correlations that can provide insights into certain coarse details. During comparison, we filtered out data when the air humidity exceeded 70%. While these conditions do not necessarily imply inaccurate data, the manufacturer does not guarantee the predicted accuracy of 10% under such circumstances.

The results of daily concentrations for PM10 and PM2.5 are summarized in Table **??** The average daily concentration of PM2.5 is 11.7 g/m3 with a range from 2 to 136 g/m3. During the study period, 13.2% of the daily concentrations of either PM10 or PM2.5 did not meet the WHO air quality guidelines (50 g/m3 for PM10 and 25 g/m3 for PM2.5). There were some gaps in the data from the EEA for several hours to several days (due to power supply issues, malfunctions, maintenance), and some of the sources for hospital admissions had missing data during weekends (Saturday, Sunday, official holidays). Therefore, an analysis of the gaps was conducted to exclude the possibility of bias.

Table 2.2: Daily concentration of PM10 and PM2.5, divided into working and non-working days

Day from the week	**N**					
PM10 Working days	522	34.17	27.5	5	336	20
PM10 Non-working days	230	37.1	26	6	290	18
PM2.5 Working days	522	11.15	9	2	104	4
PM2.5 Non-working days	230	12.93	8	2	113	3

The correlations observed in the data for hospitalized patients with more severe conditions are summarized in Tables 2.3 and 2.4. An increase (compared to background levels) in respiratory and pulmonary diseases is noted from the 1st to the 3rd day following an exceedance of PM10 by 120%, as well as heart failure during the same period by 19%. Exceedances of PM2.5 are associated with a 59% increase in the frequency of pulmonary embolism on days 2 and 3.

The results show an increase in emergency medical service contacts by 11% on days with exceedances of PM10 and 13.5% on days with exceedances of PM2.5, for a period of at least 3 days. The average increase (compared to background levels) in neuroses on day 2 following an exceedance of PM10 is 1%, for myocardial infarctions on day 3 following an exceedance it's 8%, for strokes immediately following an exceedance it's 9%, continuing until day 3 post-exceedance. The same applies to hypertensive conditions, which increase by 5% immediately around the exceedance and up to the third day after the event.

From the analyzed data, a significant correlation is observed between cardiovascular and cerebrovascular diseases, fully coinciding with the rapid increase in risk at low levels

Table 2.3: Comparison of nonparametric data from hospitalized patients on days with exceedances of WHO standards for PM10

Morbidity	ICD-10	no lag		1-day lag		2-days lag		3-days lag	
		Z	p	Z	p	Z	p	Z	p
Resp. system	C30-C39	-1,504	0,133	-2,211	0,027	-2,870	0,004	-2,862	0,004
Heart attack	I50	-1,729	0,084	-3,656	0,001	-3,475	0,001	-2,821	0,005

Table 2.4: Comparison of non-parametric data from hospitalized patients in severe conditions on days with exceedances of WHO standards for PM2.5

Morbidity	ICD-10	no lag		1-day lag		2-days lag		3-days lag	
		Z	p	Z	p	Z	p	Z	p
Resp. system	C30-C39	-0,361	0,718	-1,270	0,204	-1,265	0,206	-2,119	0,034
Pulm. embolism	I26	-1,601	0,109	-2,302	0,021	-2,894	0,004	-1,852	0,064
Heart attack	I50	-1,586	0,113	-2,831	0,005	-3,195	0,001	-2,799	0,005

Table 2.5: Comparison of non-parametric data from outpatient examinations in hospitals with WHO standards for mean daily concentrations of PM10

Morbidity	ICD-10	no lag		1-day lag		2-days lag		3-days lag	
		Z	p	Z	p	Z	p	Z	p
Respiratory system	C30-C39	−1.515	0.130	−2.950	0.003	−2.941	0.003	−2.423	0.015
Ear diseases	H65,H81	−0.754	0.451	−0.976	0.329	−1.580	0.114	−2.322	0.020
Myocardial infarct.	I20-I25	−0.244	0.807	−1.393	0.164	−1.876	0.061	−2.104	0.035
Heart deficiency	I50	−0.771	0.440	−2.840	0.005	−2.559	0.010	−2.972	0.003
Upper respir. tract	J00-J06	−3.218	0.001	−3.507	0.000	−3.910	0.000	−4.605	0.000
Bacterial pneum.	J13-J18	−3.171	0.002	−3.395	0.001	−2.556	0.011	−3.683	0.000
Acute bronchiolitis	J20-J21	−3.096	0.002	−2.848	0.004	−1.606	0.108	−1.588	0.112
Lower resp. tract	J40-J47	−3.988	0.000	−4.049	0.000	−2.842	0.004	−3.481	0.001
Skin infections	L00-L08	−2.552	0.011	−1.379	0.168	−0.267	0.789	−0.489	0.625
Mycoplasma pneum.	J20.0	−2.330	0.020	−2.327	0.020	−0.430	0.667	−0.431	0.667
Acute bronchitis	J20.9	−2.496	0.013	−1.449	0.147	−0.039	0.969	−1.292	0.197
Asthma, allergic	J45.0	−2.426	0.015	−2.808	0.005	−3.577	0.000	−3.643	0.000
Angina	I20.8	−1.971	0.049	−0.763	0.445	−0.773	0.440	−0.674	0.500

of PM2.5 pollution fractions, consistent with similar findings in other studies [74], [63], [69].

But the highest correlation is observed in acute upper respiratory tract infections, increasing by 47%, and notably pneumonia by 60%. Chronic obstructive pulmonary disease (COPD) increases by 36% after one day. Regarding asthma, records of allergic asthma increase more on days with increased pollution compared to non-allergic asthma. Ambulatory hospital visits for mild conditions that do not require hospitalization logically show an increased presence of respiratory diagnoses (Table 2.5).

2.5 Conclusion

The analyses and tests conducted have proven that acute illness increases on days when the concentrations of fine particulate matter exceed the WHO air quality standards. This was established by comparing hospital admissions and emergency cases on days with and without exceedances of PM standards in Sofia.

Respiratory diseases show higher and faster development when air pollution exceeds healthy limits. Other diseases related to air pollution, such as cardiovascular and cerebrovascular diseases, also increase.

This is among the first studies of its kind in Bulgaria and could be valuable for healthcare professionals, environmental scientists, and policymakers.

The exposure-effect method is complex, and achieving precise measurements is chal-

lenging, even in exposure chambers. However, with the help of epidemiological research methods and population statistics, it becomes evident that air pollution contributes to expected changes in certain disease indicators. In conclusion, similar to data from published studies in cities in Europe, America, and Asia, elevated levels of air pollution are associated with and lead to higher levels of diagnosed diseases. With few exceptions, short-term health effects are measured using averaged concentrations of air pollution across the entire city as exposure indicators. This can lead to misclassification of exposure and thus bias.

The use of extensive data from multiple low-cost laser sensors can effectively enrich the information provided by the high-precision official stations of the Executive Environment Agency, which, besides being few in number, often have missing data for certain periods. This approach leads to improved measurements of the links between health and air pollution by providing a more comprehensive exposure to pollutants.

Chapter 3

Calibrating Low Cost Sensors through Machine Learning

To understand how we can improve the data obtained from laser sensor stations for PM, we need to measure their error under real conditions. For this purpose, we place laser sensor stations in close proximity to the stations of the Ministry of Environment and Water (MOEW) for PM2.5 measurement. We compare their data and create a calibration model based on machine learning, which enhances the data from the inexpensive laser stations.

Traditional air quality monitoring stations, equipped with complex and expensive instruments, are the primary source of data for assessing air pollution levels. However, their high cost, limited spatial coverage, and logistical challenges have led to the emergence of low-cost air quality sensors as an alternative monitoring solution. These sensors offer the potential to expand air quality monitoring networks by providing data with higher resolution and enabling more localized assessments. The inexpensive laser sensors operate on the principle of light scattering, detecting and quantifying the concentration of PM in the air. They are much more cost-effective than traditional monitoring stations, but their measured data lack reliability due to their sensitivity to temperature and relative humidity. These weaknesses are particularly evident when dealing with high levels of PM in combination with high relative humidity - then the measured PM2.5 values are artificially elevated. Calibration is crucial to ensuring the accuracy and consistency of the measurements from the inexpensive laser sensors.

In this chapter, we present a new two-step model for calibrating inexpensive laser sensors for air quality monitoring, using supervised and unsupervised machine learning (ML) techniques. Our proposed calibration method aims to improve the accuracy of individual laser sensors by utilizing data from official air quality monitoring stations as references. The first step of our calibration model involves supervised ML, where we train a predictive model using data collected from both inexpensive laser sensors and reference air quality monitoring stations. The supervised ML algorithm learns the relationship between the sensor readings and the corresponding reference measurements, enabling calibration of the sensor data. By leveraging the comprehensive and accurate data from the reference stations, we can overcome the discrepancy between the readings of the inexpensive sensors and the true measurements. The second step of our model employs unsupervised ML techniques to further enhance the calibration process. Un-

supervised learning algorithms, such as clustering and anomaly detection, help identify and correct deviations and inconsistencies in the sensor data. By detecting and rectifying irregularities, we can improve the overall accuracy and reliability of the inexpensive laser sensors.

3.1 Air quality monitoring

This study utilizes five air quality monitoring stations employing traditional measurement methods as a reference. To calculate PM10 concentrations, automatic monitoring stations from the Executive Environment Agency (EEA) are used. These stations utilize gravimetric methods for collecting PM samples on filters and weighing the filters to determine the mass of collected PM. These methods are similar to beta attenuation but do not rely on radiation detection. Gravimetric methods are highly accurate but require specialized equipment and trained personnel for sample collection and analysis. The monitoring stations used in the study are located in Sofia, specifically in the districts of Mladost, Druzhba, Nadezhda, Hipodruma, and Krasno Selo. Only one of these stations measures PM2.5, so PM2.5 is not used as a reference in this study.

The PM sensor used in this study is the NovaFitness SDS011 laser particle sensor. These types of sensors operate based on the principle of laser diffraction. The laser illuminates captured particles, causing scattered light waves to be detected at specific angles as air passes through the sensor's photosensitive area. A particle size continuum is created by classifying these pulse signals into different particle size intervals to measure the mass concentration of particles [37].

This study utilized the wireless sensor network (WSN) by Luftdaten. It is a citizen-driven network comprising 300 stationary sensors covering Sofia. Each sensor is installed in a plastic tube that can be mounted on walls, balconies, streetlight poles, and other structures. Specific guidelines have been developed to achieve the best representation of particulate matter emissions in the city with the fewest possible sensors. The WSN employs fixed sensors mounted in a 1 km grid to ensure that the majority of the city center is covered with appropriate density.

3.2 Data Calibration Model

This section describes the two-step calibration model for inexpensive laser sensors, which utilizes a combination of supervised and unsupervised machine learning techniques. This model uses data from measurements from inexpensive sensors and standard air monitoring stations as a reference tool.

Other researchers have shown that machine learning correction models of sensor data yield good results when trained against a reference instrument [15].

3.2.1 Introduction to Supervised Machine Learning

Supervised machine learning is a type of artificial intelligence (AI) where algorithms are trained on a labeled dataset and used to make predictions or decisions based on this training. The goal of supervised learning is to find the relationship between the input features of a dataset and their corresponding outcomes and use this relationship to make predictions for new, unseen data.

Supervised learning algorithms are trained on a labeled dataset, which contains both input features and their corresponding outputs. These algorithms use this training data to learn the relationship between the input features and the outputs, and then use this relationship to make predictions for new data. The accuracy of the predictions depends on the quality of the training data and the choice of algorithm.

3.2.2 Methodology of the Study

The methodology used to calibrate the concentration of PM10 in Sofia involves several steps:

1. Data cleaning: Data from citizen stations are further refined by limiting PM10 concentration levels based on official hourly measurements. Stations with less than a predetermined threshold (90 days) are also removed from the dataset;

2. Deviation correction: The study begins by correcting the data from citizen measurements obtained from luftdaten. These measurements are compared with the official measurement stations of the Executive Environment Agency (IAOS) to remove any instrumental deviations;

3. Data preparation: Distances between station pairs are calculated, and groups of stations within a certain distance are created. For each pair, differences in measurements are calculated, and stations with high differences in measurements are removed from the dataset;

4. Analysis of factors and characteristics: A exploratory data analysis is conducted using IAOS's official data. The significance of meteorological variables and engineering characteristics influencing air pollution levels is examined;

5. Calibration model: A two-step machine learning model has been developed, incorporating additional factors from meteorological parameters and topographic data. The characteristics used for calibration include temperature, humidity, pressure, PM10 concentration from the previous day, wind speed, and various dummy variables. More information on the two-step machine learning method will be presented in this chapter;

6. Evaluation and validation: The accuracy of predictive models is assessed using quadratic error. The models are compared with each other, as well as with the initial data before calibration, showing a significant improvement in accuracy.

3.2.3 Data and Model

The model utilizes 5 standard air quality monitoring stations located in Sofia and 5 low-cost laser sensors with the specifications shown in the previous section. The small size of the low-cost laser stations allows them to be placed directly adjacent to the official stations of the IAOS.

The input variables consist of relative humidity (RH), atmospheric pressure (AP), and temperature (temp). The parameter we are calibrating the data for is PM10. These input data are obtained from both standard instruments and low-cost sensors. Since the temporal resolution of the sensor differs from that of the conventional instrument, the hourly average value is used for calculation and assessment.

As temperature, atmospheric pressure, and humidity affect the values of many air quality sensors, measurements of these variables are often recorded on-site and used in the calibration model, as done by other researchers [30], [79].

3.2.4 Brief Description of the Five ML Techniques Used

Here we briefly describe the five supervised machine learning techniques used in the first step of the calibration model. These techniques are: linear regression (LR), decision tree (DT), gradient boosted decision tree (GBDT), random forest (RF), and artificial neural network (ANN).

Linear regression is a simple and widely used supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables. Decision tree is a model that uses a tree-like structure of decisions and their potential consequences, including the consequences of random events [58]. GBDT is an iterative decision tree algorithm composed of multiple decision trees, where the decisions of all trees are summed to obtain the final answer [38]. RF is a hybrid tree-based predictor where each tree is based on the values of a random variable chosen independently and with the same distribution for all trees in the forest [7]. Artificial neural networks are mathematical models that simulate the behavior of neurons and are automatically adjusted by backpropagation of errors [21].

3.2.5

3.2.6 Model Design

Figure 3.1 demonstrates the modeling of the two-step method for calibrating data from air monitoring stations. In the first stage of the model, 5 supervised machine learning techniques are used and their results are evaluated. The most effective of these five methods is then used again in the second step, which includes anomaly detection to clean deviations. At each step, we evaluate whether this improves the model's results and by how much.

Anomaly detection and removal is the second step in the model and it can improve the data, but it is not guaranteed. This is a method for detecting unusual objects or events in datasets that are beyond the ordinary [61]. The anomaly detector is used to remove deviations only from the training dataset. Since this is unsupervised learning, an evaluation has been made with the same ANN settings before and after cleaning the dataset to determine if unsupervised learning is suitable for this dataset.

The process of training the model is divided into two stages: training and testing. The raw data is randomly divided into two datasets, with 80% for training and 20% for testing. The model is first trained using the training data and then its output is evaluated by the testing dataset.

Anomaly detection and removal is the second step in the model and it can improve the data, but it is not guaranteed. This is a method for detecting unusual objects or events in datasets that are beyond the ordinary [61]. The anomaly detector is used to remove deviations only from the training dataset. Since this is unsupervised learning, an evaluation has been made with the same ANN settings before and after cleaning the dataset to determine if unsupervised learning is suitable for this dataset.

The process of training the model is divided into two stages: training and testing. The raw data is randomly divided into two datasets, with 80% for training and 20%

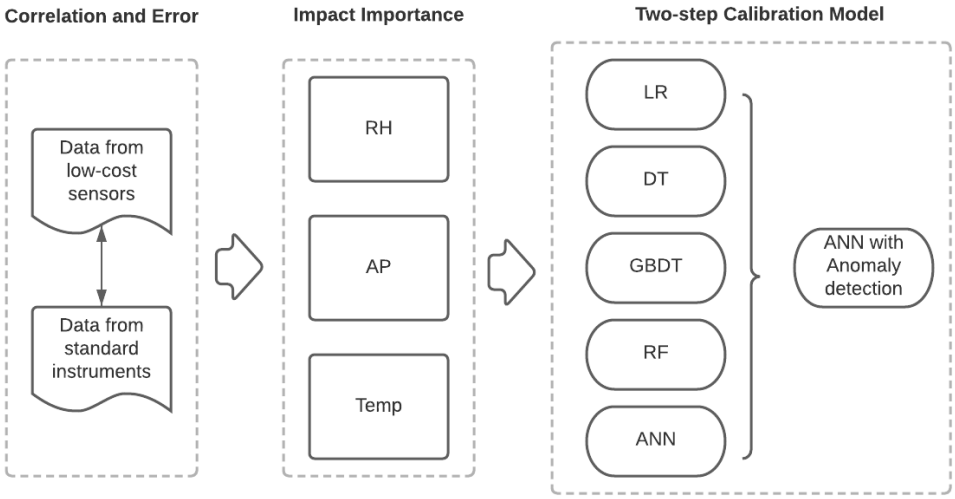

Figure 3.1: Two-step calibration model using artificial neural network and anomaly detection

for testing. The model is first trained using the training data and then its output is evaluated by the testing dataset.

3.3 Application of the Model on Wireless Sensor Network

After calibrating the 5 sensors located precisely next to the standard instruments, we apply the same model to other sensors from the Luftdaten network. In the analysis, data from measurements from 13 other sensors are added, which are in proximity to 500 meters to the regulatory stations. The purpose of this correction is to predict the "real" concentration of PM10 on-site by reproducing the PM10 concentration as accurately as possible.

For the purposes of this study, we take data from inexpensive sensors that are installed and operated by citizens. Therefore, there are missing data. The percentage of missing data among the inexpensive sensors is 9.4

The aforementioned five supervised ML techniques are used in the first step of the analysis: LR, DT, GBDT, RF models, and ANN. Their effectiveness is then investigated. To evaluate the models, the coefficient of determination for the mean squared error R^2 is used as an indicator of effectiveness in the selection and evaluation of our models. Regarding variable selection, we only chose variables in the validation stage that are used to improve the findings.

The wireless sensor network consists of sensors mounted at different heights. To assess the measurement of pressure from the inexpensive sensors and identify any deviations, we use the barometric formula and the Pearson correlation coefficient. Both values from inexpensive PM sensors and reference instruments are used as input data. The mounting

Table 3.1: Results from supervised ML models

MODEL TYPE	**MAE**	**MSR**	R^2
LR:	11,19	288,12	0,77
DT:	8,89	170,03	0,86
GBDT:	8,68	145,22	0,89
RF:	7,96	125,57	0,90
ANN:	6,27	83,90	0,94

Table 3.2: The importance of the parameters based on RF.

Content	**Parameter**
RH:	46,62%
Temperature:	30,71%
Atmospheric Pressure:	22,67%

height for each inexpensive sensor is known and added to the dataset. Therefore, we can also accurately identify the height difference between each sensor and analyze the pressure measurements.

3.4 Results and Evaluation

In Table 3.1, the results of the five supervised ML techniques are presented. For a better assessment of the inexpensive sensors, along with the coefficient of determination (R-squared), the mean absolute error (MAE) and the mean squared error (MSE) have been calculated.

The average mean squared error between the particulate matter sensors and the standard instruments before calibration is 0.62. The linear regression model showed the worst result after calibration with an average R^2 value of 0.77.

The best correlation for PM10 was obtained from the artificial neural network model. The average R^2 value is 0.94 (PM10), which aligns with findings from previous studies [55], [54].

For long-term comparison, the inexpensive sensor and the conventional instrument were placed at the same location, which is a common approach for sensor evaluation in previous studies [25], [30], [33].

3.4.1 Assessment of Results for Relative Humidity

Table 3.2 shows that RH, followed by temperature, are considered the most significant factors influencing the performance of the particulate matter sensor. Previous studies have demonstrated that high RH acts as a catalyst for inducing hygroscopic growth of particles and altering their optical properties, leading to significant disturbances for laser-based PM sensors [40]. This is corroborated in our RF and ANN models for PM10 values, where RH emerged with the highest importance.

Unfortunately, standard devices cannot be compared for the PM2.5/PM10 ratio since, as mentioned above, out of the five standard devices, only one measures PM2.5.

Table 3.3: Comparison of One-Step and Two-Step Models

MODEL TYPE	MAE	MSE	R^2
ANN Model:	6.27	83.90	0.94
ANN Model with anomaly detection:	5.62 (↑5.16%)	65.37 (↑14.69%)	0.95

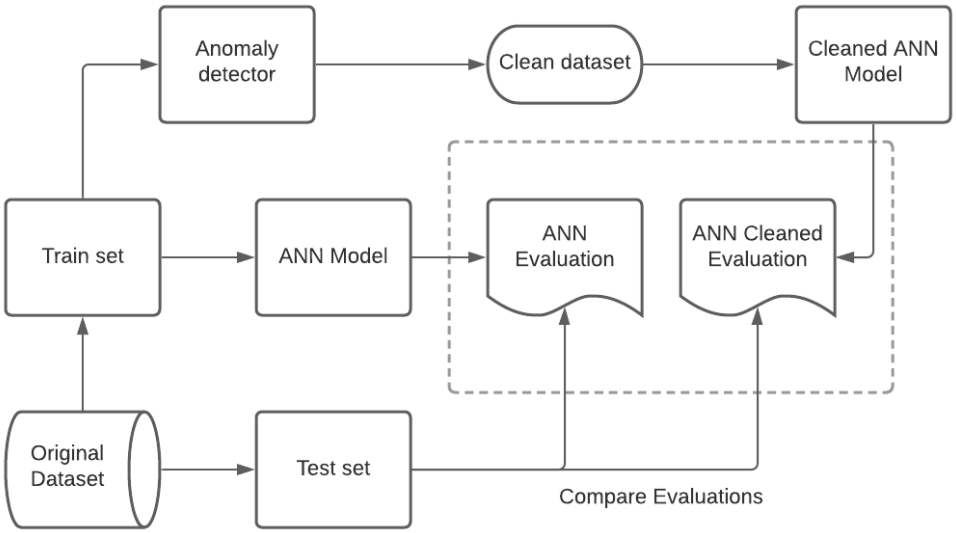

Figure 3.2: Artificial Neural Network Model with Anomaly Detection

3.4.2 Results from the Calibration Model

Table 3.2 presents the statistical results from testing each model, where the mean absolute error (MAE), mean squared error (MSE), and R^2 are determined

The results from the remaining five individual models show that the ANN model performed the best. The RF model showed slightly worse results. The R^2 for PM10 increased from 0.62 to 0.9 and 0.94 respectively for RF and ANN. The ANN model performed best out of the five models, slightly better than the RF model, which is why it was chosen to be used for the second step - anomaly detection. Therefore, the use of unsupervised learning in this study is considered useful and improves the result of calibration.

In conclusion, the final ANN model (ANN with anomaly detection) has the best calibration result, with the smallest error and the best correlation, indicating that the proposed two-step model is more accurate than the single model in calibrating the low cost sensor model.

3.5 Conclusion

The effectiveness of particulate matter sensors was measured by comparing standard instruments using the wireless sensor network. A two-step process was developed for

calibrating the stationary sensors, and the results of the model were evaluated.

The results from the two-step model are encouraging. The R^2 for fixed PM10 sensors increased from 0.62 to 0.95. The ANN model had the strongest impact among the five independent models, followed by the RF model, while the LR model was ineffective.

Anomaly detectors can serve as an uncontrolled alternative to classifiers in an unbalanced dataset. They identify undesirable sensor behavior and remove it from the dataset. Anomaly detection improved the final results in this study.

Relative humidity proved to be the most important factor for calibration results. This was expected, as high humidity is the condition under which cheap sensors exhibit the most data quality weaknesses. This was confirmed in this study, with humidity having a higher impact factor than temperature and atmospheric pressure.

Atmospheric pressure values from standard stations and sensors were evaluated using calculations with the barometric formula. The correlation was strong, indicating that cheap sensors can be considered a good source for modeling air pollution in vertical planning in further studies.

This type of calibration, using controlled and uncontrolled ML, demonstrates the potential to improve results from cheap sensors. Furthermore, it can also be used to assess deviations. Additional studies can be conducted to determine whether these deviations are due to faulty sensors, improperly installed ones, or they are functioning well but there are hyperlocal changes in atmospheric conditions and air quality.

Chapter 4

Development of a software model for bicycle route selection with respect to PM pollution

Most of the specialized and segregated bicycle lanes in cities are located near busy traffic arteries, which can pose a significant health risk to cyclists due to increased pollutant intake associated with higher ventilation rates and physical activity levels [80], [26], and high levels of physical activity [6], [51]. Researchers have focused on assessing actual exposure levels of cyclists along predetermined routes using personal samplers [16] or deriving personal exposure from street-level pollution measurements using mobile laboratories or stationary air monitoring stations [50]. Many studies also attempt to link specific physiological reactions to cyclists' exposure to air pollution and find evidence that short-term exposure can lead to adverse health effects [70]. One study even found that cyclists absorb a greater portion of fine particles PM2.5 and black carbon than drivers of motorized transport [29].

4.1 Formulation of the Problem

Here, the issue of the lack of quantitative determination of exposure of urban cyclists to PM in areas with poor air quality is addressed. In the city of Sofia, known for its high levels of air pollution, there is a lack of a comprehensive method for assessing and quantitatively determining the exposure levels of cyclists. This knowledge gap hinders the development of effective strategies to minimize exposure and provide healthier cycling routes.

To address this problem, we propose the development and evaluation of a software model that utilizes a modified route determination method reflecting the expected inhalation of PM by the cyclist, providing a healthier route. By incorporating data on air quality, pollutant concentrations, and traffic patterns into the software model, we can quantify cyclists' exposure to PM and provide recommendations for routes with minimal exposure. This innovative approach aims to fill the existing gap in quantitative exposure

determination and provide a practical tool for urban cyclists in polluted cities like Sofia.

The software implements the decision-making system outlined in this chapter. It is deployed in a cloud infrastructure, and Python, Django, and GraphQL (for the database) were used for its development. It combines GIS data with traffic and air quality data to apply the modified $k - SPwLO$ algorithm to find the most suitable bicycle route and display it on the OpenStreet map.

The input parameters include: PM measurements from fixed sensors; PM measurements from mobile sensors; traffic data; temperature and humidity from stationary sensors; start and end points of the journey; terrain elevation; heart rate.

4.2 Research study

This study examines the creation of a software tool aimed at selecting the optimal cycling route that minimizes the inhalation dose of PM for a cyclist traveling from point A to point B.

The inhalation dose (ID) of air pollutants is a variable that depends on pollutant concentrations, time, and ventilation rate (VR) [min]. We calculate the inhalation dose by incorporating PM exposure for each cyclist with biomarkers such as heart rate and the time required to travel each route into the model. The next subsection will provide more detailed information on the calculation methods used in the model.

To calculate the cyclist's ventilation rate (VR) (in l/min), we use the equation from the model [18], which is based on the heart rate (HR) [min] (Equation 4.1).

$$VR = 0,00070724 \times HR^{2,17008537} \tag{4.1}$$

To determine the quantities of particulate matter affecting cyclists, we use Equation 4.2 [18] to calculate the inhalation dose of PM for each segment:

$$_{inh} = \mathrm{PM}_{conc} \times VR \times time \tag{4.2}$$

where PM_{inh} (μg) is the mass of pollutants that enter the cyclists' respiratory tract during travel (in both directions); PM_{conc} ($\mu g/m^3$) is the average exposure to pollutants.

The formulas for calculating the inhalation dose (Equations 4.1 and 4.2) lead to the following hypothesis: If we want to create a tool that reduces the inhalation dose of PM, it should select a route that is fast and short. The shorter the time, the smaller the PM_{inh}; it requires less effort. HR increases during uphill climbs and high speeds. Look for routes with low elevation gain. Pass through areas with lower concentrations of PM; avoid heavily trafficked roads. Prefer smaller streets and park alleys.

The decision-making system is calibrated using data on HR, IR, cycling speed, PM measurements from mobile and stationary sensors, and traffic data. The software integrates traffic and air quality data from multiple sources with specific bicycle route data, including exact location, duration of ride, and elevation gain.

To calculate the least inhalation of polluted air, this software first needs to compute the shortest path in linear time [4], [71]. Additionally, it should present options that are longer than the shortest path but have different desired characteristics, such as lower vehicle traffic and minimal elevation gain. The k-Shortest Paths problem is a simple method for computing alternative routes [44].

Therefore, in this study, we decided to use an alternative routing, specifically k-Shortest Paths with Limited Overlap (k-SPwLO), which was introduced earlier in [11].

The k-SPwLO algorithm searches for paths that are: as short as possible; sufficiently different from each other.

Despite the method's better performance compared to the baseline solution, which lists paths in increasing order of their length, OnePass is not useful even for moderately sized road networks [12]. For this purpose, we use MultiPass - a more precise method that, by adding a second pruning criterion, extends and improves OnePass. MultiPass traverses the network k-1 times, but evaluates and expands only the most promising paths, unlike OnePass, which traverses the road network once and expands only those paths that meet a similarity criterion. Pruning is performed for each path that cannot lead to a solution.

Let P be a set of routes in the road network G connecting nodes s and t. A path p' (s $\rightarrow$ t) is called an "alternative" in P when p' is sufficiently different from every path p P. Formally, the overlap coefficient between p' and p determines how similar they are:

$$Sim(p', p) = \frac{\sum_{(n_x, n_y) \in p' \cap p} w_{xy}}{l(p)}, \tag{4.3}$$

where $p' \cap p$ denotes the set of edges that are common to p' and p. Given the similarity threshold θ, the route p' is an alternative to the set P if $Sim(p', p) \leq \theta$.

Given a source node s and a target node t, the k-SPwLO query provides a collection of k routes from s to t, arranged in increasing order of length, such that: the shortest route $p'(s \rightarrow t)$ is always included; all k routes are pairwise dissimilar with respect to the similarity threshold θ; all k routes are as short as possible.

The final result of the `k-SPwLO-modified` function is a single path. This path represents the shortest path from the source node to the target node, taking into account certain constraints such as avoiding major boulevards and locations with high levels of $PM_{2.5}$ or heavy traffic. The algorithm traverses through the neighbors of each node, considering their relevant characteristics based on the specified criteria, and selects the path with the lowest cost. If the algorithm successfully reaches the target node, it returns this single path.

4.2.1 Measurements from fixed sensors

The software utilizes an aggregation tool that extracts air quality data from both standard environmental monitoring instruments by IAOS and inexpensive sensor networks such as Luftdaten, Smog, Openaqi, and many others. It receives, records, cleans, and calibrates air quality data from fixed inexpensive sensors.

To mitigate the aforementioned limitations of fixed inexpensive sensors, we calibrate the data obtained from them. This is done by examining data from both environmental monitoring stations and the Luftdaten network of inexpensive sensors. To calibrate the data from the inexpensive sensors and enhance their reliability, we employ a two-step calibration method [75], which utilizes artificial neural networks and anomaly detection.

4.2.2 Measurement of PM exposure using mobile sensors

Portable low-cost pollution sensing devices are environmental monitoring devices that people can carry or wear while going about their daily activities. Since they detect pollution levels directly and in real-time, they can enable healthcare providers and researchers to monitor exposure at an individual level and empower citizens to manage their personal pollution exposure beyond what regulatory monitoring stations can achieve [72].

The particle sensor in the Airbeam2 - PMS 7003 is the seventh generation in the PMSx003 series developed by Plantower. The PMS7003 is a digital and universal particle concentration sensor used to measure the quantity of suspended particles in the air and output the results through a digital interface. It can be used with various instruments to provide accurate real-time concentration data. It provides a solid particle counting efficiency of 98% and can operate within a temperature range of -10 to 60 degrees Celsius.

4.2.3 PM Measurements and Calibration of the Decision-Making Model

As previously discussed, the software is programmed to search for a route where: the shortest path $p'(s \to t)$ is always included; the routes are pairwise dissimilar with respect to the similarity threshold θ; the routes are as short as possible.

The computational model relies on the previously mentioned formulas for calculating ventilation and inhalation rates, which are used to develop the weights for the search algorithm.

Additionally, the algorithm is further enhanced through route assessment, combining traffic data with air quality data from mobile and stationary sensors. The integration aids in more accurately forecasting the computation of the optimal route. Stationary sensors positioned near bicycle routes establish the baseline concentrations of PM pollutants in the area - $PM_{baseline}$. Meanwhile, traffic data provides information to estimate the PM concentrations from vehicles and transportation - $PM_{traffic}$. The expected exposure to PM pollutants PM_{exp} along the bicycle route is determined by combining these two variables.

$$PM_{exp} = PM_{baseline} + PM_{trafic} \tag{4.4}$$

As the bicycle route approaches main roads, the influence of traffic pollution on overall PM concentrations increases. Therefore, the variable $PM_{traffic}$ increases. The time of travel also affects this variable - passing during rush hours and near heavily congested streets and boulevards increases PM concentrations, as noted further in the chapter. Actual PM concentrations are estimated using measurements from mobile sensors. Mobile sensors provide highly accurate measurements as they are attached to the bicycle itself, transmit PM measurements every minute via a mobile phone, and reflect ultra-local pollution, such as being directly behind a bus or large truck. Later in the field studies, this assessment process is demonstrated.

The characteristics for evaluating bicycle routes are highly dependent on the location and features of the specific area. This is why we conducted a very detailed field study, examining two bicycle routes described in the next section.

4.2.4 Field Tests and Investigated Routes

In this section, the software model is evaluated through real-world field testing in Sofia, examining two bicycle routes. Both investigated routes start from the Vasil Levski National Stadium and end at "Petite Kiosque." These two locations are situated in the center of Sofia and experience active bicycle traffic.

The two selected routes were chosen based on the following criteria: they have the same starting and ending points; one route is suggested by the software while the other follows dedicated bicycle lanes; both routes are actively used by cyclists; both routes

pass near stationary sensors from the Luftdaten network; and both routes are deemed safe for regular cyclists.

On the following images, you can see the two paths used for evaluation: Segment A (Fig.4.1), which is the optimized bicycle route proposed by the software tool, passing through the shortest path via small central streets and parks; while Segment B (Fig.4.2) utilizes the developed bicycle infrastructure (mainly cycling near major road arteries with heavy traffic), is longer in distance, and is proposed by the navigation software as it follows dedicated bicycle lanes.

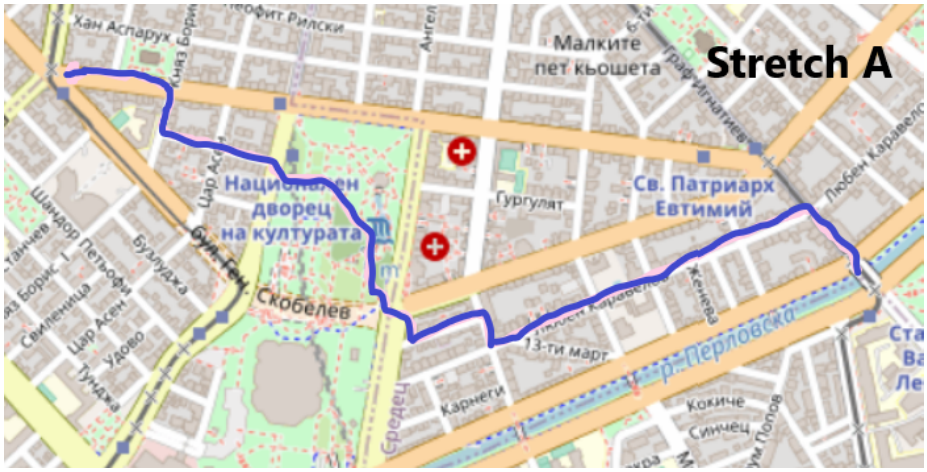

Figure 4.1: Section A - suggested by our software looking for the smallest inhalation dose

Figure 4.2: Section B - offered by most navigation software, which follows a designated cycle lane

Field tests were conducted with 10 participants in the study. The mobile equipment was attached to the front part of each bicycle, allowing the sampling lines to capture pollutants without obstruction; it was also secured from below to reduce vibrations.

We selected 10 individuals (aged 27-41) through contact with the local cycling community (8 males and 2 females). Before participating in the study, participants completed a pre-screening questionnaire. Exclusion criteria from the group included respiratory (including asthma), cardiovascular, or other chronic diseases, as well as tobacco smoking (current or recent). Only individuals who commuted daily by bicycle in Sofia were included. These factors were used to minimize the risk of injury due to unfamiliarity with Sofia's streets, lack of cycling experience, and adverse health consequences. Additionally, participants were asked to abstain from alcohol and caffeine for 48 hours prior to the tests.

Circular journeys were conducted on weekdays during rush hours (HT) (8:00-9:30 a.m.) and during off-peak hours (LT) (10:30 a.m.-12:00 p.m.) and on non-working days (NWD): weekends and holidays. The results from the three scenarios are evaluated separately in the next section. Field tests were conducted in April and May when the weather was moderate, and air pollution from domestic heating did not influence the data. Therefore, increases in particulate matter concentrations were mainly due to traffic. The ten participants performed circular journeys on both segments during HT, LT, and NWD.

4.3 Results and discussion

4.3.1 PM1 and PM2.5 concentrations from mobile measurements

In Table 4.1, the minimum, maximum, and mean (median) values of PM10 and PM2.5 concentrations for the two surveyed segments during working days under heavy traffic (HT) are shown. The concentrations of ultrafine particles with a diameter below 1 micron (PM10) ranged between 8 and 24 μg/m3 (average 11 μg/m3) for segment A and between 8 and 41 μg/m3 (average 14 μg/m3) for segment B. The concentrations of fine particles below 2.5 microns (PM2.5) ranged between 12 and 29 μg/m3 (average 15 μg/m3) for segment A and between 12 and 45 μg/m3 (average 19 μg/m3) for segment B.

Table 4.1: PM1 and PM2.5 concentrations from mobile measurements on weekdays during heavy traffic (HT)

		1			2,5	
	min	max	mean	min	max	mean
Stretch	8	24	11	12	29	15
Stretch B	8	41	14	12	45	19

In Table 4.2, the concentrations of PM10 and PM2.5 for segments A and B during working days under light traffic (LT) are presented. Measurements from segment A show lower concentration levels for both PM10 and PM2.5. The concentrations of ultrafine particles with a diameter below 1 micron (PM10) ranged between 6 and 19 μg/m3 (average 10 μg/m3) for segment A and between 5 and 34 μg/m3 (average 12 μg/m3) for segment B. The concentrations of fine particles below 2.5 microns (PM2.5) ranged between 10 and 24 μg/m3 (average 13 μg/m3) for segment A and between 9 and 38 μg/m3 (average 15 μg/m3) for segment B.

Table 4.3 illustrates the minimum, maximum, and median values of PM10 and PM2.5 concentrations for both surveyed segments during weekends and holidays.

Table 4.2: Concentrations of PM1 and PM2.5 from mobile measurements on weekdays with low traffic (LT)

	1			2,5		
	min	max	mean	min	max	mean
Stretch	6	19	10	10	24	13
Stretch B	5	34	12	9	38	15

Measurements for segment A show almost identical low concentration levels for PM10 and PM2.5. PM10 concentrations range between 3 and 7 μg/m3 (average 4 μg/m3) for segment A and between 3 and 8 μg/m3 (average 4 μg/m3) for segment B. PM2.5 concentrations range between 4 and 11 μg/m3 (average 7 μg/m3) for segment A and between 5 and 11 μg/m3 (average 7 μg/m3) for segment B.

Table 4.3: Concentrations of PM1 and PM2.5 from mobile measurements on non-working days

	1			2,5		
	min	max	mean	min	max	mean
Stretch	3	7	4	4	11	7
Stretch B	3	8	4	5	11	7

Despite the fact that the dedicated bicycle route is mostly exposed, the high volume of cars, buses, and trucks in this corridor is the main reason for the elevated pollution concentrations, as increased traffic raises the levels of PM. As a result, PM2.5 measurements are almost identical during weekends but with nearly 20% higher concentrations between the two routes during weekdays.

4.3.2 Ventilation Rate

To measure the ventilation rate (VR) during breathing, we conducted measurements of heart rate (HR), oxygen saturation (SpO2), and respiratory rate (RR) for each participant in the project. The average value of each parameter has low variability, and the final results for VR are presented in Table 4.4. No differences are observed in these three factors depending on whether the route is traversed on a workday or a non-workday, as they are not directly influenced by traffic or different levels of short-term air pollution exposure.

Table 4.4: Averaged indicators from field measurements

	Stretch	Stretch B
VR (L/min)	10,14	11,06
HR (beats/min)	82,30	85,66
Denivelation (m)	3	15
Speed (km/h)	14,3	14,7

Despite having the same starting and ending points, the distances are different, with

Segment A being shorter - 3.8 km compared to 4.4 km for Segment B, and it is faster - 15:52 minutes compared to 18:04 minutes for Segment B. The time taken to traverse both segments remains relatively constant on workdays and non-workdays. The specialized bike route had longer and straighter corridors with fewer intersections and crossings, leading to higher maximum and average speeds. Additionally, Segment B has a greater elevation change - 15 m compared to 3 m in Segment A. All these data result in increased cyclist speeds in Segment B - 11.06 km/h compared to 10.14 km/h for Segment A.

Table 4.5 shows the results of the inhalation dose of cyclists for PM10 and PM2.5. They are calculated based on the average exposure value of PM during the ride measured by mobile sensors, along with the time taken for the ride and the ventilation rate for each cyclist.

Table 4.5: Inhalation Dose of PM10 and PM2.5 during the Circuits

Period	Pollutant	Stretch A(optimized)	Stretch B(bike lanes)
Heavy traffic	PM 1_{inh}	29.74	46.45 (↑56%)
	PM2.5_{inh}	40.56	63.04 (↑55%)
Low traffic	PM 1_{inh}	27,04	39,82 (↑47%)
	PM2.5_{inh}	35.15	49.77 (↑42%)
Non-working days	PM1_{inh}	10.82	13.27 (↑23%)
	PM2.5_{inh}	18.93	23.23 (↑23%)

Despite having the same starting and ending points, Segment B shows an increased inhalation dose of particulate matter (PM) due to the longer duration of the route, requiring more effort, and exposing the cyclist to higher concentrations of PM due to vehicle exhaust gases. Even on non-workdays, when PM concentrations are similar across both routes, Segment B exhibits higher inhalation doses due to the longer duration and higher ventilation rate (due to higher elevation change and higher average speed).

Measurements for SpO2 and RR do not show any significant short-term health effects. This was expected since the study participants, selected for safety reasons, were non-smokers without chronic conditions and were regular cyclists. However, this does not necessarily mean that people with chronic conditions and sensitivity to air pollution may not experience symptoms or irritations, as observed in some studies.

4.3.3 Visualization of Bicycle Routes by Incorporating Data from Stationary Sensors

We will utilize the network of inexpensive stationary sensors from Luftdaten to demonstrate the technology for our fixed sensors in our software. Figure 4.3 displays the concentrations for Segment B during LT, where measurements from stationary sensors within 200 m or closer to the route are applied. The black-colored line indicates that in this part of the route, there is no fixed sensor closer than 200 m, while the green and yellow lines represent the concentrations measured by the fixed sensors nearby. Green color is used for PM2.5 concentration values between 0 and 12, while yellow indicates PM2.5 values between 12 and 35. These color categories are inspired by the EPA Air Quality Index and are the same for visualizing Aircasting routes measured with portable sensors. Thanks to the dense network of inexpensive sensors in Sofia, 5 fixed sensors are used for Segment A, and 6 sensors are used for Segment B, as they meet the selection criteria.

The software has continuous access to particulate matter (PM) concentrations from fixed sensors, while traffic data can dynamically change the input values for the decision-

Figure 4.3: Biking Route with Exposure from Fixed Sensors

making mechanism. The tool also has the functionality to display the same route in different time frames and accordingly show different air pollution concentrations.

4.3.4 Refinement of the model and evaluation of data from mobile and stationary sensors

The average values for PM2.5 concentrations from stationary sensors, located near the surveyed routes (within 200 meters), are compared with PM2.5 measurements from mobile sensors, as shown in Table 4.6 below. We find a strong correlation between the average values of $PM2.5_{fixed}$ and the minimum values for $PM2.5_{mobile}$, partly proving the assumption that the concentration from nearby fixed sensors can be used as an output for calculating the overall PM concentration estimate.

Data from the fixed sensors closely match the minimum concentrations along the route. Typically, this part of the route has negligible traffic, and its value often approximates $PM_{baseline}$, measured by the fixed sensors. Meanwhile, the maximum PM concentrations along the bicycle route are observed when passing by significant sources of pollution such as trucks, vehicles with old diesel engines, mopeds, or queues of cars waiting at traffic lights. As expected, the difference between the minimum and maximum PM concentrations is more significant during HT.

4.4 Conclusion

This chapter presents the development of a software tool that optimizes bicycle routes based on algorithms predicting the least harmful air pollutants. The algorithm utilizes a modified implementation of the k-shortest path with limited overlap method. It relies on experimental data and equations that calculate the total inhaled pollutant dose for cyclists. Additionally, the study evaluates two bicycle routes: Route A, suggested by the newly developed software, which traverses small streets in Sofia, and Route B, suggested by navigation apps, which follows a dedicated bicycle lane.

Table 4.6: PM2.5 concentrations from fixed sensors near the route and mobile measurements

		$2,5_{fixed}$		$2,5_{mobile}$	
Period	Route	mean	min	max	mean
WD-HT	Stretch A	10	12	29	15
	Stretch B	10	12	45	19
WD-LT	Stretch A	8	10	24	13
	Stretch B	8	9	38	15
NWD	Stretch A	4	4	11	7
	Stretch B	4	5	11	7

In the field test, ten cyclists make circular trips on both routes over three periods: (1) during heavy traffic, (2) during light traffic on workdays, and (3) on non-workdays. Based on the data collected during the study from a mobile sensor, cyclists' exposure and potential inhalation doses of PM2.5 and PM10 along both routes are calculated.

Exposure concentrations along dedicated bicycle lanes were found to be higher than exposure levels along the optimal route, especially on workdays. Even when average concentrations were nearly equal, the inhalation dose for cyclists was always higher on the bicycle lane route, as it took longer in terms of time and distance, had higher elevation differences, and required more intense cycling. Choosing the optimized bicycle route reduced the inhalation dose of PM2.5 by 23% on non-workdays and up to 55% during heavy traffic on workdays. Results during workdays show an additional health risk for cyclists using the investigated bicycle lanes due to pollutants directly associated with traffic. Exposure to PM concentrations in the investigated bicycle lanes was closely linked to vehicular traffic, as the study was not conducted during the heating season.

The findings of this study build upon previous observations [3], which suggest that redesigning streets for low-speed multimodal traffic without barriers is a more sustainable and pragmatic approach than constructing bicycle infrastructure on heavily trafficked roadways. There is a significant difference between studies conducted during light traffic and heavy traffic conditions. Bicycle lanes without physical barriers between the bicycle route and the road have higher exposure levels, a conclusion also noted in another study [36]. Our findings contribute to a better understanding of Sofia's air pollution issues, particularly related to vehicular traffic.

Chapter 5

Software System for Visualization of Air Pollution

In order to utilize all the research described up to this point and to convey information on how air pollution affects health in an accessible way, it is necessary to develop a comprehensive software system that allows for the collection, processing, and modeling of data.

The architecture of the software is modular, defining how different components interact, how data is processed, and how results are visualized. Figure 5.1 illustrates its appearance. The software architecture consists of the following components: API connection module; Central cloud repository; Data processing module; Forecasting module; Module for measuring PM concentration along a specified route; Data visualization and dissemination.

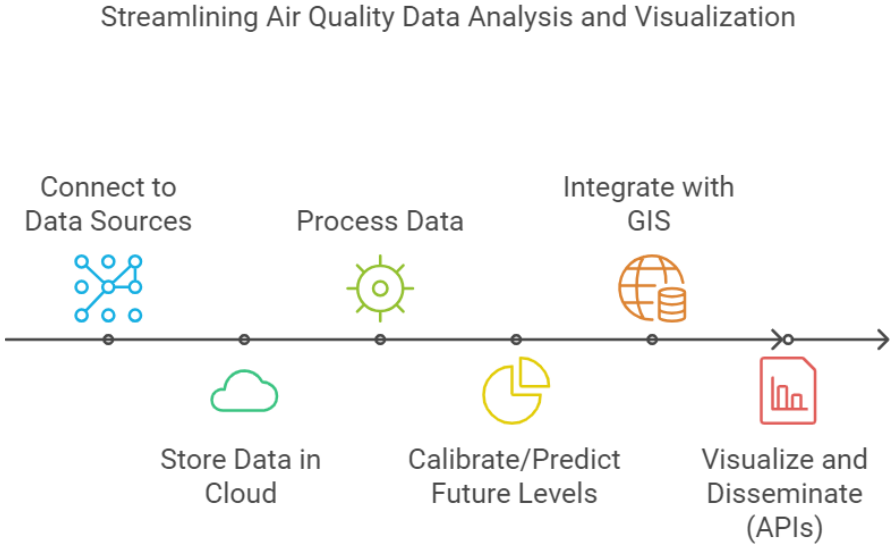

Figure 5.1: Software architecture

Ensuring the accuracy and reliability of our decision-making system is paramount to the success of our research. To achieve this, our software relies on a series of input parameters. These parameters include: measurements of PM concentration from stationary laser sensors; measurements of PM concentration from official monitoring stations; measurements of PM concentration from portable sensors; data on temperature and humidity from stationary sensors; traffic data.

The inclusion of these parameters is contingent upon their significance for assessing air quality and their potential impact on human health. Each parameter contributes to a comprehensive understanding of air quality, enabling our decision-making system to make informed forecasts and assessments.

To achieve this, we are committed to creating a modern air quality monitoring system that collects data from various sources.

The output data include: real-time air quality data by location, geographic radius, or city; station locations with their values on geographic maps; air quality forecast; calibrated data from specific stations; selection of the most suitable route based on PM pollution and traffic.

5.1 Data Collection Module

Access to data from national and civilian air quality monitoring networks is facilitated through public APIs (such as Luftdaten, aqicn, openaq, etc.). The built system periodically connects to these APIs over a specified period. These data provide us with a comprehensive view of air quality in a given area.

- Air Pollutants: We receive information about various air pollutants, including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO). Each pollutant contributes to the overall air quality index.

- Traffic Data: Real-time updates on road traffic help us understand how vehicle activity affects air quality. This data source helps us identify congestion points and emissions.

- Temperature: Knowledge of temperature is essential as it affects the dispersion and concentration of pollutants in the air. Higher pollution levels are often observed during the winter heating season and temperature inversions.

- Humidity: Humidity levels influence air quality by affecting the concentration and dispersion of particles and pollutants in the air. High humidity can improve air quality, while low humidity can exacerbate issues. Humidity is also an important factor when comparing and processing pollution data from different measurement devices. Laser measuring stations show higher error levels when humidity is high.

- Air Pressure: Air pressure data helps in forecasting meteorological conditions that may impact air quality. This is an important factor in understanding atmospheric stability and modeling air pollution forecasts.

5.2 Cloud Data Storage Module

The system for monitoring air quality and its impact on human health is based on an integrated approach that involves collecting and processing diverse data. We gather data on air pollutants, road traffic, temperature, humidity, and pressure from various sources and centralize and store them in the cloud infrastructure. This approach ensures scalability and flexibility, allowing us to meet the growing demands of the system.

By collecting data on air pollutants, road traffic, temperature, humidity, and pressure, we can provide a comprehensive view of air quality. This approach not only enhances our understanding of air quality but also enables us to make informed decisions and take timely actions to protect public health and the environment. Traffic data is integrated as it provides information on vehicle emissions and their impact on air quality. By combining traffic data with air quality measurements, we can identify areas with high pollution due to congestion and develop strategies to reduce pollution.

The data is updated in real-time, allowing us to provide users with up-to-date information on air quality. Integrating data from various sources makes analysis and information extraction easier and more efficient, providing a unified platform for data processing. Users with different privileges have easy access to information from anywhere with an internet connection.

5.3 Module Analysis, Forecasting, and Air Pollution Measurement

The Data Processing Module is the heart of the system and performs several key functions. Figure 5.2 provides a comprehensive architectural overview of the air quality monitoring system. It visually outlines the complex data flow between the data processing, forecasting, and pollution measurement modules. The diagram clearly demonstrates how data is processed, used for forecasting, and measured in real-time. Additionally, it emphasizes the integration of geographic context through GIS data, facilitating a holistic understanding of the system's functionalities.

First and foremost, this module provides basic data preprocessing, including cleaning and structuring. This means that data obtained from various sources and APIs undergo analysis and processing to ensure their accuracy and integrity. Cleaning involves removing invalid or corrupted data, while structuring them in a suitable format enables easier navigation and analysis of the information. It is important to note that the module includes a two-step machine learning-based calibration method, which enhances the accuracy of the data and prepares it for analysis.

5.4 Used technologies

In this section, we will focus on the reason for choosing Python and the related libraries for data modeling within our software system for measuring air quality.

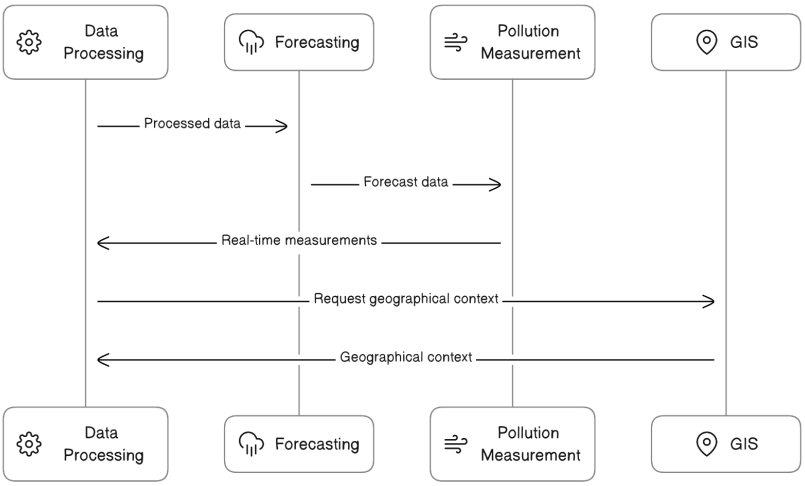

Figure 5.2: Architectural diagram showing the flow of data between the data processing modules.

We will provide information on how Python and these libraries integrate with the system and what key functionalities they provide.

Python is used as the primary programming language for developing the system, providing flexibility and power in data processing. Its readability and rich set of libraries make it a suitable choice for data processing and analysis.

Using Python and modeling libraries brings several significant advantages to our software system. Python provides a convenient environment for developing and testing models, allowing for rapid prototyping and efficient data modeling.

Python provides an extensive and flexible ecosystem of libraries that are crucial tools for data processing, including data obtained from financial and physical measurements. These libraries, such as NumPy, pandas, and SciPy, offer powerful tools for analysis, manipulation, and processing of numerical data. For instance, NumPy provides capabilities for efficient operations on multidimensional arrays and matrices, which are crucial when working with data obtained from physical measurements where the data structure is complex and requires fast and efficient traversal.

Django provides a powerful and robust framework for developing web applications. Built on top of Python, Django offers comprehensive tools for rapid and convenient development of web-based applications. An important aspect of this toolkit is the built-in data processing and the capability for storing and processing information in databases.

Combining Django with data processing tools like NumPy and pandas can be extremely useful for developing web applications that require data manipulation and analysis. Integrating these tools helps create powerful web-based applications that work with data from various sources and provide valuable analytical capabilities.

Additionally, Django provides capabilities for building public APIs, which is extremely useful for creating web applications that provide access to data through

standardized interfaces. This facilitates sharing and integrating data between different systems and applications.

GraphQL is a powerful and flexible query language for working with data, used to interact with various data sources including databases, web services, and others. By employing this approach to data management, we achieve significant flexibility and control over data queries, allowing for precise extraction of only the data required for a specific task or application.

GraphQL not only simplifies data handling but also facilitates efficient interaction between the client and server sides of the application by providing optimized queries and responses. This abstract and flexible approach to data management, combined with data processing tools like NumPy and pandas, can be crucial in creating applications that require complex and dynamic data queries.

Using GraphQL has helped us build applications that not only provide user satisfaction but also facilitate data extraction and analysis, which is crucial in today's world of information and data.

REST (Representational State Transfer) is an architectural style for creating web services and APIs that provides a secure and efficient way to communicate between different systems and applications. REST APIs offer a lightweight and straightforward method for accessing resources through common HTTP requests, such as GET (for retrieving data), POST (for creating new data), PUT (for updating data), and DELETE (for deleting data).

REST APIs provide an easy way to access and manage data through standardized and intuitive requests, making it suitable for integration into various applications and platforms. This structured and transparent approach to data handling ensures flexibility and scalability of applications.

The cloud repository serves as a central storage for data coming from various sources. This includes information on pollutant concentrations, meteorological data, traffic, and many other factors affecting air quality. The cloud infrastructure provides secure storage and management of vast amounts of data.

Utilizing cloud infrastructure is crucial for processing large volumes of data and delivering services that require flexibility and scalability. Cloud platforms such as Amazon Web Services (AWS) and Microsoft Azure are used, providing resources and services for storing, processing, and analyzing data on a global scale. Combined with Python, Django, GraphQL, and REST APIs, along with cloud infrastructure, we provide tools for analyzing air quality and its impact on health.

5.5 Example of Using the Software Infrastructure: The AirLief Mobile Application

At the core of our software's functionality lies its ability to integrate data from GIS, measurements of air quality from stationary monitoring stations, and traffic data. This integration is achieved through advanced data merging techniques, allowing us to gain a comprehensive understanding of air quality on various spatial and temporal scales. GIS data provides essential geographic context, enabling us to precisely determine air quality measurements. By overlaying air quality and traffic

data onto GIS layers, we can identify spatial patterns, hotspots, and correlations that would otherwise go unnoticed. This integration is crucial for our research as it allows us to assess the impact of traffic patterns and geographic characteristics on air quality levels.

The software has the capability to visualize air quality information on popular mapping platforms such as Google Maps and OpenStreetMap. This feature transforms raw data into a visually interpretable format, facilitating the dissemination of important information to a wide audience. We can analyze variations in air quality over time, identifying trends and anomalies to serve as the basis for our research findings. This temporal analysis provides insight into how air quality changes throughout different times of the day, seasons, or under certain meteorological conditions.

Integrating these diverse data sources into a cohesive framework enables us to draw significant conclusions about air quality and its interaction with the environment, contributing to the depth and comprehensiveness of our research.

AirLief is a mobile air quality application developed for iOS and Android platforms. It is a free, open-access application that shows users the air pollution levels around them based on their geolocation. AirLief retrieves data from the software system using the API created for data communication outside the system.

Air quality and atmospheric data are distributed from the software system to the mobile application through purpose-built infrastructure endpoints. To facilitate data transmission to our system, location information (GPS coordinates), geographic radius, or city inputs are entered, and real-time air quality data corresponding to the query is outputted. Three infrastructure endpoints have been created for the AirLief application:

- Nearest Point Measurements (input data: GPS coordinates): This endpoint allows users to receive results from the nearest monitoring station located near the specified GPS coordinates. It provides information about the current air quality at a specific location.

- Location-Based Measurements (input data: GPS coordinates, radius in km): This endpoint allows users to obtain the arithmetic mean values of measurements from all monitoring stations within a radius of specified kilometers around the given GPS coordinates. This gives a generalized overview of air quality in a particular area.

- City-Based Measurements (input data: city name): This endpoint enables users to obtain the arithmetic mean values of measurements from all monitoring stations in a specific city or populated place. It provides information about air quality within the boundaries of the specified city or populated place.

These infrastructure endpoints provide flexibility and convenience in extracting air pollution data, contributing to the extensive functionality of the air quality monitoring and management software system. They are presented to users in the mobile application in an easy and accessible manner.

The AirLief app retrieves and delivers air pollution information in an accessible way for over 100 countries from around 15,000 stations. The application itself is used in 70 countries by over 70,000 users, as shown in Figure 5.3. It consists of a dashboard that presents the current state of the air in an accessible and easy-to-understand manner; a map displaying values from stationary stations; the option to add favorite stations, and personalized tips for protection against air pollution. Figure 5.4 illustrates how the mobile application transforms raw data obtained from the software system into an Air Quality Index (AQI). The index, which directly correlates with the impact of air pollution on the human body, includes not only PM concentrations but also other pollutants such as ozone pollution, nitrogen dioxide, and others.

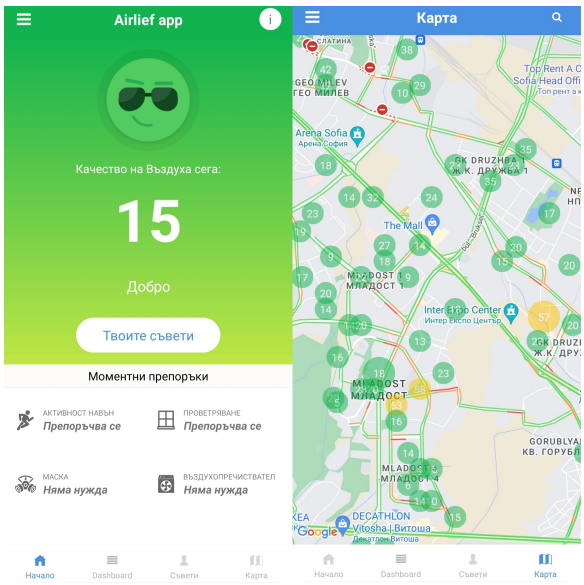

Figure 5.3: AirLief App is one of the applications developed from the software system, displaying real-time air pollution data from stationary sensors.

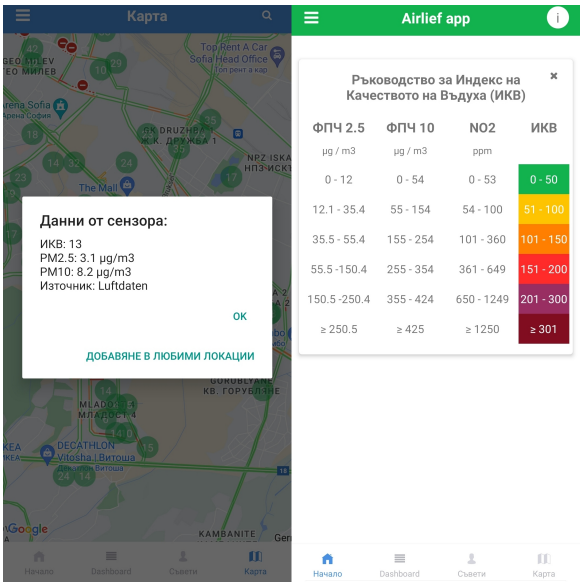

Figure 5.4: AirLief App - the mobile application that translates raw data into understandable Air Quality Index (AQI).

Chapter 6

Conclusion

The dissertation defines a goal and four tasks for achieving it. Initially, a detailed analysis was conducted on the data regarding air pollution and acute illnesses from two hospitals and emergency services in Sofia. The benefits of combining pollution data from official monitoring stations and laser sensor stations (low-cost IoT devices deployed by citizens) have been identified.

The dissertation develops and implements a solution for calibrating particulate matter (PM) data from laser stations using machine learning, which corrects errors and anomalies by comparing with data from a standardized station. Table 3.3 shows the improvement in laser station results, where the use of a combination of artificial neural networks and anomaly detection methods increases the R-squared value from 0.62 to 0.95 after calibration.

A route selection method based on PM pollution was developed, using a customized approach. Table 4.5 shows that in a field test with 10 cyclists, the routes generated by the software resulted in 67% better outcomes compared to other bicycle route-finding solutions.

All these tasks and solutions were developed into a scalable software system with a modular structure for collecting and processing data from various stations and IoT devices. This system facilitates effective integration with diverse hardware devices for measuring PM, utilizing different communication protocols. Practical results from the system's operation are visualized in a user interface, serving as a point of interaction between users and the system. The system is capable of processing and calibrating data as well as generating air pollution maps, and it is utilized in different projects, such as a mobile application accessed by thousands of users weekly.

The scientific and practical research presented in this dissertation provides the following contributions:

6.1 Scientific and Applied Contributions

1. A statistical correlation analysis was performed between air pollution and hospital and emergency admissions, identifying important cause-and-effect

relationships.

2. An algorithm for calibrating laser sensors was developed using a two-step method with artificial neural networks and anomaly detection. This innovative method supports more reliable and accurate measurements from laser sensors and has the potential to be used in a wide range of sensor applications.

3. An algorithm was created to evaluate the influence of humidity, altitude, and atmospheric pressure on air pollution data from laser sensors, allowing a more comprehensive and accurate understanding of the factors affecting air quality. It can also be applied in vertical planning.

4. An algorithm was developed to calculate the optimal cycling route based on PM concentration. This routing method promotes a healthy lifestyle and supports sustainable urban mobility.

6.2 Practical Contributions

1. Software for calibrating PM data from laser sensors was developed, using reference data from official monitoring stations. This software enables air quality control organizations to use more reliable data for decision-making and pollution management.

2. A software solution was developed to account for the influence of humidity, altitude, and atmospheric pressure, improving PM data from laser sensors. This software can be used by urban authorities and health organizations to monitor and manage air quality in real-time.

3. A software tool for finding the optimal cycling route based on PM inhalation was created. This software can be useful for individual cyclists and city infrastructures that promote cycling.

4. An IoT platform for aggregating and analyzing sensor data on air quality was developed. This platform supports automation and facilitates data integration from various sources, enhancing the efficiency and accuracy of air quality monitoring. In addition to meteorological data, it integrates traffic and GIS data. The platform visualizes maps, complemented by measurements from air quality sensors and real-time traffic data.

In light of the research and results presented in this dissertation, we can conclude that this project provides important scientific and practical contributions, influencing the development and improvement of air pollution measurement systems and related technologies.

6.3 Publications on the dissertation topic

- Zhivkov, P., & Simidchiev, A. (2024). Software Tool for Optimizing Cycling Route by Defining Cyclist Air Pollution Exposure
  Studies in Computational Intelligence, 1158 SCI, pp. 152–170, SJR 0.208 (2023)

  https://www.scopus.com/record/display.uri?eid=2-s2.0-85190651936&origin=resultslist&sort=plf-f&src=s&sid=3eca39362dd1584d77f2fb4c5901e3b2&sot=aut&sdt=a&sl=18&s=AU-ID%2857221769514%29&relpos=0&citeCnt=0&searchTerm=

- Zhivkov, P., & Simidchiev, A. (2022). Development of Software Tool for Optimization and Evaluation of Cycling Routes by Characterizing Cyclist Exposure to Air Pollution Annals of Computer Science and Information Systems, 32, 105-112.

  https://d1wqtxts1xzle7.cloudfront.net/99417790/230-libre.pdf?1677962458=&response-content-disposition=inline%3B+filename%3DDevelopment_of_Software_Tool_for_Optimiz.pdf&Expires=1725292377&Signature=QB4TSFkNK4~abulkZUncp_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

- Zhivkov, P. Optimization and Evaluation of Calibration for Low-cost Air Quality Sensors: Supervised and Unsupervised Machine Learning Models Annals of Computer Science and Information Systems, Volume 25, FedCSIS 2021, 2021, pp. 255–258

  https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117793346&doi=10.15439%2f2021F95&partnerID=40&md5=2f0e02da6b0d78aaa6523df4f48308c9

- Zhivkov, P., & Simidchiev, A. (2020). Relationship between Particulate Matter and Health Indicators for Acute Morbidity in Sofia.
  In Proceeding of 1st international Conference on Environmental Protection and Disaster Risks (pp. 180-193).

  https://www.ceeol.com/search/chapter-detail?id=926651

6.4 Declaration of originality

I declare that the dissertation contains original results, obtained during research conducted by me with the support and assistance of my supervisor.

The results that have been obtained, described and/or published by other scientists are correctly and extensively cited in the bibliography.

This dissertation has not been applied for a degree at another college, university or research institute.

Signature:

Bibliography

[1] Norfazillah Ab Manan, Azimatun Noor Aizuddin, and Rozita Hod. Effect of air pollution and hospital admission: a systematic review. *Annals of global health*, 84(4):670, 2018.

[2] Stacey E Alexeeff, Noelle S Liao, Xi Liu, Stephen K Van Den Eeden, and Stephen Sidney. Long-term pm2. 5 exposure and risks of ischemic heart disease and stroke events: review and meta-analysis. *Journal of the American Heart Association*, 10(1):e016890, 2021.

[3] Emma Barnes and Marc Schlossberg. Improving cyclist and pedestrian environment while maintaining vehicle throughput: Before-and after-construction analysis. *Transportation research record*, 2393(1):85–94, 2013.

[4] Hannah Bast, Daniel Delling, Andrew Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F Werneck. Route planning in transportation networks. In *Algorithm engineering*, pages 19–80. Springer, 2016.

[5] M.L. Bell, J.M. Samet, and F. Dominici. Time-series studies of particulate matter. *Annu. Rev. Public Health*, 25:247–280, 2004.

[6] Alexander Y Bigazzi and Miguel A Figliozzi. Review of urban bicyclists' intake and uptake of traffic-related air pollution. *Transport Reviews*, 34(2):221–245, 2014.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] Iain M Carey, Richard W Atkinson, Andrew J Kent, Tjeerd Van Staa, Derek G Cook, and H Ross Anderson. Mortality associations with long-term exposure to outdoor air pollution in a national english cohort. *American journal of respiratory and critical care medicine*, 187(11):1226–1233, 2013.

[9] N. Castell et al. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment international*, 99:293–302, 2017.

[10] J.W. Cherrie et al. Effectiveness of face masks used to protect beijing residents against particulate air pollution. *Occupational and environmental medicine*, 75(6):446–452, 2018.

[11] Theodoros Chondrogiannis, Panagiotis Bouros, Johann Gamper, and Ulf Leser. Alternative routing: k-shortest paths with limited overlap. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–4, 2015.

[12] Theodoros Chondrogiannis, Panagiotis Bouros, Johann Gamper, and Ulf Leser. Exact and approximate algorithms for finding k-shortest paths with limited overlap. In *20th International Conference on Extending Database Technology: EDBT 2017*, pages 414–425, 2017.

[13] L. Clancy et al. Effect of air-pollution control on death rates in dublin, ireland: an intervention study. *The lancet*, 360(9341):1210–1214, 2002.

[14] Aaron J Cohen, Michael Brauer, Richard Burnett, H Ross Anderson, Joseph Frostad, Kara Estep, Kalpana Balakrishnan, Bert Brunekreef, Lalit Dandona, Rakhi Dandona, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *The Lancet*, 389(10082):1907–1918, 2017.

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

[16] Audrey De Nazelle, Mark J Nieuwenhuijsen, Josep M Antó, Michael Brauer, David Briggs, Charlotte Braun-Fahrlander, Nick Cavill, Ashley R Cooper, Hélène Desqueyroux, Scott Fruin, et al. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. *Environment international*, 37(4):766–777, 2011.

[17] CE Delft. Health costs of air pollution in european cities and the linkage with transport. *CE Delft: Delft, The Netherlands*, 2020.

[18] Inês D Do Vale, Ana S Vasconcelos, and Gonçalo O Duarte. Inhalation of particulate matter in three different routes for the same od pair: A case study with pedestrians in the city of lisbon. *Journal of Transport & Health*, 2(4):474–482, 2015.

[19] F. Dominici, L. Sheppard, and M. Clyde. Health effects of air pollution: a statistical review. *International Statistical Review*, 71(2):243–276, 2003.

[20] Jiyuan Dong, Yurong Liu, and Hairong Bao. Revalue associations of short-term exposure to air pollution with respiratory hospital admissions in lanzhou, china after the control and treatment of current pollution. *International Journal of Hygiene and Environmental Health*, 231:113658, 2021.

[21] Muriel Gevrey, Ioannis Dimopoulos, and Sovan Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264, 2003.

[22] M. Gochfeld and J. Burger. Disproportionate exposures in environmental justice and other populations: the importance of outliers. *American Journal of Public Health*, 101(S1):S53–S63, 2011.

[23] T. Guan et al. The effects of facemasks on airway inflammation and endothelial dysfunction in healthy young adults: a double-blind, randomized, controlled crossover study. *Particle and fibre toxicology*, 15(1):30, 2018.

[24] A. Gábelová et al. Genotoxicity of environmental air pollution in three european cities: Prague, košice and sofia. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 563(1):49–59, 2004.

[25] David H Hagan, Gabriel Isaacman-VanWertz, Jonathan P Franklin, Lisa MM Wallace, Benjamin D Kocar, Colette L Heald, and Jesse H Kroll. Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. *Atmospheric Measurement Techniques*, 11(1):315–328, 2018.

[26] Walter Ham, Abhilash Vijayan, Nico Schulte, and Jorn D Herner. Commuter exposure to pm2. 5, bc, and ufp in six common transport microenvironments in sacramento, california. *Atmospheric Environment*, 167:335–345, 2017.

[27] Ghassan B Hamra, Neela Guha, Aaron Cohen, Francine Laden, Ole Raaschou-Nielsen, Jonathan M Samet, Paolo Vineis, Francesco Forastiere, Paulo Saldiva, Takashi Yorifuji, et al. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environmental health perspectives*, 2014.

[28] A.J. Hedley et al. Cardiorespiratory and all-cause mortality after restrictions on sulphur content of fuel in hong kong: an intervention study. *The lancet*, 360(9346):1646–1652, 2002.

[29] Mario A Hernández, Omar Ramírez, John A Benavides, and Juan F Franco. Urban cycling and air quality: Characterizing cyclist exposure to particulate-related pollution. *Urban Climate*, 36:100767, 2021.

[30] David M Holstius, A Pillarisetti, KR Smith, and EJAMT Seto. Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in california. *Atmospheric Measurement Techniques*, 7(4):1121–1131, 2014.

[31] Jalil Jaafari, Kazem Naddafi, Masud Yunesian, Ramin Nabizadeh, Mohammad Sadegh Hassanvand, Mansour Shamsipour, Mohammad Ghanbari Ghozikali, Hamid Reza Shamsollahi, Shahrokh Nazmara, and Kamyar Yaghmaeian. The acute effects of short term exposure to particulate matter from natural and anthropogenic sources on inflammation and coagulation markers in healthy young adults. *Science of The Total Environment*, 735:139417, 2020.

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

[33] Wan Jiao, Gayle Hagler, Ronald Williams, Robert Sharpe, Ryan Brown, Daniel Garver, Robert Judge, Motria Caudill, Joshua Rickard, Michael Davis, et al. Community air sensor network (cairsense) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern united states. *Atmospheric Measurement Techniques*, 9(11):5281–5292, 2016.

[34] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.

[35] K. Katsouyanni et al. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of Epidemiology & Community Health*, 50(Suppl 1):S12–S18, 1996.

[36] Christine M Kendrick, Adam Moore, Ashley Haire, Alexander Bigazzi, Miguel Figliozzi, Christopher M Monsere, and Linda George. Impact of bicycle lane characteristics on exposure of bicyclists to traffic-related particulate matter. *Transportation research record*, 2247(1):24–32, 2011.

[37] Kirsten A Koehler and Thomas M Peters. New methods for personal exposure monitoring for airborne particles. *Current environmental health reports*, 2(4):399–411, 2015.

[38] Suk Wah Kwok and Chris Carter. Multiple decision trees. In *Machine Intelligence and Pattern Recognition*, volume 9, pages 327–335. Elsevier, 1990.

[39] P. Li et al. The acute effects of fine particles on respiratory mortality and morbidity in beijing, 2004–2009. *Environmental Science and Pollution Research*, 20(9):6433–6444, 2013.

[40] Xiaoting Liu, Rohan Jayaratne, Phong Thai, Tara Kuhn, Isak Zing, Bryce Christensen, Riki Lamont, Matthew Dunbabin, Sicong Zhu, Jian Gao, et al. Low-cost sensors as an alternative for long-term air quality monitoring. *Environmental research*, 185:109438, 2020.

[41] Feng Lu, Dongqun Xu, Yibin Cheng, Shaoxia Dong, Chao Guo, Xue Jiang, and Xiaoying Zheng. Systematic review and meta-analysis of the adverse health effects of ambient pm2. 5 and pm10 pollution in the chinese population. *Environmental research*, 136:196–204, 2015.

[42] Lisha Luo, Yunquan Zhang, Junfeng Jiang, Hanghang Luan, Chuanhua Yu, Peihong Nan, Bin Luo, and Mao You. Short-term effects of ambient air pollution on hospitalization for respiratory disease in taiyuan, china: a time-series analysis. *International journal of environmental research and public health*, 15(10):2160, 2018.

[43] Kamal Jyoti Maji, Wei-Feng Ye, Mohit Arora, and SM Shiva Nagendra. Pm2. 5-related health and economic loss assessment for 338 chinese cities. *Environment international*, 121:392–403, 2018.

[44] Ernesto QV Martins and Marta Pascoal. A new implementation of yen's ranking loopless paths algorithm. *Quarterly Journal of the Belgian, French and Italian Operations Research Societies*, 1(2):121–133, 2003.

[45] L. Morawska et al. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment international*, 116:286–299, 2018.

[46] Christopher JL Murray. The global burden of disease study at 30 years. *Nature medicine*, 28(10):2019–2026, 2022.

[47] National Statistical Institute of Bulgaria (NSI). Population by cities https://www.nsi.bg/en/content/2981/population-towns-and-sex, 2019.

[48] World Health Organization. *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index*. 1978.

[49] World Health Organization. International classification of diseases (icd) information sheet. 2018.

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

[51] Hye-Youn Park, Susan Gilbreath, and Edward Barakatt. Respiratory outcomes of ultrafine particulate matter (ufpm) as a surrogate measure of near-roadway exposures among bicyclists. *Environmental Health*, 16(1):1–7, 2017.

[52] S. Pattenden, B. Nikiforov, and B. Armstrong. Mortality and temperature in sofia and london. *Journal of Epidemiology & Community Health*, 57(8):628–633, 2003.

[53] C.A. Pope, D.W. Dockery, and J. Schwartz. Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation toxicology*, 7(1):1–18, 1995.

[54] Xiaoliang Qin, Lujian Hou, Jian Gao, and Shuchun Si. The evaluation and optimization of calibration methods for low-cost particulate matter sensors: Inter-comparison between fixed and mobile methods. *Science of The Total Environment*, 715:136791, 2020.

[55] Aakash C Rai, Prashant Kumar, Francesco Pilla, Andreas N Skouloudis, Silvana Di Sabatino, Carlo Ratti, Ansar Yasar, and David Rickerby. End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of The Total Environment*, 607:691–705, 2017.

[56] Sanjay Rajagopalan, Sadeer G Al-Kindi, and Robert D Brook. Air pollution and cardiovascular disease: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 72(17):2054–2070, 2018.

[57] A.M. Rendón et al. Effects of urbanization on the temperature inversion breakup in a mountain valley with implications for air quality. *Journal of Applied Meteorology and Climatology*, 53(4):840–858, 2014.

[58] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[59] J. Samet and D. Krewski. Health effects associated with exposure to ambient air pollution. *Journal of toxicology and environmental health, Part A*, 70:227–242, 2007.

[60] J.M. Samet et al. The national morbidity, mortality, and air pollution study. part ii: morbidity and mortality from air pollution in the united states. *Res Rep Health Eff Inst*, 94(pt 2):5–79, 2000.

[61] David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, and Qingmai Wang. Anomaly detection in online social networks. *Social Networks*, 39:62–70, 2014.

[62] J. Shi et al. Cardiovascular benefits of wearing particulate-filtering respirators: a randomized crossover trial. *Environmental health perspectives*, 125(2):175–180, 2017.

[63] S. Shin et al. Ambient air pollution and the risk of atrial fibrillation and stroke: A population-based cohort study. *Environ Health Perspect*, 127(8):87009, 2019.

[64] L.A. Tallon et al. Cognitive impacts of ambient air pollution in the national social health and aging project (nshap) cohort. *Environ Int*, 104:102–109, 2017.

[65] Yan Tao, Shengquan Mi, Shuhong Zhou, Shigong Wang, and Xiaoyun Xie. Air pollution and hospital admissions for respiratory diseases in lanzhou, china. *Environmental pollution*, 185:196–201, 2014.

[66] G.D. Thurston et al. A joint ers/ats policy statement: what constitutes an adverse health effect of air pollution? an analytical framework. *European Respiratory Journal*, 49(1), 2017.

[67] H. Traboulsi et al. Inhaled pollutants: the molecular scene behind respiratory and systemic diseases associated with ultrafine particulate matter. *International Journal of Molecular Sciences*, 18(2):243, 2017.

[68] J. Vermylen et al. Ambient air pollution and acute myocardial infarction. *Journal of Thrombosis and Haemostasis*, 3(9):1955–1961, 2005.

[69] W. Wang et al. Particulate air pollution and ischemic stroke hospitalization: How the associations vary by constituents in shanghai, china. *Sci Total Environ*, 695:133780, 2019.

[70] Scott Weichenthal, Ryan Kulka, Aimee Dubeau, Christina Martin, Daniel Wang, and Robert Dales. Traffic-related air pollution and acute changes in heart rate variability and respiratory function in urban cyclists. *Environmental health perspectives*, 119(10):1373–1378, 2011.

[71] Lingkun Wu, Xiaokui Xiao, Dingxiong Deng, Gao Cong, Andy Diwen Zhu, and Shuigeng Zhou. Shortest path and distance queries on road networks: An experimental evaluation. *arXiv preprint arXiv:1201.6564*, 2012.

[72] Sherrie Xie and Blanca E Himes. Personal environmental monitoring. In *Precision in Pulmonary, Critical Care, and Sleep Medicine*, pages 305–320. Springer, 2020.

[73] Antonella Zanobetti, Francesca Dominici, Yun Wang, and Joel D Schwartz. A national case-crossover analysis of the short-term effect of pm2. 5 on hospitalizations and mortality in subjects with diabetes and neurological disorders. *Environmental Health*, 13(1):1–11, 2014.

[74] R. Zhang et al. Acute effects of particulate air pollution on ischemic stroke and hemorrhagic stroke mortality. *Front Neurol*, 9:827, 2018.

[75] Petar Zhivkov. Optimization and evaluation of calibration for low-cost air quality sensors: Supervised and unsupervised machine learning models. In *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 255–258. IEEE, 2021.

[76] Petar Zhivkov and Alexander Simidchiev. Development of software tool for optimization and evaluation of cycling routes by characterizing cyclist exposure to air pollution. *Annals of Computer Science and Information Systems*, 32:105–112, 2022.

[77] Petar Zhivkov and Alexander Simidchiev. Software tool for optimizing cycling route by defining cyclist air pollution exposure. In *The Workshop on Computational Optimization*, pages 152–170. Springer, 2024.

[78] N. Zimmerman et al. Closing the gap on lower cost air quality monitoring: Machine learning calibration models to improve low-cost sensor performance. *Atmos. Meas. Tech. Discuss*, pages 1–36, 2017.

[79] Marina Zusman, Cooper S Schumacher, Amanda J Gassett, Elizabeth W Spalt, Elena Austin, Timothy V Larson, Graeme Carvlin, Edmund Seto, Joel D Kaufman, and Lianne Sheppard. Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. *Environment international*, 134:105329, 2020.

[80] Moniek Zuurbier, Gerard Hoek, Marieke Oldenwening, Virissa Lenters, Kees Meliefste, Peter Van Den Hazel, and Bert Brunekreef. Commuters' exposure to particulate matter air pollution is affected by mode of transport, fuel type, and route. *Environmental health perspectives*, 118(6):783–789, 2010.