UNIWERSYTET WSB MERITO W GDAŃSKU
WYDZIAŁ INFORMATYKI I NOWYCH TECHNOLOGII

Szymon Bartoszewicz
Nr albumu: 65788

MULTIVARIATE REGRESSION METHOD FOR AIR POLLUTION PREDICTION IN A GIVEN AREA

Praca magisterska
na kierunku Informatyka

Praca napisana pod kierunkiem
Prof. dr. hab. inż. Antoniego Wilińskiego

Gdańsk, 2023

OŚWIADCZENIE AUTORA PRACY

Ja, niżej podpisany student Wyższej Szkoły Bankowej w Gdańsku oświadczam, że:

1. Wersja elektroniczna przedkładanej pracy dyplomowej jest wersją ostateczną przedstawioną do egzaminu dyplomowego w Wyższej Szkole Bankowej w Gdańsku;

2. Praca ta stanowi mój utwór i nie narusza majątkowych i osobistych praw autorskich innych osób.

3. Udzielam nieodpłatnie Wyższej Szkole Bankowej w Gdańsku licencji na umieszczanie ww. pracy w elektronicznym archiwum prac oraz do zwielokrotniania i udostępniania tej pracy w zakresie koniecznym do ochrony mojego prawa do autorstwa lub praw osób trzecich, w tym w systemach antyplagiatowych;

4. Udzielam nieodpłatnie Wyższej Szkole Bankowej w Gdańsku licencji na udostępnianie mojej pracy w archiwum prac Wyższej Szkoły Bankowej w Gdańsku bez ograniczeń czasowych, terytorialnych i ilościowych;

5. Nie udzielam nieodpłatnie Wyższej Szkole Bankowej w Gdańsku licencji na udostępnianie mojej pracy w sieci Internet bez ograniczeń czasowych, terytorialnych i ilościowych;

6. Oświadczam, że praca dyplomowa nie zawiera informacji podlegających ochronie na podstawie przepisów o ochronie informacji niejawnych.

Data ..... Podpis autora pracy .....

TABLE OF CONTENTS

Abstract

Title: Multivariate Regression Method for Air Pollution Prediction in a Given Area

The increasing air pollution is causing significant risks to public health and the global economy. It caused premature deaths to seven million people (about fourteen times the population of Gdansk) in 2019. The issue emphasizes the need for reliable forecasting models to be developed. This thesis investigates a novel algorithm, combining multivariate regression, the sliding window approach, and the Moore-Penrose pseudoinverse to predict air pollution levels in Warsaw – a city with significant pollution. The research identifies a promising algorithm utilizing a 30-day sliding window for forecasting air pollution. This approach delivers encouraging results when integrated with the Least Absolute Shrinkage and Selection Operator (LASSO) method for predictor selection. The primary application of this model involves its integration into software applications, providing an accessible tool for individuals to plan outdoor activities, thereby mitigating health risks associated with high pollution levels. This research contributes significantly to environmental data science and enhances public access to vital air quality information by offering an innovative solution to a crucial global concern.

Introduction

In the modern world, there are a myriad of environmental challenges. Among many, air pollution is a significant concern. It has a detrimental impact on health, ecosystems, and the economy. Considering those, the need for developing reliable predictive, accurate models for air pollution becomes increasingly critical. The public and people in power should receive timely and accurate information about air pollution. This thesis aims to address this need by exploring an algorithm combining the multivariate regression approach, sliding window, and Moore-Penrose pseudoinverse. The method forecasts air pollution levels in Warsaw – the city where the issue is of substantial concern.

The inspiration to undertake this subject was born because of the topic's usefulness. It is paramount and relevant in today's society. Influenced by a complex mix of factors, air pollution poses a significant challenge in developing an accurate and reliable predictive model. This research's originality lies in employing a combination of multivariate regression, a sliding window technique, and the Moore-Penrose inverse, providing a nuanced understanding of the complex interactions at play.

The main study objective is to develop a multivariate regression algorithm to predict main pollutants levels – PM2.5 and PM10 levels. Additionally, the research aims to evaluate the performance of various feature selection methods on prediction accuracy and to investigate the impact of different time periods of the year on forecast accuracy. The paper will also evaluate the developed algorithm's performance using error metrics such as MAE, RMSE, R-squared, and MAPE.

The study's framework is as follows: data acquisition, preprocessing, feature selection, and model implementation. The results are then presented and discussed, highlighting the model's performance, the implications of the findings for air pollution prediction and management, and potential improvements and limitations of the study.

The research is significant in multiple ways. Firstly, one of the main contributions of this study is to provide a more thorough understanding of how various factors relate to levels of air pollution. This knowledge adds to the existing body of literature on the subject. Secondly, the algorithm used in this paper can find its use case as implementation in a web or mobile application offering air pollution forecast. That would assist in planning everyday activities for parents, mothers with kids or older adults. Thirdly, the findings could assist environmental agencies and policymakers in Warsaw to make informed

decisions about air pollution management and public health planning. Finally, the approach utilized in this research could apply to other cities or areas, expanding the understanding of air pollution patterns and forecasting.

No work can be done without some barriers and difficulties. For example, data acquisition and preprocessing proved challenging due to the accessibility and complex nature of environmental data. Nonetheless, overcoming these challenges provided valuable insights and learning experiences. This endeavor has emphasized the significance of precise and dependable predictive models in guiding society and environmental management.

The structure of the thesis develops as follows: the first chapter delivers an overview of air pollution, its effects, classifications, and prediction methods. Chapter 2 outlines the methodology, describing the data acquisition, preprocessing, feature selection, and implementation process. Finally, the third chapter delivers the study's results, while chapter 4 discusses the findings' implications and potential improvements.

1. Air Pollution Issue

1.1. Air Pollution and Its Effects

All creatures on the earth have their survival instinct – a willingness to stay alive. Human is not different in that aspect. A human has three primary external sources: food, water, and air. With a lack of any of those, it dies – just without food but with proper hydration — after 2-3 months (Chalela & Lopez, 2013), without water – within a week (WMA, 2006), and without air – within minutes (Suggitt, 2021). Regarding food, more specifically, the human body needs essential nutrients: carbohydrates, lipids, proteins, vitamins, minerals, and water (Snell et al., 2022).

Nevertheless, when considering factors such as longevity and healthy living, not only the amount of those compounds counts. Besides the quantity of the mentioned ingredients, another crucial factor is the quality of those components – how healthy or unhealthy they might be. Nowadays, one can observe soaring consumption of so-called ultra-processed food. This type of food mainly consists of "cheap industrial sources of dietary energy and nutrients plus additives, using a series of processes" (Monteiro et al., 2018). It is not unknown that they have a negative influence on human health. A recent study shows that increased ultra-processed food intake may be associated with a higher mortality risk (Schnabel et al., 2019). Water is not different. Everyone can agree that when it comes to water and its quality, drinking water, compared to water used for cleaning, laundry, and bath, should be the cleanest one – have the highest quality of itself. Drinking water should contain specific minerals and a specific amount to maintain human health (Dore et al., 2021). Considering just nutrition, it is not uneasy to spot that the human body needs specific conditions to function correctly. The third resource that is invaluable for living is oxygen from the air. The earth's atmosphere comprises multiple gases like nitrogen, oxygen, argon, carbon dioxide, neon, etc.

However, in the atmosphere, besides indispensable for living organisms O_2, there are also so-called pollutants that consist of solid particles, liquid droplets, or gases that are a hazard to organisms and the environment. **Air pollution** is interpreted as those substances' existence in the atmosphere (Zehnder et al., 2018). The quality of the breathing air depends on the number of pollutants that lie in it. For the simple reason that we cannot stop breathing, air pollution is of paramount importance nowadays. It is no

trifling issue as it impacts every aspect of the modern world – health, the economy, and the environment.

Health Effects of Air Pollution

Health is undoubtedly one of the most important values. Some statistics can describe the significant impact of air pollution on health.

In 2019 almost 7 million people died prematurely due to ambient and household air pollution (Egerstrom et al., 2022).

- 89% of premature deaths in 2019 occurred in low- and middle-income countries, mainly in WHO South-East Asia and Western Pacific Regions (WHO, 2022c).
- In 2019, nearly 500 000 infants died in their first month due to air pollution (Health Effects Institute, 2020b).
- There is a 1% reduction of cognitive function for every 10μg/m^3 of PM2.5 (Cedeño Laurent et al., 2021).

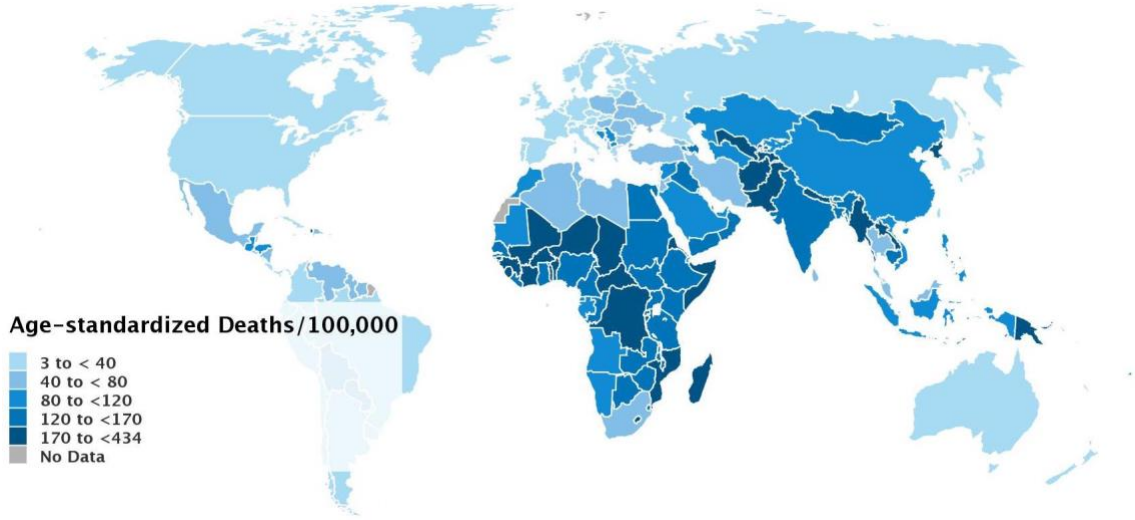

Figure 1: Age-standardized Deaths/100,000 Attributable to Air Pollution in 2019 (Health Effects Institute, 2020a)

As data suggests, there is confirmed evidence of the negative impact of air pollution on various aspects of health – starting with short-term exposure and immediate effects like a drop in cognitive function of the brain, through skin damage, to shorten life expectancy and development of chronic diseases which lead to death. The young, the elderly, and people with pre-existing conditions are especially vulnerable to those effects. These adverse health consequences also have an impact on the economy.

Economic Impacts of Air Pollution

The effects of air pollution do not just stop with health. Poor air quality directly impacts the economy, costing $8.1 trillion (6.1 percent of global GDP) in 2019 alone (World Bank, 2022). The economic damages from air pollution are primarily concentrated within a few specific economic sectors. For instance, four significant sectors in the American economy – agriculture, utilities, manufacturing, and transportation – were responsible for more than 75% of all air pollution related damage while contributing just under 20% of the GDP (Tschofen et al., 2019). Most of the economic cost is due to the health impacts, including illness and death. Over the years, studies have shown a direct link between short- and long-term exposure to poor air quality and mortality of different kinds (Huangfu & Atkinson, 2020; Orellano et al., 2020). It is estimated that nearly 7 million people can die due to air pollution related causes (Egerstrom et al., 2022), which leaves a significant burden on the economy. The situation leads to increased healthcare costs. It can also devastate the individuals sickened by air pollution or their families if someone dies due to poor air quality. The enormous costs can be seen in that way.

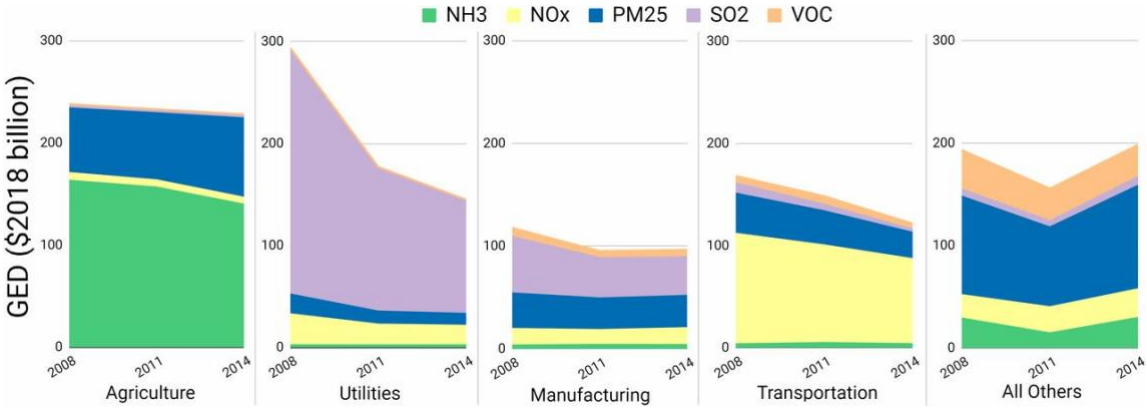

Figure 2: GED (in $2018) attributable to economic sectors and their respective precursor pollutants (NH3, NOx, primary PM2.5, SO2, and VOCs). GED was calculated for the three most recent NEI years: 2008, 2011, and 2014. (Tschofen et al., 2019)

However, the budget does not stop on the expenditures due to the health consequences of air pollution. On top of everything, countries spend additional sums to clean up the air. Over the years, the Clean Air Act in the US has made a significant improvement in the quality of air. Cleaning up the air costs an additional estimated $65 billion (about $200 per person in the US) yearly (EPA, 2011).

1.2. Air Pollution Classification

Air pollution is no trifling matter in the modern world. Taking into consideration places where one can experience air pollution, it can be categorized as indoor (household) and ambient (outdoor) pollution. Considering how pollutants originated, two categories are primary and secondary pollutants.

Indoor (household) air pollution is, as the name suggests, a kind of air pollution that one meets indoors, e.g., in houses, public buildings, workplaces, and means of transport. Sources involve household products (e.g., detergents, office materials, bleaches, cleaners, and polishes), tobacco smoke, chemicals, and the out-gassing of building materials (Fisher, 2021). Many people in the world use in their houses solid fuels (such as coal, charcoal, wood, dung, or crop residues) as a primary fuel for cooking. In 2017, an estimated 3 billion people (39% of the global population) were in that situation. Those kinds of conditions outcome in a dangerous level of pollutants (e.g., PM2.5, carbon monoxide) in the air (WHO, 2014). It resulted in 3.2 million deaths per year in 2020, of which 237 thousand consist of children under the age of 5 (WHO, 2022b). However, this pollution is not only a threat in developing countries. The study conducted in China, India, Mexico, the United Kingdom, Thailand, the United States of America shows that poor indoor air quality diminishes brain cognition (Cedeño Laurent et al., 2021). That case is especially relevant in those days as the employee's productivity or students' focus count. It is indeed worth noting because nowadays, most of the time is spent indoors: in homes, commuting by car, public transport, in offices, etc., where the hazard of indoor air pollution is high.

Nonetheless, the most often discussed type of air pollution is **ambient** (**outdoor) air pollution**. It is estimated that 99% of all people on the planet live in places where ambient air pollution is exceptionally hazardous (WHO, 2022a). Particulate matter (or particle pollution – PM), ground-level ozone, sulfur oxides (SO_2), carbon monoxide (CO), nitrogen oxides (NO_2), and lead (Pb) are commonly distinguished pollutants. There are many air pollution sources, which depend on the context. One way to classify them is as follows (Fisher, 2021).

- Mobile sources – these are the sources that are able to move on their own power supply, e.g., cars, scooters, trains, ships, etc.
- Stationary sources (point sources) – immobile objects such as power plants, factories, sewage treatment, etc.

- Area sources – consist of a set of small sources where collective pollution is substantial, e.g., cities and housing developments.
- Agricultural sources – they come from human agrarian activities such as livestock and fertilizer.
- Natural sources – sources that are not connected with human activities, for instance, storms, volcanic eruptions, and wildfires.

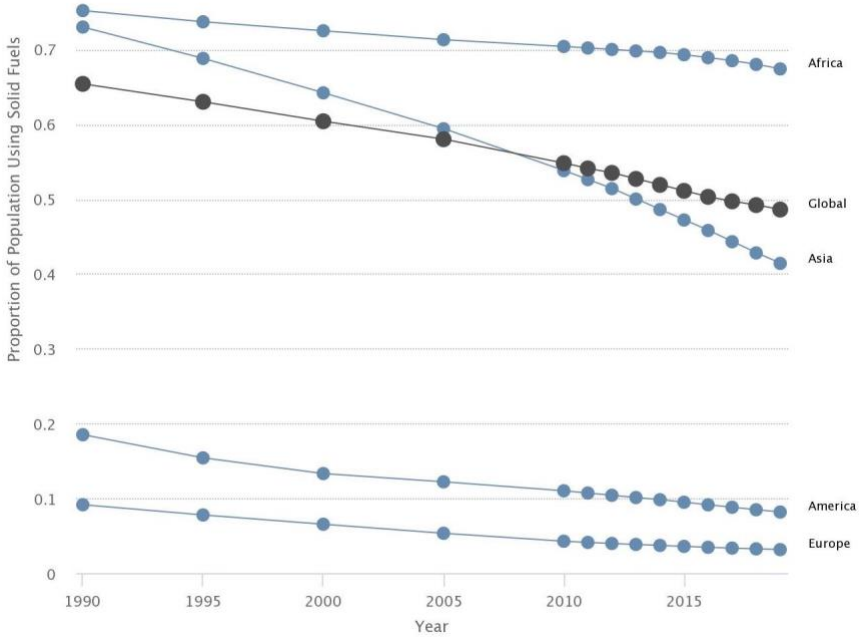

Figure 3: Proportion of Population Using Solid Fuels (Health Effects Institute, 2020a)

The other way to classify pollutants is by how they were made. If a pollutant is produced directly, comes straight from the source, it is called a primary pollutant. If pollutants are the outcome of chemical reactions between primary pollutants and other atmospheric mix, they are called as secondary pollutants. A good example is a ground-level ozone. The result of mixing volatile organic compounds (VOCs) and oxides of nitrogen in the sunlight is **ground-level (tropospheric) ozone.** (Zehnder et al., 2018). It is hazardous to both human health and vegetation (Health Effects Institute, 2020b).

Particulate Matter (PM)

The most hazardous types of pollution are particles and ground-based pollutants (smog). The first one, particle matter, also known as particle pollution (PM), is composed of liquid droplets and small particles such as soil, organic chemicals, metals, dust particles, acids.

Health issues associated with inhalation of those are different, depending on the size of the particles (EPA, 2022a). The most often differentiated sizes are 10µm and 2.5µm, called PM10 and PM2.5 respectively. Their diameter is less than 10µm and 2.5µm. They are worth special attention because of their ability of reaching bronchi and trachea area (Brunelli et al., 2007a).

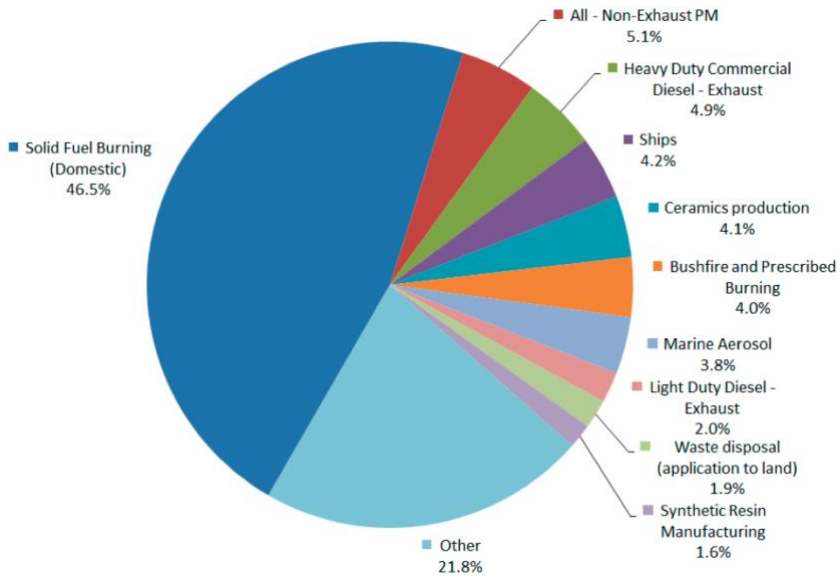

Figure 4: Top 10 PM2.5 sources in Sydney (NSW & EPA, 2013)

A particulate pollution of a thickness less than 10µm and larger than 2.5µm is called **ambient respirable particle (inhalable coarse particle, PM10)** (EPA, 2022a; Piao et al., 2018; Sarkar et al., 2010). One can differentiate two types of PM10 sources: anthropological and natural. Heating, traffic, incinerator, and construction work are part of the first group. Volcanic eruptions, dust storms, marine salts, vegetation, and forest fires are examples from the second group. (Brunelli et al., 2007b; Ni et al., 2012; Vautard et al., 2005). Extremely concerning is its impact on infants and fetuses. (Glinianaia et al., 2004; Šrám et al., 2005).

The next pollutant, **fine particulate matter (PM2.5)** is especially hazardous as it applies adverse biological effects human body on several major organs, including the lung (Liu et al., 2017), cardiovascular system (Du et al., 2016), nervous system (Wang et al., 2017), immune system (J. Zhao et al., 2013), or even a skin (Piao et al., 2018). PM2.5, next to ozone pollution, is one of the most monitored pollutants. According to studies, PM2.5 contributes to annually 4.1 million premature deaths worldwide. The Western Pacific

Region, where one-quarter of the population lives, contributes to one-third to total deaths due to PM2.5 ambient air pollution (Health Effects Institute, 2020b).

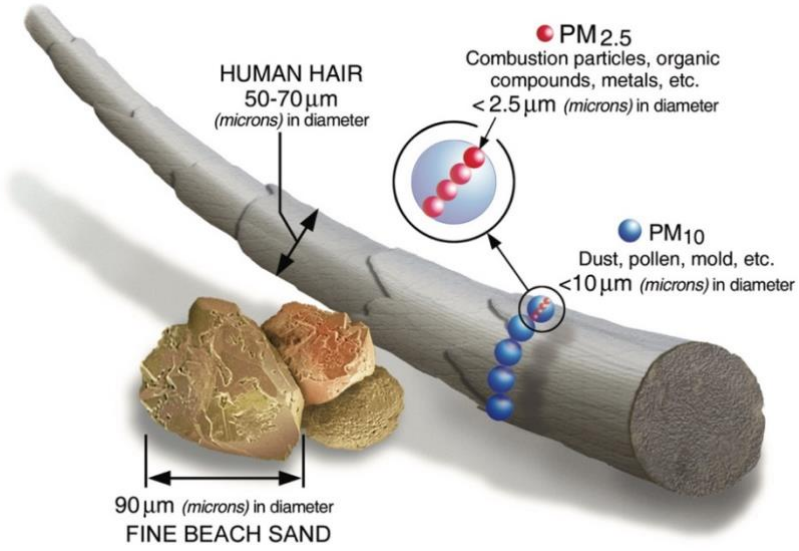

Figure 5. Size comparison for PM particles (EPA, 2022b)

1.3. Regulatory Frameworks for Managing Air Pollution

As one can see, air pollution is a serious global challenge, posing threats to human health and the environment. Many international, regional, and national authorities have developed regulatory frameworks and guidelines to manage, mitigate and combat air pollution. The World Health Organization's Air Quality Guidelines (AQGs) provide the key aspects of air quality management. When it comes to legal frameworks, these can be found in the European Union, the United States, and China. The attempts of these global players can be perceived as valuable models for other countries to fight against air pollution.

Air Quality Guidelines

The World Health Organization (WHO) has developed air quality guidelines (AQGs) (WHO, 2021) to provide recommendations based on evidence on the levels of key air pollutants that pose health risks. These guidelines are designed to help governments establish and implement air quality standards that protect joint health. The guidelines aim to minimize the adverse health effects associated with air pollution, such as respiratory

and cardiovascular diseases, and also premature death, by providing recommended concentration limits for these pollutants.

European Union

The EU has embraced an integrated approach to air pollution management, merging command-and-control and market-based instruments. The principal legislation, the National Emission Ceilings (NEC) Directive, sets maximum annual emission levels for pollutants in each Member State. The Directive covers four air pollutants: nitrogen oxides (NOx), sulfur dioxide (SO2), non-methane volatile organic compounds (NMVOCs), and ammonia (NH3) (The European Parliament et al., 2016). The EU has also adopted the Industrial Emissions Directive (IED), which sets emission limit values for large industrial plants (the European Parliament et al., 2010).

The command-and-control mechanisms have been integrated by the European Union Emissions Trading System, also known as EU ETS. Approximately 45% of the EU's greenhouse gas emissions is covered by this system. By allowing companies to trade emission allowances, the EU ETS incentivizes them to emission reductions. (Directive 2003/87/EC, 2003).

United States

A technology-driven approach is present in the US air quality management framework. The critical legislation, the Clean Air Act (CAA), sets up National Ambient Air Quality Standards (NAAQS) for six major air pollutants. Under the CAA, the US Environmental Protection Agency (EPA) establishes emission standards for given industrial categories based on the performance of the best available control technology (BACT) (The Clean Air Act, 1963).

The US has also introduced the Acid Rain Program, a cap-and-trade system targeting SO2 and NOx emissions from power plants. This approach has successfully reduced emissions by allowing companies to trade allowances, stimulating innovation in pollution abatement technologies (the EPA, 2021).

China

In China, one can observe a command-and-control regulatory framework with regional flexibility. The primary legislation, the Air Pollution Prevention and Control Law, establishes national ambient air quality standards and emission limits for specific

pollutants. However, regional governments have flexibility in designing and implementing air pollution control policies. For example, it involves a Total Emission Control (TEC) target for a pollutant (Law of the People's Republic of China on the Prevention and Control of Atmospheric Pollution, 2018).

China has also experimented with market-based instruments, such as the pilot emissions trading systems in selected cities and provinces. These initiatives encourage industries to adopt cleaner technologies and reduce emissions cost-effectively (Jin et al., 2016).

Managing air pollution is a vital and complex task that requires implementing comprehensive and well-coordinated strategies. The World Health Organization's Air Quality Guidelines serve as a valuable reference for governments worldwide. The regulatory frameworks in the European Union, the United States, and China demonstrate various approaches to tackling air pollution, each with strengths and weaknesses.

Measuring Air Quality

It is essential to recognize that understanding and managing air pollution effectively necessitates accurate and reliable measurement methods. In order to help fight against air pollution and observe long-term mode, policymakers need access to pollution measurements (Directive 2008/50/EC, 2008). To change anything, one needs a tool to measure something and compare the results. By accurately measuring air pollution, policymakers and stakeholders can make better-informed decisions and contribute to more effective air quality management strategies.

Ground-based monitoring and remote sensing networks are examples of the different techniques and technologies used to measure air pollution. Therefore, data collection and analysis are crucial for informed decision-making and policy formulation.

Ground-Based Monitoring

Stationary air quality monitoring stations monitor air pollution in urban and industrial locations. These stations are equipped with sophisticated tools that continually measure pollutant concentrations, such as particle matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3). Data from these stations are used to assess air quality, track trends, and enforce regulatory standards (e-Gminy.pl, n.d.).

Mobile monitoring units offer greater flexibility in air pollution measurement, as they can be deployed in various locations, as necessary. These units are often used to investigate

pollution hotspots, validate air quality models, or supplement data from fixed measurement stations (M. Adams & Corr, 2018).

Remote Sensing Technologies

Satellite-based remote sensing allows large-scale air pollution monitoring through substantial geographical areas. By analyzing data from satellite sensors, researchers can track the movement and dispersion of pollutants, identify sources of pollution, and assess emission control measures' effectiveness (*Remote Sensing*, n.d.).

LIDAR (Light Detection and Ranging) consists of an optical remote sensing technology. It uses laser pulses to measure atmospheric pollutant concentrations. LIDAR systems can provide high-resolution vertical profiles of pollutants, enabling researchers to study the distribution of pollutants within the air column (Wasser, 2022).

1.4. Air Pollution Prediction Methods

Public health officials, policymakers or urban planners should have an understanding and the possibility to manage air pollution quality. Monitoring is just one step – it is just observing the data. The next step is the ability to predict and implement appropriate measures. This allows to the mitigation of adverse consequences of air pollution. For example, accurate air pollution prediction can help identify pollution hotspots, develop effective control strategies, and raise public awareness about air quality.

Statistical Methods

Ease of implementation, simplicity, and interpretability make statistical methods widely used for air pollution prediction. Time series analysis, linear regression or multivariate regression are commonly known examples.

Historical data makes time series analysis focus on identifying patterns and trends. This allows us to predict future pollution levels. **Autoregressive Integrated Moving Average** (**ARIMA**) and **Seasonal Decomposition of Time Serie**s (**STL**) (Amato et al., 2019) are techniques often used for this purpose. Time series analysis can capture temporal patterns, but it may not fully account for the impact of external factors or spatial dependencies.

The relationship between air pollution levels and other variables can also be modeled by **linear regression**. When one independent variable is taken to the model, it is called simple linear regression, when more -- multiple linear regression. However, because of

the simplicity of those method, the may not accurately capture nonlinear or complex relationships between conditions taken to the model.

Multivariate regression could be understood as extension of the multiple linear regression. Considering multiple dependent variables simultaneously allows that. Some techniques include Partial Least Squares Regression (PLSR) or Principal Component Regression (PCR). They are used to improve prediction accuracy and reduce multicollinearity issues. Despite its better accounting for the interdependence of various pollutants, capturing complex, nonlinear relationships and interactions between variables may be a limitation (Boncelet, 2019).

Statistical methods may be considered a good foundation for air pollution forecasting. They may not be sufficient though for modeling complex, nonlinear relationships or capturing the interaction between variables. A solution could be using machine learning techniques and hybrid methods aiming to address these limitations.

Machine Learning Techniques

There are some reasons why machine learning (ML) techniques have gained popularity in air quality prediction, e.g., the ability to model intricate, nonlinear relationships and adapt to new data.

One of the most commonly known examples of ML techniques is **Artificial Neural Networks (ANNs)** (Cabaneros et al., 2019). Their name comes from the function and structure of neural networks in biology and are composed of interconnected layers of nodes or neurons that can learn complex patterns in the data. Auspicious results (such as effectively catching temporal and spatial dependencies) in terms of air pollution prediction were given by the ANNs, especially deep learning (DL) architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Classification and regression tasks can both be done by the **support vector machines (SVMs)** (Leong et al., 2020). They are compelling for modeling nonlinear relationships through the use of kernel functions. As a result, SVMs have shown strong performance in air pollution forecasts, especially when dealing with small or noisy data sets.

Probably the most known instance of ML is **random forests**. They consist of an ensemble learning method (Araujo et al., 2020) – it constructs multiple decision trees and combines their forecasts to overcome the bias of one single tree. This technique prevents overfitting. Being easily understandable, they can handle large datasets, high-dimensional feature spaces, and missing data.

Another exciting and also ensemble method is **Gradient Boosting Machines (GBMs)** (Sharma et al., 2022). In a stage-wise manner, it constructs multiple weak learners, such as decision trees. They focus on minimizing prediction errors by iteratively refining the model. That results in strong predictive performance in air pollution prediction tasks (Grigorev, 2021).

Hybrid Methods

Hybrid methods consist of combinations of machine learning approaches with statistical ones to leverage both techniques' strengths and improve overall forecast accuracy.

ARIMA-ANN combines ARIMA's time series modeling functions with the nonlinear pattern recognition of ANNs. Integration of both allows for better capture of linear and nonlinear relationships in air pollution data (Shahriar et al., 2021).

Wavelet Transform-Based Methods can decompose time series data to various frequency components. For example, machine learning models like ANNs or SVMs can improve prediction accuracy. Then, they can be applied to these decomposed components to capture complex models and improve prediction accuracy (X. Zhao et al., 2022).

In **ensemble learning techniques** (stacking and bagging), higher prediction accuracy is achieved by combining the predictions of multiple base models (e.g., statistical models and machine learning models). The use of these techniques can minimize the weakness of individual models. Furthermore, it enhances the overall performance in prediction tasks (Pinheiro & Patetta, 2021).

Machine learning methods and hybrid approaches have shown encouraging results in prediction by dealing effectively with the limitations of traditional statistical methods. For example, they can pick up complicated, nonlinear relationships and interactions among variables, leading to more accurate and reliable predictions.

Spatial and Temporal Prediction Methods

The spatial and temporal prediction methods could explain the spatiotemporal dynamics of air pollution. However, pollution levels may differ across space and time due to meteorological conditions, topography, and emissions sources. For example, spatiotemporal Regression, Kriging, and Space-Time Models may be notable prediction methods.

The data's spatial and temporal dependencies are considered in **Geographically Weighted Regression (GWR)** (Fotheringham et al., 2019) and **Bayesian**

Spatiotemporal Regression. They provide more accurate predictions by capturing the changing relationships between air pollution levels and predictor variables.

Kriging is a geostatistical interpolation technique that guesses values at unobserved locations based on spatially correlated observed data. For example, one can observe space-time kriging (Park, 2016) or Bayesian kriging (Vicedo-Cabrera et al., 2013) to forecast air quality levels in unsampled areas and predict their temporal evolution.

A combination of temporal and spatial information in the form of space-time models, such as the Space-Time Autoregressive Integrated Moving Average (STARIMA) and Dynamic Linear Models (DLMs), provide a prediction of air pollution levels.

Space-time models, such as **Dynamic Linear Models (DLMs)** (Sánchez-Balseca & Pérez-Foguet, 2020), combine temporal and spatial information to predict air pollution levels. These models can capture complex spatiotemporal dependencies and provide more reliable forecasts.

Factors Affecting Prediction Accuracy

Prediction not rarely lacks accuracy. Several factors influence it, e.g., data quality, sampling frequency, model selection, and model validation.

High-quality data with the adequate temporal and spatial resolution is critical for accurate forecasting. One can get suboptimal model performance with noisy, incomplete, or biased data (Carmichael et al., 2008).

The sampling frequency affects the ability to capture seasonal patterns and short-term variations. Generally, higher sampling frequency leads to better prediction accuracy.

One should choose an appropriate model, which means a model that captures the underlying relationships and dependencies in the data. When selecting a model, interpretability, complexity, and computational requirements are desired factors.

The other thing that helps improve foresting accuracy is including relevant predictor variables and excluding redundant or irrelevant features. For example, feature selection techniques and domain knowledge may help identify the most critical variables (Dominici et al., 2010).

Germane model validation, cross-validation, hold-out validation, or other validation methods can be used to evaluate model performance on unseen data and prevent overfitting and underfitting (M. D. Adams et al., 2020).

Each of the methods, i.e., traditional statistical methods, machine learning techniques, hybrid approaches, and spatial and temporal prediction models, has its strengths and weaknesses, and the choice of the appropriate method depends on the specific context, data availability, and desired level of accuracy. Researchers can develop more reliable and effective air pollution prediction models by considering factors affecting prediction accuracy and employing suitable techniques.

Multivariate Regression Techniques

Because of their ability to concurrently model multiple dependent variables and capture the interdependence of various pollutants, multivariate regression techniques are often used in air pollution prediction. Examples include multiple linear regression, principal component regression (PCR), partial least squares regression (PLSR), canonical correlation analysis (CCA), redundancy analysis (RDA), regularization techniques, and nonlinear multivariate regression methods. The traditional linear regression method is extended to account for the complicated relationships between air quality levels and multiple predictor variables.

Multiple Linear Regression

Multiple linear regression, in opposing to the simple version, models the relationship between a single output variable and more than one explanatory variable.

The general form is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where Y is predicted variable, X_1, X_2, ..., X_k are the explanatory variables, β_0 is the intercept, β_1, β_2, ..., β_k are the regression coefficients, and ε is the random error term.

Taking into consideration air pollution forecast, it is possible to give different meanings to variables. *Y* equals the pollution level. The predictor variables ($X_1, X_2, \ldots, X_k$) may be interpreted as factors affecting air quality (like meteorological factors, industrial emissions, traffic data, etc.).

Multicollinearity is the assumption of the multiple linear regression. It means that predictor variables cannot be highly correlated with each other. Also, the connection between output and input variables must be linear. These assumptions, though, may not always hold in practice. It leads to potential issues lie unstable parameter estimates and poor predictive performance (Alexopoulos, 2010).

Notwithstanding its limitations, multiple linear regression is the basis for more advanced multivariate regression techniques that address these issues.

Principal Component Regression (PCR)

Principal Component Regression (or PCR) is the other multivariate method. Principal component analysis (or PCA) is integrated with linear regression in order to cope with reduction of dimensionality. PCR converts the original set of correlated predictors into a smaller set of uncorrelated features (principal components). It explains most of the deviation in the data.

The first step involves performing PCA on the predictors to get the principal components. Then, a linear regression model is fitted by use of a chosen number of principal components as the new independent variables. Multicollinearity is shrunk by the use of principal orthogonal components. Additionally, it may improve the stability of regression estimates. The PCR's restrictions are merely focused on explaining the variance in the predictors and do not take into consideration the relationship with the output variable. In effect, some essential predictors may be rejected unless they contribute significantly to the variance in the predictors. Moreover, the primary components may be demanding to describe as for the original variables. (Jolliffe, 1986).

Partial Least Squares Regression (PLSR)

Partial Least Square Regression (or PLSR) is considered an other possibility to PCR when it comes to addressing multicollinearity and dimensionality reduction. A smaller set of uncorrelated variables (latent variables) is created. Those latent variables are linear mix of original predictors. PLSR differs from PCR in that it identifies latent variables that account for the variance in the predictors and maximize the covariance with the output variable. The PLSR approach tries to find directions that help explain the response and the predictors (James et al., 2013).

PCR and PLSR can offer valuable multicollinearity and dimensionality reduction solutions in multivariate regression. Therefore, the objectives of the analysis should lead to the choice between these two methods.

Canonical Correlation Analysis (CCA) and Redundancy Analysis (RDA)

The following multivariate statistical techniques are Canonical Correlation Analysis (CCA) and Redundancy Analysis (RDA). They are used to analyze the relationships

between multiple sets of variables. This can be particularly useful for air pollution prediction tasks involving complex interactions between multiple pollutants and environmental factors.

Canonical Correlation Analysis (CCA) looks for linear combinations of variables from two sets to maximize the correlation between these linear combinations. Germane air pollution prediction, CCA may be used to investigate relationships between pollutant concentrations and various predictor variables, such as meteorological factors, land-use characteristics, and emission sources. CCA can help select the most appropriate features for building predictive models by identifying the strongest correlations between these variables.

Redundancy Analysis (RDA) explores the linear relationship between a set of explained variables and predictor variables while accounting for the influence of other covariates.

In air pollution studies, it can detect changes in variables best explained by environmental variables (González et al., 2003). In addition, RDA can provide valuable insight into feature selection and model development, enabling researchers to build more accurate and interpretable air quality forecast models.

Regularization Techniques

Regularization techniques are helpful methods in regression analysis. They help prevent overfitting and improve model generalization. An objective function is being added a penalty term to. For example, they may be particularly beneficial in air pollution prediction tasks where the number of predictor variables is significant, and multicollinearity might be present.

Ridge Regression (L2 Regularization), as an example of a regularization technique, adds an L2-norm penalty term to the least squares objective function, which is proportional to the sum of squared coefficients. It is particularly desirable to be used in air pollution studies, where data is not obtained from a well-designed or controlled experiment. It is a promising method as it helps avoid distortions (McDonald, 2009).

An L1-norm penalty term is incorporated in **Lasso Regression (L1 Regularization)**. It is proportional to coefficients' sum of the absolute values. (Lello et al., 2018). The penalty term leads to sparse solutions, effectively carrying out feature selection by driving some coefficients to zero. This method is favored for improving the accuracy of prediction and solvable output (Sethi & Mittal, 2021).

Incorporating CCA, RDA, and regularization techniques into developing air pollution forecast models allows one to understand better the complicated relationships between

multiple pollutants and explanatory variables and build more accurate and interpretable models.

Nonlinear Multivariate Regression

Nonlinear multivariate regression consists of an extension of the concept of the linear multivariate regression. It can explain the nonlinear relationships between multiple dependent and independent variables.

Generalized Additive Models (GAMs) provide a flexible framework that expands on a standard linear model by enabling the inclusion of nonlinear functions for each variable (James et al., 2013). Because it captures nonlinear relationships, it potentially provides more accurate predictions. However, the main limitation of GAMs is that the model must be additive which may cause missing some essential interactions.

When it comes to **polynomial regression**, it expands the linear model by adding extra predictors. Extra predictors comes from the original predictors which are raised to a power. This method provides a straightforward way to provide a nonlinear fit to data. Mentioned techniques, i.e., CCA, RDA, regularization techniques, or nonlinear multivariate regression, provide a better understanding of the complex relationships between predictors (e.g., multiple pollutants) and predictor features and allow to build of more accurate and interpretable prediction models. The choice of appropriate techniques depends on the researcher and the specific context, data availability, and desired level of accuracy.

A Sliding Window Technique

An essential algorithmic technique used in various domains such as signal processing, data analysis or networking is the sliding window approach. By processing data structures in smaller, overlapping segments it optimizes computational complexity and improve efficiency of algorithms (Zivot & Wang, 2006).

It involves a variable-sized or fixed window that slides through a data structure like array or string, processing elements within the window at each step. A crucial parameter is window size. It can impact the applicability and performance of the technique. Variable window size is more appropriate for problems that require the longest or the shortest contiguous subarray meeting specific conditions. A constant window size is well-suited for fixed-size substructure problems. The sliding step determines how the window moves through the data, often one position at a time.

A vital task is choosing an appropriate data structure for implementing the sliding window approach effectively. Ques, double-ended queues, and arrays are often used. This technique optimizes computational complexity by reducing the time complexity of certain algorithms. It can transform quadratic time complexities $O(n^2)$ to linear time complexities O(n) (Thakral, 2023).

The Moore-Penrose Pseudoinverse

Another powerful mathematical tool that finds its use in various fields including signal processing, statistics and machine learning is the Moore-Penrose Pseudoinverse (the generalized inverse, pseudoinverse, or Moore-Penrose inverse) (Weisstein, n.d.). A special use-case is dealing with systems of linear equations that do not have unique solutions.

The concept is used where there is a non-square matrix, or a square matrix that is singular (i.e., it does not have an inverse). Having in mind that the traditional concept of matrix inversion applies only to no-singular square matrices, there is a need for the pseudoinverse.

The generalized inverse is defined for any matrix and provides a solution to a system of linear equations that may not have a unique solution. When it comes to regression analysis, the Moore-Penrose Pseudoinverse is used to solve the normal equations which is the method of least squares for approximating the unspecified parameters in a linear regression model (Laub, 2012).

The Moore-Penrose inverse of a matrix A, denoted as A^+, is used to find an approximate solution to the matrix equation AX = B, when A is not invertible, or the system is over- or under-determined (Smoktunowicz & Wróbel, 2012).

Dealing with multivariate regression makes a need for the pseudoinverse. Having a non-invertible matrix, or when the matrix of predictors has more predictors than observations, the equation cannot be solved using the normal inverse. Instead, the pseudoinverse is used to solve the system.

However, the most often use case for pseudoinverse is solving least squares systems. The residuals' sum of squares is minimalized, making it an optimal solution in the context of regression analysis (MacAusland, 2014).

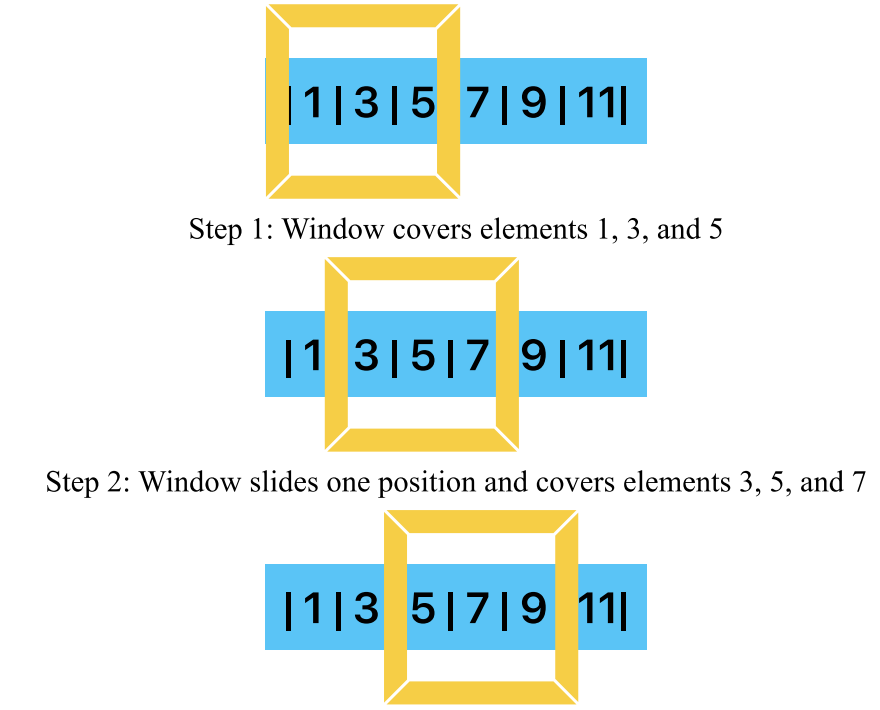

Figure 6. Visualization of the Sliding Window Technique. Based on a figure from Hota et al., 2017

This part underlined the criticality and complexity of predicting air pollution levels. The severe health and economic impacts of air pollution necessitate accurate and efficient prediction models. As the literature reveals, multiple methods, and techniques, ranging from simple statistical methods to sophisticated machine learning models, have been applied to this problem. Among them, multivariate regression techniques hold significant promise due to their ability to model complex systems with multiple outputs.

Each of the brought techniques, including multiple linear regression, principal component regression, partial least squares regression, canonical correlation analysis (CCA), redundancy analysis (RDA), and nonlinear multivariate regression brings unique strengths to the table as well as their limitations. The choice of method dependent on the specific characteristics of the data and research specific aim.

Furthermore, the sliding window approach was explored, a technique that allows taking advantage of temporal dependencies in time series data, and the Moore-Penrose pseudoinverse, a mathematical tool used when systems of linear equations do not have unique solutions.

This part of the thesis explored the theoretical foundations that form the basis for the practical analysis that follows in the next chapter. The aim is to leverage techniques and approaches to build a robust and accurate model for predicting air pollution levels in Warsaw, Poland. The next chapters will focus on the specific methods, tools, technologies and data used in this study, the obtained results and discussion of them.

2. Methodology

This chapter will explore the methodology used to investigate and tackle air pollution. Each section serves a distinct, crucial role in the research process.

The first part outlines the data sources and collection process. The following section dives into overview of the nature and composition of the data. Next, the techniques used to clean and prepare the data for analysis are described.

"Feature Selection" identifies key variables impacting air pollution levels. They are crucial for building accurate prediction models. "Implementation" discusses applying multivariate regression techniques, a sliding window technique, and the Moore-Penrose Pseudoinverse.

Finally, "Error Metrics" explains the metrics that measure the performance of these prediction models. By providing an in-depth overview of these methodologies, this chapter lays a solid foundation for the following results and discussions.

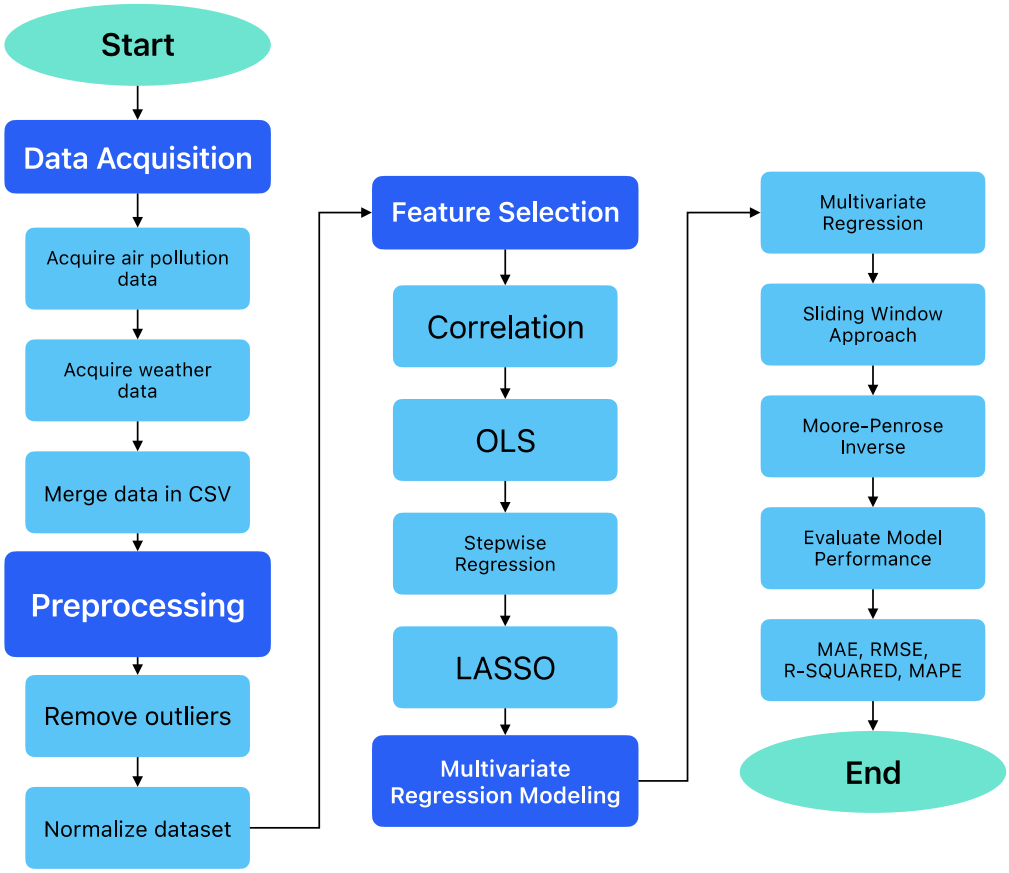

Figure 7. Study's Structure Flowchart

2.1. Data Acquisition

In order to conduct this study, the acquisition of relevant, accurate and timely data was crucial. This step was divided into two. It involved data related to air pollution and weather conditions for the capital city of Poland – Warsaw, from 2nd April 2022 till 31st March 2023.

Due to the extensive amount and type of data needed, an automated and programmatic method was utilized for data gathering. To achieve this, Python was applied. Python is a programming language. It is widely used in data science due to its user-friendly nature and extensive range of data processing libraries.

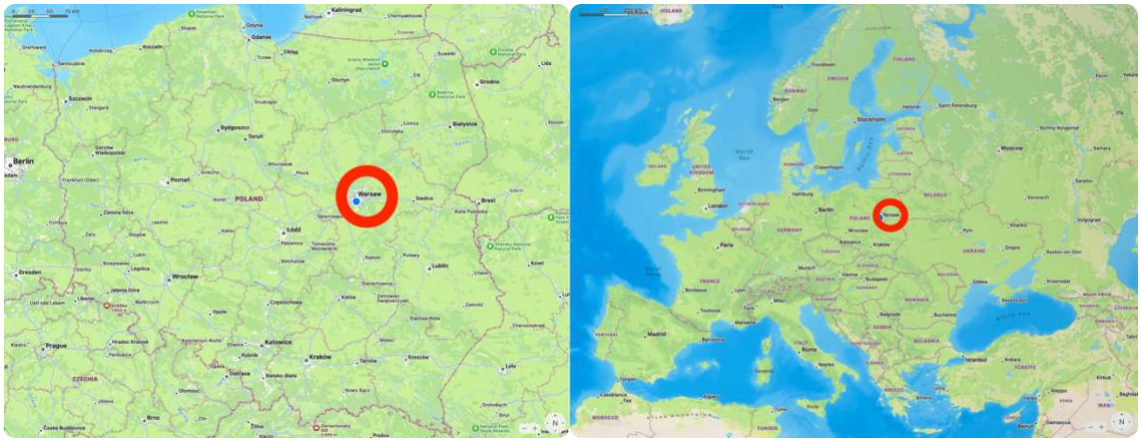

Figure 8. Location of Warsaw in Poland and in Europe

Two Python scripts were developed – one fetching weather conditions data and the other – air pollution data. OpenWeather is a platform that offers access to weather data for each point on Earth (OpenWeather, 2023). Convenience of the use OpenWeather's API and broad spectrum of available data decided for choosing this site as a source of data for the study.

The first Python program was developed to acquire air pollution data (Figure 10), the second one – weather data. The first one was designed to fetch air pollution data including PM2.5, PM10, O3, NO2, SO2, and CO levels. The second script was responsible for obtaining weather data, specifically temperature, humidity, windspeed, pressure, and cloud cover. OpenWeather offers accessibility to historical data up to one year back (Student's subscription). Historical air pollution as well as weather data is available in hourly frequency. In each of the scripts fetched data was averaged so that results from the entire day, i.e., 24 hours, were averaged. Both programs after successful fetching and averaging save data into CSV file. CSV, short for Comma Separated Values, is a simple

file format that is widely used for storing tabular data and is compatible with a wide range of applications, including MATLAB, the primary tool used for the analysis in this study.

After careful investigation of both files, it was observed that pollution data file lacks results for three days, in different months. To overcome this, data from the previous and the following day for those days were averaged and filled in. That operation was done manually. The last step consisted of merging those two files into one which became a source file for the MATLAB script. This was done using spreadsheet software (Apple Numbers), ensuring that the data from the two sources was correctly aligned based on timestamps. The final CSV file, therefore, contained a comprehensive and chronological set of both air pollution and weather data for the specified period.

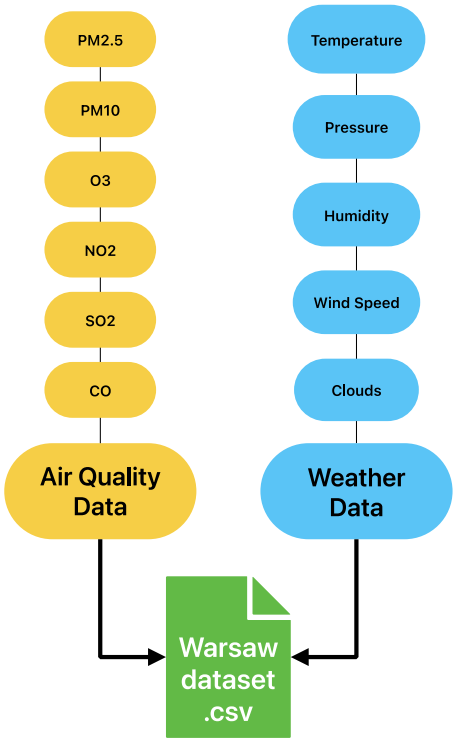

Figure 9. Python Programs' Responsibilities

This data collection and preparation process set the foundation for the subsequent analysis. By automating the data collection process through Python and OpenWeather's API, a large volume of relevant data was efficiently gathered and structured, paving the way for applying multivariate regression for air pollution prediction.

2.2. Dataset Description

To conduct a successful data-driven study, it is crucial to thoroughly comprehend the dataset's composition, scale, and characteristics. It serves as the foundation of the thesis. This study relies on a comprehensive dataset that spans one year, from 2nd April 2022 to 31st March 2023, covering Warsaw – the capital city of Poland.

The dataset has been assembled from two distinct sources, both obtained with use of OpenWeather service, with the data collection process powered by Python. It encapsulates a broad spectrum of variables related to air pollution and weather conditions.

Air pollution data includes measurements of several critical pollutants that impact health and the environment. These include Particulate Matter (PM2.5 and PM10), Ozone (O3), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), and Carbon Monoxide (CO). All of them are given in $\mu g/m^3$.

```
1. Import libraries
2. Set API key and city coordinates
3. Define a function to fetch air pollution data
   from the API
4. Define a function to aggregate daily pollution
   data
5. Define a function to save data to a CSV file
6. Define a main function to:
• Initialize an empty list for data
• Calculate start and end timestamps and datetimes
• Loop through the cities and:
   - Fetch air pollution data
   - Aggregate daily data
   - Add the data to the final data list
   - Save the data to a CSV file
7. Execute the main function
```

Figure 10. Pseudocode of the Program Fetching Air Pollution Data

In addition to the air pollution data, weather data forms a crucial part of the dataset. These variables are known to influence the dispersion and concentration of pollutants, making them valuable inputs for the prediction model. The weather-related variables in the dataset include temperature (kelvins), pressure (hPa), humidity (%), wind speed (m/s), and cloud cover (%).

Each row in the dataset corresponds to a specific timestamp, providing a chronological record of Warsaw's air pollution levels and weather conditions over the specified period.

The dataset thus forms a multivariate time series, with multiple dependent and independent variables recorded over time.

This dataset serves as the empirical foundation for the analysis and modeling of the thesis. Its breadth and granularity enable a detailed exploration of the multivariate relationships among the variables, facilitating the development and testing of a robust air pollution prediction model.

Table 1. The Beginning of the CSV File of Air Pollution and Weather Data for Warsaw, Poland

date	pm2_5	pm10	o3	no2	so2	co	temp	pressure	humidity	wind_speed	clouds
2022-04-02	4.76	5.49	73.52	9.02	10.76	302.35	274.55	1006.00	61.00	6.17	0.00
2022-04-03	11.50	13.56	71.64	13.65	16.19	349.36	274.36	1012.00	63.00	4.12	0.00
2022-04-04	7.76	9.54	77.28	10.43	10.16	303.19	277.75	995.00	58.00	7.20	0.00

2.3. Data Preprocessing

Data Preprocessing involved cleaning and transforming the dataset for the study. It engages several steps, including applying the Interquartile Range (IQR), normalization, and specific transformations for skewed features.

Initially, the dataset was subjected to a thorough exploring process. It started by “summary” MATLAB function and drawing boxplots. In statistics, boxplots are a graphical tool used to show the distribution of numerical data visually. The data's minimum, first quartile (25th percentile), median (50th percentile), third quartile (75th percentile), and maximum are summarized. Boxplots can help identify any outliers or skewness present in the data. It was helpful to use this simple and efficient method for obtaining an overview of the data. The graphs clearly indicate the necessity of removing outliers.

Next, the variables were all subjected to the Interquartile Range (IQR). This approach was utilized to control and lessen the impact of outliers, which could otherwise unfairly sway the results of the subsequent analysis.

The IQR method is a statistical analysis technique that helps measure the spread of data and identify potential outliers. In this study, the IQR method is applied in the main MATLAB script as a separate function in a different file.

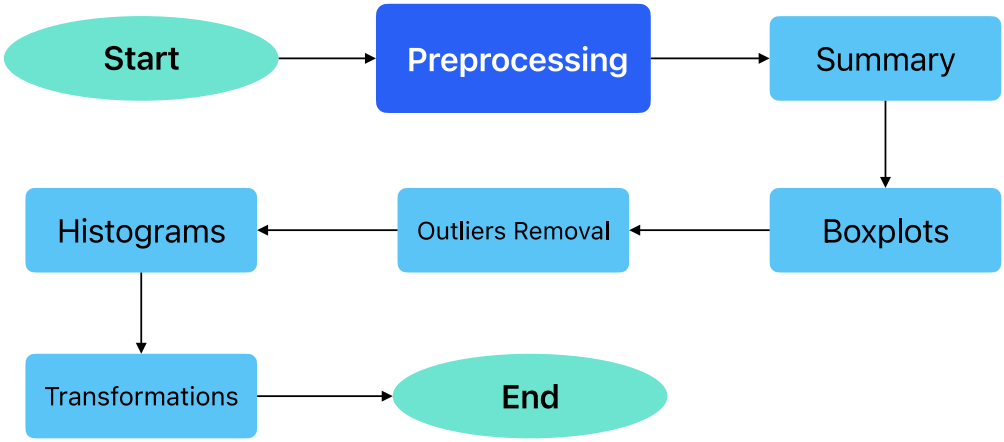

Figure 11. Preprocessing Flowchart

The IQR is computed as the difference between a dataset's third quartile (Q3) and the first quartile (Q1). These quartiles represent the 75th and 25th percentile of the data, respectively, effectively describing the middle 50% of observations when ordered from lowest to highest. The IQR, therefore, provides a measure of statistical dispersion, or how spread out the values in the dataset are.

Outliers in a dataset can be identified using the IQR. Outliers are data points that are either under Q1 - 1.5*IQR or over Q3 + 1.5*IQR. This is based on the assumption that most of the data in a distribution lies within this range. Any data point outside this range could be due to variability or may indicate an experimental error.

The IQR method was essential to ensure the reliability of the results, as outliers can significantly impact the outcomes of statistical analyses and potentially lead to erroneous conclusions. Furthermore, it enhanced the study's robustness and validity. To ensure the technique worked correctly, boxplots were drawn once again.

The next step introduced drawing histograms. Histograms are graphical representations that show how data is distributed. They are represented as bars that show data points. Each bar's height corresponds to the frequency or the count of data points within each bin. Looking at the new plots gave an idea of applying normalization to most features.

Normalization involved scaling all the variables to a standard range, which helped to reduce the potential bias caused by variables measured at different scales and to enhance the efficiency and accuracy of the multivariate regression model.

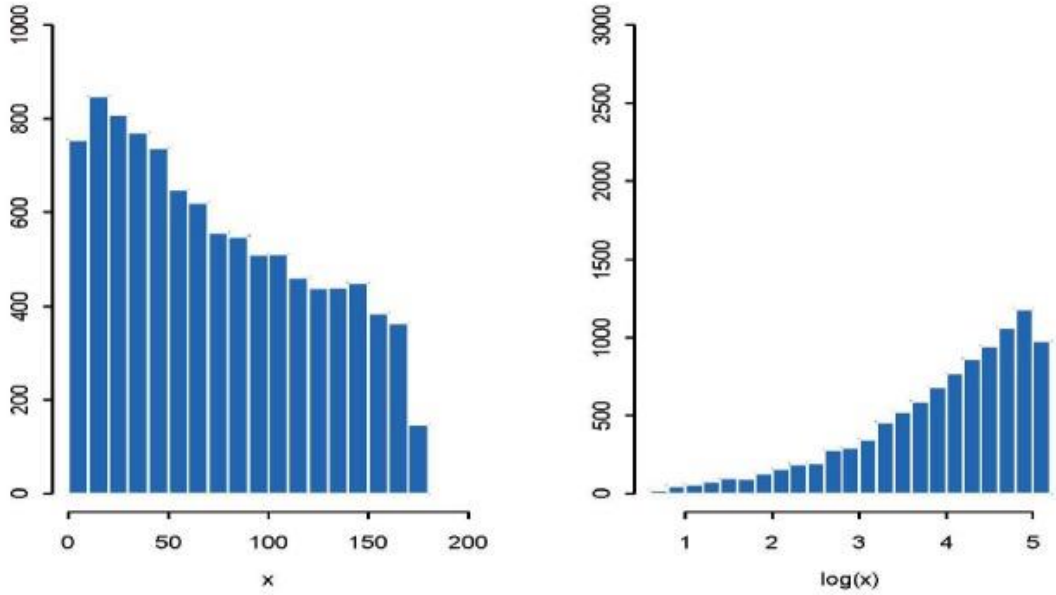

Figure 12. Histograms. On the left side – with original data. On the right side – with log-transformed data. From (FENG et al., 2014)

Regarding the skewed features in the dataset, specific transformations were applied. A log transformation was implemented for the right-skewed features. This transformation helped to reduce the skewness and make the distribution more symmetrical. A square root transformation was also utilized for a feature, which resisted the log transformation. A cube transformation was applied in the case of the left-skewed feature. This transformation helped to correct the skewness and make the distribution more symmetrical.

Prior to preprocessing, several variables exhibited high skewness and a wide range of values, indicating a non-normal distribution. Post-preprocessing, these variables showed significantly reduced skewness and a smaller range of values, signifying a more symmetric and normal-like distribution. This transformation procedure is crucial for improving the validity of subsequent statistical analyses, such as linear regression, which assume the normality of the data.

Overall, the preprocessing steps have refined the dataset, ensuring it is more suitable and robust for the subsequent feature selection.

2.4. Feature Selection

This research employed four methods for selecting features: correlation analysis, Ordinary Least Squares (OLS), stepwise regression, and also the Least Absolute Shrinkage and Selection Operator (LASSO). Each of the methods provides unique advantages and capabilities. In order to predict significant pollutants robustly, the need for correct factors affecting them is essential.

Correlation Analysis

Correlation analysis was the first method applied. It is a popular and reliable method for feature selection in machine learning. It measures how much two variables in a dataset affect each other. This interaction can be direct or inverse. A positive correlation demonstrates that the variables increase together. A negative correlation suggests that one variable decreases as the other increases and vice versa.

The correlation coefficient, which ranges from -1 to +1, indicates the magnitude and direction of the relationship. If a coefficient is near +1 or -1, it means there is a strong relationship. However, if the value is closer to zero, it indicates a weak or non-existent relationship.

Regarding feature selection, correlation analysis is essential for two main reasons: identifying predictive features and reducing redundancy.

The model's output is significantly influenced by features that strongly correlate with the target variable. These features are often reliable predictors. During the research, the correlation of each feature with PM2.5 was calculated. Those with a high absolute correlation coefficient (greater than 0.7 in the script) were considered highly predictive.

When two features have a high correlation, they have similar information, which is redundant. This redundancy can result in overfitting and make the model complex without improving its predictive power. To avoid this, redundant features can be identified and eliminated, a process called dimensionality reduction. By doing so, it can make the model simpler and more efficient without compromising its predictive power.

The study calculated a correlation matrix for all the features and displayed it using a heatmap. The heatmap delivered a clear visual representation of relationships between features. Also, a partial correlation coefficients heatmap was computed to account for the influence of all other variables when determining the correlation between two specific variables.

Figure 13. Sample Correlation Plot (own elaboration)

Nonetheless, it is crucial to note that correlation does not imply causation. A strong correlation between two features does not necessarily mean that one causes the other to change. Therefore, while correlation is a practical tool for feature selection, it should not be used as the sole determinant.

Ordinary Least Squares

Next, Ordinary Least Squares (OLS) regression was performed. Ordinary Least Squares (OLS) regression is a popular method for estimating unknown parameters in a regression model. It reduces the sum of the squared residuals, thus the term least squares. The estimates produced by OLS are known to be unbiased, efficient, and consistent, given that the linear regression model's assumptions hold true.

When it comes to selecting features, OLS regression provides several insights that can help in determining the relevance of a given feature for a model:

- Coefficient Estimates: The estimates for the coefficients in an OLS model indicate the expected change in the response variable for a one-unit change in the predictor, all other predictors being equal. Larger absolute values of the coefficient estimates suggest that the predictor strongly affects the response variable.
- P-values: The p-value associated with each predictor tests the null hypothesis that the predictor's coefficient is zero, given that all other predictors are in the model.

A low p-value (typically less than 0.05) indicates one can reject this null hypothesis, suggesting that the predictor is statistically significant.

- R-squared and Adjusted R-squared: R-squared measures the variance's share in the response variable that can be described by the predictors. However, regardless of their effectiveness, R-squared always increases when new predictors are added. The adjusted R-squared accounts for the number of predictors in the model, penalizing the addition of uninformative predictors.

The predictors were chosen based on their statistical significance in the OLS model and used in the primary model forecasting PM2.5 and PM10.

A Stepwise Regression

Subsequently, a stepwise regression was conducted. This method was chosen because it combines both forward selection and backward elimination techniques, providing a more thorough examination of potential feature combinations. In this process, features were added and removed based on specified criteria until the optimal model was found.

Stepwise regression is a statistical technique that helps choose the most influential input variables in a multiple regression model. Features are added or removed through an iterative process based on their statistical importance. The primary goal of stepwise regression is to determine the simplest model that describes the most variance in the response variable.

The stepwise regression algorithm begins with a null model with no predictors, then adds or removes explanatory variables depending on their importance. Stepwise regression can be classified into two types: forward and backward. In forward stepwise regression, predictors are added one at a time based on their contribution to the model. In contrast, in backward stepwise regression, the algorithm starts with all predictors and removes them individually. Finally, the algorithm checks how the F-statistic and p-value change at every step to decide whether to add or remove a predictor. The best predictors that have a strong influence on PM2.5 concentration in the Warsaw air quality dataset were chosen and taken to the following research step.

The Least Absolute Shrinkage and Selection Operator (LASSO)

Lastly, the Least Absolute Shrinkage and Selection Operator (LASSO) regression method was utilized. This method was selected for its ability to perform variable selection and

regularization. Reducing overfitting and selecting the most relevant features helped to improve the accuracy and interpretability of the model's predictions.

To prevent overfitting and enhance the generalization of the model, the dataset was first divided into two sets: 70% of the data was assigned to the training set and the remaining 30% to the testing set. The LASSO regression model was constructed using the training set, and its performance was validated using the testing set.

Before fitting the model, the “date” and “pm10” columns were removed from the dataset. PM10 has already proved highly correlated with PM2.5, so it should not be present in the analysis. Also, the “date” column does not include pollution or weather information. The left columns were considered possible predictors, and the “pm2_5” column was taken as the outcome variable.

A 10-fold cross-validation was performed on the LASSO regression to improve the model’s performance. Cross-validation is a reliable strategy used to assess the effectiveness of models. It involves dividing the training set into smaller subsets, or “folds,” and continually training and validating the model on various subsets. With this process, one can discover the best-performing model on data that has not been previously seen. This ensures that the model is both strong and versatile.

LASSO regression's lambda parameter determines how much the coefficients of the predictors are shrunk. Through cross-validation, the optimal lambda value was selected that minimized the mean cross-validated error. Then the LASSO coefficients corresponding to this optimal lambda value were obtained.

The LASSO regression method reduces the coefficient of less significant predictors to zero, effectively removing them from the model. The most relevant predictors for predicting PM2.5 concentration were determined by identifying the ones with non-zero coefficients at the optimal lambda value.

Seasonal Distinction

Each of these methods was applied not only to the entire dataset but also to two specific segments: the summer period (21.06.2022 — 23.09.2022) and the winter period (22.12.2022 — 21.03.2023). The reason for that is severalfold. Analyzing the dataset separately for summer and winter periods allows a more nuanced understanding of the factors affecting air pollution levels during these seasons.

Air pollution levels can vary significantly between summer and winter due to meteorological conditions, changes in human activities, and energy consumption

patterns. For example, during the winter months in Warsaw, the increased use of residential heating systems can lead to higher emissions of pollutants, such as PM2.5 and PM10, from the combustion of solid fuels like coal, wood, and other biomass.

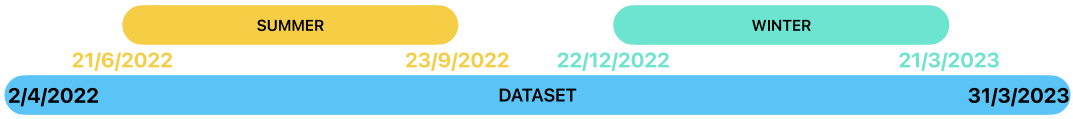

Figure 14. Dataset Split Visualization

Weather conditions, such as temperature, wind speed, and humidity, play a crucial role in the dispersion and concentration of air pollutants. In Warsaw, as in many other regions, meteorological conditions differ significantly between the summer and winter. For instance, winter months are characterized by colder temperatures, which can lead to the formation of temperature inversions that trap pollutants near the ground. On the other hand, summer months are typically warmer and have more frequent rainfall events, which can help disperse pollutants. Analyzing the data separately for summer and winter periods makes it possible to understand better the influence of meteorological factors on air pollution levels.

In Poland, coal is still a significant energy source for electricity generation and residential heating. The reliance on coal and other solid fuels for heating is exceptionally high during the winter months when heating demand is at its peak. This leads to increased emissions of pollutants from the combustion of these fuels, particularly in residential areas. The impact of different energy consumption patterns and sources on air pollution levels can be better assessed by analyzing the dataset separately for summer and winter periods.

People's activities and behavior can also vary between summer and winter months, which can, in turn, affect air pollution levels. For example, during the winter months, people tend to spend more time indoors, leading to a grow in indoor air pollution sources, such as cooking and heating. Additionally, road transportation can be affected by winter weather conditions, leading to increased vehicle emissions due to factors like increased idling and cold-start emissions. By examining the summer and winter periods separately, it is possible to account for the effect of seasonal human activities and behavior on air pollution levels.

Analyzing the dataset separately for summer and winter periods allows for a more comprehensive understanding of the factors affecting air pollution levels. This approach

allowed for an exploration of whether different features might be more relevant during different times of the year, potentially improving the accuracy of the air pollution predictions.

2.5. Implementation

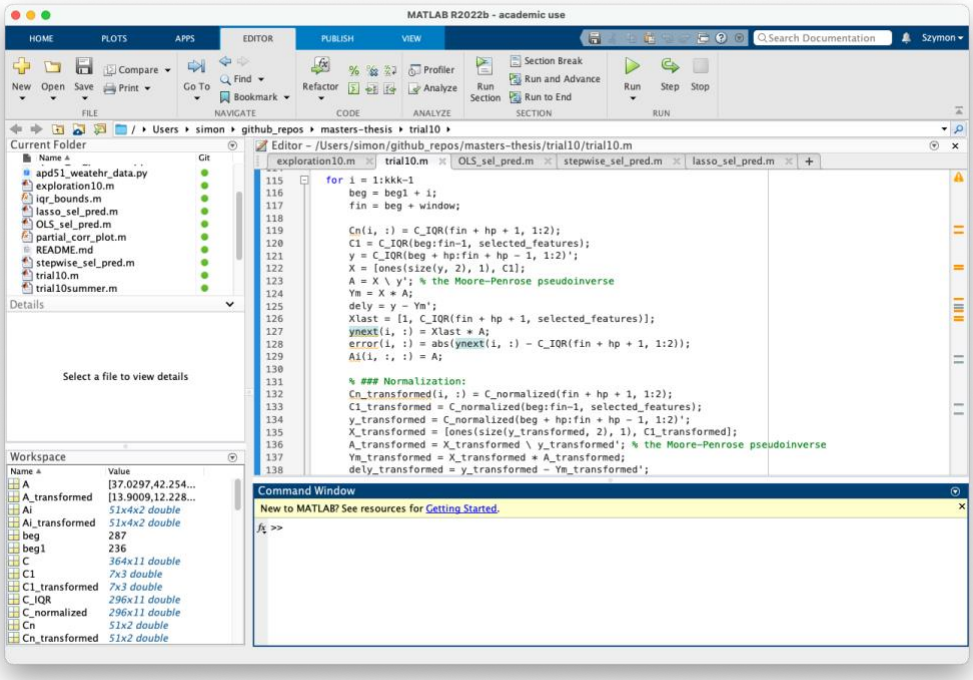

Figure 15. MATLAB's Workspace

In the study, the MATLAB programming environment was extensively used to develop and implement the air pollution prediction models. MATLAB, which stands for MATrix LABoratory, is a powerful programming language designed to offer a versatile environment for computational and algorithmic purposes. The chosen approach to analyze the data was multivariate regression combined with the sliding window method and Moore-Penrose Inverse. MATLAB's comprehensive computational and visualization capabilities made it an excellent choice for this task. The MATLAB environment was chosen for implementing the regression model due to its robust numerical computation capabilities and ease of use.

Multivariate Regression

The primary method used in this research is multivariate regression. This technique uses statistics to predict the value of a target variable(s) by analyzing the values of two or more

predictor variables. It allowed the forecast of PM2.5 and PM10 levels concurrently based on selected predictors.

In MATLAB, the multivariate regression can be implemented using the "fitlm" function. However, instead of using it, a different approach was taken in the form of a mathematical formula.

The multivariate regression was implemented by formulating a design matrix, *X*, and a response vector, *y*. Each row in the design matrix corresponds to an individual observation, and each column represents a predictor variable. The response vector, *y*, consists of observed outcomes for each observation.

To solve for the coefficient vector, denoted by *A*, MATLAB's "mldivide" function was utilized, invoked using the backslash operator ("\"). This operator in MATLAB provides a least squares solution in the case of an overdetermined system, which is a common occurrence in regression analyses. It effectively calculates the Moore-Penrose pseudoinverse of the design matrix, *X*, and multiplies it with the response vector, *y*. In the script, it is represented as:

$$A = X \setminus y';$$

where *A* is the vector of coefficients. In the context of linear regression, these coefficients represent the weight of each predictor variable in predicting the response variable.

Once the coefficients were approximated, they were used to predict new data and evaluate the model's performance. It is worth noting that this approach assumes that the relationship between predictor variables and the response is linear and that all relevant predictors have been included in the model.

A Sliding Window Approach

A sliding window approach is implemented for the multivariate regression analysis. A sliding window is a type of approach where a subset of the data is selected, a model is fit to this subset, and then the window "slides" along to the following subset of data, and the process is repeated. The sliding window technique is a commonly used method in time series analysis. The model is trained on a subset of data (*window*) that moves forward in time, enabling the prediction of future values based on past observations.

Several critical parameters are defined in the script:

- *selected_features* – an array of indices that selects specific features from the original dataset,

- *window* – the size of the sliding window, representing the number of past data points to consider when making predictions,
- *kkk* – the maximum number of windows that can be used, and
- *hp* – the prediction horizon.

After defining the necessary parameters, a loop starts. It runs for *kkk-1* iterations. In each iteration of the loop, a subset of the data is selected, starting from *beg* and ending at *fin*, creating a sliding window of size defined by *window*.

For the selected *window* of data, multivariate linear regression is performed using the Moore-Penrose pseudoinverse method. The response variables (*y*) are the data points at the current window plus the prediction horizon. The regressor matrix (*X*) is constructed by augmenting the selected features with a column of ones to account for the intercept term in the linear regression.

Then, the regression coefficients are calculated as $A = X \setminus y'$, using the Moore-Penrose pseudoinverse, which minimizes the sum of squares of the residuals, providing the least squares solution. The predicted responses (*Ym*) are then calculated by multiplying *X* with the coefficients *A*. Finally, the prediction error (*dely*) is calculated as the difference between the actual and predicted responses.

The process is then repeated for the next window of data, moving one step forward in the dataset. The final predictions for the next time step (*ynext*) are calculated using the last row of the current window and the derived coefficients *A*. This process continues until all windows are processed, resulting in a sequence of predictions and associated errors.

Using the sliding window methodology, one can use information from specific parts of the dataset to make predictions. This technique can help capture more intricate and possibly changing relationships in the data.

To provide a broader spectrum of research results, the same process is repeated twice – once for the original data and once for the normalized data. The results of both regression models are stored for subsequent comparison and analysis.

The Moore-Penrose Pseudoinverse

The pseudoinverse, also called the Moore-Penrose inverse, is used when dealing with not-invertible matrices due to being either non-square or singular.

The already described multivariate regression model utilizes the generalized inverse. This study uses the Moore-Penrose inverse to calculate the least-squares solution for linear

regression coefficients. This solution minimizes the sum of squared residuals between the observed and predicted response variables.

The backslash operator (\) in MATLAB is commonly used when implementing solutions for linear regression. The backslash operator is used to resolve a system of linear equations, $Ax = b$, where A is the matrix of predictors, x is the vector of unknown coefficients, and b is the vector of response variables. The system will be solved using the most suitable method, which the operator will automatically detect. In the context of this study, the backslash operator is applied as follows:

$$coef = X \setminus y;$$

Here, X is the predictor matrix, constructed by augmenting the selected features with a column of ones to account for the intercept term in the linear regression, and y is the response variable vector. The resulting vector *coef* contains the estimated coefficients of the linear regression model.

This approach provides an elegant and concise way to implement the multivariate linear regression model with the sliding window approach, as demonstrated in this study.

2.6. Error Metrics

The performance of the multivariate regression models can be gauged by use of several error metrics. Those that were used in the study consist of the popular ones in the field of predictive modeling and regression analysis. They give numerous ways of measuring dissimilarity between predicted and actual values.

The Mean Absolute Error (MAE) calculates the average absolute difference between the predicted and actual values. It indicates the overall accuracy of the predictions, regardless of the direction of the errors. A lower MAE indicates a better-performing model.

The Mean Squared Error (or MSE) is a method for calculating errors similar to the MAE, but it squares the differences before averaging them. This means that larger errors are punished more than smaller ones, making it more sensitive to outliers than MAE. The **Root Mean Squared Error (or RMSE)** is the square root of the MSE. By taking the square root of the MSE, the error metric is returned to the same unit as the target variable, making it easier to interpret than the MSE.

The coefficient of determination, also known as **R-squared**, determines how much of the output variable's variance can be predicted by the independent variables. It shows how closely the regression predictions match the actual data points. For example, an R-squared

of 100% indicates that changes in the dependent variable can be fully explained by changes in the independent variable(s).

The Mean Absolute Percentage Error (or MAPE) is a percentage-based measure of the accuracy of predictions, making it helpful in comparing predictions across different datasets or models. This error metric was calculated for normalized and not-normalized datasets in the script and the results were printed. By analyzing these metrics, one can better understand the magnitude and nature of prediction errors, which can help improve models' performance.

Another useful tool is the utilization of plots. In the program, four plots were generated after calculating the errors.

- Model vs Real for PM2.5 and PM10 (Non-normalized). The plots depict a comparison between the actual and predicted values for PM2.5 and PM10 separately. The actual values are represented by green, while the predicted values are shown in red within their respective ranges.
- Model vs Real for PM2.5 and PM10 (Normalized and Non-normalized). There are two distinct plots available - one for PM2.5 and another for PM10. These plots show a comparison between the real values and both the non-normalized and normalized predicted values. The actual values are depicted in green, the non-normalized predicted values in red, and the normalized predicted values in blue.
- Error Metrics Comparison for PM2.5 and PM10. The following bar plots show a comparison of different error metrics for PM2.5 and PM10. These metrics include MAE, MAE_norm (norm -- from "normalized"), RMSE, RMSE_norm, R-squared, R-squared_norm, MAPE, and MAPE_norm. Each metric is represented by a separate bar on the x-axis, while the y-axis indicates the corresponding error value.

Visual plots are valuable tools for evaluating a model’s performance. For example, one can determine the model's accuracy by comparing actual and predicted values. Additionally, comparing error metrics can provide a more comprehensive evaluation of the model's performance based on various criteria.

This section has comprehensively described this study's methods, techniques, technologies, and tools to predict air pollution levels using a multivariate regression model. The data collection process, which involved Python programs to acquire air pollution and weather data, and later manual merge, ensured the accuracy and reliability

of the dataset. The data preprocessing steps, including outliers removal, normalization, and various transformations for skewed features, prepared the data for the subsequent analyses and modeling.

The variable selection methods – correlation, OLS, stepwise regression, and LASSO – were essential in identifying the most suitable features for the prediction model. Applying these methods to different parts of the dataset (in summer, winter, and the whole year) determined a robust set of predictors. The development of the multivariate regression model, which incorporated a sliding window approach and the Moore-Penrose Inverse, facilitated accurate predictions of PM2.5 and PM10 levels in the study area.

1. **Clear workspace**
2. **Set input parameters and read data**
   - Set filename
   - **Read the data** from the file into a table
   - Extract relevant columns
   - Set other required parameters (e.g., window, constant)
3. **Preprocess data**
   - **Remove outliers** based on IQR bounds
   - **Normalize selected columns** (log, square root, cubic)
4. Initialize variables for loop
5. **Loop over data with a sliding window**
   - **Calculate the model coefficients** for the current window
   - **Make predictions** for the next data point
   - Calculate errors
   - Repeat for both original and transformed data
6. **Calculate error metrics** for the predictions (MAE, RMSE, R-squared, MAPE)
7. Print error metrics
8. **Plot predictions** and real values for visual comparison
9. Calculate error metrics for transformed predictions
10. Print transformed error metrics
11. **Plot predictions and real values for transformed data**
12. **Compare original and transformed error metrics**
13. **Save figures** to a directory
14. **Create a comparison table** of error metrics
15. **Save comparison table** as a .csv file

Figure 16. Main MATLAB Script's Pseudocode

Total Number of Conducted Studies

Overall, 28 studies were conducted, according to the formula in Figure 17.

As already mentioned, there are three periods (P) determined: the entire year (02.04.2022 - 31.03.2023), summer (22.06.2022 – 23.09.2022), and winter (23.12.2022 – 21.03.2023).

The sliding window (W) has three sizes: 7 days (week), 30 days (month), and 92 days (quarter). There are a few of reasons for that choice. First, seven days represent the duration of the week. People live and function in those seven days periods. The behaviors like going to work on weekdays and doing leisure activities on weekends are repetitive. People also tend to do things once a month (30 days). Ninety-two days represent one quarter, a more extended time.

This study has four methods for selecting the best predictors (F): stepwise regression, OLS, LASSO, and correlation. Each of the methods is used in every instance.

Using 92 days in case of short periods does not make sense. Summer and winter periods last approximately ninety days. This is why the number of possible window sizes (3) is not multiplied in the equation. Therefore, the number 4 must be added as the entire year (A) amount, the whole dataset period for each method.

$$P * W * F + A = S$$

where
P -- number of periods (3)
W -- number of window sizes (2)
F -- number of feature selection methods (4)
A -- number of the whole dataset periods for each method (4 * 1)
S -- total number of studies (28)

Figure 17. A Formula for Total Number of Studies

By combining various methods, techniques, and tools, a strong foundation for the research has been established, which has enabled a comprehensive analysis of the dataset. The subsequent sections of the thesis will present and discuss the results obtained through applying these approaches, showcasing the effectiveness of the chosen methods in addressing the research problem.

3. Results

This section shows the results acquired from the study, which are organized into three subsections. The first part discusses the outcomes of the exploratory data analysis conducted on the dataset, highlighting the essential findings and patterns observed. Next, the selected features for each of the four feature selection methods (correlation, OLS, stepwise regression, and LASSO) are presented. Differences and similarities are analyzed. Lastly, the evaluation of the performance of multivariate regression models by reporting the error metrics computed (MAE, RMSE, R-squared and MAPE). These results provide insights into the effectiveness of the proposed approach in predicting PM2.5 and PM10 levels and offer a basis for the subsequent discussion section, where these findings will be interpreted and contextualized.

3.1. Exploratory Data Analysis

Boxplot Analysis

In this section, the analysis of the boxplots for each variable is presented. It visually represents the central tendency, dispersion, and presence of outliers in the data.

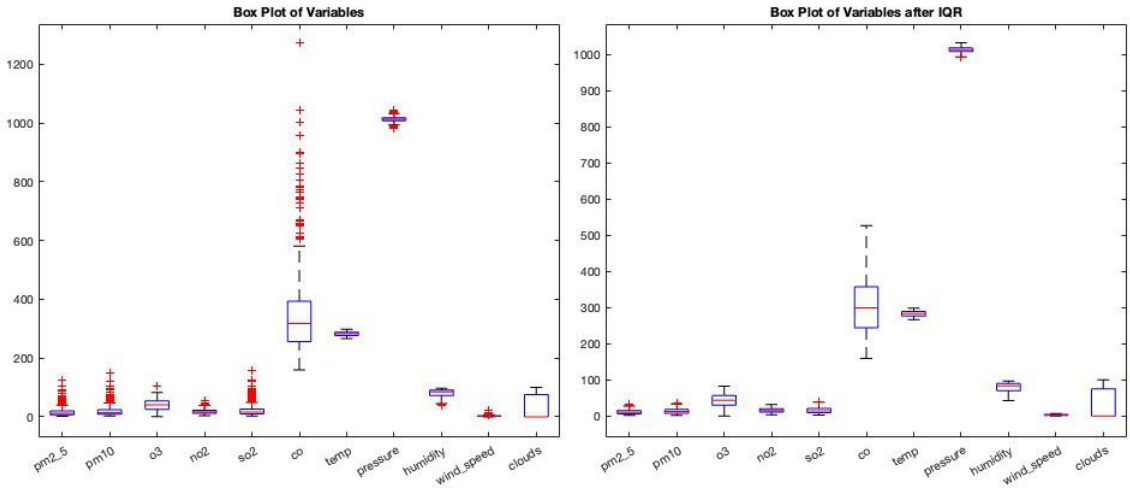

Chart 1. Boxplots Before and After Outliers Removal (IQR)

The **PM2.5** boxplot reveals a median of 11.81, Q1 at 6.98, and Q3 at 20.09. The minimum and maximum values are 1.83 and 122.88, respectively. In total, thirty-seven outliers were detected, ranging from 40.26 to 122.88.

The **PM10** boxplot gives a median of 14.25, Q1 at 8.58, and Q3 at 23.70. The minimum value equals 2.40, and the maximum value is 151.32. Thirty-seven outliers were observed, with values ranging from 47.03 to 151.32.

The **O3** boxplot exhibits a median of 40.7, Q1 at 24.76, and Q3 at 53.76. The minimum and maximum values are 0.28 and 105.41, respectively. A single outlier was recorded at 105.41.

The **NO2** boxplot displays a median of 16.71, Q1 at 11.56, and Q3 at 22.53. The minimum value equals 3.91, and the maximum value is 56.92. Five outliers were identified, with values ranging from 40.37 to 56.92.

The **SO2** boxplot shows a median of 16.15, Q1 at 10.23, and Q3 at 26.68. The minimum value is 2.7, and the maximum value equals 158.75. Twenty-seven outliers were detected, ranging from 52.23 to 158.75.

The **CO** boxplot indicates a median of 317.55, Q1 at 256.11, and Q3 at 393.24. The minimum value equals 159.17, and the maximum value is 1272.84. Thirty-three outliers were observed, with values ranging from 604.43 to 1272.84.

The **temperature** boxplot presents a median of 282.75, Q1 at 275.98, and Q3 at 288.1. The minimum value is 266.56, and the maximum value equals 298.41. No outliers were detected for this variable.

The **pressure** boxplot exhibits a median of 1013, Q1 at 1008, and Q3 at 1018. The minimum value is 982, and the maximum equals 1042. Thirteen outliers were identified, ranging from 986 to 1042.

The **humidity** boxplot displays a median of 85, Q1 at 72.5, and Q3 at 91. The minimum value is 38, and the maximum value equals 97. Three outliers were detected at 38 and 43 (twice).

The **wind speed** boxplot shows a median of 3.09, Q1 at 2.06, and Q3 at 4.12. The minimum value is 0, and the maximum value equals 21.01. Seven outliers were identified, ranging from 7.72 to 21.01.

The **cloud cover** boxplot exhibits a median value, with Q1, the minimum value at 0%, and Q3 at 75%. The maximum value is 100%. No outliers were detected for this variable.

Applying the interquartile range (IQR) method for outlier removal significantly reduced the number of outliers in the dataset. For example, for PM2.5, the number of outliers decreased from 36 before IQR to 9 after IQR, removing 27 outliers. Similarly, for PM10, the outliers reduced from 37 to 7, with 30 outliers removed.

For O3, only one outlier was identified before IQR, and after applying the IQR method, no outliers remained, resulting in the removal of one outlier. In the case of NO2, five outliers were identified before IQR, and after IQR, none remained, leading to the removal of 5 outliers. For SO2, the number of outliers decreased from 27 to 1, with 26 outliers removed after applying the IQR method.

Table 2. Outliers Summary

	Before_IQR	After_IQR	Outliers Removed
pm2_5	36	9	27
pm10	37	7	30
o3	1	0	1
no2	5	0	5
so2	27	1	26
co	33	0	33
temp	0	0	0
Pressure	13	2	11
humidity	3	0	3
wind_speed	7	0	7
clouds	0	0	0

For CO, 33 outliers were detected before IQR, and none remained after IQR, resulting in removing all 33 outliers. No outliers were identified for temperature, pressure, humidity, wind speed, and cloud cover before and after applying the IQR method.

In summary, the IQR method effectively reduced the number of outliers in the dataset, particularly for PM2.5, PM10, SO2, and CO, enhancing the reliability of the multivariate regression model for air pollution prediction in the given area.

Histogram Analysis

The **PM2.5** histogram reveals a right-skewed distribution, with a concentration of values between 0 and 20. The mean equals 17.11, the median is 11.81, and the standard deviation is 16.81.

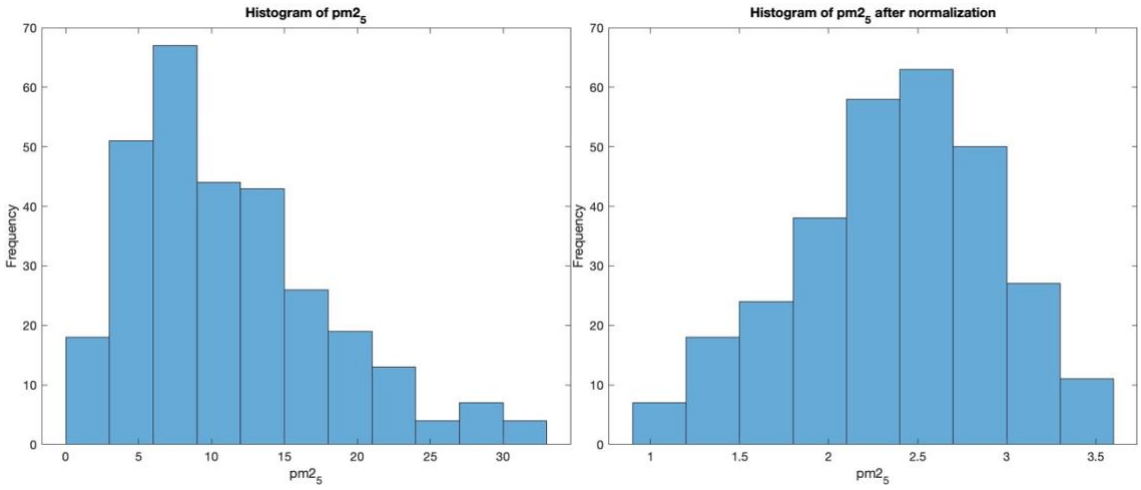

Chart 2. Histogram of PM2.5 Before and After Normalization

The histogram for **PM10** also exhibits a right-skewed distribution, predominantly featuring values between 0 and 30. The mean value is 20.47, the median equals 14.25, and the standard deviation is 19.43.

The **O3** histogram displays a slightly right-skewed distribution, with most values ranging from 20 to 50. It has a mean of 40.04, a median of 40.70, and a standard deviation of 19.77.

The **NO2** histogram demonstrates a right-skewed distribution, primarily consisting of values between 5 and 20. The mean equals 17.78, the median is 16.71, and the standard deviation is 8.26.

The **SO2** histogram presents a right-skewed distribution, with values mostly between 0 and 30. The mean is 22.39, the median equals 16.15, and the standard deviation is 20.77.

The **CO** histogram exhibits a right-skewed distribution, with most values ranging from 100 to 300. The mean is 358.75, the median is 317.55, and the standard deviation is 162.78.

The **temperature** histogram features a more symmetrical distribution, with values primarily falling between 273 and 285. The mean is 282.65, the median is 282.75, and the standard deviation is 7.69.

The **pressure** histogram shows a slightly right-skewed, almost symmetrical distribution, with most values ranging between 1000 and 1015. The mean equals 1013.10, the median is 1013, and the standard deviation is 8.97.

The **humidity** histogram displays a left-skewed distribution, with a majority of values between 80 and 95. The mean equals 80.33, the median is 85, and the standard deviation is 13.16.

The **wind_speed** histogram demonstrates a right-skewed distribution, primarily featuring values between 1 and 5. The mean is 3.24, the median equals 3.09, and the standard deviation is 2.01.

The **clouds** histogram exhibits a left-skewed distribution, with most values at 0. The mean is 31.45, the median is 0, and the standard deviation equals 38.25.

To summarize, the histograms for the air pollutants (PM2.5, PM10, O3, NO2, SO2, and CO) and meteorological variables (temperature, pressure, humidity, wind speed, and cloud cover) show varying degrees of skewness and data spread. Most pollutant histograms exhibit right-skewed distributions, indicating that lower concentrations are more common. Meteorological variables display a mix of symmetrical, left-skewed, and right-skewed distributions, suggesting diverse patterns in the data. The standard deviations reveal different spread levels for each variable, with some exhibiting a widespread and others a more moderate or narrow spread.

Transformations

The analysis of the dataset revealed that several variables exhibited skewed distributions. To address this, appropriate transformations were applied to achieve better normality. The study analyzed air quality data for various pollutants such as PM2.5, PM10, SO2, CO, O3, and NO2, as well as environmental factors like temperature, pressure, wind speed, humidity, and cloud cover. Some of these variables had skewed distributions, which could impact the accuracy of our analysis. Therefore, different transformation techniques were applied to improve the data's normality and enhance the results' reliability. Specifically, logarithmic transformations were utilized for PM2.5, PM10, SO2, and CO, while O3, NO2, temperature, pressure, and wind speed were normally distributed and required no transformations. For humidity, which was left-skewed, we used a cube transformation. Finally, a combination of logarithmic and square root transformations to the clouds variable with a right-skewed distribution was applied. As a result of these transformations, significant improvements in the normality and distribution properties of the dataset were observed, particularly for the pollutants that

had skewed distributions initially. The mean, median, and standard deviation values for these variables also showed adjustments that reflect the transformations.

Table 3. Data Transformation Summary

	Distribution	Transformation
pm2_5	Right-skewed	Log
pm10	Right-skewed	Log
o3	Normal	-
no2	Normal	-
so2	Right-skewed	Log
co	Right-skewed	Log
temp	Normal	-
pressure	Normal	-
humidity	Left-skewed	Cube
wind_speed	Normal	-
clouds	Right-skewed	Log & square root

For O3, NO2, temperature, and pressure, which already had normal distributions, only minor changes were observed in their mean, median, standard deviation, and skewness values. These variables required no transformations and were maintained in their original form.

Humidity, which had a left-skewed distribution, underwent a cube transformation. The result of this operation is not auspicious, as the change in measured parameters is minor. With an initially normal distribution, wind speed displayed moderate changes in its mean, standard deviation, and skewness values, while its median remained constant.

Lastly, the clouds variable, which required a combination of logarithmic and square root transformations, showed improvements in its skewness value and changes in its mean, median, and standard deviation values. Overall, the preprocessing of the dataset led to enhanced normality and distribution properties, which are essential for the accuracy and reliability of the subsequent analysis.

3.2. Feature Selection

When creating a multivariate regression model, selecting the right features is essential. This step helps to find the most important predictor variables while reducing noise and

multicollinearity. This study employed four feature selection methods: correlation, ordinary least squares (OLS), stepwise regression, and LASSO, to determine the optimal set of features for predicting air pollution levels in Warsaw, Poland. These analyses were performed on the whole dataset. The results obtained will be used while conducting the research for summer, winter, and the whole study period.

Table 4. Descriptive Statistics: Before and After Preprocessing

	Mean	Median	Std dev	Skeweness	Mean processed	Median processed	Std dev processed	Skeweness processed
pm2_5	17.10	11.81	16.81	2.54	2.37	2.42	0.55	-0.25
pm10	20.47	14.25	19.43	2.55	2.56	2.60	0.54	-0.26
o3	40.04	40.70	19.77	0.07	43.60	43.25	17.97	-0.02
no2	17.78	16.71	8.26	0.84	15.61	15.06	6.13	0.27
so2	22.39	16.15	20.77	2.77	2.70	2.75	0.53	-0.32
co	358.75	317.55	162.78	2.02	5.69	5.70	0.26	-0.05
temp	282.65	282.75	7.69	0.08	283.40	283.61	7.61	0.01
pressure	1013.14	1013.00	8.97	0.04	1012.59	1013.00	7.87	-0.14
humidity	80.33	85.00	13.16	-1.00	544776.66	592704.00	223119.46	-0.45
wind_speed	3.24	3.09	2.01	2.64	3.19	3.09	1.56	0.37
clouds	31.45	0.00	38.25	0.60	3.99	1.00	3.71	0.55

Correlation

The correlation matrix reveals the following relationships between variables in the dataset.

PM2.5 and **PM10** have a strong positive correlation (0.99). **O3** has negative correlations with PM2.5 (-0.52), PM10 (-0.52), NO2 (-0.60), SO2 (-0.49), and CO (-0.53). **NO2** shows positive correlations with PM2.5 (0.80), PM10 (0.81), SO2 (0.86), and CO (0.87). **SO2** and **CO** exhibit positive correlations with PM2.5 (0.77, 0.85), PM10 (0.79, 0.86), and each other (0.88). The **temperature** has negative correlations with PM2.5 (-0.25), PM10 (-0.24), NO2 (-0.41), SO2 (-0.41), and CO (-0.51), and positive with O3 (0.35) and pressure (0.03). **Humidity** displays a positive correlation with clouds (0.54) and weak correlations with PM2.5 (0.19) and O3 (-0.53).

Wind speed reveals weak or negligible correlations with most variables, with the highest correlation being with clouds (0.24).

These observed correlations provide an overview of the dataset's relationships between pollutants and meteorological factors. The interpretation of these relationships and their potential implications will be discussed further in the Discussion of the thesis.

The script contained a formula showing strong and weak correlations with PM2.5. Variables highly correlated with PM2.5 (more than 0.7) are PM10, **NO2**, **SO2**, and **CO**. The ones that shew a low correlation (less than 0.2) are pressure, humidity, and wind speed. The first group (without PM10 due to high correlation with PM2.5) will be used in the subsequent analysis and was taken as predictors of PM2.5 and PM10 levels.

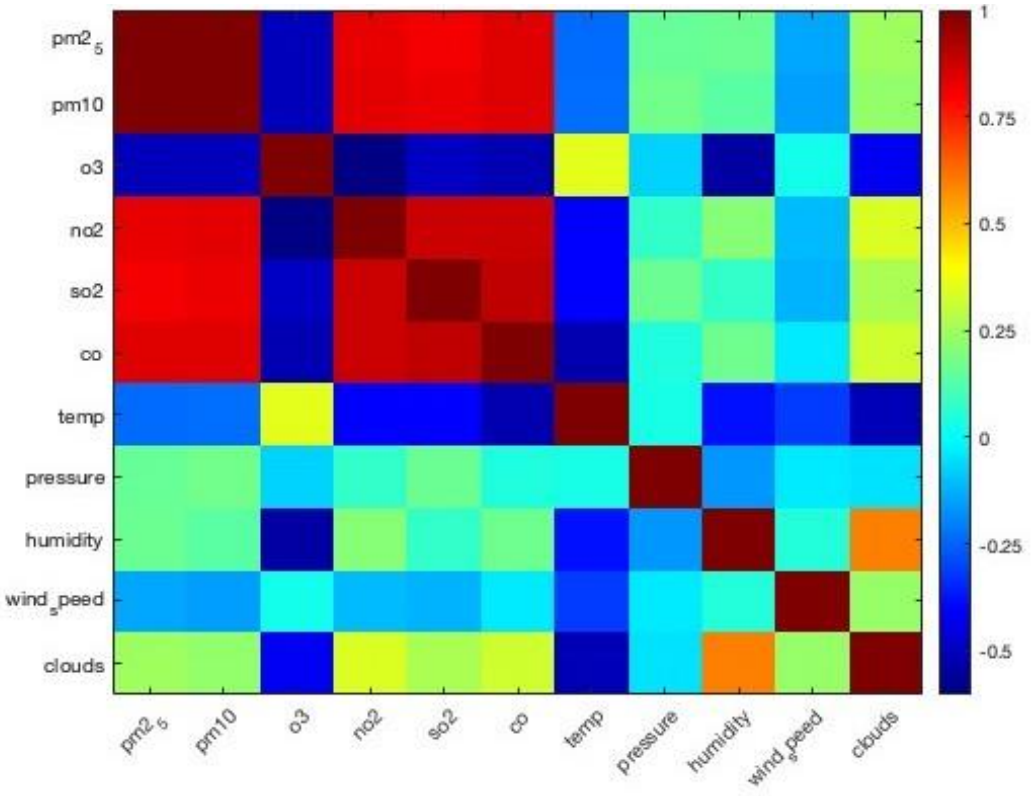

Chart 3. Correlation Heatmap

Ordinary Least Squares (OLS)

The results of the ordinary least squares (OLS) regression analysis for predicting PM2.5 concentrations based on various predictors are as follows:

The linear regression model used is:

pm2_5 ~ 1 + o3 + no2 + so2 + co + temp + pressure + humidity + wind_speed + clouds

The model has an R-squared value of 0.93 and an adjusted R-squared value of 0.93, indicating that the model explains approximately 93% of the variation in PM2.5 concentrations.

The coefficients of the predictors resulted as follows:

- O3: 0.02 (p-value: 0.29)
- NO2: -0.08 (p-value: 0.19)
- SO2: -0.06 (p-value: 0.04)
- CO: 0.12 (p-value: <0.001)
- Temp: 0.46 (p-value: <0.001)
- Pressure: 0.12 (p-value: <0.001)
- Humidity: 0.10 (p-value: <0.001)
- Wind_speed: 0.25 (p-value: 0.05)
- Clouds: 0.02 (p-value: 0.01)

The model shows that CO, temperature, pressure, and humidity have significant positive associations with PM2.5 concentrations, while SO2 and clouds have significant negative associations. Wind speed has a borderline significant positive association, and O3 and NO2 show non-significant associations.

This summary overviews the relationships between various predictors and PM2.5 concentrations. Only the significant features were chosen for the study's next steps: **CO, temperature, pressure,** and **humidity**. These findings will be further reviewed in the Discussion of the thesis.

Stepwise regression

The stepwise algorithm was employed to identify the most relevant predictors for the model predicting PM2.5 and PM10 concentrations. The algorithm selected the following predictors based on their statistical significance:

1. CO (p-value: <0.001)
2. Temperature (p-value: <0.001)
3. Humidity (p-value: <0.001)
4. Pressure (p-value: 0.001)
5. SO2 (p-value: 0.014)
6. Wind_speed (p-value: 0.008)
7. Clouds (p-value: 0.015)

The interaction terms between CO and humidity, SO2 and CO, SO2 and humidity, CO and clouds, CO and temperature, and wind_speed and clouds were also significant.

The resulting model, which includes the selected predictors and interaction terms, achieved an adjusted R-squared value of 0.94, indicating that the model explains approximately 94% of the variation in PM2.5 and PM10 concentrations. These selected predictors will be used to further analyze the final predictive model.

LASSO

The optimal lambda value, which determines the level of regularization in the LASSO regression, was found to be 0.94. The LASSO coefficients for the optimal lambda value were as depicted in Table 5.

Table 5. LASSO Coefficients for the Optimal Lambda Value

Predictor	Coefficient
o3	0
no2	0
so2	0
co	0.09
temp	0.12
pressure	0
humidity	0.03
wind_speed	0
clouds	0

Based on the LASSO coefficients, the selected predictors for the multivariate regression model were:

- **CO** with a coefficient of 0.09
- **Temperature** with a coefficient of 0.12
- **Humidity** with a coefficient of 0.03

These predictors were selected by the LASSO method as they had non-zero coefficients, indicating their significance in the model.

The adjusted R-squared value for the multivariate regression model using the LASSO-selected predictors was 0.90. This indicates that the model accounts for approximately 90% of the variability in the air pollution data, which suggests a solid predictive capability.

The LASSO method resulted in a multivariate regression model with three significant predictors: CO, temperature, and humidity. The model had an adjusted R-squared value of 0.90, demonstrating its strong performance in predicting air pollution levels in the given area.

Table 6. Table of Selected Features

Method	Selected Features
Correlation	NO2, SO2, CO
OLS	CO, temp, pressure, humidity
Stepwise Regression	SO2, CO, temp, pressure, humidity, wind speed, clouds
LASSO	CO, temp, humidity

The selected features across the different methods and periods showed some similarities and differences. For instance, all methods consistently selected CO for the whole dataset, indicating its strong influence on air pollution levels in Warsaw. On the other hand, temperature and humidity were selected by OLS, stepwise regression, and LASSO for the whole dataset, suggesting their potential importance as well.

Based on these findings, all the selected features will be used in the implementation of the multivariate regression model. Then, the performance in predicting air pollution levels in Warsaw will be evaluated.

Prediction Accuracy of Feature Selection Methods

The breakdown of all 28 studies (in Appendices 1 and 2) regarding the prediction accuracy of the feature selection methods revealed exciting facts. After analyzing the OLS and LASSO methods, both showed strong performance, especially during summer. The selected predictors for these models were Carbon Monoxide (CO), temperature, and humidity, including pressure in the OLS model. These methods yielded low error rates

and high R-squared values, indicating their predictive solid accuracy and explanatory power.

Models that used the correlation feature selection method showed inconsistent performance, as they had higher error rates and lower R-squared values when predicting on a yearly basis. This suggests that the performance of the correlation method is more inconsistent, highlighting the crucial role of careful feature selection.

When looking at the relationship between window size and the number of predictors, an intriguing trend was noticed. Models employing larger window sizes and a more extensive set of predictors, such as those using the stepwise regression and LASSO methods over a window size of 92, demonstrated increased error rates and lower R-squared values. This pattern implies the potential issue of overfitting in these models.

Lastly, during the winter months, the LASSO method was notably less effective, resulting in lower performance from the models overall. This issue was manifested through higher error rates and low R-squared values, suggesting the increased complexity of air pollution prediction during winter.

These findings highlight that the OLS and LASSO techniques are effective when used with a specific group of predictors, emphasize the significance of proper feature selection, draw attention to the risk of overfitting with larger window sizes, and more predictors shed light on the intricacy of predicting winter air pollution.

Seasonal Distinction

By looking at Chart 4, it is hard to spot any trends in PM2.5 pollution. However, one can identify that in May, June, July, and August, there are just two days with pollution levels above 40 $\mu g/m^3$. The pikes are present in April, November, and February. Higher values of PM concentrations are observed in winter months.

3.3. Model Performance

The following section presents the calculated error metrics for 28 studies (full results are listed in Appendices 1 and 2) that used different combinations of periods, sliding window sizes, and feature selection methods. These studies assessed how well multivariate regression models perform under varying conditions.

For each of the 28 studies, the following error metrics were calculated: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared, and Mean Absolute

Percentage Error (MAPE). The error metrics evaluate how accurate and dependable the multivariate regression models are in forecasting air pollution.

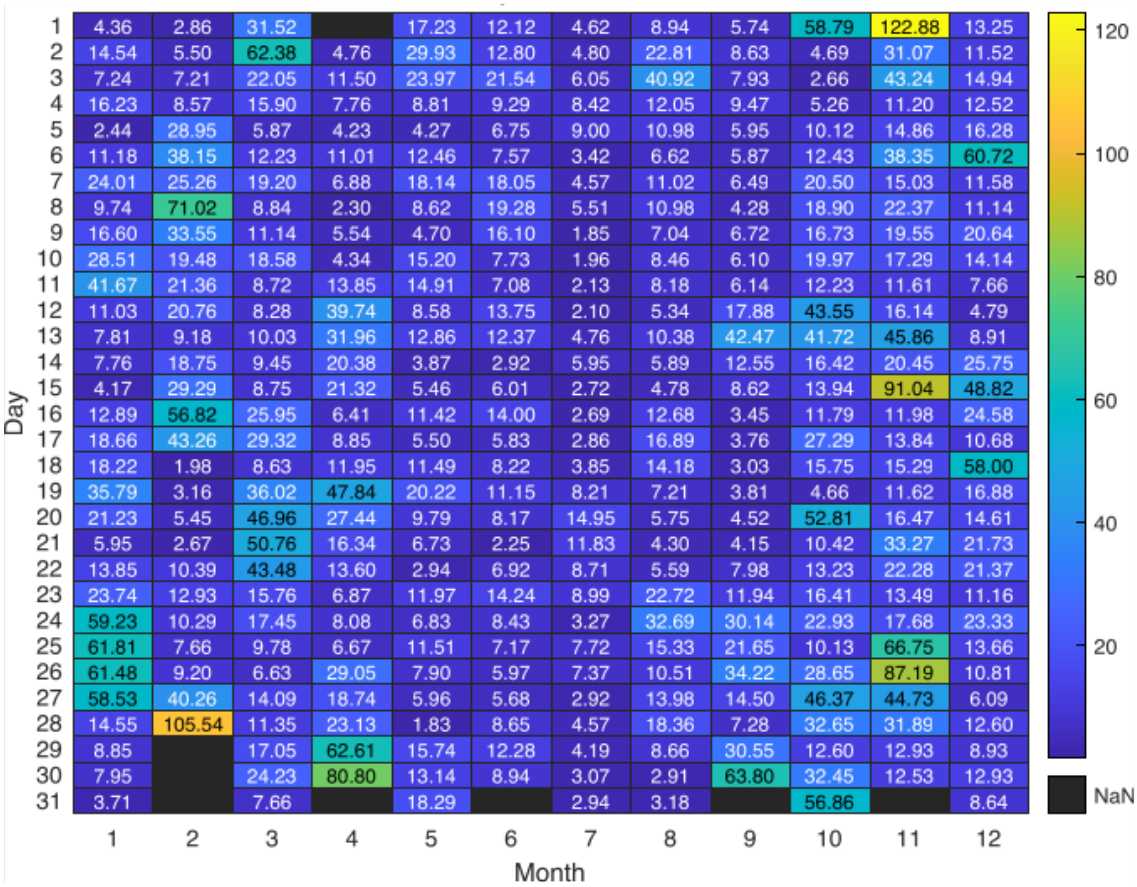

Chart 4. Calendar Heatmap of PM2.5 Values in 2022 and 2023, Combined.

The following tables 7, 8 and 9 provide the metric error values for chosen studies. Appendices 1 and 2 contain a comprehensive table of most the study findings.

The implications of these results for air pollution prediction and management will be addressed in the upcoming Discussion chapter, providing a thorough interpretation and discussion.

In summary, exploratory data analysis revealed critical insights about the nature and distribution of the dataset. Skewed features were identified, and appropriate transformations to normalize them were applied, providing a cleaner and more manageable dataset for modeling.

Table 7. Error Metrics for the Entire Year with a 30-Day Sliding Window

Period	window	Method	Y	R^2 norm.	MAPE norm.	Study No.
year	30	LASSO	PM10	0.54	7.69	29
year	30	LASSO	PM2_5	0.49	8.45	29

Table 8. Error Metrics for The Entire Year with a 92-Day Sliding Window

Period	window	Method	Y	R^2 norm.	MAPE norm.	Study No.
year	92	correlation	PM10	0.32	11.56	3
year	92	stepwise regression	PM10	0.21	11.48	21

Table 9. Error Metrics for The Summer Period with a 30-Day Sliding Window

Period	window	Method	Y	R^2 norm.	MAPE norm.	Study No.
summer	30	OLS	PM10	0.71	6.3	14
summer	30	LASSO	PM10	0.7	6.29	32

The feature selection process, involving correlation analysis, OLS, stepwise regression, and LASSO, identified key predictors for both PM2.5 and PM10. Notably, the optimal features differ depending on the dataset's selection method and time period, suggesting that air pollution levels are affected by various factors that may vary throughout the year.

The multivariate regression model demonstrated respectable performance in predicting PM2.5 and PM10 levels, as evidenced by the computed error metrics. Nonetheless, room for improvement remains, as the model's performance can feasibly be enhanced with more sophisticated methods or additional data.

These results provide a foundation for a more in-depth discussion and interpretation of the findings, which will be addressed in the following Discussion. The aim is to delve deeper into the implications of these results and explore potential avenues for future research.

4. Discussion

The purpose of this chapter is to give background information and explain the research findings. It focuses on how the findings align with the broader context of air pollution prediction and management, providing insights into their significance and implications.
Next, the significance of the findings will be explained concerning the research questions and objectives. "Implications of Findings for Air Pollution Prediction and Management" will explore the practical significance of the results. It will discuss how these findings could be applied in real-world scenarios, potentially informing policy-making and strategic planning in air pollution control. Finally, the last part will acknowledge the constraints encountered during the research process and suggest areas for future research. This chapter aims to underscore the paper's relevance in the broader field of air pollution studies and highlight the potential for applying the findings in real-world contexts.

4.1. Interpretation of Results

Boxplot Analysis

The descriptive statistics provided through boxplots revealed the distributions of air pollutants (PM2.5, PM10, O3, NO2, SO2, and CO), temperature, pressure, humidity, wind speed, and cloud cover in Warsaw. Notably, PM2.5, PM10, SO2, and CO had several outliers, which could impact the reliability and generalizability of the multivariate regression model.

By applying the interquartile range (IQR) method for outlier removal, the number of outliers in the dataset was significantly reduced, particularly for PM2.5, PM10, SO2, and CO. This step was crucial in enhancing the reliability of the multivariate regression model for air pollution prediction. For other variables (O3, NO2, temperature, pressure, humidity, wind speed, and cloud cover), the IQR method was optional or had little impact on the number of outliers.

Removing outliers can improve the robustness of the multivariate regression model and lead to more accurate predictions of air pollution levels in the given area. However, it is essential to consider potential reasons for the presence of these outliers, such as exceptional events or measurement errors, to ensure that the model captures relevant phenomena and remains valid in real-world applications.

Additionally, further analysis could explore the relationships between the variables and their respective contribution to air pollution prediction. Investigating these variables' individual and combined effects can provide valuable insights into the underlying factors influencing air pollution levels in the given area, informing policy decisions and mitigation strategies.

In conclusion, the IQR method effectively reduced the number of outliers in the dataset, particularly for PM2.5, PM10, SO2, and CO. This enhanced the reliability of the multivariate regression model for air pollution prediction. Future work should consider the potential reasons for outliers and explore the relationships between the variables to understand better the factors influencing air pollution levels.

Histogram Analysis

Analyzing the collected data through a histogram provides essential information about the distribution and features of each variable. This includes air pollutants such as PM2.5, PM10, O3, NO2, SO2, and CO, as well as meteorological parameters like temperature, pressure, humidity, wind speed, and cloud cover. This understanding is essential to the robustness of the multivariate regression models utilized in the study for air pollution prediction.

Most air pollutants showed a right-skewed distribution, indicating that lower concentrations are more frequently observed. This skewness can affect the accuracy of the regression models, as these models assume a normal distribution of the residuals. Therefore, the observed skewness necessitated the investigation of suitable data transformation techniques such as log transformation, square root transformation, or cube transformation to ensure the best fit. Future research might explore the impact of these transformations on the model's prediction accuracy.

The standard deviations in the dataset showed a wide range of variability. For example, certain variables, such as CO, exhibited a greater spread of values, which could disproportionately affect their contribution to the regression models. Normalization and data transformations were used to bring all variables to a similar scale to ensure that no single variable dominated the model solely due to its scale.

The diverse distribution patterns observed in meteorological variables, including symmetrical, left-skewed, and right-skewed distributions, revealed an intricate interplay of environmental factors. This complexity underscores the need for comprehensive

multivariate approaches, like the ones utilized in this study, for accurate air pollution prediction.

Additionally, the variability and skewness inherent in the dataset could affect the validity of the regression model assumptions, mainly the normality and homoscedasticity of residuals. Therefore, conducting diagnostic tests on the residuals and applying remedial measures where necessary to achieve unbiased and efficient estimates was crucial.

These observed data distribution characteristics also underscored the importance of careful feature selection. The scale or spread of variables could influence their significance in the model. Thus, the feature selection process was adjusted to account for these aspects, ensuring that the selected features indicated the underlying connections and not merely a result of scale or spread.

Lastly, the observed skewness and variability had implications for understanding error metrics. For example, given the possibility of asymmetric forecast errors, robust error metrics like Mean Absolute Error (MAE) could provide a more reliable evaluation of model performance in some cases, compared to Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

In summary, the histogram analysis delivered valuable insights into the dataset's attributes, which directly informed the choice of preprocessing steps, the selection of features, and the interpretation of model performance. This highlights the significance of a thorough exploratory data analysis in air pollution prediction studies. Further studies could explore the influence of these data characteristics on other prediction methods and their potential implications for air pollution management.

Transformations

The analysis identified the need for transformations for better normal distribution. Several variables showed skewed distributions. These transformations were a crucial part of preprocessing steps, which enhanced the accuracy and reliability of the multivariate regression analysis for predicting air pollution.

The variables PM2.5, PM10, SO2, and CO, which exhibited right-skewed distributions, underwent logarithmic transformations. This resulted in a substantial reduction in their skewness and significant adjustments in their mean, median, and standard deviation values, demonstrating the effectiveness of the transformations in addressing skewness.

However, the cube transformation applied to the left-skewed humidity variable did not substantially reduce skewness, despite causing significant changes in mean, median, and standard deviation values. This suggests that the cube transformation may have been less effective for this variable. This is a significant limitation of the study and indicates that alternative transformation methods might be more suitable for such data in future research.

Interestingly, O3, NO2, temperature, and pressure, which initially displayed normal distributions, required no transformations. The minor changes observed in their descriptive statistics can be attributed to the overall effects of preprocessing the dataset.

Despite a combination of logarithmic and square root transformations, the clouds cover also did not show a significant reduction in skewness. This indicates that these transformations may not have been as effective for this variable, another study limitation.

Despite these limitations, the overall transformation process significantly enhanced the normality and distribution properties of most of the dataset. This was essential for improving the accuracy and reliability of the multivariate regression analysis.

Nevertheless, it is essential to remember that transformations can introduce complexity and potentially alter the regression coefficients' interpretability. Therefore, caution is indicated when interpreting the results of the multivariate regression analysis post-transformation.

Future research could explore other transformation techniques, such as Box-Cox transformations, and compare their effects on the regression results. By conducting such efforts, one can gain better insights into the most effective preprocessing techniques for using multivariate regression to predict air pollution.

Feature Selection Results

The feature selection process plays a critical role in developing a predictive model. It helps identify the most relevant predictors to improve the model's performance. To select features from the dataset, the study employed four methods – correlation analysis, Ordinary Least Squares (OLS), stepwise regression, and Least Absolute Shrinkage and Selection Operator (LASSO). Notably, these methods were applied to three distinct periods: the whole year, the summer period, and the winter period.

During this process, several features considered crucial by each method were identified. However, it is essential to note that there was some variation in the selected features depending on the selection method and the time period of the dataset. Such variation offers a captivating look into air pollution dynamics and its predictors, suggesting that different factors may exert influence at different times of the year.

This section delves deeper into the feature selection results. An extensive analysis of the methods and periods will be conducted, highlighting their similarities and differences. Furthermore, the impact of these findings on the accuracy of air pollution prediction will be described.

Correlation

The results of the correlation matrix revealed intriguing associations between the variables within the dataset, which can be instrumental in enhancing understanding of the complex interplay between various air pollutants and meteorological factors.

A strong positive correlation (0.99) between PM2.5 and PM10 was expected, as these variables typically originate from similar sources and undergo similar atmospheric processes. This correlation indicates the possibility of using either of these variables as a proxy for the other in future air pollution prediction models.

Contrary to PM2.5 and PM10, O3 was found to have negative correlations with almost all other pollutants. This is particularly interesting as it could signify the different nature of ozone pollution, which is mainly driven by photochemical reactions and tends to peak during periods of high solar radiation. These findings also reflect the well-established phenomenon of ozone titration, where high levels of NO2, SO2, and CO (which have shown positive correlations with each other) contribute to the reduction of O3 in the atmosphere (Ngarambe et al., 2021).

The negative correlation between temperature and most pollutants, particularly NO2, SO2, and CO, could indicate enhanced pollutant dispersion during warmer periods or the influence of anthropogenic heating sources during colder periods. This might suggest seasonal variations in air pollution levels, which would be an essential consideration for future models.

The weak correlations of pressure, humidity, and wind speed with PM2.5 present an exciting avenue for further research. While these meteorological factors showed low linear correlation, their impacts on air pollution dispersion and transformation processes

could be more complex and non-linear. This highlights the importance of exploring non-linear relationships and interactions between variables in future studies.

Finally, the correlations identified in this study were used to inform the selection of features for the multivariate regression models. Variables with a high correlation to PM2.5 were included as predictors, while those with a low correlation were excluded. Nevertheless, it is essential to note that correlation does not allude to causation. These relationships should be further investigated using more complex analyses that can account for potential confounding factors and causal relationships. It is also vital to remember that these findings are specific to this study's dataset and geographic location (Warsaw) and may not be generalizable to other areas or periods.

The study's strong and weak correlations can be a reliable foundation for developing air pollution prediction models. However, they also highlight the complexity of air pollution dynamics and the need for more comprehensive and nuanced approaches in future research.

OLS

Utilized Ordinary Least Squares (OLS) regression analysis, which aimed to predict PM2.5 concentrations based on various predictors, produced noteworthy findings. This analysis reveals which environmental factors have the most significant impact on PM2.5 levels, giving valuable knowledge about air pollution dynamics.

The model achieved a high R-squared value of 0.93, suggesting that the selected predictors could explain approximately 93% of the variation in PM2.5 concentrations. This high value underscores the model's predictive solid power and the selected factors' relevance in explaining PM2.5 concentrations.

The findings showed a clear positive correlation between PM2.5 levels and CO, temperature, pressure, and humidity. This suggests that increases in these factors may lead to heightened PM2.5 levels. The implications of this spotting could be critical for pollution management strategies, as these factors could be controlled or mitigated to manage PM2.5 levels effectively.

The analysis showed a significant negative association between SO2 and clouds with PM2.5 concentrations. Higher SO2 levels or cloudiness might be associated with lower PM2.5 concentrations. This counterintuitive finding warrants further investigation, as it could reveal novel insights into air pollution dynamics and potentially be leveraged to devise innovative pollution reduction strategies.

Wind speed showed a borderline significant positive association with PM2.5 concentrations, indicating that this factor might also play a role in determining PM2.5 levels. However, the exact relationship is more complex than with other factors.

Interestingly, O3 and NO2 did not significantly correlate with PM2.5 concentrations in the model. While these factors are known to be essential components of air pollution, their lack of significant association with PM2.5 concentrations in the study suggests that they may not be significant predictors of PM2.5 levels, at least in the context of this study area and the specific period under investigation.

Summing up, the OLS analysis revealed complex interrelationships between various environmental factors and PM2.5 concentrations, highlighting the multifaceted nature of air pollution dynamics. These findings suggest that considering multiple factors, a comprehensive approach is necessary for effective air pollution prediction and management.

Stepwise Regression

Through the stepwise regression method, seven predictors were identified as having a significant impact on the concentration of PM2.5. These predictors were CO, temperature, humidity, pressure, SO2, wind speed, and clouds. The selection of these variables indicates their substantial importance in predicting air pollutant concentrations.

Notably, CO was found to be a significant predictor with a p-value of less than 0.001, emphasizing its critical role in air pollution models. The pivotal role of CO aligns with previous studies that have also identified CO as a significant contributor to air pollution (Alberts, 1994).

Temperature and humidity, each with a p-value of less than 0.001, were also identified as significant predictors. This suggests that atmospheric conditions, particularly temperature, and humidity, play a substantial role in determining the concentration of air pollutants. These findings corroborate existing literature highlighting atmospheric conditions' influence on air pollution levels (Zhang & Ding, 2017).

Moreover, the stepwise regression method identified interaction terms as significant predictors. This implies that the relationship between PM2.5 and PM10 concentrations and these predictors is linear and involves complex interactions. For instance, the interaction between CO and humidity and CO and temperature suggests that these variables impact air pollution levels.

The resulting model achieved an impressive adjusted R-squared value of 0.94, which signifies that the model can explain 94% of the variation in PM2.5 and PM10 concentrations. This high explanatory power underscores the robustness of the stepwise regression method in selecting relevant predictors for air pollution prediction models.

It is essential to understand that the stepwise method creates a robust model but relies heavily on the data given. For instance, it may exclude variables with less apparent, yet still relevant, impact on the dependent variables. This potential limitation points towards the importance of careful feature selection, which could be supplemented by other methods such as LASSO or correlation analysis.
The identified predictors and interaction terms provide significant insights for predicting air pollution levels, which could inform more effective air quality management strategies. However, further research is suggested to explore the potential influence of other variables and interaction terms not identified in this study.

LASSO

The LASSO method was crucial in selecting the most significant predictors for the multivariate regression model: carbon monoxide (CO), temperature, and humidity. The chosen predictors were determined by LASSO based on their non-zero coefficients, using an optimal lambda value of 0.94.
The non-zero coefficients indicate the importance of these predictors in the model. Specifically, CO had a coefficient of 0.09, a temperature coefficient of 0.12, and a humidity coefficient of 0.03. Other predictors, including ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), pressure, wind speed, and cloudiness, had coefficients of zero, suggesting their limited impact on the model's predictive accuracy.
The strength of the model constructed using these selected predictors is additionally demonstrated by the high adjusted R-squared value of 0.90. This suggests that the model, built upon the three significant predictors, was able to account for approximately 90% of the variability in the air pollution data. The model's predictive power underscores the utility of the LASSO method in feature selection for air pollution prediction.
It is worth noting that the LASSO method resulted in a model that factors in temperature and humidity - both weather variables - along with CO. This supports the idea that air pollution levels are influenced not just by direct pollutant emissions but also by weather

conditions. It also suggests that future research and modeling efforts should consider various factors, including pollutant and weather variables.

However, the LASSO method has limitations, such as potential bias in coefficient estimates and possible sensitivity to the choice of lambda value. Future research could explore other feature selection techniques or hybrid methods that combine LASSO with other techniques to enhance the model's prediction accuracy further.

To summarize, the LASSO method successfully identified the key predictors for the multivariate regression model of air pollution, indicating its potential usefulness for this research. Further studies are recommended to verify these findings and explore additional ways to refine the model and improve its predictive power.

Model Performance Results

Analyzing the performance metrics of multivariate regression models is crucial in gaining insights about their effectiveness in predicting air pollution levels. The study employed four feature selection methods (stepwise regression, OLS, LASSO, and correlation) across different periods and sliding window sizes.

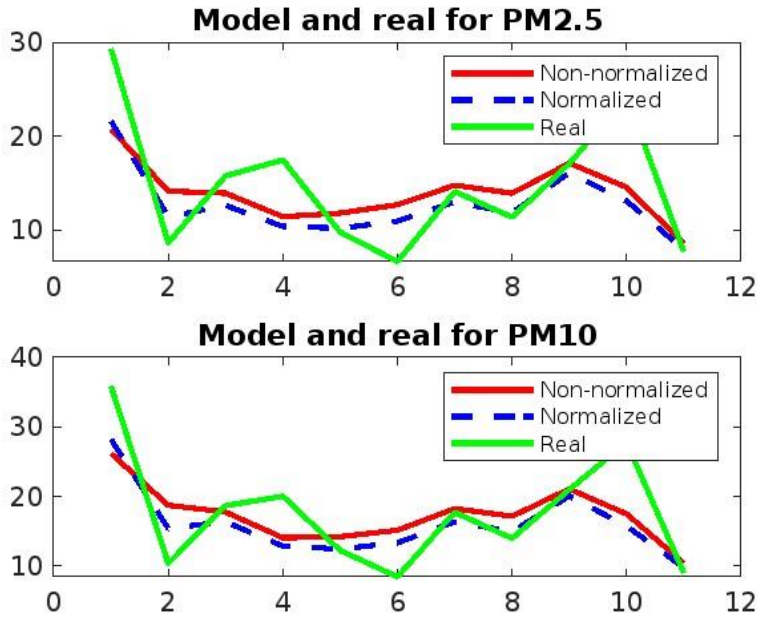

Chart 5. Study 29th Results

Feature Selection Method Performance

Observations from Tables 7 through 9 exhibit that different feature selection methods have varying performance levels depending on the period and the sliding window size.

For instance, the LASSO method yielded the highest R-squared value (0.54 and 0.49) for both PM10 and PM2.5, respectively, for the entire year with a 30-day sliding window (Table 7). This shift suggests that the LASSO feature selection method may more effectively capture the relationships between the independent and dependent variables (PM10 and PM2.5) when used with a 30-day sliding window over an entire year.

In contrast, for the same period but with a 92-day sliding window, the correlation method and stepwise regression exhibited lower R-squared values (0.32 and 0.21, respectively, for PM10, Table 8). This could imply that these methods may be less effective for longer sliding windows. The reasons for this may be many, including potentially less sensitivity to short-term variations in the data due to the more extended window size.

Sliding Window Size Impact

It is important to note that the sliding window size also appears to influence the model's performance. For the entire year period, the models with a 30-day sliding window perform better than those with a 92-day sliding window (as seen in Tables 7 and 8). This difference suggests that shorter sliding windows capture air pollution dynamics more effectively, possibly because they can accommodate more rapid changes in pollutant levels.

Seasonal Variation

The results also hint at potential seasonal variations in air pollution prediction accuracy. For example, the OLS and LASSO methods showed relatively high R-squared values for the summer period with a 30-day sliding window (0.71 and 0.7 for PM10, respectively, Table 9). This variation suggests that these methods might be particularly effective for predicting summer air pollution levels. Several factors, including changes in weather patterns, human activities, and regulatory measures, can influence seasonal variations in air pollution levels. Further research could explore these seasonal effects in more detail.

4.2. Implications of Findings for Air Pollution Prediction and Management

The regression models' results give a clear view that creating an air pollution prediction tool for the public requires a careful balance between prediction accuracy and practical usability.

The study results show a range of prediction accuracies, with R-squared values spanning from 0.21 to 0.71 and MAPE values between 6.29 and 11.56 (Tables 7 – 9). While higher

accuracy is naturally desirable for a more reliable prediction, it is crucial to consider the use-case scenarios in which such a prediction model will be employed.

The general public, who may use this tool for planning outdoor activities or gaining a general understanding of air quality in their vicinity, may require a different level of precision than would be needed for scientific research or policy-making purposes. In these contexts, a model with moderate errors, such as the LASSO regression model with a 30-day sliding window for the entire year, might be quite adequate. This model offers reasonable accuracy while maintaining applicability throughout the year and make it suitable for a broad audience in Warsaw.

In addition to these technical considerations, it is crucial to focus on enhancing the user experience to promote the adoption and utility of such a tool. This promotion could include presenting the air quality data intuitively and easily understandable. Visual representations of the data, simple interpretations of air quality levels, and practical advice tailored to different air quality conditions could significantly enhance the user experience.

Finally, it is essential to communicate the limitations of the prediction model to users. Regardless of the accuracy achieved, users must understand that the model provides an estimate and might only sometimes perfectly match actual conditions. This transparency not only enhances user trust but also ensures the responsible use of the tool.

The model can be further improved by gathering more data and feedback, improving its prediction accuracy. Therefore, this approach offers a promising avenue for the practical application of air pollution prediction in Warsaw and other similar urban environments.

4.3. Limitations of the Study and Potential Improvements

The paper brought some interesting results, but it does not lack limitations and improvements that could be explored in future research.

Limited Period and Location

The study is based on data from Warsaw, Poland, from April 2022 to April 2023. It is crucial to note that the findings may not apply universally to other places or time periods due to differences in environmental factors, sources of pollutants, and local regulations. Future studies could include data from multiple locations over several years, not just one, to ensure the model's robustness and applicability in different contexts.

Dependency on the OpenWeather Service

The trustworthiness of the results depends on the accuracy of the data provided by OpenWeather. Any inaccuracies in this data could affect the model's performance. To address this issue, researchers in future studies could gather data from various trustworthy sources instead of solely relying on OpenWeather. This move would help in verifying the accuracy and comprehensiveness of the data.

Manual Merging of Data

The data from the two Python programs were manually merged, which could introduce errors or inconsistencies. Future research could automate this process using scripting or data integration tools to reduce the potential for manual errors and increase efficiency.

Feature Selection Methods

While the study evaluated four feature selection methods, many other techniques were not explored, which might yield different results. Future studies could investigate the performance of other feature selection methods or combinations of methods to improve the model's prediction accuracy.

Single Predictive Model and Linearity

The study employed a multivariate regression model. While this model provided valuable insights, other predictive models might offer different strengths or perform better under certain conditions. Future research could explore using other predictive models, such as machine learning, deep learning, or hybrid models that combine multiple methods. The performance of these models could be compared with that of the multivariate regression model to identify the most effective prediction method for different scenarios. Also, the study did not account for potential non-linear relationships between independent variables and air pollution levels, which could be explored using more complex machine learning models.

Limitation of the Number of Pollutants

Additionally, the paper focused only on two pollutants (PM10 and PM2.5). However, they are not only two hazardous mixtures. There are plenty of others. Future research could extend this work by including other pollutants, like nitrogen dioxide (NO2) and sulfur dioxide (SO2) and examining their prediction accuracy using multivariate regression models or other methods.

Conclusion

Air pollution is a significant issue that poses numerous challenges, including health hazards, environmental damage, and economic consequences. Therefore, it is vital to have a broad comprehension and precise anticipation of pollution levels to tackle this issue effectively. This thesis aimed to create and test a multivariate regression algorithm that can accurately predict levels of PM2.5 and PM10 in Warsaw. Additionally, the thesis aspired to evaluate the effectiveness of different feature selection methods and how different periods can affect the accuracy of the predictions.

The primary objective of developing a multivariate regression algorithm for air pollution prediction was successfully achieved. Of all 28 conducted studies, one of them appeared to be acceptable in terms of prediction error. It uses predictors chosen by the LASSO method and a 30-day sliding window on an all-year period. The model offers reasonable accuracy for a broad audience in Warsaw, planning their everyday activities. However, MAPE at a level of 8.45 and R-squared of 0.49 prove the model may not be accurate for law decision-making purposes.

The study assessed different feature selection methods and how they affect prediction accuracy. The Ordinary Least Squares (OLS) and LASSO regression models were the most effective feature selection methods for predicting PM2.5 and PM10 levels, specifically during the summer period, with a window size of 30. These models used carbon monoxide (CO) predictors, temperature, pressure, and humidity. Finally, the findings provide insights into how different factors interact and influence air pollution levels, contributing to better understanding of this complex environmental issue.

The results of the research have various practical implications. For instance, the LASSO regression model, with a 30-day sliding window for the entire year, might be quite adequate for the general public. After implementing the algorithm with the above configuration in web or mobile software applications, it could be a tool for planning outdoor activities or gaining a general understanding of air quality.

Like any research, there were limitations and opportunities for improvement that could be explored. Future research could include data from multiple locations over several years to ensure the model's robustness and applicability in different contexts. Also, for future studies, researchers need to collect information from reliable sources beyond the one utilized in this paper. This move would help in verifying the accuracy and

comprehensiveness of the data. Next, future research could automate this process of collecting and merging data. It would help optimize the number of possible human errors and speed up the process. Forthcoming studies could investigate the performance of other feature selection methods or combinations of methods to improve the model's prediction accuracy. The study could also consider incorporating more advanced machine learning techniques or other predictive methods to enhance the model's predictive power. Other environmental factors or different types of pollutants could also be included to provide a more comprehensive prediction model.

In summary, this research provides valuable insights into predicting air pollution and lays the foundation for next studies. Specifically, future research could focus on improving the model's year-round prediction capabilities and exploring other feature selection methods and periods. Furthermore, the algorithm from this research could be leveraged as a bedrock to develop a mobile app, helping plan everyday activities for the general public.

References

Adams, M., & Corr, D. (2018). A Mobile Air Pollution Monitoring Data Set. *Data*, *4*(1), 2. https://doi.org/10.3390/data4010002

Adams, M. D., Massey, F., Chastko, K., & Cupini, C. (2020). Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction. *Atmospheric Environment*, *230*, 117479. https://doi.org/10.1016/j.atmosenv.2020.117479

Alberts, W. M. (1994). Indoor air pollution: NO, NO2, CO, and CO2. *Journal of Allergy and Clinical Immunology*, *94*(2), 289–295. https://doi.org/10.1053/ai.1994.v94.a56007

Alexopoulos, E. (2010). Introduction to Multivariate Regression Analysis. *Hippokratia*, *14*, 23–28.

Amato, F., Laib, M., Guignard, F., & Kanevski, M. (2019). *Analysis of air pollution time series using complexity-invariant distance and information measures*. https://doi.org/10.1016/j.physa.2020.124391

Araujo, L. N., Belotti, J. T., Alves, T. A., Tadano, Y. de S., & Siqueira, H. (2020). Ensemble method based on Artificial Neural Networks to estimate air pollution health risks. *Environmental Modelling & Software*, *123*, 104567. https://doi.org/10.1016/j.envsoft.2019.104567

Boncelet, C. (2019). *Probability, Statistics, and Random Signals*. Oxford University Press.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007a). Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. *Atmospheric Environment*, *41*(14), 2967–2995.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007b). Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. *Atmospheric Environment*, *41*(14), 2967–2995.

Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*, *119*, 285–304. https://doi.org/10.1016/j.envsoft.2019.06.014

Carmichael, G. R., Sandu, A., Chai, T., Daescu, D. N., Constantinescu, E. M., & Tang, Y. (2008). Predicting air quality: Improvements through advanced methods to integrate models and measurements. *Journal of Computational Physics*, *227*(7), 3540–3571. https://doi.org/10.1016/j.jcp.2007.02.024

Cedeño Laurent, J. G., MacNaughton, P., Jones, E., Young, A. S., Bliss, M., Flanigan, S., Vallarino, J., Chen, L. J., Cao, X., & Allen, J. G. (2021). Associations between acute exposures to PM $_{2.5}$ and carbon dioxide indoors and cognitive function in office workers: a multicountry longitudinal prospective observational study. *Environmental Research Letters*, *16*(9), 094047. https://doi.org/10.1088/1748-9326/ac1bd8

Chalela, J. A., & Lopez, J. I. (2013). Medical Management of Hunger Strikers. *Nutrition in Clinical Practice*, *28*(1), 128–135. https://doi.org/https://doi.org/10.1177/0884533612462896

Dominici, F., Peng, R. D., Barr, C. D., & Bell, M. L. (2010). Protecting Human Health From Air Pollution. *Epidemiology*, *21*(2), 187–194. https://doi.org/10.1097/EDE.0b013e3181cc86e8

Dore, M. P., Parodi, G., Portoghese, M., Errigo, A., & Pes, G. M. (2021). Water Quality and Mortality from Coronary Artery Disease in Sardinia: A Geospatial Analysis. *Nutrients*, *13*(8). https://doi.org/10.3390/nu13082858

Du, Y., Xu, X., Chu, M., Guo, Y., & Wang, J. (2016). Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence. *Journal of Thoracic Disease*, *8*(1), E8–E19. https://doi.org/10.3978/j.issn.2072-1439.2015.11.37

Egerstrom, N., Rojas-Rueda, D., Martuzzi, M., Jalaludin, B., Nieuwenhuijsen, M., So, R., Lim, Y.-H., Loft, S., Andersene, Z. J., & Cole-Huntere, T. (2022). Health and economic benefits of WHO air quality guidelines, Western Pacific Region. *Bulletin of the World Health Organization*. https://cdn.who.int/media/docs/default-source/bulletin/online-first/blt.22.288938.pdf?sfvrsn=c975d9ff_1

e-Gminy.pl. (n.d.). *EKO-SŁUPEK*. Retrieved March 17, 2023, from https://comross.pl/index.php/ekobollards/index

the Clean Air Act, 42 U.S.C. ch. 85 (§§ 7401-7671q) (1963). https://www.epa.gov/clean-air-act-overview/clean-air-act-text

EPA. (2011). *The Benefits and Costs of the Clean Air Act from 1990 to 2020*. https://www.epa.gov/sites/default/files/2015-07/documents/summaryreport.pdf

EPA. (2022a, March 29). *What is Particulate Matter?* https://www3.epa.gov/region1/eco/uep/particulatematter.html

EPA. (2022b, July 18). *Particulate Matter (PM) Basics*. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

FENG, C., WANG, H., & LU, N. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, *26*(2). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/

Fisher, M. R. (2021). *Environmental Biology* (M. R. Fisher, Ed.). Open Oregon Educational Resources. https://openoregon.pressbooks.pub/envirobiology/

Fotheringham, A. S., Yue, H., & Li, Z. (2019). Examining the influences of air quality in China's cities using multi-scale geographically weighted regression. *Transactions in GIS*, *23*(6), 1444–1464. https://doi.org/10.1111/tgis.12580

Glinianaia, S. v, Rankin, J., Bell, R., Pless-Mulloli, T., & Howel, D. (2004). Particulate air pollution and fetal health: a systematic review of the epidemiologic evidence. *Epidemiology*, 36–45.

González, C. M., Pignata, M. L., & Orellana, L. (2003). Applications of redundancy analysis for the detection of chemical response patterns to air pollution in lichen. *Science of The Total Environment*, *312*(1–3), 245–253. https://doi.org/10.1016/S0048-9697(03)00253-5

Grigorev, A. (2021). *Machine Learning Bookcamp—Build a Portfolio of Real-Life Projects*. Manning Publications Co.

Health Effects Institute. (2020a). *State of Global Air 2020*. Global Burden of Disease Study 2019. https://www.stateofglobalair.org/data/#/air/plot

Health Effects Institute. (2020b). *State of Global Air 2020. Special Report.* . https://www.healthdata.org/sites/default/files/files/policy_report/2021/soga-2020-report-10-26_0.pdf

Hota, H. S., Handa, R., & Shrivas, A. K. (2017). Time series data prediction using sliding window based RBF neural network. *International Journal of Computational Intelligence Research*, *13*(5), 1145–1156.

Huangfu, P., & Atkinson, R. (2020). Long-term exposure to NO2 and O3 and all-cause and respiratory mortality: A systematic review and meta-analysis. *Environment International*, *144*, 105998. https://doi.org/10.1016/j.envint.2020.105998

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jin, Y., Andersson, H., & Zhang, S. (2016). Air Pollution Control Policies in China: A Retrospective and Prospects. *International Journal of Environmental Research and Public Health*, *13*(12), 1219. https://doi.org/10.3390/ijerph13121219

Jolliffe, I. T. (1986). Principal Components in Regression Analysis. In I. T. Jolliffe (Ed.), *Principal Component Analysis* (pp. 129–155). Springer New York. https://doi.org/10.1007/978-1-4757-1904-8_8

Laub, A. J. (2012). *The Moore-Penrose Pseudoinverse*. https://www.math.ucla.edu/~laub/33a.2.12s/mppseudoinverse.pdf

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., & Hsu, S. D. H. (2018). Accurate Genomic Prediction of Human Height. *Genetics*, *210*(2), 477–497. https://doi.org/10.1534/genetics.118.301267

Leong, W. C., Kelani, R. O., & Ahmad, Z. (2020). Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, *8*(3), 103208. https://doi.org/10.1016/j.jece.2019.103208

Liu, Q., Xu, C., Ji, G., Liu, H., Shao, W., Zhang, C., Gu, A., & Zhao, P. (2017). Effect of exposure to ambient PM2.5 pollution on the risk of respiratory tract diseases: a meta-analysis of cohort studies. *Journal of Biomedical Research*, *31*(2), 130–142. https://doi.org/10.7555/JBR.31.20160071

MacAusland, R. (2014). The Moore-Penrose Inverse and Least Squares. In *MATH 420: Advanced Topics in Linear Algebra*. http://buzzard.ups.edu/courses/2014spring/420projects/math420-UPS-spring-2014-macausland-pseudo-inverse.pdf

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93–100. https://doi.org/10.1002/wics.14

Law of the People's Republic of China on the Prevention and Control of Atmospheric Pollution, (2018). https://www.mee.gov.cn/ywgz/fgbz/fl/201811/t20181113_673567.shtml

Monteiro, C. A., Cannon, G., Moubarac, J.-C., Levy, R. B., Louzada, M. L. C., & Jaime, P. C. (2018). The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutrition*, *21*(1), 5–17. https://doi.org/10.1017/S1368980017000234

Ngarambe, J., Joen, S. J., Han, C.-H., & Yun, G. Y. (2021). Exploring the relationship between particulate matter, CO, SO2, NO2, O3 and urban heat island in Seoul, Korea. *Journal of Hazardous Materials*, *403*, 123615. https://doi.org/https://doi.org/10.1016/j.jhazmat.2020.123615

Ni, T., Han, B., & Bai, Z. (2012). Source apportionment of PM10 in four cities of northeastern China. *Aerosol and Air Quality Research*, *12*(4), 571–582.

NSW, & EPA. (2013). *Managing particles and improving air quality in NSW*. https://www.epa.nsw.gov.au/-/media/epa/corporate-site/resources/air/epamngprtclsimprovairqualnswoct2014.pdf

OpenWeather. (2023). *OpenWeather*. https://openweathermap.org

Orellano, P., Reynoso, J., Quaranta, N., Bardach, A., & Ciapponi, A. (2020). Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment International*, *142*, 105876. https://doi.org/10.1016/j.envint.2020.105876

Park, N.-W. (2016). Time-Series Mapping of PM $_{10}$ Concentration Using Multi-Gaussian Space-Time Kriging: A Case Study in the Seoul Metropolitan Area, Korea. *Advances in Meteorology*, *2016*, 1–10. https://doi.org/10.1155/2016/9452080

Piao, M. J., Ahn, M. J., Kang, K. A., Ryu, Y. S., Hyun, Y. J., Shilnikova, K., Zhen, A. X., Jeong, J. W., Choi, Y. H., Kang, H. K., Koh, Y. S., & Hyun, J. W. (2018). Particulate matter 2.5 damages skin cells by inducing oxidative stress, subcellular organelle dysfunction, and apoptosis. *Archives of Toxicology*, *92*(6), 2077–2091. https://doi.org/10.1007/s00204-018-2197-9

Pinheiro, C. A. R., & Patetta, M. (2021). *Introduction to Statistical and Machine Learning Methods for Data Science*. SAS Institute Inc.

Remote Sensing. (n.d.). https://www.mdpi.com/journal/remotesensing/special_issues/4MC80VQW3D

Sánchez-Balseca, J., & Pérez-Foguet, A. (2020). Modelling hourly spatio-temporal PM2.5 concentration in wildfire scenarios using dynamic linear models. *Atmospheric Research*, *242*, 104999. https://doi.org/10.1016/j.atmosres.2020.104999

Sarkar, S., Khillare, P. S., Jyethi, D. S., Hasan, A., & Parween, M. (2010). Chemical speciation of respirable suspended particulate matter during a major firework

festival in India. *Journal of Hazardous Materials*, *184*(1–3), 321–330. https://doi.org/10.1016/j.jhazmat.2010.08.039

Schnabel, L., Kesse-Guyot, E., Allès, B., Touvier, M., Srour, B., Hercberg, S., Buscail, C., & Julia, C. (2019). Association Between Ultraprocessed Food Consumption and Risk of Mortality Among Middle-aged Adults in France. *JAMA Internal Medicine*, *179*(4), 490–498. https://doi.org/10.1001/jamainternmed.2018.7289

Sethi, J. K., & Mittal, M. (2021). An efficient correlation based adaptive LASSO regression method for air quality index prediction. *Earth Science Informatics*, *14*(4), 1777–1786. https://doi.org/10.1007/s12145-021-00618-1

Shahriar, S. A., Kayes, I., Hasan, K., Hasan, M., Islam, R., Awang, N. R., Hamzah, Z., Rak, A. E., & Salam, M. A. (2021). Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for Atmospheric PM2.5 Forecasting in Bangladesh. *Atmosphere*, *12*(1), 100. https://doi.org/10.3390/atmos12010100

Sharma, M., Kumar, N., Sharma, S., Jangra, V., Mehandia, S., Kumar, S., & Kumar, P. (2022). Assessment of Fine Particulate Matter for Port City of Eastern Peninsular India Using Gradient Boosting Machine Learning Model. *Atmosphere*, *13*(5), 743. https://doi.org/10.3390/atmos13050743

Smoktunowicz, A., & Wróbel, I. (2012). Numerical aspects of computing the Moore-Penrose inverse of full column rank matrices. *BIT Numerical Mathematics*, *52*(2), 503–524. https://doi.org/10.1007/s10543-011-0362-0

Snell, E. E., Carpenter, K., & Truswell, A. S. (2022, September 5). *nutrition*. Encyclopedia Britannica. https://www.britannica.com/science/nutrition

Šrám, R. J., Binková, B., Dejmek, J., & Bobak, M. (2005). Ambient air pollution and pregnancy outcomes: a review of the literature. *Environmental Health Perspectives*, *113*(4), 375–382.

Suggitt, C. (2021). *56-year-old freediver holds breath for almost 25 minutes breaking record*. Guinness World Records Limited. https://www.guinnessworldrecords.com/news/2021/5/freediver-holds-breath-for-almost-25-minutes-breaking-record-660285

Thakral, K. (2023). *Window Sliding Technique*. GeeksforGeeks. https://www.geeksforgeeks.org/window-sliding-technique/

the EPA. (2021). *Power Sector Programs Progress Report*. https://www3.epa.gov/airmarkets/progress/reports/

Directive 2008/50/EC, (2008). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:02008L0050-20150918

The European Parliament, The Council Of The European Union, & The European Commission. (2016). Directive (EU) 2016/2284. *Official Journal of the European Union*. http://data.europa.eu/eli/dir/2016/2284/oj

Directive 2003/87/EC, (2003). http://data.europa.eu/eli/dir/2003/87/2020-01-01

the European Parliament, the Council, & the European Commission. (2010). Directive 2010/75/EU. *Official Journal of the European Union*. http://data.europa.eu/eli/dir/2010/75/oj

Tschofen, P., Azevedo, I. L., & Muller, N. Z. (2019). Fine particulate matter damages and value added in the US economy. *Proceedings of the National Academy of Sciences*, *116*(40), 19857–19862. https://doi.org/10.1073/pnas.1905030116

Vautard, R., Bessagnet, B., Chin, M., & Menut, L. (2005). On the contribution of natural Aeolian sources to particulate matter concentrations in Europe: testing hypotheses with a modelling approach. *Atmospheric Environment*, *39*(18), 3291–3303.

Vicedo-Cabrera, A. M., Biggeri, A., Grisotto, L., Barbone, F., & Catelan, D. (2013). A Bayesian kriging model for estimating residential exposure to air pollution of children living in a high-risk area in Italy. *Geospatial Health*, *8*(1), 87. https://doi.org/10.4081/gh.2013.57

Wang, Y., Xiong, L., & Tang, M. (2017). Toxicity of inhaled particulate matter on the central nervous system: neuroinflammation, neuropsychological effects and neurodegenerative disease. *Journal of Applied Toxicology : JAT*, *37*(6), 644–667. https://doi.org/10.1002/jat.3451

Wasser, L. A. (2022). *The Basics of LiDAR - Light Detection and Ranging - Remote Sensing*. https://www.neonscience.org/resources/learning-hub/tutorials/lidar-basics

Weisstein, E. W. (n.d.). *Moore-Penrose Matrix Inverse.* MathWorld--A Wolfram Web Resource. Retrieved April 28, 2023, from https://mathworld.wolfram.com/Moore-PenroseMatrixInverse.html

WHO. (2014). *WHO guidelines for indoor air quality: household fuel combustion*. World Health Organization.

WHO. (2021). *What are the WHO Air quality guidelines?* https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines#:~:text=The%20WHO%20Air%20quality%20guidelines%20are%20a%20set,the%20latest%20global%20version%20was%20published%20in%202005.

WHO. (2022a). *World health statistics 2022: monitoring health for the SDGs, sustainable development goals*. https://www.who.int/publications/i/item/9789240051157

WHO. (2022b, November 28). *Household air pollution*. https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health

WHO. (2022c, December 19). *Ambient (outdoor) air pollution*. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

WMA. (2006). WMA declaration of Malta: A background paper on the ethical management of hunger strikes. *World Medical Journal*, *52*(2).

World Bank. (2022). *The Global Health Cost of PM2.5 Air Pollution: A Case for Action Beyond 2021*. The World Bank. https://doi.org/10.1596/978-1-4648-1816-5

Zehnder, C., Manoylov, K., Mutiti, S., Mutiti, C., VandeVoort, A., & Bennett, D. (2018). *Introduction to Environmental Science* (2nd ed.). Biological Sciences Open Textbooks. 4. . https://oer.galileo.usg.edu/biology-textbooks/4

Zhang, J., & Ding, W. (2017). Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. *International Journal of Environmental Research and Public Health*, *14*(2), 114. https://doi.org/10.3390/ijerph14020114

Zhao, J., Gao, Z., Tian, Z., Xie, Y., Xin, F., Jiang, R., Kan, H., & Song, W. (2013). The biological effects of individual-level PM(2.5) exposure on systemic immunity and inflammatory response in traffic policemen. *Occupational and Environmental Medicine*, *70*(6), 426–431. https://doi.org/10.1136/oemed-2012-100864

Zhao, X., Barber, S., Taylor, C. C., Nie, X., & Shen, W. (2022). Spatio-temporal forecasting using wavelet transform-based decision trees with application to air quality and covid-19 forecasting. *Journal of Applied Statistics*, 1–19. https://doi.org/10.1080/02664763.2022.2064976

Zivot, E., & Wang, J. (2006). *Modeling Financial Time Series with S-PLUS* (Second Edition).

List of Tables

List of Charts

List of Figures

Appendices

Appendix 1. Studies' Details

Study number	Period	window size	kkk	feature selector	Y
1	year	7	52	correlation	PM2_5
1	year	7	52	correlation	PM10
2	year	30	12	correlation	PM2_5
2	year	30	12	correlation	PM10
3	year	92	3	correlation	PM2_5
3	year	92	3	correlation	PM10
4	summer	7	13	correlation	PM2_5
4	summer	7	13	correlation	PM10
5	summer	30	3	correlation	PM2_5
5	summer	30	3	correlation	PM10
7	winter	7	13	correlation	PM2_5
7	winter	7	13	correlation	PM10
8	winter	30	3	correlation	PM2_5
8	winter	30	3	correlation	PM10
10	year	7	52	OLS	PM2_5
10	year	7	52	OLS	PM10
11	year	30	12	OLS	PM2_5
11	year	30	12	OLS	PM10
12	year	92	3	OLS	PM2_5
12	year	92	3	OLS	PM10
13	summer	7	13	OLS	PM2_5
13	summer	7	13	OLS	PM10
14	summer	30	3	OLS	PM2_5
14	summer	30	3	OLS	PM10
16	winter	7	13	OLS	PM2_5
16	winter	7	13	OLS	PM10
17	winter	30	3	OLS	PM2_5
17	winter	30	3	OLS	PM10
19	year	7	52	stepwise regression	PM2_5
19	year	7	52	stepwise regression	PM10
20	year	30	12	stepwise regression	PM2_5
20	year	30	12	stepwise regression	PM10
21	year	92	3	stepwise regression	PM2_5
21	year	92	3	stepwise regression	PM10
22	summer	7	13	stepwise regression	PM2_5
22	summer	7	13	stepwise regression	PM10
23	summer	30	3	stepwise regression	PM2_5
23	summer	30	3	stepwise regression	PM10
25	winter	7	13	stepwise regression	PM2_5

25	winter	7	13	stepwise regression	PM10
26	winter	30	3	stepwise regression	PM2_5
26	winter	30	3	stepwise regression	PM10
28	year	7	52	LASSO	PM2_5
28	year	7	52	LASSO	PM10
29	year	30	12	LASSO	PM2_5
29	year	30	12	LASSO	PM10
30	year	92	3	LASSO	PM2_5
30	year	92	3	LASSO	PM10
31	summer	7	13	LASSO	PM2_5
31	summer	7	13	LASSO	PM10
32	summer	30	3	LASSO	PM2_5
32	summer	30	3	LASSO	PM10
34	winter	7	13	LASSO	PM2_5
34	winter	7	13	LASSO	PM10
35	winter	30	3	LASSO	PM2_5
35	winter	30	3	LASSO	PM10

Appendix 2. Studies' Outcomes

Study number	MAE	MAE_norm	RMSE	RMSE_norm	R-squared	R-squared_norm	MAPE	MAPE_norm
1	10.95	0.83	19.07	1.1	-5.92	-2.87	123.39	37.73
1	13.07	0.81	22.23	1.07	-5.96	-3.11	105.58	32.08
2	5.42	0.39	6.68	0.49	0.03	-0.35	36.82	14.76
2	6.45	0.37	7.96	0.45	0.02	-0.2	35.97	13.05
3	5.34	0.42	6.95	0.48	0.3	0.2	26.06	14.48
3	6.27	0.36	7.75	0.44	0.32	0.32	29.1	11.56
4	3.95	0.57	5.03	0.73	-2.76	-3.92	74.76	31.73
4	4.47	0.5	5.8	0.67	-2.28	-3.39	62.99	24.34
5	2.07	0.23	2.51	0.31	-0.72	-0.27	46.11	13.91
5	2.57	0.24	2.94	0.3	-0.79	-0.33	42.33	12.41
7	7.92	0.88	10.08	1.26	-2.18	-0.86	68.55	37.25
7	9.35	0.74	11.78	1.06	-2.05	-0.59	67.03	28.03
8	7.34	0.34	8.17	0.35	0.1	-6.04	42.52	14.88
8	8.76	0.35	10.09	0.35	0.03	-2.39	38.14	13.99
10	13.7	1.09	17.18	1.42	-4.62	-5.44	143.43	46.23
10	15.96	1.05	20.06	1.34	-4.67	-5.45	130.21	40.36
11	4.83	0.28	6.08	0.34	0.2	0.34	31.62	10.14
11	5.6	0.28	6.96	0.33	0.25	0.38	31.62	9.57
12	6.85	0.4	8.82	0.53	-0.13	0.01	33.99	12.88
12	7.73	0.4	9.68	0.5	-0.06	0.1	35.01	12.75
13	6.81	0.85	8.54	1.07	-9.87	-9.6	131.62	45.8
13	7.67	0.79	9.68	0.97	-8.13	-8.12	104.99	36.84
14	1.2	0.16	1.59	0.2	0.31	0.47	15.93	7.83
14	1.2	0.14	1.22	0.14	0.69	0.71	16.06	6.3
16	13.38	2.01	18.93	2.46	-10.23	-6.07	124	103.93
16	15.06	1.85	21.17	2.28	-8.83	-6.29	116.12	80.89
17	7.05	0.81	7.17	0.84	0.31	-38.93	48.92	35.03
17	8.5	0.75	8.8	0.79	0.26	-16.68	44.4	29.52
19	25.31	2.71	45.35	7.4	-38.14	-174.18	257.77	115.53
19	29	2.58	52.81	6.87	-38.26	-169.62	222.19	100.08
20	7.1	0.46	8.64	0.51	-0.62	-0.49	49.88	17.41
20	8.65	0.44	10.3	0.5	-0.65	-0.44	50.72	15.7
21	6.67	0.42	8.32	0.51	-0.01	0.1	35.08	13.94
21	7.5	0.37	9.12	0.47	0.06	0.21	35.72	11.48
22	37.42	5.58	79.06	15.68	-930.38	-2254.73	536.63	248.61
22	50.13	6.04	108.13	17.54	-1138.1	-2992.76	529.92	240.29
23	2.77	0.34	3.47	0.43	-2.29	-1.38	38.64	15.88
23	2.23	0.23	3.1	0.32	-0.99	-0.55	22.91	9.56
25	32.53	3.83	48.95	6.41	-74.08	-46.98	270.88	164.21
25	36.22	3.8	53.44	6.55	-61.66	-59.32	252.08	141.9
26	7.83	0.48	7.86	0.55	0.17	-16.01	62.37	21.55
26	9.31	0.4	9.31	0.46	0.17	-4.95	54.94	16.76

28	10.23	0.84	13.49	1.11	-2.47	-2.96	133.51	38.41
28	12.22	0.81	15.83	1.07	-2.53	-3.14	116.48	32.8
29	3.99	0.23	5.09	0.3	0.44	0.49	30	8.45
29	4.47	0.22	5.78	0.28	0.48	0.54	28.85	7.69
30	7.23	0.43	8.46	0.5	-0.04	0.11	42.44	14.47
30	8.19	0.43	9.3	0.48	0.02	0.18	43.78	14.28
31	2.11	0.35	2.79	0.4	-0.16	-0.47	35.9	18.31
31	3.07	0.35	3.52	0.41	-0.21	-0.6	38.78	16.17
32	1.19	0.16	1.57	0.22	0.33	0.4	15.97	7.68
32	1.2	0.14	1.2	0.14	0.7	0.7	16.48	6.29
34	6.19	1.26	8.5	1.62	-1.26	-2.06	55.66	67.25
34	7.18	1.08	10.11	1.42	-1.24	-1.84	53.19	48.82
35	7.31	0.15	7.67	0.15	0.2	-0.27	47.03	6.55
35	8.82	0.16	9.48	0.18	0.14	0.11	42.8	6.87

Streszczenie

Tytuł: Metoda regresji wielowymiarowej do przewidywania zanieczyszczenia powietrza w danej lokalizacji

Obecny świat doświadcza rosnących trudności w zakresie zdrowia publicznego oraz problemów ekonomicznych. Niezmiernie istotny na to wpływ niesie zanieczyszczenie powietrza. Druzgocze fakt o niemalże siedmiu milionach przedwczesnych zgonów spowodowanych zanieczyszczeniem powietrza, w jednym tylko roku. Problem ten motywuje do rozwijania rzetelnych modeli prognozujących owe zanieczyszczenie. Praca ta eksploruje innowacyjny algorytm łączący cechy regresji wielowymiarowej, kroczącego okna oraz pseudoinwersji Moore'a-Penrose'a. Model przewiduje zanieczyszczanie powietrza dla Warszawy – miasta borykającego się z owym problem. Wynik badań stanowi obiecujący algorytm wykorzystujący trzydziestodniowe okno. Jest on szczególnie interesujący z wykorzystaniem metody LASSO do wyboru najefektywniejszych czynników przewidujących. Implementacja takiego modelu w programie komputerowym gwarantuje niezwykle dostępne narzędzie, z którego mogliby korzystać rodziny pragnące zaplanować swój czas na zewnątrz. Unikając bowiem dni o wysokim zanieczyszczeniu powietrza zmniejszamy prawdopodobieństwo negatywnych skutków zdrowotnych. Poprzez oferowane innowacyjnego narzędzia, który stanowi model, praca ta wnosi do literatury światowej oraz zwiększa dostęp do rzetelnych informacji o zanieczyszczeniu powietrza, będącym istotnym światowym problemem.