Journal of Economic Behavior and Organization 221 (2024) 544–569

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research Paper

How to better predict the effect of urban traffic and weather on air pollution? Norwegian evidence from machine learning approaches

Cong Cao [a,b,*]

[a] Center for Science, Society, and Public Policy, Division of the Humanities and Social Sciences, California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, USA
[b] Department of Economics, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway

ARTICLE INFO

JEL Classification:
C53
C52
Q5
R4

Keywords:
Machine Learning
Urban Traffic
Air Pollution
Transportation Policy

ABSTRACT

This paper uses machine learning approaches to predict the association between traffic volume, air pollution, and meteorological conditions. A key focus is on the interaction between these factors. The paper does this using hourly traffic volume, NO_x, $PM_{2.5}$, and weather data for Oslo, Norway. I considered a total of six datasets of the 2019 whole-year data to verify the prediction accuracy of the models. I find that the autoregressive integrated moving average model with exogenous input variables, and the autoregressive moving average dynamic linear model outperform the machine learning models in predicting air pollution. At the same time, I also explored the effect of sampling weather subsets on prediction accuracy. Finally, my study makes optimal policy recommendations for reducing air pollution from traffic volume, after considering the interaction and lagged effects of meteorology, time variables, traffic, and air pollution.

1. Introduction

Air pollution caused by traffic, and its resulting health effects, have become increasingly recognized as a source of public concern (Currie et al., 2005 & 2009; Pasquier and André, 2017; Kendrick et al., 2015). Poor urban air quality poses a significant risk to the environment and human health: It increases the incidence of respiratory diseases, especially among those living near major traffic routes and highways (Font & Fuller, 2016; Moretti et al., 2011; Bai et al., 2018). Across the globe, more than 5.5 million people die prematurely every year because of air pollution (Amos, 2016). In addition, traffic-related air pollution drains more public hospital care resource usage as well as personal health costs. It also influences people's behavior. As an example, the extreme air pollution experienced in Beijing has led to the demand for air filtration equipment, air freshening equipment, and regular precautionary hospital visits for respiratory and lung examinations, which increases the cost of personal medical care. At the same time, residents need to wear $PM_{2.5}$ disposable masks outside as a protective measure during winter in Beijing; here the $PM_{2.5}$ represents particulate matter with a diameter of less than, or equal to, 2.5 microns. Sustainable transport is one of the sustainable development goals of the UN 2030 Agenda (Kurz et al., 2014), and many policies have been suggested and implemented aimed at improving urban transportation and curbing air pollution (Parry et al., 2007). These include low-emission zones, restrictions on urban vehicle use, and congestion pricing (Bjørgen & Ryghaug, 2022; Green et al., 2016, 2020; Green & Krehic, 2022).

Effective policies to address these externalities rely on a clear understanding of the links between traffic volume and air pollution. One problem is that the mechanism between traffic volume and air pollution is complicated due to confounders such as meteorological

* Corresponding author.
E-mail address: ccao4541@gmail.com.

https://doi.org/10.1016/j.jebo.2024.03.018
Received 20 July 2023; Received in revised form 6 March 2024; Accepted 13 March 2024
Available online 10 April 2024

conditions. As an example, while some researchers have demonstrated that increased traffic volumes exacerbate airborne PM_{10}[1] and $PM_{2.5}$ concentrations (Srimuruganandam et al., 2010; Kendrick et al., 2015; Conte et al., 2018-), while others have found that $PM_{2.5}$ and PM_{10} concentrations are not the main indicators of traffic-related air pollution (Brugge et al., 2007), therefore, what the main pollutants brought by traffic are and the relationship between them is not clear (Gualtieri et al., 2015; Briggs et al., 1997; Luecken et al., 2006). This could, for instance, reflect the role of meteorological factors such as precipitation, air temperature, and humidity that affect the transformation of emissions into pollutants in the air. In this sense, the effect of a given level of emission on air quality can vary markedly under different meteorological conditions (Kamińska et al., 2018; Qu et al., 2019; Gryech et al., 2020). Along these lines, Wærsted et al. (2022) show that NOx concentrations from emissions are highly air temperature dependent. As highlighted by Aldrin et al. (2005), approaches to estimating the effect of, for instance, traffic volume on air pollution typically relies on regression models. Aldrin et al. (2005) provide a good estimate of the relationship between air pollution, traffic, and meteorological variables through a generalized additive model, but ignore their underlying interaction. This interplay complicates the relationship between the three and merits further investigation.

A second problem is that traditional approaches for predicting air pollutant concentration are prone to overfitting when faced with high-dimensional, small-sample data, and where there are nonlinear relationships. Machine learning approaches are known to be able to handle high-dimensional and nonlinear nonseparable problems. However, it's important to acknowledge that machine learning models also cannot avoid overfitting. Despite their capability to memorize noise within the training data when learning complex relationships in the data, this can lead to reduced generalization when use the trained algorithm to new datasets. Overfitting in machine learning might happen if the training data is insufficient, the methodologies used are incorrect, or the hyperparameters are not properly adjusted. To mitigate the risk of overfitting, especially considering the abundant data and high frequency of observations in this study, the dataset will be partitioned into distinct training and test sets. This approach will robustly evaluate the model's performance on unseen data, represented by the test set. This division can help with assessing the potential for overfitting. Machine learning approaches have the potential to address these problems.

This paper uses detailed Norwegian data to study the relationship between traffic volume, weather, and air pollution. Norway provides an advantageous focus due to the availability of high-quality, high-frequency data on air pollution and traffic volume. I estimate the relationship between traffic volume and air pollutant concentrations, where a key focus is allowing for complex meteorological influences. I use high-frequency hourly data, to examine air pollution due to traffic volume and meteorological factors. I exploit machine learning approaches, specifically Support Vector Machine (SVM), Random Forest (RF), Neural Networks (NN) and Decision Tree (DT) and examine whether they exhibit superior performance to traditional approaches, regarding air pollution prediction, the traditional approaches I use are Autoregressive Moving Averages with exogenous input variables (ARMAX) model and Autoregressive Moving Average dynamic linear (ARDL) model.

I apply existing machine learning models to analyze traffic and air pollution in Oslo, Norway. Oslo frequently exceeds the European Space Agency (ESA) standard for NOx concentration (Santos et al., 2020). I consider the interaction between weather, air pollution, and traffic variables, with a focus on the performance of machine learning approaches. An improved prediction has the potential to provide policymakers with superior information and, through this, an improved policy design aimed at mitigating air pollution damage. I provide the first evidence of this type from Norway but stress that the results have implications for other jurisdictions.

Apart from NOx and $PM_{2.5}$, automobile emissions also include contaminants like PM_{10}, O_3, or nitrogen dioxide (NO_2). The complete hourly O_3 data for 2019 is unavailable; NOx includes various compounds such as nitrous oxide, nitric oxide, NO_2, dinitrogen pentoxide, etc., among which only NO and NO_2 remain stable and are not easily decomposed in the air. As such, NOx is assessed independently. $PM_{2.5}$, also known as the particulate matter that can enter the lungs and pose health risks. Since $PM_{2.5}$ particle size is smaller than PM_{10}, and $PM_{2.5}$ is more likely to stay in the bronchus and lungs and causing heightened health hazards. Hence, my focus for evaluation is on $PM_{2.5}$. Additional details can be found in Appendix 3.

I predict NOx and $PM_{2.5}$ with traffic volume and weather as prediction factors. This paper makes two main contributions. First, the paper provides a comprehensive analysis of the association between traffic volume and air pollution, by considering the interaction and lagged effects of meteorological factors, time variables, traffic volume, and air pollution. Second, by re-sampling the whole year's data into five subsets based on air temperature and snowfall, as well as temporal variables. I explored the impact of seasonal and meteorological subset division methods on improving prediction accuracy. Exploring optimal prediction models and evaluating predictive factors that affect prediction accuracy is important. Together, I aim to provide cleaner estimates of the association between traffic volume and air pollution, which, as discussed above, is critical for the development of appropriate policy.

The rest of the structure in this paper is as follows. Section 2 presents the conceptual framework. Section 3 describes the methods and data used in the paper, and section 4 provides the results. Finally, the policy suggestions and conclusions of this work are provided in Section 5.

2. Conceptual framework and hypotheses on the effects of weather and traffic on air pollution

There exists a large literature on predicting traffic-related air pollution (Kleine Deters et al., 2017; Chen et al., 2021). The literature uses a range of approaches from traditional statistical methods to machine learning; it remains controversial whether the prediction performance of traditional approaches or machine learning models is better.

[1] PM_{10} refers to particulate matter with an aerodynamic equivalent diameter of less than, or equal to, 10 microns in ambient air.

Grange et al. (2018) uses random forests to predict PM_{10} trends in Switzerland through surface meteorology, time variables, synoptic scales, etc. They find that poor dispersion conditions caused by weather led to elevated PM_{10} concentrations. At the same time, they show that random forests are more effective than traditional standard statistical analysis methods due to lower model uncertainty since traditional statistical models need to meet strict assumptions, while this is not necessary with random forest approaches. However, other research has found that seasonal autoregressive composite moving average (SARIMA) models outperform neural networks in predicting traffic volumes on urban highways (Williams et al., 2003). Martín-Baos et al. (2022) proposed a management system for monitoring traffic flow and air quality index (AQI), thereby collecting diverse feature data and estimating AQI based on temperature, pressure, humidity, and PM levels. By comparing the AQI estimation accuracy of two traditional statistical models: linear regression (LR), Gaussian process regression (GPR) and a machine learning model: random forest (RF), the study revealed varying performance of the three models across different cities; Therefore, further research is needed to determine whether traditional models or machine learning algorithms are superior. Kimbrough et al. (2013) investigate the impact of pollutants generated by traffic on air quality. They found that traffic flow in Las Vegas has seasonal changes and is affected by wind direction, wind speed, and traffic volume simultaneously, which will lead to elevated air pollution concentrations along specific routes and emphasizing the significance of local meteorology in the context of road-related air pollution. However, the study employed traditional statistical methods for descriptive statistical analysis. Therefore, using more advanced machine learning algorithms to highlight the interactions among traffic, meteorology, and air pollution might reveal additional insights.

Support vector machine (SVM) approaches have been shown to exhibit superior performance when predicting air quality. Shaban et al. (2016) use SVM to predict the concentration of air pollutants and a neural network model to explore changes in air quality. They find that, when including meteorological factors as independent variables, SVM exhibits better performance than an artificial neural network in predicting air quality. This reflects the SVM's superior adaptability to high-dimensional data. Janarthanan et al. (2021) use the support vector regression and long short-term memory methods to examine the impact of occupational meteorological factors on air quality. They compared the predictive accuracy of these algorithms for Air Quality Index (AQI) prediction, demonstrating improved accuracy. The study suggested that deep learning algorithms could be applied to air quality control mechanisms for enhancing air quality. Moazami et al. (2016) use pollutant data including PM_{10}, NOx, and ozone from northern Tehran, and meteorological variables such as air pressure, air temperature, and relative humidity to predict carbon monoxide concentrations and find that SVM can reduce the uncertainty of the air quality prediction model, and its uncertainty is lower than that of neural networks (NN), and adaptive neuro-fuzzy inference systems. Ameer et al. (2019) compare the prediction performance of different existing machine-learning methods for air pollution. The models included DT, RF, multilayer perceptron, and gradient boosting, and MAE and RMSE are used as the prediction evaluation standards. They find that RF has the best prediction performance among the four algorithms in terms of predicting air pollution. Hence, I choose SVM, DT, RF, and NN as the choice of machine learning approaches in this study.

One model designed specifically for calculating emissions from road transport is the European emissions inventory model COPERT[2]. The main purpose of the model is to develop emissions inventories and to use them as an environmental policy assessment tool. In addition to considering traffic factors, inputs to the model include environmental conditions (such as temperature and humidity) and fuel type. The model's output covers information on various air pollutants and energy consumption. However, it is important to note that the model does not consider additional meteorological factors. Consequently, future air prediction models should comprehensively consider major meteorological factors and the interactive effects of both traffic and meteorology on air pollution.

There exists a small literature that focuses on traffic-related air pollution in Norway. Aldrin et al. (2005) analyze meteorological variables, traffic volume variables, and air pollutant concentrations in Oslo. By using generalized additive modeling, they find that traffic volume has a substantial impact on air pollution, especially for NOx, while meteorological variables also have an impact on air pollution. Mignone et al. (2022) conducted research and analysis on air pollution and traffic flow in Oslo, Norway, and proposed a model capable of anomaly detection in historical data. By adopting such a model, it helps to reduce the systematic errors introduced during the data collection process, thereby improving the quality of air pollutant and traffic flow data. However, these papers have not considered interactions between different predictors, for example, likely interactions between wind direction and wind speed, traffic flow and air pollution. Wærsted et al. (2022) seek to quantify the dependence of NOx emission on ambient temperature, using Norwegian road traffic as the emission source, and find changes in NOx concentrations across different air temperature ranges. These are then used to adjust expected air pollution levels from given levels of road traffic emissions; However, this paper does not consider+66 the relationship between other meteorological factors and NOx emissions. Yildirim et al. (2021) used machine learning algorithms to analyze the impact of temperature, air pressure and humidity on chronic lung infections in southern Norway, nevertheless, the interaction between air pollution and meteorological factors was not considered.

There is a range of challenges in accurately predicting air pollution (Aldrin et al, 2005). For example, even if traffic volumes are relatively stable over time, but meteorological factors are uncertain, then the overall prediction model has uncertainty. The question then is how the model prediction accuracy can be improved when faced with this uncertainty. When Santos et al. (2020) assess the impact of traffic control policies on Norway's air quality policy, they also propose that in Oslo, as a city with great seasonal and climate differences, the wind direction has a significant impact on air pollution concentrations. They suggest that adding meteorological variables when collecting data might improve the model. This, however, also complicates the model (Gauderman et al., 2007), which will introduce more challenges in providing accurate predictions.

[2] https://web.jrc.ec.europa.eu/policy-model-inventory/explore/models/model-copert/

The development of effective transport policies aimed at improving air quality remains challenging. Bigazzi and Rouleau (2017) study whether traffic management policies improve urban air quality and their effects on exposure and health. They note that there is currently a lack of ex-post evaluations of policies and therefore insufficient research on population exposure outcomes. More research is necessary to gain a deeper understanding of the impact of traffic management policies on air quality. Santos et al. (2020), using a traffic model, emissions model, and urban air quality diffusion model, discussed the policy and economic difficulties of traffic control policy in practice and concluded that the most effective permanent measures are to create low-emission zones and increase parking fees, and the most effective temporary traffic control measure is a ban on diesel vehicles. However, these policy proposals do not always appear to work. For instance, Wærsted et al. (2022), in a study on the impact of Norway's speed limit policy on local air pollution, conclude that lower vehicle speeds did not reduce the concentration of NOx and particulate matter.

2.1. Research questions and hypotheses

Therefore, I propose two research questions: (1) Under the interaction of traffic, weather, and air pollution, what is the impact of traffic and weather on air pollution? (2) Which approach can better predict traffic-related air pollution, machine learning or traditional statistical approaches? Fig. 1 describes the analysis process for the first research question.

The hypothesis is (1) The interaction terms of weather and traffic have different effects on air pollution; (2) The second hypothesis is that the predictive power of machine learning is superior to a traditional statistical method.

I choose urban traffic because urban cities are expected to generate more traffic volumes than rural areas, and thus potentially contribute to more air pollution. From Fig. 1, the meteorological variables I include are air temperature, air pressure, wind direction, mean wind speed, relative air humidity, and snow depth. The air pollutants I choose to study include $PM_{2.5}$ and NO_x . The lines and arrows in the figure represent the interaction between them. I focus on the interaction between traffic volume, air pollution, meteorological factors, and personal behavior. Finally, I hope to provide corresponding traffic and air quality policies, as well as personal behavior travel model suggestions.

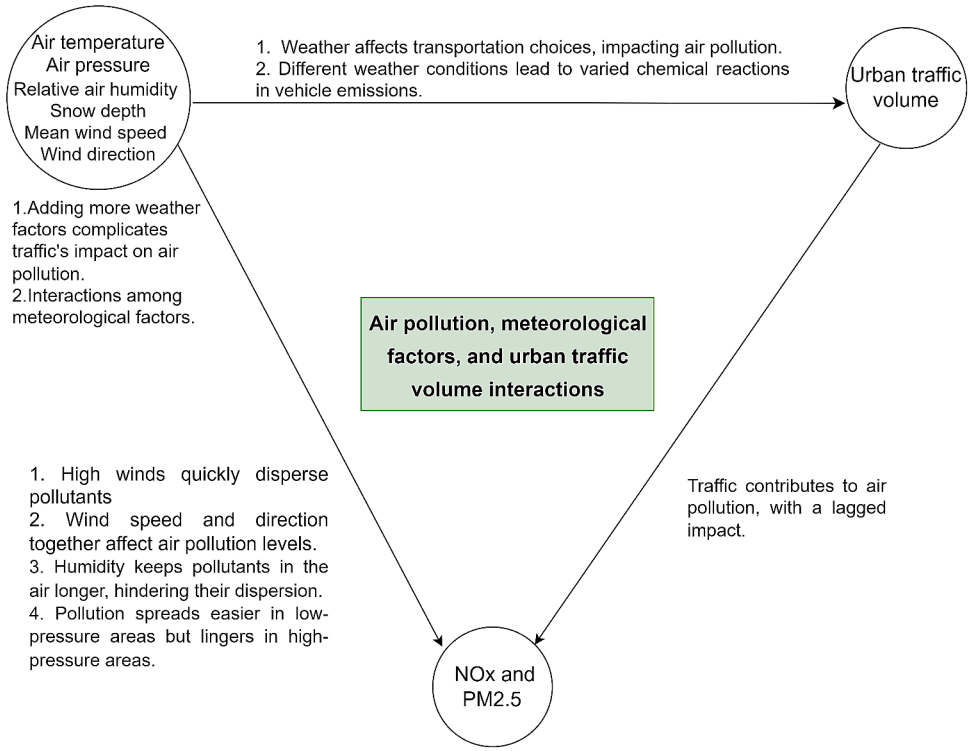


Fig. 1. Figure of the First Research Question.

3. Methods and data

This paper aims to examine the relative performance of machine learning and traditional time series approaches in predicting traffic-related air pollution. It uses hourly traffic volumes, pollutant concentration, and meteorological factors as inputs, and focuses on pollutants concentrations as the main output. The focus is on estimating the effect of traffic on air pollution at an hourly level. Here is my empirical approach:

$$P_t = f\left(T_t, M_{t,j}\right) \tag{1}$$

Here P_t is a pollutant, NOx or $PM_{2.5}$, T_t is traffic volumes. $M_{t,j}$ are meteorological variables, subscript j $\epsilon\{1, 2\ldots J\}$ is the metrological variables number. It includes (1) air temperature, (2) wind direction, (3) mean wind speed, (4) snow depth, (5) relative air humidity, (6) air pressure. Subscript t ϵ {1, 2..., T}, represents the time where the unit is an hour. Appendix 1 gives the interpretation and measurement of these variables.

The autoregressive integrated moving average model (ARIMA) represents a standard approach to time series data prediction. ARIMA models are denoted by ARIMA (p, d, q), where p represents the number of lags or the autoregressive (AR) term; d represents the degree of difference to obtain stationarity; and q represents the number of lags of the prediction error, is also called the moving average (MA) term. To answer the first question of my study, that is, on the analysis of the relationships between air pollution (NOx and $PM_{2.5}$), traffic volume, and meteorological conditions, I add the traffic volume and all the meteorological variables to the ARIMA model. Appendix 4 shows that the time series of NOx is stationary, and since I study many independent variables. I use an autoregressive integrated moving average model with an exogenous input variables (ARMAX) model instead. I also adopt the Autoregressive Dynamic Linear (ARDL) approach, together to answer question one in this study.

I estimate the models as follows[3]:

ARMAX model:

$$P_t = f\left(T_t, M_{t,j},\ \text{lag}\right) \tag{2}$$

ARDL model:

$$P_t = f\left(T_t, M_{t,j},\ \text{Interaction},\ \text{lag}\right) \tag{3}$$

$$\text{Lag} = \sum_{i=0}^{k} \psi_i T_{t-i} + \sum_{i=0}^{K} \xi_i M_{j,t-i} \tag{4}$$

$$\text{Interaction} = \sum_{j=1}^{J} \delta_t M_{t,j} * T_t + \sum_{j=1}^{J} \theta_t M_{t,j} * M_{t,j+1} \tag{5}$$

T_{t-i} represent *i* hours lagged traffic volumes, $M_{j,t-i}$ are *i* hours lagged meteorological factors, *i* is the lag number, from 1, 2…I. The explanation of the equation term is in Table 1.

There exist several complications to estimating the model. First, because the inclusion of more weather variables complicates the impact of traffic on air pollution, there will be higher demands on the model when estimating air pollution. Second, vehicle emissions undergo chemical reactions in the air. This, in part, is affected by weather insofar as under different meteorological conditions, vehicle emissions have different chemical reactions. This makes the link between emissions and air quality less clear. For example, if the wind speed is high, the dilution and diffusion of pollutant is fast, and concentration changes quickly. In practice, these effects can be complex and interactive. For instance, the synergistic effect of wind speed and wind direction also affects the degree of air pollution. As another example, air humidity can prolong the residence time of pollutants suspended in the air, which is not easy for the diffusion and dilution of pollutants. In terms of air pressure, air pollution diffuses more easily in low-pressure areas, while it is less likely to disperse in high-pressure areas. Third, the weather can have a direct impact on an individual's transportation decisions, which in turn affects air pollution levels.

I choose rush hour as a subset. Because I expect rush hours to have higher traffic volumes, and thus possibly more air pollutants relative to the whole dataset. As it is during this period that the relationship between traffic and air pollution is likely to be most acute. The rush hour is from 7:00 to 9:00 and from 13:00 to 16:00. The correlation analysis results are presented in Table 2.

I find that air temperature is positively correlated with traffic volume, which tends to be lower on days of low air temperatures, as well as more relative air humidity. Traffic volume is positively correlated with NOx concentration. In regard to the correlations between weather variables, a negative correlation is shown between wind speed and air pressure, along with a negative correlation between air temperature, snow depth, and humidity, a positive correlation between air temperature and wind direction, and a negative correlation between wind speed and air pressure and humidity. The correlation between traffic and weather variables, and the correlation between weather, complicates estimating the impact of traffic volume on air pollution. All these correlation coefficients are

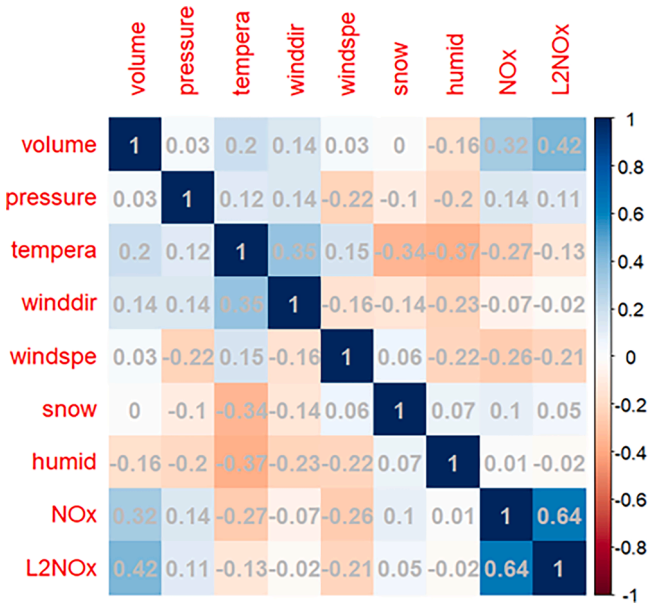
[3] For both models I include a dummy variable, which is the holiday when the traffic volumes are expected to be low. Since I use data from the year 2019, these holidays include January 1 as the new year, April 14–22 as the Easter holiday, May 1 as Labor Day, and May 17 as the Constitution Day of Norway. Additionally, May 30, June 10, and December 25 – 26 are holidays in Norway.

Table 1
Equation (5)'s Equation Term Explanation.

Equation term	Explanation
$\mathbf{M_{t,j}}^{*}\mathbf{T_t}$	The interactions between meteorological factors $M_{t,j}$ and traffic volumes T_t
$\mathbf{M_{t,j}}^{*}\mathbf{M_{t,j+1}}$	The interactions between two meteorological factors $M_{t,j}$ and $M_{t,j+1}$

Note: Equation (5) systematically traverses all interactions between variables

Table 2
The Correlation Coefficients between Traffic Volume, Meteorological Factors, and Air Pollution during Rush Hours. The Rush Hours I Select Are from 7:00 to 9:00 and from 13:00 to 16:00.



Note: volume = traffic volume, pressure = air pressure, tempera = air temperature, winddir = wind speed, windspe = mean wind speed, snow = snow depth, humid = relative air humidity, L2NOx = lag of 2 hours NOx. The first column represents the correlation coefficient between traffic volume and other variables. For example, the second row of the first column indicates that the correlation coefficient between traffic volume and air pressure is 0.03.

small. The datasets collected are linearly inseparable eigenspace and complex, which means only a few feature variables can represent most of the information, and other features are considered noise. As a result, when training a model, the model can be prone to overfitting.

The above can be summarized in terms of two problems. The first is multicollinearity. There are correlations between the variables, for example, a higher wind speed can lead to a lower air pollution level. Thus, a variable can be explained by a linear combination of other independent variables. Linear regression and machine learning have different approaches to addressing multicollinearity. To select the key independent variable, the standard method of linear regression is to increase the sample size, variable elimination, or stepwise regression. Machine learning approaches use principal component analysis methods to select principal components, or models with regularization terms, which makes it easy to shrink or delete collinear elements. The prediction accuracy of the final linear model or machine learning model is evaluated by using different model evaluation methods.

The second difficulty is that due to there being many variables or features, there are high-dimensional eigenspace and the problem of non-linear inseparability. Machine learning has the potential to address these problems.

The following approaches are adopted to predict air pollution: ARDL, ARMAX; and two machine learning algorithms: support vector machine (SVM) and decision tree (DT).

3.1. Methods

3.1.1. Time series approaches: ARMAX

ARIMA works by using a model to describe a time series and then identifying the model to derive prediction values from past and present values of the time series. In my setting, I seek to capture traffic and weather effects which contain many independent variables, so I have chosen the multivariate time series method, which is the ARMAX model. The difference between ARIMA and ARMAX is that

ARIMA only contains one single explanatory variable, while ARMAX could use many explanatory variables. The details of ARMAX can be found in Appendix 4, Table A and B.

3.1.2. ARDL approaches

The ARDL model adopts autoregression, which is the AR part, that is, in the model, it uses the past value of the dependent variable as the lagged variable, and combines other independent variables as the input variables, to estimate the current value of the dependent variable. Thus, the dependent variable depends on its lag value and other independent variables. ARDL models can be used for the analysis of multivariate time series.

3.1.3. Machine learning (ML) approaches

While ARMAX can provide predictions from past values of a time series, it requires the data to be stationary, otherwise, the data needs to be differentiated until the time series is stationary before modeling. At the same time, ARIMA cannot the patterns of nonlinear relationships (Zhang, 2003). When there is a large amount of training data, ARMAX displays poor performance and is prone to over-fitting. Machine learning approaches have advantages in solving big data and nonlinear problems. I use four specific approaches, SVM, NN, RF and DT. In theory, there are other alternative machine learning approaches, such as long short-term memory algorithms, etc., which are also worthy of further exploration in the future.

The reasons for choosing these are, first, they are the most widely used machine learning algorithms, as I stated in the literature review in Section 2, which reflects their advantages in terms of efficiency and prediction accuracy; second, the traditional regression model requires the value of the loss function to be 0, which means the predicted value and the real value have to be the same, while the ML allows an error between the predicted value and the true value. That is, only when the distance interval between the true value and the predicted value is large enough, will it be considered a loss. Therefore, this relaxes the restrictions of many traditional models. A brief introduction to these machine-learning methods can be found in Appendix 4. A comprehensive introduction to random forests and neural networks (Rigatti et al., 2017; Islam et al., 2019).

My main approaches to model evaluation are Mean Absolute Error (MAE) and Mean Squared Error (MSE). I calculated MAE and MSE by comparing the predicted values obtained by using the models with the actual pollutant concentrations values in the test dataset. MAE and MSE have been the commonly used model evaluation indicators, both of which are suitable for comparing relative errors. First, when using MSE to calculate the loss model, it is calculated in the direction of reducing the error of the outlier; the outlier here represents, in the data, one or several values that differ greatly from other values. Thereby, the outliers sacrifice the error of the remaining samples, reducing the overall performance of the model. Therefore, MSE is suitable for models that need to detect outliers, and outliers are important information for the model, while MAE is suitable for models that need to remove outliers. This paper estimates the relationship between traffic volume and air pollution, and the data used are of high quality as there are few outliers, so I pay more attention to the results of MAE since, in this situation, MAE has a better absolute performance evaluation than MSE. Adjusted R-squared (R^2) are the fundamental standard of model evaluation indicators. I also add the Root Mean Square Error (RMSE) method for more comprehensive model evaluation information.

3.2. Data

Three sources of data are used: traffic volume data, air pollution data, and meteorological data. They are obtained from the Norwegian Public Road Administration (SVV), the Norwegian Institute for Air Research (NILU), and the Norwegian Meteorological Institute (MET), respectively. These three administrative institutions are responsible for the monitoring stations from which the data were collected. The time interval used is the 2019 calendar year, Oslo, hourly data.

I focus on the capital of Norway, Oslo, which generally experiences a humid continental climate. Air temperatures vary widely throughout the year. Summers are warm with convective rain, while winters are cold and severe with little rain and low humidity. Oslo is Norway's largest and most populous city, has high economic growth, and is the country's industrial and shipping hub.

The three sets of data are merged into one dataset with 8760 observations and 13 variables, which represents the whole-year hourly data for 2019. Appendix 1 provides a list of all collected variables included in the data set. Because of measurement errors, the data collected by air pollutant monitoring stations sometimes have some changes around zero, and even small negative values, ranging from 0 to -5, are taken as effective values. Values of -9900 are considered missing values. Meanwhile, a value with a traffic volume of 0 is considered a missing value, because the traffic monitoring station is on a busy road section and usually has vehicles passing by.

Data preprocessing includes missing values or outliers, which are mainly caused by the failure of the equipment due to changes in the external environment. If a small number of outliers occurs in a short time, they can be directly excluded, but if a large percentage of data is missing, it needs to be imputed. The data only have a small number of outliers here, so the outliers are removed. I find that the percentage of missing values for the whole year dataset is 4.24 %. Since SVM is sensitive to missing values, so I performed missing value imputation at the very beginning.

3.2.1. Traffic data

The traffic volume data from SVV is measured as the number of approved vehicle registrations during the relevant hour. Fig. 2 exhibits Statens vegvesen's traffic registration maps. The monitoring stations have different geographic locations. Therefore, the monitoring stations need to contain both traffic and air pollution data, and considering this, I choose the closest station Oslo-Blindern to obtain the meteorological data. In theory, meteorological data from a monitoring station can be estimated more precisely by employing the interpolation method between two weather stations. However, this method has its limitations, assuming a continuous

and uniform variation of meteorological fields. Boke et al. (2017) have compared the limitations of various spatial interpolation methods. Consequently, I opted for the Oslo-Blindern station, which already serves as the nearest station to Oslo-Manglerud station, providing the required complete meteorological data and variables.

Fig. 2 shows the traffic registration map from Statens vegvesen, the traffic and air pollution monitoring station used is Oslo-Manglerud. The triangle represents the geographic location of Oslo-Blindern and the circular icon of Oslo-Manglerud. The distance between the two stations is between 5 and 10 km.

Fig. 3 shows the daily variation in traffic volumes. These increase from 6:00, with the first peak at 7:00. The traffic volumes also

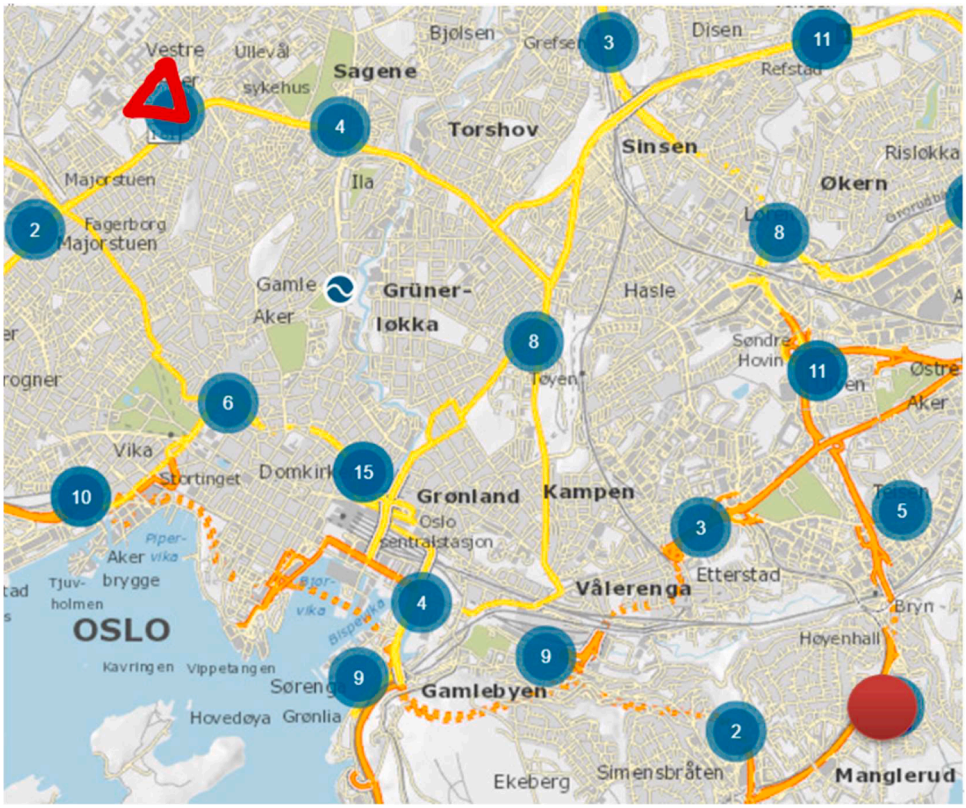


Fig. 2. Traffic registration map with data monitoring stations.

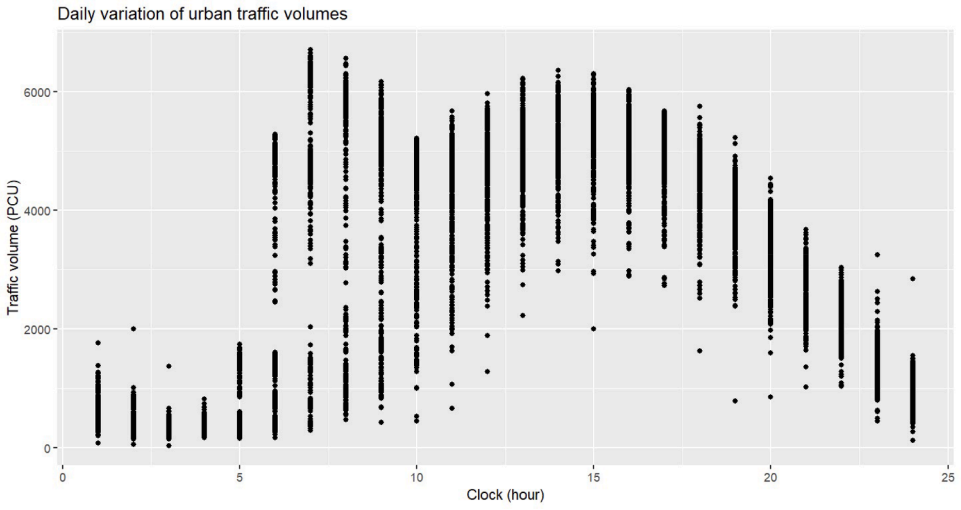


Fig. 3. Daily Variation in Urban Traffic Volumes, Oslo, 2019. Note: The x-axis represents 24 hours a day, and the y-axis shows the traffic volumes. Each dot represents the traffic volumes passing through a monitoring station each hour.

increase from 10:00, and the second peak is at 14:00. The rush hour is from 7:00 to 9:00, and from 13:00 to 16:00.

3.2.2. Meteorological data

The meteorological variables include Air pressure (qnh), Air temperature (Celsius), Wind direction (degrees), Mean wind speed (m /s), Snow depth (cm), and Relative air humidity (%). I add a column of variables to convert the wind direction from degrees to four angles, with 90° separations, i.e., north (N), south (S), west (W), and east (E). Precipitation is thought to be important, but data are not available. Some monitoring stations have precipitation data for certain days, others have data for other days, and no monitoring station has complete precipitation data for 2019. Ideally, could capture data for very short periods as a subset, so that precipitation data could be included for future exploration. Fig. 4 and Appendix 2 show the monthly variation of the metrological factors and traffic volumes. Fig. 4 Panel A shows that in Oslo, from January to March 2019 the daily snow depth is the deepest, and there is almost no snow from April to October. The snow depth in November and December is close to 10 cm, which is the less snow depth.

From Fig. 4 Panel B, the daily air temperature in Oslo is above 20 degrees Celsius from April to September, which I define as warm months here. Other months with temperatures below 20 degrees Celsius are defined as cold months. Fig. 4 Panel A shows that in Oslo, from January to March 2019, the daily snow depth is the deepest, and there is almost no snow from April to October. The snow depth in November and December is close to 10 cm, with the least amount of snow.

As shown in Panel C in Fig. 4, there is not much seasonality in the daily wind speed in the Oslo area, except for January. The other meteorological variables' seasonal variation throughout the year is depicted in Appendix 2, and I find that the other meteorological variables and traffic volume do not reflect seasonal differences.

3.2.3. Air pollution data

The air pollution data were obtained from automatic air pollution monitoring stations from the Norwegian Institute for Air Research (NILU). These monitoring stations are located near roads, and they are set up in cooperation between the Norwegian Public Road Administration (SVV), and NILU to measure traffic-related air pollution. These monitoring stations collect data every hour. All air pollution data are automatically manually calibrated, which means more accurate measurements are obtained by correcting for measurement errors and manually calibrating air pollution levels (Folgerø et al., 2020). Similarly, NILU contains many pollutants, such as PM_{10}, $PM_{2.5}$, O_3, and NO_2, etc. To prepare for subsequent modeling, it is necessary to select suitable input variables and reduce concerns about the existence of multicollinearity of independent variables. By studying the sources of different pollutants and their reaction mechanisms in the air (see Appendix 3 for more detail), $PM_{2.5}$ and NOx was selected as the target pollutant variable to continue exploration.

Appendix 2 depicts the monthly and daily variation of air pollution. NOx appears to be seasonal, which may relate to meteorological factors. This further provides a basis for my exploration of the impact of meteorological factors on air pollution. Appendix 1 also provides a summary of statistics of the raw data.

3.2.4. Generation of the datasets

Based on the snow depth shown in Fig. 4 Panel A, I re-sample the data into three subsets: more snowfall, less snowfall, and no snowfall. In accordance with Fig. 4 Panel B on air temperature, I split the data into two subsets: warm and cold months. Thus, five subsets are created to validate the performance of these models. The five subsets are re-sampled according to meteorological and temporal variables in Norway, and together with the 2019 whole-year dataset, I have a total of six datasets (see Appendix 1). I will use the six datasets to compare the predictive accuracy of the traditional statistical and machine learning approaches.

I re-sampled each data set, encompassing both subsets and the entire dataset, into two parts, 75 % of which was served as a training data set, while the remaining 25 % as a separate test data set. The training data set was used to train the machine learning model, followed by evaluation and prediction testing on the separate test dataset. This test dataset essentially served as an unseen data set to assess the model's predictive capabilities in the context of new test data. Initially, all data are standardized using max-min normalization, $x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$, which converts the original data to the range [0, 1].

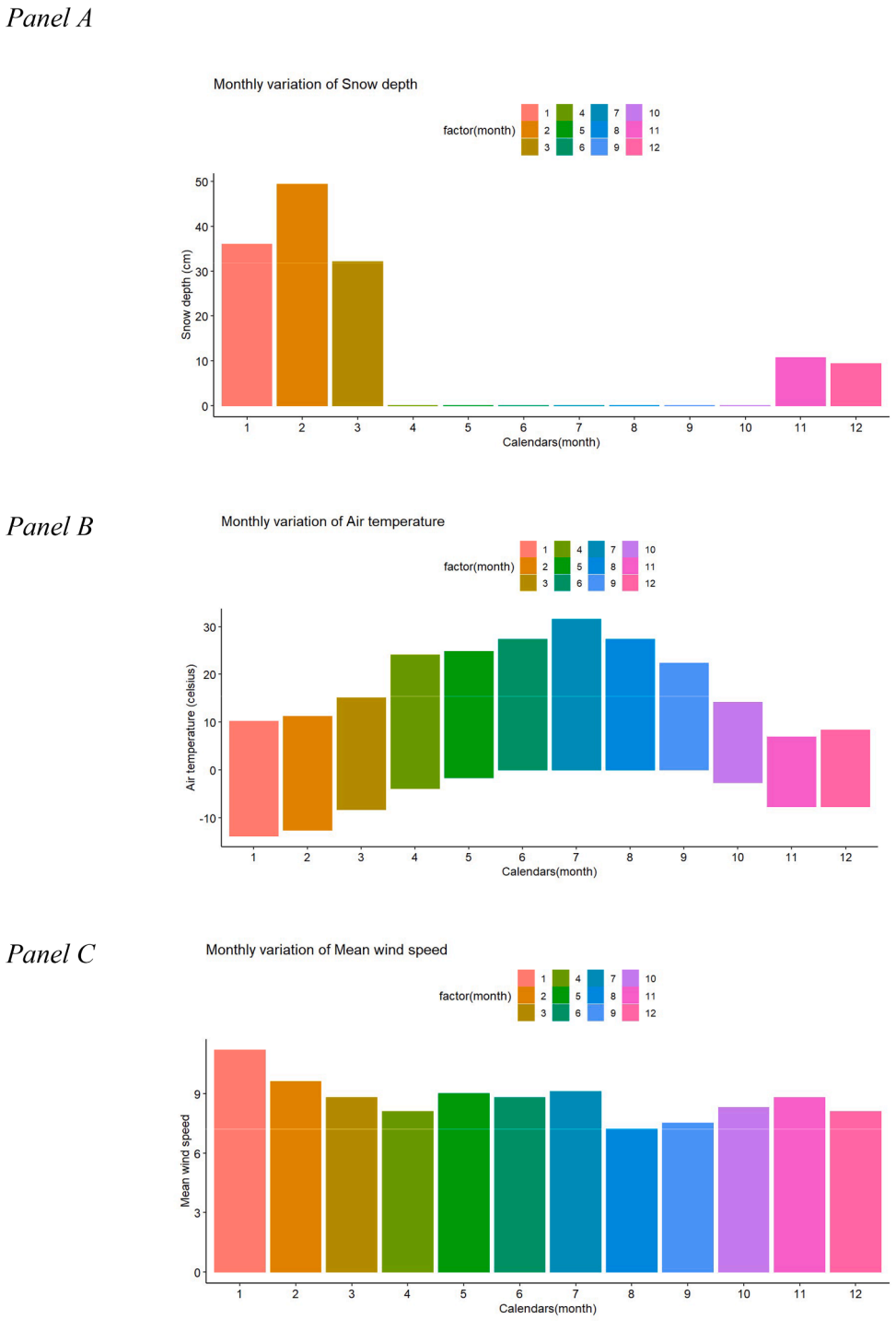


Fig. 4. The Meteorological Variables with Seasonality. Note: These demonstrate that the snow depth and air temperature have seasonality, while mean wind speed doesn't have seasonal differences.

4. Results

4.1. Results of the ARDL and ARMAX model

My initial step is to estimate an ARDL model and ARMAX model, in an attempt to explore the effect of traffic and weather on air pollution. This is estimated on hourly data for the whole year of 2019. Estimates are provided in Tables 3 and 4.

Table 3
Determinants of Air Pollution from ARDL model. This table has two pages.

Observations **Adjusted R^2**	**Variables** **8760** **0.74 (NOx as the dependent variable)** **0.81 ($PM_{2.5}$ as the dependent variable)**	**NOx**	**$PM_{2.5}$**
AR part	**Lag of 1 hour NOx/$PM_{2.5}$**	**0.7620*** **	**0.8190*** **
		(0.0108)	**(0.0109)**
	Lag of 2 hours NOx/$PM_{2.5}$	-0.0211	**0.0638*** **
		(0.0136)	**(0.0140)**
Single factors	Air pressure	0.2830*	0.0204
		(0.150)	(0.0568)
	Air temperature	**-0.1380*** **	**-0.0635*** **
		(0.0443)	**(0.0167)**
	Wind direction	0.0150	-0.000734
		(0.0158)	(0.0060)
	Mean wind speed	-0.0419	-0.0106
		(0.0303)	(0.0115)
	Snow depth	0.0029	0.0099
		(0.0215)	(0.0081)
	Relative air humidity	0.0217	0.0167**
		(0.0181)	(0.0069)
	Traffic Volume	**0.1280*** **	**0.0381*** **
		(0.0179)	**(0.0067)**
The lagged effect of single factors	Lag of 1 hour Air pressure	-0.4120	-0.0334
		(0.2650)	(0.1000)
	Lag of 2 hours Air pressure	0.0651	0.0545
		(0.2650)	(0.1000)
	Lag of 1 hour Wind direction	-0.0004	0.0019*
		(0.0030)	(0.0011)
	Lag of 2 hours Wind direction	-0.00301	-0.0008
		(0.0029)	(0.0011)
	Lag of 1 hour Air temperature	0.0240	0.0183
		(0.0603)	(0.0228)
	Lag of 2 hours Air temperature	0.0049	0.0210
		(0.0603)	(0.0228)
	Lag of 1 hour Relative air humidity	-0.0268	-0.0030
		(0.0163)	(0.0061)
	Lag of 2 hours Relative air humidity	**0.0348****	-0.0045
		(0.0163)	(0.0061)
	Lag of 1 hour Mean wind speed	0.0046	0.0021
		(0.0089)	(0.0034)
	Lag of 2 hours Mean wind speed	-0.0021	-0.0014
		(0.0089)	(0.0034)
	Lag of 1 hour Snow depth	-0.0027	-0.0030**
		(0.0038)	(0.0015)
	Lag of 2 hours Snow depth	-0.0020	0.0007
		(0.0038)	(0.0015)
Interaction between weather factors and traffic volume	**Air pressure * Traffic volume**	**-0.0058*** **	0.0010
		(0.0135)	(0.0051)
	Air temperature * Traffic volume	**-0.0721*** **	**-0.0335*** **
		(0.0167)	**(0.0063)**
	Wind direction * Traffic volume	0.0149*	0.0064**
		(0.0077)	(0.0029)
	Mean wind speed * Traffic volume	**-0.0394*** **	**-0.0162*** **
		(0.0153)	**(0.0058)**
	Snow depth * Traffic volume	0.0094	0.0033
		(0.0106)	(0.0040)
	Relative air humidity * Traffic volume	-0.0084	-0.0045
		(0.0105)	(0.0040)
Interaction between weather factors	Wind direction * Air pressure	-0.0187	-0.0005
		(0.0137)	(0.0052)
	Air pressure * Snow depth	0.0261	0.0041
		(0.0189)	(0.0071)
	Air pressure * Air temperature	0.0839**	0.0296**
		(0.0333)	(0.0126)
	Air temperature * Wind direction	0.0200	-0.0098
		(0.0162)	(0.0061)
	Mean wind speed * Air pressure	**-0.0929*** **	**-0.0362*** **
		(0.0276)	**(0.0104)**

(*continued on next page*)

Table 3 (*continued*)

Observations **Adjusted R^2**	**Variables** **8760** **0.74 (NOx as the dependent variable)** **0.81 ($PM_{2.5}$ as the dependent variable)**	**NOx**	**$PM_{2.5}$**
	Air pressure * Relative air humidity	0.0104	-0.0021
		(0.0177)	(0.0067)
	Mean wind speed * Air temperature	**0.201*** **	**0.0840*** **
		(0.0334)	**(0.0127)**
	Snow depth * Air temperature	**-0.0367****	**-0.0400*** **
		(0.0184)	**(0.0071)**
	Wind direction * Mean wind speed	-0.0191	-0.0028
		(0.0153)	(0.0058)
	Snow depth * Wind direction	-0.0062	0.0009
		(0.0111)	(0.0042)
	Wind direction * Relative air humidity	**-0.0241****	-0.0005
		(0.0102)	(0.0038)
	Mean wind speed * Snow depth	0.0110	0.0146*
		(0.0215)	(0.0081)
	Mean wind speed * Relative air humidity	-0.0182	-0.0148*
		(0.0206)	(0.0078)
	Snow depth * Relative air humidity	-0.0085	-0.00425
		(0.0137)	(0.0052)
	Holiday	**-0.0103*** **	0.0003
		(0.0033)	(0.0013)
	Constant	0.0590***	0.0152**
		(0.0194)	(0.0074)

Notes: This table contains the statistical results of time series analysis with NOx and $PM_{2.5}$ as dependent variables, and weather and traffic volume as independent variables, as well as their interaction terms and lagged effects, with the ARDL model. The period is the whole year of 2019. *, **, and *** indicate statistical significance at the $P < 0.05$, $P < 0.01$, and $P < 0.001$ levels, respectively; the Mean wind speed * Air temperature represents the interaction of Mean wind speed and Air temperature.

Table 4
Determinants of Air Pollution, from ARMAX model.

Observations **Adjusted R^2**	**Variables** **8760** **0.73 (NOx as the dependent variable)** **0.78 ($PM_{2.5}$ as the dependent variable)**	**NOx**	**$PM_{2.5}$**
AR part	Lag of 1 hour NOx/$PM_{2.5}$	1.550***	1.070***
		(0.040)	(0.020)
	Lag of 2 hours NOx/$PM_{2.5}$	-0.590***	-0.160***
		(0.030)	(0.020)
	L2.ar	-0.030***	0.020***
	L.ma	(0.007)	(0.008)
		-0.770***	-0.230***
		(0.040)	(0.020)
Single factors	Air pressure	0.280**	0.040
		(0.120)	(0.050)
	Air temperature	-0.110***	-0.050***
		(0.03)	(0.010)
	Wind direction	-0.003	**-0.003*** **
		(0.002)	**(0.0009)**
	Mean wind speed	**-0.030*** **	-0.006**
		(0.008)	(0.003)
	Snow depth	0.002	-0.002
		(0.004)	(0.002)
	Relative air humidity	0.010	**0.010*** **
		(0.010)	**(0.004)**
	Traffic Volume	0.070***	0.020***
		(0.005)	(0.002)
The lagged effect of single factors	Lag of 1 hour Air pressure	-0.570**	-0.090
		(0.230)	(0.100)
	Lag of 2 hours Air pressure	0.290**	0.050
		(0.120)	(0.050)

(*continued on next page*)

Table 4 (*continued*)

Observations **Adjusted R^2**	**Variables** **8760** **0.73 (NOx as the dependent variable)** **0.78 ($PM_{2.5}$ as the dependent variable)**	**NOx**	**$PM_{2.5}$**
	Lag of 1 hour Wind direction	0.002	0.003
		(0.003)	(0.001)
	Lag of 2 hours Wind direction	-0.001	-0.001
		(0.002)	(0.001)
	Lag of 1 hour Air temperature	0.150**	0.040**
		(0.007)	(0.020)
	Lag of 2 hours Air temperature	-0.050	0.008
		(0.040)	(0.010)
	Lag of 1 hour Relative air humidity	-0.030	-0.005
		(0.020)	(0.005)
	Lag of 2 hours Relative air humidity	-0.005	-0.007*
		(0.008)	(0.004)
	Lag of 1 hour Mean wind speed	-0.003	0.004
		(0.006)	(0.004)
	Lag of 2 hours Mean wind speed	0.0009	-0.003
		(0.004)	(0.003)
	Lag of 1 hour Snow depth	-0.003	-0.002
		(0.006)	(0.002)
	Lag of 2 hours Snow depth	0.0009	0.002
		(0.004)	(0.002)
	Holiday	0.0007	0.0007
		(0.001)	(0.001)
	Constant	0.010***	0.060***
		(0.003)	(0.0002)

Notes: This table contains the statistical results of time series analysis with NOx and $PM_{2.5}$ as dependent variables, and weather and traffic volume as independent variables, as well as their lagged effects, with the ARMAX model. The period is the whole year of 2019. *, **, and *** indicate statistical significance at the $P < 0.05$, $P < 0.01$, and $P < 0.001$ levels, respectively; the Mean wind speed * Air temperature represents the interaction of Mean wind speed and Air temperature.

I focus primarily on estimating statistically significant levels at *** $p < 0.01$. Table 3 demonstrates a range of patterns that are consequential for understanding both the links between traffic volume and air pollution and how this is influenced by weather conditions. For instance, while traffic volume has a direct statistically significant impact on both pollutants, including, for example, in the AR part of the model, I see that for $PM_{2.5}$, the regression estimates the value of lag of 1 hour, and lag of 2 hours gradually drops; for NOx, there is also an overall downward trend in the value, so traffic volume leads to pollutant concentration up to two hours later after heavy traffic.

Regarding the single factors, I find that air temperature alone has a direct statistically significant effect on both NOx and $PM_{2.5}$, and it shows a statistically negative significant effect, which means that the concentration of these two pollutants decreases when the air temperature rises. Meanwhile, traffic volume has a direct statistically significant impact on both pollutants. More traffic volume leads to a higher concentration of these two pollutants.

Considering the lagged effects of the single factors, there is a statistically significant effect on NOx from relative air humidity two hours earlier, meaning that NOx concentrations increase when the relative air humidity increases.

Regarding the interactions between meteorological variables and traffic volume. I find a statistically significant interaction effect between air pressure and traffic volume for NOx, not for $PM_{2.5}$. As well as a statistically significant interaction effect between mean wind speed and traffic volume, between air temperature and traffic volume, on both pollutants. In addition, all of them are negative effects.

Interactions between meteorological variables also resulted in statistically significant effects on both pollutants. Except that the interaction of mean wind speed and air temperature will increase air pollution, all other interaction terms reduce air pollution. For example, the interaction of mean wind speed and air pressure, and the interaction of snow depth and air temperature. This further emphasizes the moderating role of weather factors in the impact of traffic volume on air pollution. The interaction of wind direction and relative air humidity will also decrease NOx concentration.

I use the ARMAX model to explore more. The results are in Table 4. In this model, I only include a single variable and its lagged effects. I focus on estimating statistically significant levels at *** $p < 0.01$.

I find that both the wind direction, as well as relative air humidity, have statistically significant effects on $PM_{2.5}$. When the wind blows from north to south or when the relative humidity is lower, the $PM_{2.5}$ concentration decrease; meanwhile, when the mean wind speed increase, the NOx concentration decrease.

Taken together, for $PM_{2.5}$ and NOx, I find that on colder days, traffic volume increase, and relative air humidity increase, increasing concentrations of both pollutants. All traffic volume and meteorological variables interactions reduce air pollution, this, in addition to showing the moderating effect of weather on air pollution when there is traffic volume, also emphasizes the role of the interaction term.

4.2. Model prediction performance evaluation

I use the ARMAX model, the ARDL model, and the machine learning algorithms to predict air pollution concentrations, and then compare their prediction accuracy. I employ the default settings for the DT and SVM approaches, implementing ten-fold cross-validation. 10-fold cross-validation involves training ten models simultaneously and taking the average, which is equivalent to using the same model for different tests on a dataset. The diversity in each training set helps expand the dataset and enhance the generalization ability of the model. As for neural networks, fine-tuning is executed through a random search strategy. Random search selects a set of hyperparameters from the hyperparameter space for evaluation. The rationale behind this choice lies in the versatility of random search across various machine learning and deep learning problems. Its flexibility in exploration and computational efficiency stands out, as it doesn't necessitate traversing all possible hyperparameter combinations. This approach is particularly well-suited for high-dimensional search spaces, as highlighted by Bergstra et al. (2012) and Javeed et al. (2019). The optimal values of these approaches are in Appendix 5.

Handling relatively small datasets might result in the underperformance of machine learning models. This is because insufficient sample size in small datasets, which preventing the models from acquiring generalization capabilities, consequently causing overfitting. In this paper, the data is re-sampled into several subsets, each subset still contains thousands of samples, and the simplest machine learning model structure is used to improve the generalization ability and performance of the model.

Fig. 5 presents the evaluation results of the four models. These provide prediction results for NOx and $PM_{2.5}$, respectively. First, I use the ARMAX model to compare the machine learning algorithms. In the six datasets, the ARMAX model has the smallest MAE, MSE, and RMSE, and the largest adjusted R-squared (R^2), which means that ARMAX exhibits the best performance regarding air pollution prediction. The adjusted R^2 represents the proportion of the independent variable that can explain the dependent variable, which means the ability of traffic and weather factors to explain air pollution concentrations.

When I compare these two traditional statistical models, the ARDL model, and the ARMAX model, I find that when predicting both NO_x concentration and $PM_{2.5}$ concentration, the ARMAX model has a similar MAE, MSE, and RMSE to the ARDL model in most cases. The MAE, MSE, and RMSE measure the gap between the predicted value and the actual value. The MSE is often used as a loss function in machine learning, particularly in regression tasks. Together these show that when predicting air pollution, the prediction power of the ARMAX model and ARDL model is nearly the same.

In Fig. 5 Report A, for the NO_x concentration prediction, I find that in all the datasets, compared with the ARMAX model, the ARDL model has a larger adjusted R^2. In Fig. 5 Report B, for the prediction of $PM_{2.5}$ concentration, I find that in the warm months, no snowfall, and less snowfall, except for those subsets, the ARDL model has a larger adjusted R^2. Considering that ARIMA has only single variables and the lagged effect of single variables, ARDL has more variables than ARMAX, such as the interaction terms, so the reason why ARDL has a larger adjusted R^2 than ARMAX may be that ARDL has to overfit.

According to Report A, using the ARMAX model as an example, I observed that cold months exhibit higher adjusted R^2 compared to warm months. There isn't a substantial difference between the various months regarding snowfall. Report B, also employing the ARMAX model as an example, like Report A, cold months demonstrate higher adjusted R^2 values than warm months. Additionally, months with more snowfall display larger adjusted R^2 values.

Both reports A and B share the finding that cold months consistently yield higher adjusted R^2 than warm months, suggesting the model has stronger explanatory power during the cold months. The rise in heating during colder months, often through wood burning for heating in Norway, might contribute to increased air pollution. Furthermore, the settling of $PM_{2.5}$ into snow during snowfall may decrease the concentration of $PM_{2.5}$ in the air. Therefore, with increased snowfall over time, air quality tends to improve.

Model prediction accuracy was evaluated using MAE, MSE, RMSE and adjust R^2. Fig. 5 report A and B show the adjust R^2 results of the six models, while report C and D shows the separate figure for each dataset, and a detailed prediction accuracy comparison table of the six models can be found in Appendix 5.

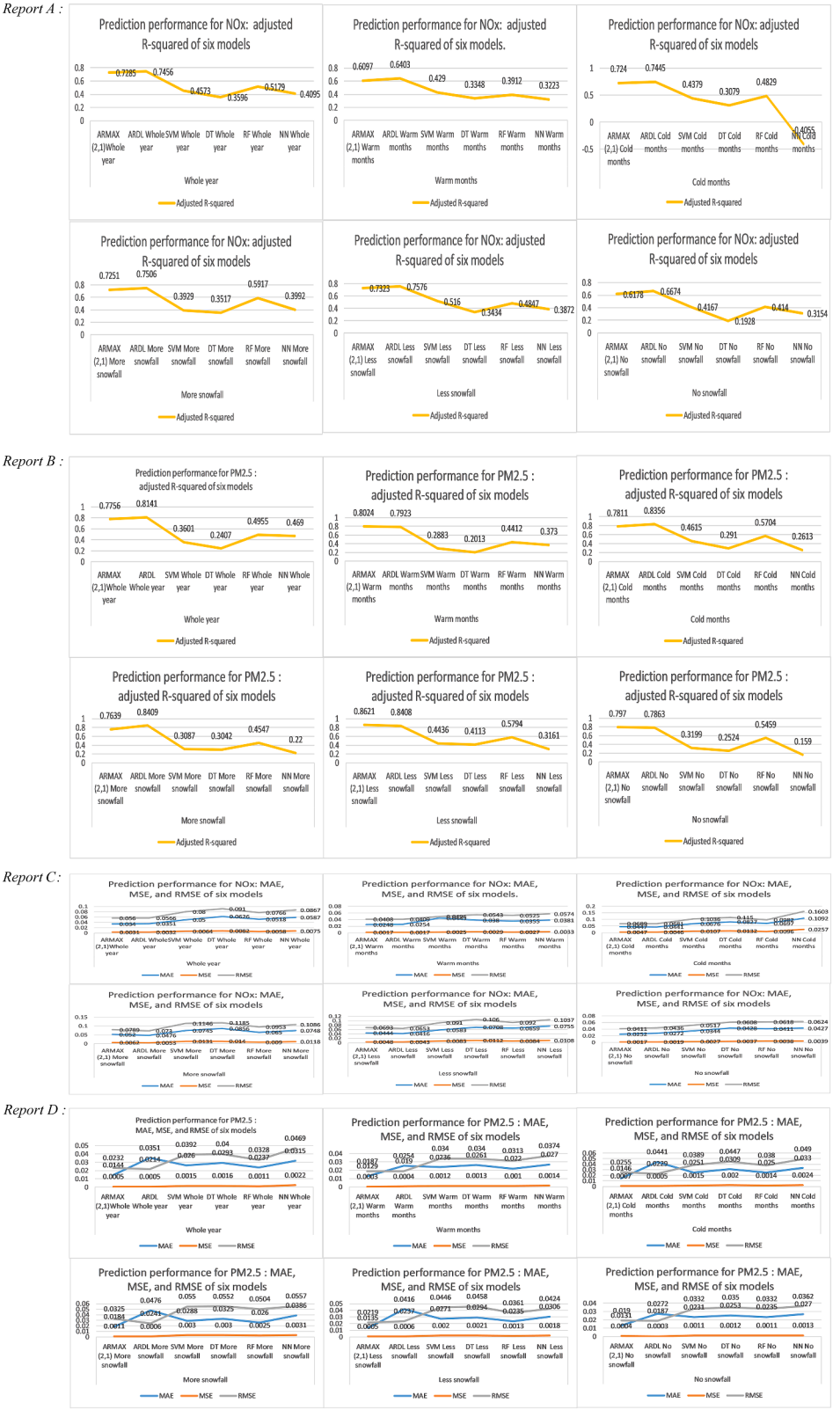


Fig. 5. Prediction Accuracy for *NOx* and $PM_{2.5}$.

5. Discussion and conclusion

To understand the complex relationship between traffic and air pollution and the intervention of meteorological factors, and to draw effective policy recommendations, I used the 2019 Norwegian hourly data.

My initial descriptive approaches demonstrate clear links between traffic volume and measured air pollution within rush hour traffic periods. I then go beyond this and seek to examine the role of meteorological factors in influencing this relationship. This is done using both traditional statistical and machine-learning approaches. I re-sample the data into five data subsets according to Norwegian meteorological and temporal variables, so the six models are evaluated six times. The ARMAX and ARDL models were always found to have the smallest MAE, MSE, and RMSE, and the largest adjusted R^2 in all six datasets. The results obtained suggest that traditional statistical models have significant advantages over these two machine learning approaches. The possible reason is that I add interaction items and lagged effects to the traditional statistical model. Such considerations will be closer to the actual situation in real life. Therefore, if the model design can better explain the actual phenomenon, it will affect the predictive accuracy.

Regression results demonstrate that weather conditions serve to change the relationship between traffic volume and air pollution. For instance, more traffic volume leads to higher air pollution levels, and colder days have more air pollutant concentrations. Similarly, mean wind speed, air temperature, and air pressure all have moderating effects on the link between traffic volume and air pollution. At the same time, there are dynamic effects of traffic volume on air pollution insofar as pollutant levels remain elevated for up to two hours after traffic surges. Taken together, this suggests complex links between traffic volume, meteorological factors, and harmful pollutants.

A number of my results differ from previous Norwegian findings (Aldrin & Haff, 2005). One possible explanation for these differences is that their paper didn't consider the interaction between different independent variables, and my results suggest that such interactions are important.

These results have policy implications. They suggest that, when formulating transportation policies, consideration should be given to weather conditions, for instance by reducing the traffic volume on days with lower air temperatures. This, for example, fits with a view that efficient road pricing should vary according to time-varying changes in road traffic externalities. This fits with earlier theoretical literature on optimal pricing (Parry et al., 2007). Specifically, the results in this paper suggest that optimal road charges should consider weather conditions. From the point of view of individual residents, depending on the weather, the two hours after the heavy traffic recommended reducing going out.

In the process of collecting hourly data for the entirety of 2019, it was identified that the data for O_3 was incomplete. Given the crucial role of O_3 in the global greenhouse effect and climate warming, its inclusion is imperative in future research endeavors. In this study, air pollutants, traffic monitoring stations, and meteorological data monitoring stations are located at distinct monitoring sites, with distances ranging from 5 to 10 km between them. Due to potential variations in meteorological conditions across different regions, the accuracy of meteorological data from more distant sites may be compromised. To enhance precision, employing an interpolation method between two weather stations can provide a more reliable estimation of meteorological data from monitoring stations. Given the limitations identified by Boke et al. (2017) concerning spatial interpolation methods, it is worthwhile to explore different interpolation strategies for meteorology in future research.

Data availability statement

The data used in this study, including traffic flow, air pollutants, and meteorological factors, are accessible to the public and can be obtained from the Norwegian Public Road Administration (SVV), the Norwegian Institute for Air Research (NILU), and the Norwegian Meteorological Institute (MET).

Declaration of competing interest

The author declare no conflict of interest.

Data availability

Data will be made available on request.

Appendix 1. The variables included in the data

Variables	Further explanation
Time	7 days per week, 24 hours per day. The period from 01.01.2019 00:00 to 01.01.2020 00:00.
Air pressure (qnh)	The air pressure is obtained by lowering the air pressure at the measuring station to the mean sea level.

(*continued on next page*)

(*continued*)

Variables	Further explanation
Mean wind speed (m/s)	Measurement of wind resources. This is measured as the mean value of the last ten minutes before the observation time.
Wind direction (degrees)	The direction the wind blows. The mean value of the last ten minutes before the observation time; 360 is north and 90 is east.
Wind direction (angles)	The wind direction is in four angles, namely North, South, West, and East.
Snow depth (cm)	Total daily snow depth. This is measured from the ground to the top of the snow cover.
Relative air humidity (%)	The ratio of absolute humidity to saturated absolute humidity in the air at the same temperature and pressure.
Air temperature (Celsius)	Ambient air temperature 2 meters above the ground and present value.
Traffic Volume (1 h)	The hourly volume number of vehicles passing through each hour, The unit is Passenger Car Unit (PCU).
$PM_{2.5}$ (1 h)[1]	Particulate matter in the atmosphere with a diameter of less than, or equal to, 2.5 microns, also known simply as "particulate matter," can enter the lungs.
PM_{10} (1 h)	Particulate matter with an aerodynamic equivalent diameter of less than or equal to 10 microns in ambient air, is known as inhalable particulate matter.
NO_x (1 h)	A chemical compound consisting only of nitrogen and oxygen, the common pollutants in the atmosphere.
NO_2 (1 h)	NO_2 is one type of NO_x, a brown-red atmospheric pollutant with a pungent odor at room temperature, a major factor in the formation of smog, and a precursor of ozone and particulate matter.
NO (1 h)	This is a colorless, odorless, insoluble gas. Its chemical properties are very active. When it reacts with oxygen, it can form NO_2.

[1] The pollutant unit $\mu g/m^3$ is one part per billion (ppb).

Raw data summary statistics

Category	Variable	Obs	Mean	Std.Dev.	Min	Max
Meteorological	Air pressure (qnh)	8,760	1010.42	11.80	970.30	1040.70
	Air temperature (celsius)	8,760	7.31	7.94	-13.8	31.50
	Wind direction (degrees)	8,760	126.65	105.52	0	360
	Wind direction (angles)	8,760	N/A	N/A	N/A	N/A
	Mean wind speed (m/s)	8,760	2.75	1.65	0	11.20
	Snow depth (cm)	8,760	13.18	11.87	0	49.34
	Relative air humidity (%)	8,760	74.22	19.75	13	100
Pollutants	NO (1 h)	8,760	31.71	42.52	-0.96	420.66
	NO_2 (1 h)	8,760	34.83	26.10	0.08	171.04
	NOx (1 h)	8,760	83.29	88.11	-0.75	787.28
	PM_{10} (1 h)	8,760	19.42	17.14	-4.29	202.18
	$PM_{2.5}$ (1 h)	8,760	7.55	4.53	-4.20	85.90
Traffic	Traffic volume (1 h)	8,760	3059.54	1984.02	43	6708
Temporal variables	Number of hours	8,760	4380.50	2528.94	1	8760
	Hours of the day	8,760	12.50	6.92	1	24
	Day of the month	8,760	15.72	8.80	1	31
	Month of the year	8,760	6.53	3.45	1	12

The six datasets generated in this paper are

Dataset 1, whole dataset, 2019 full-year data

According to the information on the monthly variation of air temperature, I extract the following two subsets:

Dataset 2, warm months, meaning air temperatures above 20 degrees Celsius, from April to September, includes 4392 observations.

Dataset 3, cold months, means the air temperature is below 20 degrees Celsius, from October to March, includes 4368 observations.

According to the information on the monthly variation of snow depth, I further select the following three subsets:

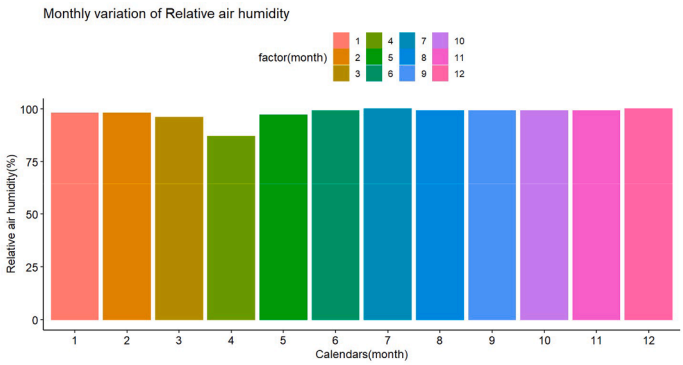
Dataset 4, months with more snowfall, when the snow depth is greater than 30 cm, from January to March, includes 2159 observations.

Dataset 5, months without snowfall, when the snow depth is less than 10 cm, from April to October, includes 5136 observations.

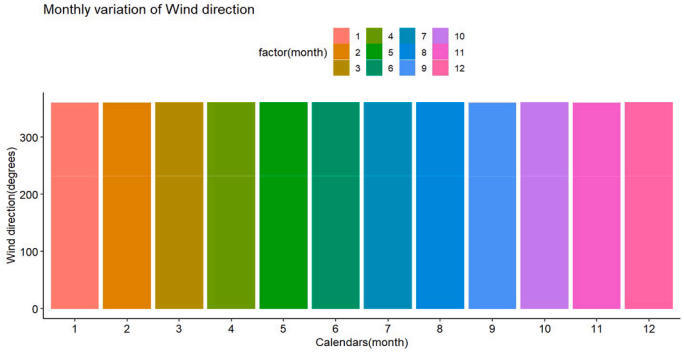
Dataset 6, months with less snowfall, when the snow depth is 10-30 cm, in November and December, includes 1465 observations.

Appendix 2. Monthly variation of meteorological factors

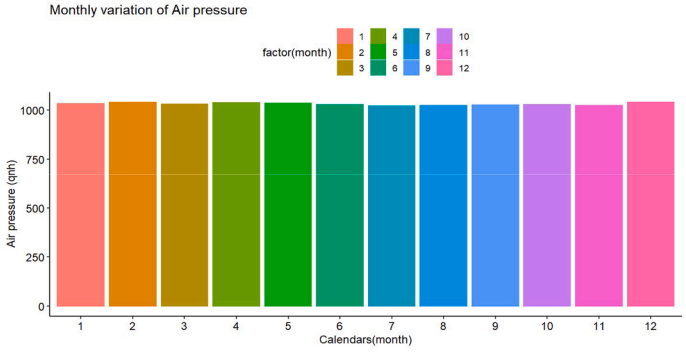
Panel A



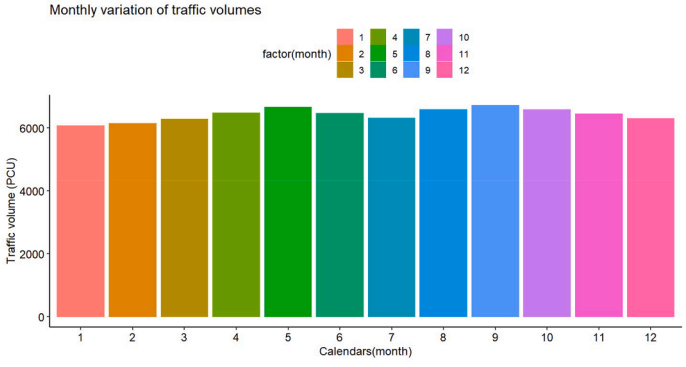
Panel B



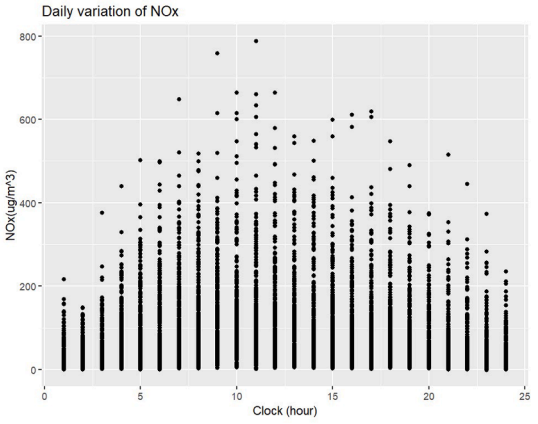
Panel C

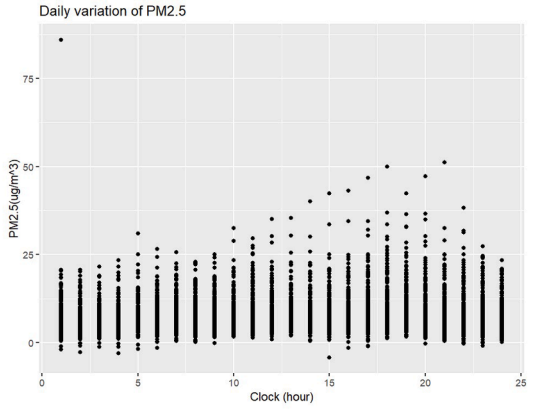


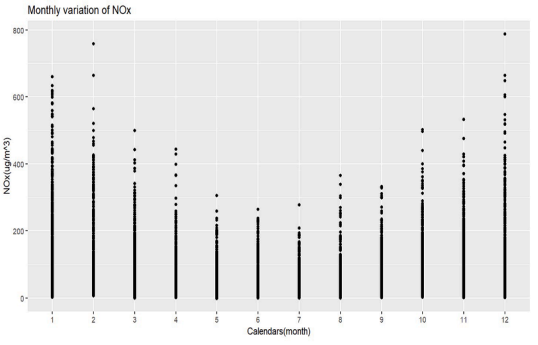
Panel D

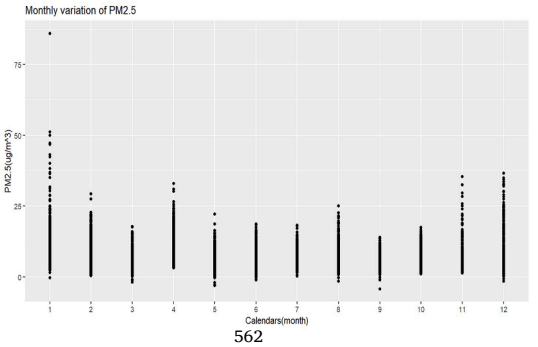


Daily and monthly variations of pollutants









Appendix 3. Sources of pollutants and their reaction mechanism's introduction

NOx is a gas mixture composed of nitrogen and oxygen. There are many kinds of NOx, such as nitrous oxide, nitric oxide (NO), nitrogen dioxide (NO_2), nitrous pentoxide, etc., but only NO and NO_2 are stable, as the other gas mixtures will decompose due to light, heat, and humidity.

The sources of NOx in the air include welding, blasting explosives, exhaust from motor vehicles, and burning coal. NO reacts with oxygen to form NO_2. The main sources of NO_2 are motor vehicle exhaust and boiler exhaust.

After entering the air, NO_X will react with common chemical substances in the air to decompose. Usually, NO_2 reacts with other chemical substances in the sun to form nitric acid, which is the main component of acid rain, or reacts with the sun to become ozone or smog. NO_2 is a greenhouse gas that can exacerbate global warming. It destroys the ozone layer and leads to the formation of ozone holes, thus causing damage to the human immune system and skin.

PM is the abbreviation for particulate matter. Both $PM_{2.5}$ and PM_{10} are particulate matter, and the main components are carbon-containing particles, sulfates, heavy metals, etc. The difference lies in the particle size. The unit is a micron. One micron is one-millionth of a meter. The value represents the aerodynamic diameter of the particle. The larger the value, the larger the particle; it indicates that the particle size is less than, or equal to, 1 micron. $PM_{2.5}$ is a particulate matter with an aerodynamic diameter of 2.5 microns or less. $PM_{2.5}$ is also known as particulate matter that can enter the lungs, and it can also be suspended in the air for a long time. PM_{10} contains $PM_{2.5}$, and $PM_{2.5}$ accounts for about 70 % of PM_{10}. $PM_{2.5}$ mainly comes from the combustion of fossil fuels, such as motor vehicle exhaust, coal, etc., in addition to some volatile organic compounds. PM_{10} mainly comes from emissions from chimneys and vehicles. At the same time, some of the sulfur oxides, NO_X, and other compounds in the air interact with each other to form fine particles. The dust raised by the wind can also increase the concentration of PM_{10}. Due to the smaller particle size of $PM_{2.5}$, it is easier for it to stay in the bronchi and alveoli and cause health hazards.

Appendix 4. Methods summary

Support vector machine (SVM)

SVM has no requirements for data stationarity and can handle interactions between nonlinear features in big data. The final decision function of SVM is only determined by a small number of support vectors, which enhances the efficiency of SVM in handling high-dimensional data. Nevertheless, the computational complexity of SVM is not entirely unrelated to dimensionality. The complexity of training SVM is usually in the range of quadratic to cubic relative to the number of samples, which can be problematic for large-scale datasets, SVM needs to weigh some aspects in practical applications, including efficiency in high-dimensional space and training complexity. The specific decision may be contingent on the particular problem and dataset.

The SVM finds the optimal decision surface with the largest interval in the eigenspace. The principle of SVM is to find a hyperplane, and this hyperplane can separate all sample points to ensure the maximum distance between the sample points and the hyperplane. The reason why it is called a "support vector" is that when determining the separation hyperplane, only the points at the extreme position are useful, so if the distance between the extreme position and the hyperplane is the largest, it is the best separation plane.

Support vector regression (SVR) is a variant of SVM in regression analysis. The principles of SVR and SVM are similar. The biggest difference is only that SVM aims to maximize the "distance" from the closest sample point to the hyperplane; SVR aims to minimize the "distance" to the farthest sample point from the hyperplane. The SVR equation I use here is:

$$f(x_i) = (w^* * x_i) + b^*$$

Where x_i stands for different traffic and weather variables. The specific implementation steps are:

Given training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

(1) Solving the quadratic programming problem:

$$\min_a \frac{1}{2} \sum_i \sum_j a_i a_j y_i y_j (x_i * x_j) - \sum_i a_i$$

$$s.t \sum_i a_i y_i = 0,\ a_i \geq 0$$

Get:

$$a^* = (a_1^*, \ldots, a_n^*)^T$$

(2) Calculating parameters w, and select a positive component a_i^* calculate b

$$w^* = \sum a_i^* \, y_i x_i, b^* = y_j - \sum a_i^* y_i (x_i * x_j)$$

(3) Constructing the decision boundary:

$$g(x) = (w^* * x) + b^* = 0,$$

From this, I will have the decision function:

$$f(x) = \text{sgn}\ (g(x))$$

After constructing the decision boundary:

$$g(x) = (w^* * x) + b^* = 0$$

I have the decision equation

$$f(x_i) = \text{sgn}\ (g(x))$$

Since the influence of traffic volume and weather on air pollution is a complex phenomenon, real-life data are usually linearly inseparable and contain a lot of noise, which appears to be challenging for prediction accuracy. SVM is good at solving the problems of small samples, nonlinearity, and high dimensionality, so they have achieved good prediction results.

The advantage of SVM over general regression models or ARMAX is that: in general, ARMAX models calculate a loss if the actual and predicted values are not equal. But for SVM, if the value is in the interval band, the SVM does not calculate the loss, unless the absolute value of the difference between the actual value and the predicted value is greater than the error term. This means SVM is more robust and flexible. Another advantage is that the way to optimize the model is different. SVM optimizes the model by maximizing the interval band and minimizing the total loss, while regression models usually optimize the regression model by calculating the mean value after gradient descent.

The main disadvantage of SVM is that when the feature dimension is much larger than the number of samples, the performance of the SVM is average. In this paper, the number of observations is 8760, and the feature dimension is 13, which is very suitable for using SVM. Second, SVM is sensitive to missing values, so I performed missing value imputation at the very beginning.

Decision tree (DT)

DT is an algorithm for solving classification or regression problems and belongs to a set of supervised machine-learning algorithms. It is formed by a tree structure that includes a root node, a leaf node, and an internal node. The root node represents the complete sample set, the internal nodes represent the judgment of feature attributes, and the leaf nodes represent the result of the decision. It makes judgments via the attribute values at the internal nodes of the tree and then selects the internal nodes of the branches according to the judgment results until it finally reaches the leaf node, which provides the result. The DT has the advantage of being easy to implement. Since both the traffic volume and the pollutant values are continuous, and the DT can be used for classification and regression, here I use a regression tree, and the tree equation is:

$$D_t = f(x_i)$$

Here D_t is air pollution, NOx or $PM_{2.5}$, and x_i are different variables from traffic volume and meteorological factors, t is time, from 1, 2, 3, …, T, and the unit is hour.

I consider air pollution as the dependent variable, with traffic and weather as the independent variables. Treating each value as a category would result in a large amount of data calculation. Therefore, based on the distribution of the continuous variables, the DT selects several feature values to classify the data, determines possible split points, and obtains:

$$R_1 = \{x_i | x_i \le s\}, R_2 = \{x_i | x_i > s\}$$

$$C_1 = \frac{1}{N_1} \sum_{x_i \in R1} y_i, C_2 = \frac{1}{N_2} \sum_{x_i \in R2} y_i$$

N_1 and N_2 are the numbers of sample points in R_1 and R_2 , respectively, and C_1 and C_2 are the mean values of the dependent variables in R_1 and R_2.

So, the regression tree $D_1(x)$ is:

$$f_1(x_i) = H_1(x_i) = \begin{cases} C_1, x_i \le s \\ C_2, x_i > s \end{cases}$$

$$f_2(x_i) = f_1(x_i) + H_2(x_i)$$

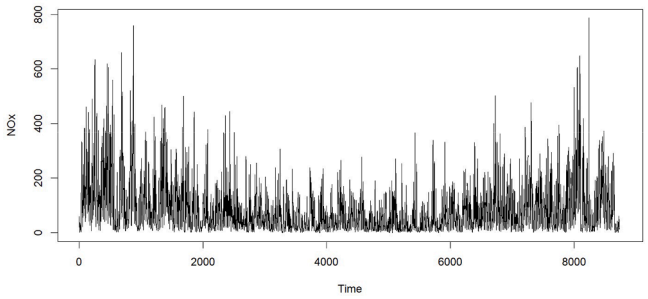
It will automatically be iterated until the sum of squared errors of the fitted training data is less than a certain threshold, then $D_t = f_m(x_i)$ is the desired regression tree. There are more than two categories of data here, and the data are continuous variables, there are more than two categories of data here, and the data are continuous variables variables; therefore, I use Classification and Regression Trees (CART) algorithm here. CART is a common algorithm for regression problems, which uses variance reduction to select the best segmentation. The leaf nodes of a regression tree contain numerical values. It chooses the best split for each node to minimize the variance of the predicted values, and then builds a regression tree.

Regression models are easy to understand, intuitive, and transparent, and are effective for small data volumes and simple relationships but have difficulties in handling highly complex data. The advantage of the DT over the regression model or ARMAX model is that it exhibits better performance for complex and nonlinear data, and the principle is easy to understand. The disadvantage of the DT is that it is easy to overfit since it usually contains a lot of subtrees. At the same time, when having a large dataset, the DT runs slowly and consumes a large amount of machine memory. In this paper, the amount of data is large, and there are correlations between different variables, and DT has the potential to solve these.

Mean absolute error (MAE) is mathematically the average absolute difference between observed and predicted results, the smaller the MAE, the better the prediction and the more reliable the prediction result. Mean squared error (MSE) refers to the mean squared error between the observed actual value and the model predicted value. The lower the MSE, the better the model performance.

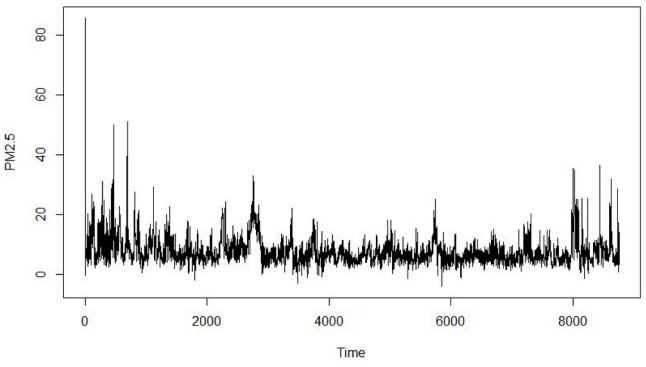
ARMAX model

As shown, a time series chart has some outliers and variance changes, but it is stationary.



Time Series Diagram of NOx.

Fig. A



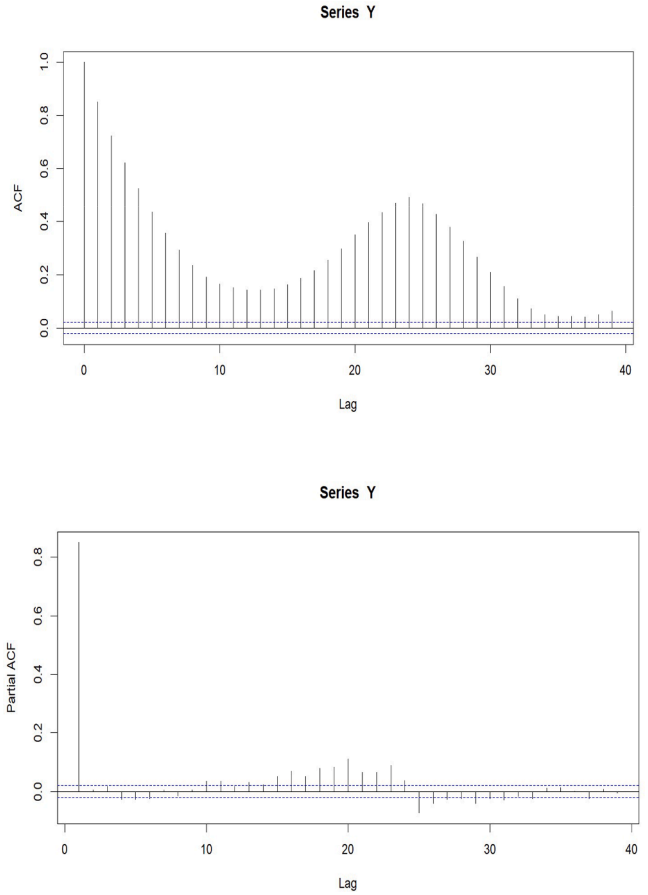
Time Series Diagram of $PM_{2.5}$.

For further verification, I use the Augmented Dickey-Fuller Test to test for the stationarity of the time series. The Augmented Dickey-Fuller Test (ADF) is a modified version of the Dickey-Fuller Test, that excludes the influence of autocorrelation. The null hypothesis is that the data are nonstationary. Set the additional lags to 0, the P-value is the ADF are all 0.01, when the absolute values of ADF are smaller than 0.01, and the null hypothesis can be rejected.

Autocorrelation Function (ACF) refers to the linear relationship between the sequence value and the lag value at any time t $(t = 1,2,3,\ldots,n)$. An ACF plot, also known as a "correlogram," refers to a plot with the lag value *i* as the x-axis and the autocorrelation coefficient as the y-axis. A correlation coefficient value between X_t and X_{t-i} is the autocorrelation coefficient. The partial autocorrelation function (PACF) is, after removing the interference, the relationship between a time series observation and previous time steps' observation. Not all shorter intervals between these observations are included in the correlation. PACF helps to identify the number of autoregressive coefficients p-values in an ARMAX model. ACF is used to confirm q values.

I present ACF and PACF graphs in the following figure. From the ACF diagram of Y, the cutoff is not obvious, and the autocorrelation coefficient of the subsequent order fluctuates irregularly, that is, it tails off, so here could take q equal to 0. From the PACF graph, after the 2nd-order cut-off, they fall within the range of two standard deviations, satisfying the short-term autocorrelation

property; thus, it can be considered that the sequence is stationary, and can take p = 2.



ACF and PACF Graph.

The best model with the smallest AIC[4] represents the best ARMAX model. AIC balances overfitting or underfitting, so if two models have the same explanatory power, the model with a smaller AIC value with fewer parameters is better. I used NOx and $PM_{2.5}$ respectively as dependent variables. I selected four models, then tried to select the model with the lowest AIC, and I find that ARMAX (2,1) has the lowest AIC.

Table A
The Results of ARMAX Models when Using NO_x as the Dependent Variable

	ARMAX (2,0,0)	ARMAX (2,0,1)	ARMAX (2,0,2)	ARMAX (2,0,3)
ar1	0.8043	1.7821	1.7467	1.0797
ar2	0.0351	-0.7823	-0.7484	-0.1699
ma1		-0.9924	-0.9644	-0.2818
ma2			-0.0060	-0.0677
ma3				-0.0179
Sigma^2	1973	1935	1935	1964
AIC	73078.7	72946.03	72947.18	73051.68

Note: According to the analysis results of the ACF and PACF graphs in Figures C, I selected the four ARMAX models most likely to have the smallest AIC values, using NO_x as the dependent variable, and I compared their AIC. I found that the AIC of ARMAX (2,0,1) = 72946.03, which is the smallest AIC value, meaning this ARMAX model is the best.

[4] AIC = (2k − 2L) /n.L = − (n /2) *ln(2*pi) − (n /2) *ln(sse /n) − n/2, where n is the number of data points in the data, SSE is the sum of squared residuals, k represents the number of independent variables, and L is likelihood.

Table B
The Results of ARMAX Models when Using $PM_{2.5}$ as the Dependent Variable

	ARMAX (2,0,0)	ARMAX (2,0,1)	ARMAX (2,0,2)	ARMAX (2,0,3)
ar1	0.8739	0.3788	0.5075	1.7501
ar2	0.0273	0.4732	0.3518	-0.7531
ma1		0.4910	0.3688	-0.8909
ma2			0.0226	-0.0282
ma3				-0.0749
$Sigma^2$	4.413	4.41	4.409	4.351
AIC	30304.85	30302.77	30303.32	30212.75

Note: According to the analysis results of the ACF and PACF graphs in Figures C, I selected the four ARMAX models most likely to have the smallest AIC values, using $PM_{2.5}$ as the dependent variable, and I compared their AIC. I found that the AIC of ARMAX (2,0,1) = 30302.77, which is the smallest AIC value, meaning this ARMAX model is the best.

Appendix 5. Table of prediction accuracy comparison of six models in six datasets

I use four machine learning algorithms, SVM, RF, NN and DT, and two statistical models, ARMAX and ARDL to predict the concentration of air pollutants. The last four columns of the table are the model evaluation results.

NOx	Model	MAE	MSE	RMSE	Adjusted R-squared
Whole year	ARMAX (2,1) Whole year	**0.034**	**0.0031**	**0.056**	0.7285
	ARDL Whole year	0.0351	0.0032	0.0566	**0.7456**
	SVM Whole year	0.05	0.0064	0.08	0.4573
	DT Whole year	0.0626	0.0082	0.091	0.3596
	RF Whole year	0.0518	0.0058	0.0766	0.5179
	NN Whole year	0.0587	0.0075	0.0867	0.4095
Warm months	ARMAX (2,1) Warm months	**0.0248**	**0.0017**	**0.0408**	0.6097
	ARDL Warm months	0.0254	0.0017	0.0409	**0.6403**
	SVM Warm months	0.0445	0.0025	0.0496	0.429
	DT Warm months	0.038	0.0029	0.0543	0.3348
	RF Warm months	0.0355	0.0027	0.0525	0.3912
	NN Warm months	0.0381	0.0033	0.0574	0.3223
Cold months	ARMAX (2,1) Cold months	0.0447	0.0047	0.0689	0.724
	ARDL Cold months	**0.0441**	**0.0046**	**0.0681**	**0.7445**
	SVM Cold months	0.0676	0.0107	0.1036	0.4379
	DT Cold months	0.0827	0.0132	0.115	0.3079
	RF Cold months	0.0697	0.0096	0.0982	0.4829
	NN Cold months	0.1092	0.0257	0.1603	-0.4055
More snowfall	ARMAX (2,1) More snowfall	0.052	0.0062	0.0789	0.7251
	ARDL More snowfall	**0.0476**	**0.0053**	**0.073**	**0.7506**
	SVM More snowfall	0.0745	0.0131	0.1146	0.3929
	DT More snowfall	0.0856	0.014	0.1185	0.3517
	RF More snowfall	0.065	0.009	0.0953	0.5917
	NN More snowfall	0.0748	0.0118	0.1086	0.3992
No snowfall	ARMAX (2,1) No snowfall	**0.0252**	**0.0017**	**0.0411**	0.6178
	ARDL No snowfall	0.0272	0.0019	0.0436	**0.6674**
	SVM No snowfall	0.0344	0.0027	0.0517	0.4167
	DT No snowfall	0.0428	0.0037	0.0608	0.1928
	RF No snowfall	0.0411	0.0038	0.0618	0.414
	NN No snowfall	0.0427	0.0039	0.0624	0.3154
Less snowfall	ARMAX (2,1) Less snowfall	0.0444	0.0048	0.0693	0.7323
	ARDL Less snowfall	0.0416	**0.0043**	**0.0653**	**0.7576**
	SVM Less snowfall	0.0583	0.0083	0.091	0.516
	DT Less snowfall	0.0708	0.0112	0.106	0.3434
	RF Less snowfall	0.0659	0.0084	0.092	0.4847
	NN Less snowfall	0.0755	0.0108	0.1037	0.3872

$PM_{2.5}$	Model	MAE	MSE	RMSE	Adjusted R-squared
Whole year	ARMAX (2,1) Whole year	**0.0144**	**0.0005**	0.0232	0.7756
	ARDL Whole year	0.0351	**0.0005**	**0.0214**	**0.8141**
	SVM Whole year	0.0260	0.0015	0.0392	0.3601
	DT Whole year	0.0293	0.0016	0.0400	0.2407
	RF Whole year	0.0237	0.0011	0.0328	0.4955
	NN Whole year	0.0315	0.0022	0.0469	0.4690

(*continued on next page*)

(*continued*)

$PM_{2.5}$		Model	MAE	MSE	RMSE	Adjusted R-squared
	Warm months	ARMAX (2,1) Warm months	**0.0129**	**0.0003**	**0.0187**	**0.8024**
		ARDL Warm months	0.0254	0.0004	0.0190	0.7923
		SVM Warm months	0.0236	0.0012	0.0340	0.2883
		DT Warm months	0.0261	0.0013	0.0340	0.2013
		RF Warm months	0.0220	0.0010	0.0313	0.4412
		NN Warm months	0.0270	0.0014	0.0374	0.3730
	Cold months	ARMAX (2,1) Cold months	**0.0146**	0.0007	0.0255	0.7811
		ARDL Cold months	0.0441	**0.0005**	**0.0229**	**0.8356**
		SVM Cold months	0.0251	0.0015	0.0389	0.4615
		DT Cold months	0.0309	0.0020	0.0447	0.2910
		RF Cold months	0.0250	0.0014	0.0380	0.5704
		NN Cold months	0.0330	0.0024	0.0490	0.2613
	More snowfall	ARMAX (2,1) More snowfall	**0.0184**	0.0011	0.0325	0.7639
		ARDL More snowfall	0.0476	**0.0006**	**0.0241**	**0.8409**
		SVM More snowfall	0.0288	0.0030	0.0550	0.3087
		DT More snowfall	0.0325	0.0030	0.0552	0.3042
		RF More snowfall	0.0260	0.0025	0.0504	0.4547
		NN More snowfall	0.0386	0.0031	0.0557	0.2200
	No snowfall	ARMAX (2,1) No snowfall	**0.0131**	0.0004	0.0190	**0.7970**
		ARDL No snowfall	0.0272	**0.0003**	**0.0187**	0.7863
		SVM No snowfall	0.0231	0.0011	0.0332	0.3199
		DT No snowfall	0.0253	0.0012	0.0350	0.2524
		RF No snowfall	0.0235	0.0011	0.0332	0.5459
		NN No snowfall	0.0270	0.0013	0.0362	0.1590
	Less snowfall	ARMAX (2,1) Less snowfall	**0.0135**	**0.0005**	**0.0219**	**0.8621**
		ARDL Less snowfall	0.0416	0.0006	0.0237	0.8408
		SVM Less snowfall	0.0271	0.0020	0.0446	0.4436
		DT Less snowfall	0.0294	0.0021	0.0458	0.4113
		RF Less snowfall	0.0235	0.0013	0.0361	0.5794
		NN Less snowfall	0.0306	0.0018	0.0424	0.3161

I optimal parameters for each algorithm are:

For DT, the minimum branch node is set to minsplit=20, the maximum tree depth is maxdepth=30, the complexity parameter is cp =0.01, and cross-validation is performed with xval=10.

For SVM, the optimal parameters are cost=1, and the kernel used is radial.

For RF, the type of random forest is set to regression. For both NOx and $PM_{2.5}$, the number of trees is 60, and the number of variables tried at each split is 2.

Neural Networks (NN)	Hidden layer NOx	Maximum steps- NOx	Hidden layer $PM_{2.5}$	Maximum steps- $PM_{2.5}$
lesssnowfall	5	35242	5	931
no snowfall	5	48213	4	7928
moresnowfall	5	68221	4	6250
warmmonth	5	21978	5	7658
coldmonth	4	43561	5	7951
whole year data	4	43561	5	8941

Note: Hidden layer: Number of hidden neurons (vertices) in each layer; Maximum number of steps, which means that reaching this value will cause the training process of the algorithm to stop for NOx.

References

Aldrin, M., Haff, I.H., 2005. Generalized additive modeling of air pollution, traffic volume, and meteorology. Atmos. Environ. 39 (11), 2145–2155. https://doi.org/10.1016/j.atmo-senv.2004.12.020.

Ameer, S., Shah, M.A., Khan, A., Song, H., Maple, C., Islam, S.U., Asghar, M.N., 2019. Comparative analysis of machine learning techniques for predicting air quality in smart Cities. IEEe Access. 7, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082.

Amos, J. (2016). Polluted air causes 5.5 million deaths a year, new research says. BBC News, February, 13.

Bai, L., Wang, J., Ma, X., Lu, H., 2018. Air pollution forecasts: An overview. Internat. J. Environ. Res. Public Health 15 (4), 780. https://doi.org/10.3390/ijerph15040780. PMID: 29673227; PMCID: PMC5923822.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 (2).

Bigazzi, A. Y., Rouleau, M., 2017. Can traffic management strategies improve urban air quality? A review of the evidence. Journal of Transport & Health 7, 111–124. https://doi.org/10.1016/j.jth.2017.08.001.

Bjørgen, A., Ryghaug, M., 2022. Integration of urban freight transport in city planning: Lesson learned. Transport. Res. Part D 107, 103310. https://doi.org/10.1016/j.trd.2022.103310. ISSN 1361-9209.

Boke, A.S., 2017. Comparative evaluation of spatial interpolation methods for estimation of missing meteorological variables over Ethiopia. J. Water. Resour. Prot. 09 (08), 945–959. https://doi.org/10.4236/jwarp.2017.98063.

Briggs, D.J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Reeuwijk, H.V., Smallbone, K., Van Der Veen, A., 1997. Mapping urban air pollution using GIS: A regression-based approach. Internat. J. Geographical Informat. Sci. 11 (7), 699–718. https://doi.org/10.1080/136588197242158.
Brugge, D., Durant, J.L., Rioux, C., 2007. Near-highway pollutants in motor vehicle exhaust: A review of epidemiologic evidence of cardiac and pulmonary health risks. Environ. Health 6, 1–12. https://doi.org/10.1186/1476-069X-6-23.
Chen, Q., Wang, Q., Xu, B., Xu, Y., Ding, Z., Sun, H., 2021. Air pollution and cardiovascular m-ortality in Nanjing, China: Evidence highlighting the roles of cumulative exposure and mort-ality displacement. Chemosphere 265, 129035. https://doi.org/10.1016/j.chemosphere.2020.129035.
Conte, M., Donateo, A., Contini, D., 2018. Characterization of particle size distributions and cor-responding size-segregated turbulent fluxes simultaneously with CO2 exchange in an urban area. Sci. Total Environ. 622–623, 1067–1078. https://doi.org/10.1016/j.scito-tenv.2017.12.040.
Currie, J., Neidell, M., 2005. Air pollution and infant health: what can we learn from California's recent experience? Q. J. Econ. 120 (3), 1003–1030. https://doi.org/10.1093/qje/120.3.1003.
Currie, J., Neidell, M., Schmieder, J.F., 2009. Air pollution and infant health: Lessons from Ne-w Jersey. J. Transp. Health. 28 (3), 688–703. https://doi.org/10.1016/j.jhealeco.2009.02.001.
Folgerø, I.K., Harding, T., Westby, B.S., 2020. Going fast or going green? Evidence from envir-onmental speed limits in Norway. Trans. Environ. 82, 102261 https://doi.org/10.1016/j.trd.2020.102261.
Font, A., Fuller, G.W., 2016. Did policies to abate atmospheric emissions from traffic have a po-sitive effect in London? Environ. Pollut. 218, 463–474. https://doi.org/10.1016/j.envpol.2016.07.026.
Gauderman, W.J., Vora, H., McConnell, R., Berhane, K., Gilliland, F., Thomas, D., Lurmann, F., Avol, E., Kunzli, N., Jerrett, M., Peters, J., 2007. Effect of exposure to traffic on lung deve-lopment from 10 to 18 years of age: A cohort study. Lancet 369 (9561), 571–577. https://doi.org/10.1016/S0140-6736(07)60037-3. PMID: 17307103.
Grange, S.K., Carslaw, D.C., Lewis, A.C., Boleti, E., Hueglin, C., 2018. Random forest meteo-rological normalization models for Swiss PM10 trend analysis. Atmos. Chem. Phys. 18, 6223–6239. https://doi.org/10.5194/acp-18-6223-2018.
Green, C.P., Heywood, J.S., Navarro, M., 2016. Traffic accidents and the London congestion c-harge. J. Public Econ. 133, 11–22. https://doi.org/10.1016/j.jpubeco.2015.10.005.
Green, C., Krehic, L., 2022. An extra hour wasted? Bar closing hours and traffic accidents in No-rway. Health Econ. 31 (8), 1752–1769. https://doi.org/10.1002/hec.4550.
Green, H., Talbot, N., Salmond, J., Dirks, K., Xie, S., Davy, P., 2020. Implications for air qualit-y management of changes in air quality during lockdown in Auckland (New Zealand) in resp-onse to the 2020 SARS-CoV-2 epidemic. Sci. Total Environ. 746, 141129 https://doi.org/10.1016/j.scitotenv.2020.141129. Epub 2020 Jul 27. PMID: 32745857; PMCID: PMC7384416.
Gryech, I., Ghogho, M., Elhammouti, H., Sbihi, N., Kobbane, A., 2020. Machine learning for air quality prediction using meteorological and traffic-related features. J. Ambient Intel-ligence Smart Environ. 12 (5), 379–391. https://doi.org/10.3233/AIS-200572.
Gualtieri, G., Crisci, A., Tartaglia, M., et al., 2015. A statistical model to assess air quality levels at urban sites. Water. Air. Soil. Pollut. 226, 394. https://doi.org/10.1007/s11270-015-2663-4.
Islam, M., Chen, G., Jin, S., 2019. An overview of neural network. American J. Neural Networks App. 5 (1), 7. https://doi.org/10.11648/j.ajnna.20190501.12.
Janarthanan, R., Partheeban, P., Somasundaram, K., Navin Elamparithi, P., 2021. A deep learning approach for prediction of air quality index in a metropolitan city. Sustain. Cities. Soc. 67 https://doi.org/10.1016/j.scs.2021.102720.
Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., Nour, R., 2019. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access 7, 180235–180243. https://doi.org/10.1109/ACCESS.2019.2952107.
Kamińska, J.A., 2018. The use of random forests in modeling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław. J. Environm. Manage. 217, 164–174. https://doi.org/10.1016/j.jenvman.2018.03.094. ISSN 0301-4797.
Kendrick, C.M., Koonce, P., George, L.A., 2015. Diurnal and seasonal variations of NO, NO2, and PM2.5 mass as a function of traffic volumes alongside an urban arterial. Atmos. Environ. 122, 133–141. https://doi.org/10.1016/j.atmosenv.2015.09.019.
Kimbrough, S., Baldauf, R.W., Hagler, G.S.W., Shores, R.C., Mitchell, W., Whitaker, D.A., Croghan, C.W., Vallero, D.A., 2013. Long-term continuous measurement of near-road air pollution in Las Vegas: Seasonal variability in traffic emissions impact on local air quality. Air Qual., Atmosph. Health 6 (1), 295–305. https://doi.org/10.1007/s11869-012-0171-x.
Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y., 2017. Modeling PM2.5 Urban pollution using machine learning and selected meteorological parameters. J. Electr. Comp. Eng. https://doi.org/10.1155/2017/5106045.
Kurz, C., Orthofer, R., Sturm, P., Kaiser, A., Uhrner, U., Reifeltshammer, R., Rexeis, M., 2014. Projection of the air quality in Vienna between 2005 and 2020 for NO2 and PM10. Urban. Clim. 10 (2014), 703–719. https://doi.org/10.1016/j.uclim.2014.03.008.
Luecken, D.J., Hutzell, W.T., Gipson, G.L., 2006. Development and analysis of air quality mo-deling simulations for hazardous air pollutants. Atmos. Environ. 40 (26), 5087–5096. https://doi.org/10.1016/j.atmosenv.2005.12.044.
Martín-Baos, J.Á., Rodriguez-Benitez, L., García-Ródenas, R., Liu, J., 2022. IoT based monitoring of air quality and traffic using regression analysis. Appl. Soft. Comput. 115 https://doi.org/10.1016/j.asoc.2021.108282.
Mignone, P., Malerba, D., Ceci, M., 2022. Anomaly detection for public transport and air pollution analysis. In: Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022, pp. 2867–2874. https://doi.org/10.1109/BigData55660.2022.10020470.
Moazami, S., Noori, R., Amiri, B.J., Yeganeh, B., Partani, S., Safavi, S., 2016. Reliable predicti-on of carbon monoxide using a developed support vector machine. Atmospheric Pollut. Res. 7 (3), 412–418. https://doi.org/10.1016/j.apr.2015.10.022.
Moretti, E., Neidell, M., 2011. Pollution, health, and avoidance behavior: Evidence from the por-ts of Los Angeles. J. Human Res. 46 (1), 154–175. https://doi.org/10.3368/jhr.46.1.154.
Parry, I.W.H., Walls, M., Harrington, W., 2007. Automobile externalities and policies. J. Econ. Literat. 45 (2), 373–399. https://doi.org/10.1257/jel.45.2.373.
Pasquier, A., André, M., 2017. Considering criteria related to spatial variabilities for the assessm-ent of air pollution from traffic. Transportat. Res. Proced. 25 (June), 3354–3369. https://doi.org/10.1016/j.trpro.2017.05.210.
Qu, H., Lu, X., Liu, L., Ye, Y., 2019. Effects of traffic and urban parks on PM10 and PM2.5 ma-ss concentrations. Energy Sources, Part A 45 (2), 0-5647. https://doi.org/10.1080/15567036.2019.1672833 n.d.
Rigatti, S.J., 2017. Random Forest. In Journal of Insurance Medicine Copyright C 2017. J. Insur. Med. (1946) 47. http://meridian.allenpress.com/jim/article-pdf/47/1/31/1736157/insm-47-01-31-39_1.pdf.
Santos, G.S., Sundvor, I., Vogt, M., Grythe, H., Haug, T.W., Høiskar, B.A., Tarrason, L., 2020. Evaluation of traffic control measures in Oslo region and its effect on current air quality policies in Norway. Transp. Policy. (Oxf) 99 (August), 251–261. https://doi.org/10.1016/j.tranpol.2020.08.025.
Shaban, B.K., Kadri, A., Rezk, E., 2016. Urban air pollution monitoring system with forecastin-g models. *In* IEEe Sens. J. 16 (8), 2598–2606. https://doi.org/10.1109/JSEN.2016.2514378.
Srimuruganandam, B., Nagendra, S.M.S., 2010. Analysis and interpretation of particulate matter–PM10, PM2. 5 and PM1 emissions from the heterogeneous traffic near an urban roadway. Atmos. Pollut. Res. 1 (3), 184–194. https://doi.org/10.5094/APR.2010.024.
Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. J. Transp. Eng. 129 (6), 664–672. https://doi.org/10.1061/(ASCE)0733-947X(2003)129.
Wærsted, E.G., Sundvor, I., Denby, B.R., Mu, Q., 2022. Quantification of the temperature depe-ndence of NOx emissions from road traffic in Norway using air quality modeling and monit-oring data. Atmos. Environ. X. 13 (x), 100160 https://doi.org/10.1016/j.aeaoa.2022.100160.
Yildirim Yayilgan, S., Bajwa, I.S., Sanfilippo, F., 2021. Intelligent Technologies and Applications (Vol. 1382). Springer International Publishing. https://doi.org/10.1007/978-3-030-71711-7.
Zhang, P.G., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0.