Doctoral theses at NTNU, 2023:250

Cong Cao

Three essays on Transportation and Environmental Economics

With a focus on Machine Learning Applications

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

Cong Cao

Three essays on Transportation and Environmental Economics

With a focus on Machine Learning Applications

Thesis for the Degree of Philosophiae Doctor

Trondheim, August 2023

Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Economics and Management
Department of Economics

ISBN 978-82-326-7200-4 (printed ver.)
ISBN 978-82-326-7199-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:250

Printed by NTNU Grafisk senter

Acknowledgments

First and foremost, I would like to thank my two supervisors, Colin Green (main) and Jan Morten Dyrstad (co), who have provided me with constant support during my Ph.D. Their advice and comments have undoubtedly improved the quality of the papers, and not just in terms of academic guidance, as they have always considered my carefully thought-out choices with an open mind and give their advice, treating every question I had seriously, which has allowed me to move forward with confidence. For example, during my research stay in the United States, they and Anne Borge Johannesen from the department always gave me timely support and help, for which I am extremely grateful! They also encouraged me to give presentations at academic conferences, and my Ph.D. journey has benefitted a lot from these precious opportunities. I also thank the other professors in the Department of Economics for their valuable comments on my presentation.

I would also like to thank Professor Hanming Fang, who hosted me at the University of Pennsylvania, for introducing me to participating in seminars organized by the Ronald O. Perelman Center for Political Science & Economics and Wharton's business school. I also thank Professor Arie Kapteyn, who hosted me at the Center for Economic and Social Research (CESR), USC, and found me suitable projects.

I want to thank the Norwegian Public Roads Administration for providing my Ph.D. research funding, and COMPutational Economics and Optimization (COMPAMA) for providing funding for my research stay in the US. I also thank

the Norwegian public health administration project members' timely help in delivering the data I need.

Furthermore, I would also like to thank Anne Larsen from the department for her efficient and fast assistance with all practical issues. The Department of Economics created a positive and friendly atmosphere to make the Ph.D. journey so enjoyable.

Finally, I thank my family; they have always tried their best to encourage me to do what I want and supported me along the way!

The Ph.D. experience was the happiest episode of my life thus far; it had its challenges, but it was most certainly exciting. Thanks to everyone I have met.

Contents

Introduction

This thesis consists of three separate chapters. It uses machine learning and traditional statistical approaches to explore air pollution, mortality, and different aspects of transportation to draw forth policy recommendations regarding climate change and sustainable development.

Urban motor vehicle exhaust, as a negative transportation externality (Moretti and Neidell 2011; Currie and Neidell 2005), has contributed to global warming and accelerated the speed of climate change, making sustainable economic and environmental development an urgent need facing the world today. The growth of the economy, the increasing sale and usage of automobiles, and the emissions caused by motor vehicles emissions are the most important sources of pollution in the urban area. Urban cities are characterized by high population density and concentrated economic activities, as well as pollutants from motor vehicle exhaust, including nitrogen oxides, particulate matter, and carbon monoxide, can all cause personal health hazards in such urban spaces (Currie et al., 2009; Chay and Greenstone 2003). Moreover, more serious is possible cause greater city-level disasters: Classic examples include the photochemical smog event in Los Angeles in the United States from 1940 to 1960, the Great Smog Event in London in 1952, and the winter smog event in Beijing, China in recent years.

Reducing air pollution and its negative impacts relies on effective policy recommendations. Concerns about environmental degradation caused by economic growth have also prompted researchers and policymakers to continuously propose new economic solutions to environmental problems. However, traffic-related air pollution has the characteristics of mobility, high emission concentration, and wide regional influence, which increases the difficulty and complexity of policy formulation and implementation.

The interaction of meteorological factors and air pollutants adds to the challenges of air pollution research. As a chemical compound, air emissions will be transformed into new pollutants under certain conditions. Meteorological factors have an impact on the concentration of air pollutants. For example, particulate matter decreases significantly with an increase in snowfall and precipitation. Pollutants in the air will be increasingly dispersed as wind speed rises, thereby reducing their concentration.

Traditional statistical methods try to explain these links, but when faced with small samples and high-dimensional complex data, they are prone to overfitting, and traditional methods are rarely used for prediction problems. We, therefore, propose several machine learning approaches as an emerging technology that has the potential to address these issues.

In the first two chapters, we study the health consequences of traffic-induced air pollution and expect to make effective policy recommendations. We do this by analyzing several administrative data from the Norwegian government, comparing the prediction performance of machine learning and traditional statistical approaches, while also making air quality and traffic control policy recommendations aimed at improving urban air quality.

We choose Norway as a focal point for a range of reasons including its varied geography, which results in significant differences in climatic conditions among its major cities. For example, Oslo, the largest industrial city in Norway, is close to the strait and borders the mainland. The climate is relatively mild throughout the year and has four distinct seasons, while Tromsø, located in the Arctic Circle, is composed of small islands and experiences low temperatures all year round, as well as snowfalls in most months; cities in central Norway have different climatic characteristics. At the same time, as a wealthy country, Norway's overall air quality is good, so research on further improving air quality on this basis is more challenging than in countries with poor air quality.

The third chapter explores the relationship between alcohol supply and traffic accidents in Norway. Drunk driving and traffic accidents are one of the focal issues of global concern, especially fatal traffic accidents. Countries have also enacted strict policies for drunk driving, including its characterization as a criminal offense. Previous studies have focused on the relationship between on-premises alcohol sales and traffic accidents, but few studies have examined the relationship between broader off-premises alcohol availability and traffic accidents. We examine this issue using Norway as an example. The Norwegian government's monopoly on its provision of high-strength alcohol industry gives us an advantage in better measuring national alcohol availability. We use 20-year monthly data of the municipalities in Norway, including various types of traffic accidents, such as the severity of accident injuries, the number of vehicles involved, etc. We divide the municipalities into two groups for research, one is the municipality without alcohol stores, and the other is the municipality with existing alcohol stores but the number of stores has expanded, and conduct a comparative study. We find that municipalities that open their first store increase traffic accident risk, however, an expansion of alcohol stores reduces traffic accident risk. More findings are from broader socioeconomic measures, and finally, we draw relevant road safety policy recommendations.

Chapter 1

The first paper uses machine learning methods to predict the relationship between traffic with air pollution, under different meteorological conditions. Air pollution from urban traffic density (Kendrick et al., 2015) increases respiratory morbidity and mortality, especially among those living near highways (Font & Fuller, 2016). Many policies aimed at improving air quality have been introduced (Green et al., 2016; Green & Krehic, 2022; Parry et al., 2007). However, the challenge for research is that effective policies depend on a better understanding of the relationship between traffic and air pollution, especially since the mechanism is complicated by the confounding issue of meteorological factors. The concentration of air pollution caused by motor vehicle exhaust (Gualtieri et al., 2015) will decrease or increase due to weather factors such as wind speed, temperature, and air pressure (Kamińska, 2018). The second challenge is that traditional prediction tools are prone to overfitting and low prediction accuracy when faced with small samples and high-dimensional data, and we use machine learning approaches to try to solve this problem.

This paper uses high-frequency hourly data from Oslo, Norway, for the whole year of 2019 to estimate the interrelationship between traffic volume, air pollution, and meteorological conditions. First, I estimate this relationship using an Autoregressive Moving Average dynamic linear (ARDL) model, considering their interactions and lagged effects. Then, I split the data into ten datasets based on seasonal variations in snow depth and temperature throughout the year, as well as the four seasons of the year. Two machine learning algorithms, a support vector machine, and a decision tree, are used to compare their prediction performance with the two traditional statistical models, respectively the Autoregressive Moving Average with exogenous input variables (ARMAX) model and the ARDL model. I also try to explore the influence of seasonal and meteorological subset division methods on improving prediction accuracy.

The results show that in ten datasets, traditional statistical models outperform machine learning in predicting air pollution, and emphasize the importance of modeling considering interaction terms, time variables, and lag terms. These results also suggest policy recommendations, such as effective road pricing, need to consider external traffic factors such as weather conditions.

Chapter 2

Air pollution and extreme meteorological conditions can lead to increased cardiovascular and respiratory disease (CPD) mortality. There exists a lack of research on the impact of the interaction of these two factors on CPD mortality. At the same time, previous studies have focused primarily on the impact of extreme temperature conditions on CPD mortality, rather than other extreme meteorological factors. Finally, previous studies on population subgroups, such as the Norwegian Young Adult Mortality Study, are also scarce (Næss et al., 2007). We suggest these effects may vary across age groups. This motivates a focus on different age subgroups. In this chapter, we explore the interactions of traffic, pollution, and meteorological factors in affecting CPD mortality. As part of this, we use a machine learning approach to predict traffic and air pollution and their impact on CPD mortality.

We focus on four cities in Norway, namely Oslo, Bergen, Trondheim, and Tromsø, using daily data from 2009 to 2018. We include a range of traffic flow, nitrogen oxides, particulate matter, and meteorological variables. We established a random forest model to explore the key factors affecting CPD mortality, and we further use the regression models to consider the interaction and lagged effects. We find that the interaction of meteorology and air pollution can reduce CPD mortality, and we demonstrate that besides air temperature and air pollution, other extreme meteorological factors can also lead to an increase in CPD mortality. We also illustrate that seasonal effects are most pronounced for older people (over 75 years old). Finally, we show that machine learning has better predictive performance for

CPD mortality. We also follow up with policy recommendations that could potentially reduce CPD mortality.

Chapter 3

There has been much empirical research on the impact of alcohol on traffic accidents. Current literature generally focuses on on-premises alcohol, i.e., alcohol sold in bars or restaurants. However, there are also significant differences in off-promise alcohol availability and alcohol store distribution, which can have a major impact on traffic safety. Alcohol availability has been considered an uncontrolled covariate when studying alcohol consumption and fatal accidents. Our research question focuses on the relationship between alcohol availability and multiple types of traffic accidents. We do this using data from Norwegian municipalities, spanning 20 years. High-strength alcohol is only available at government monopoly stores in Norway, so this provides a clean environment for measuring alcohol availability. We explore municipality variation in store openings over time. We include different types of traffic accidents according to the degree of injury, traffic accidents happening on the weekend or night, the number of vehicles involved, etc., to investigate the impact of alcohol availability on traffic safety.

We find that opening the first alcohol store is associated with the increase in traffic accidents, by slightly more than two light injuries traffic accident per year. For municipalities that already had stores, increases in the number of stores, and traffic accidents will reduce by around three-quarters per month. Both groups primarily affect light injuries traffic accidents, with little or almost no impact on serious or fatalities injuries traffic accidents. Additional findings come from broader socioeconomic measures where we include younger drivers and gender. We find no effect of population distribution on the link between alcohol availability and traffic accidents. Substantial changes in the availability of alcohol can have a substantial impact on traffic accident rates, we then provide corresponding road safety policy recommendations.

References

Chay, K. Y., & Greenstone, M. (2003). The impact of air pollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession. The Quarterly Journal of Economics, 118(3), 1121–1167, https://doi.org/10.1162/00335530360698513

Currie, J., & Neidell, M. (2005). Air pollution and infant health: What can we learn from California's recent experience? The Quarterly Journal of Economics, 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 New Jersey. Journal of Health Economics, 28(3), 688–703. https://doi.org/10.1016/j.jhealeco.2009.02.001

Font, A., & G. W. Fuller (2016). Did policies to abate atmospheric emissions from traffic have a positive effect in London? Environmental Pollution, 218, 463–474, ISSN 0269-7491,https://doi.org/10.1016/j.envpol.2016.07.026

Giacopassi, D., & Winn, R. (1995). Alcohol availability and alcohol-related crashes: Does distance make a difference? American Journal of Drug and Alcohol Abuse, 21(3), 407–416. https://doi.org/10.3109/00952999509002706

Green, C. P., Heywood, J. S., & Navarro, M. (2016). Traffic accidents and the London congestion charge. Journal of Public Economics, 133, 11–22. https://doi.org/10.1016/j.jpubeco.2015.10.005

Green, C., & Krehic, L. (2022). An extra hour was wasted. Bar closing hours and traffic accidents in Norway. Health Economics (United Kingdom). https://doi.org/10.1002/hec.4550

Gualtieri, G., Crisci, A., Tartaglia, M., Toscano, P., & Gioli, B. (2015). A statistical model to assess air quality levels at urban sites. Water, Air, and Soil Pollution, 226(12). https://doi.org/10.1007/s11270-015-2663-4

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. Journal of Environmental Management, 217, 164–174, ISSN 0301-4797,https://doi.org/10.1016/j.jenvman.2018.03.094

Kendrick, C. M., Koonce, P., & George, L. A. (2015). Diurnal and seasonal variations of NO, NO_2 and $PM_{2.5}$ mass as a function of traffic volumes alongside an urban arterial, Atmospheric Environment, 122, 133–141, ISSN 1352-2310,https://doi.org/10.1016/j.atmosenv.2015.09.019

Moretti, E., & Neidell, M. (2011). Pollution, health, and avoidance behavior: Evidence from the ports of Los Angeles. Journal of Human Resources, 46(1), 154–175. https://doi.org/10.3368/jhr.46.1.154

Næss, Ø., Nafstad, P., Aamodt, G., Claussen, B., & Rosland, P. (2007). Relation between the concentration of air pollution and cause-specific mortality: Four-year exposures to nitrogen dioxide and particulate matter pollutants in 470 neighborhoods in Oslo, Norway. American Journal of Epidemiology, 165(4), 435–443. https://doi.org/10.1093/aje/kwk016

Parry, I. W., Walls, M., & Harrington, W. (2007). Automobile externalities and policies. Journal of Economic Literature, 45(2), 373–399.

Pasquier, A., & André, M. (2017). Considering criteria related to spatial variabilities for the assessment of air pollution from traffic. Transportation Research Procedia, 25(June), 3354–3369. https://doi.org/10.1016/j.trpro.2017.05.210

Chapter 1

How to Better Predict the Effect of Urban Traffic and Weather on Air Pollution? Norwegian Evidence from Machine Learning Approaches[*]

Cong Cao

Department of Economics, Norwegian University of Science and Technology,
Høgskoleringen 1, 7491 Trondheim, Norway

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 ten 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 support vector machine and decision tree in predicting air pollution. At the same time, I also explored the effect of dividing the seasons and 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.

Keywords: Machine Learning, Urban Traffic, Air Pollution, Transportation Policy

[*] Acknowledgment: I am grateful for research funding from Statens Vegvesen (the Norwegian public roads administration). I thank Colin Green and Jan Morten Dyrstad for their comments, which helped to improve this paper substantially. I also benefited from comments that were made by participants at the Department of Economics seminar – Norwegian University of Science and Technology, the 44th meeting of the Norwegian Association of Economists (Forskermøtet), and the 92nd Annual Meeting of the Southern Economic Association (SEA) in 2022.

JEL Classification: C53, C52, Q5, R4

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 et al., 2003 & 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., 2020), 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; 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 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 & 2019), 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 rely 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 have the potential to address these problems, as well as are known to be able to handle high-dimensional and nonlinear nonseparable 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

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

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) 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.

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 dividing the whole year's data into nine 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 (Cleveland et al., 1988; Deters et al., 2020; Chen et al., 2021; Morande et al., 2022). 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.

Grange et al. (2018) use 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., 2001).

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 backpropagation 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. 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 artificial neural networks, and adaptive neuro-fuzzy inference systems. Hence, SVM provides more accurate predictions, which leads me to choose SVM as my approach of choice.

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. However, this paper does not consider interactions between different predictors, for example, likely interactions between wind direction and wind speed. 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 the relationship between other meteorological factors and NOx emissions.

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.

The development of effective transport policies aimed at improving air quality remains challenging. 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.

Research Questions and Hypotheses

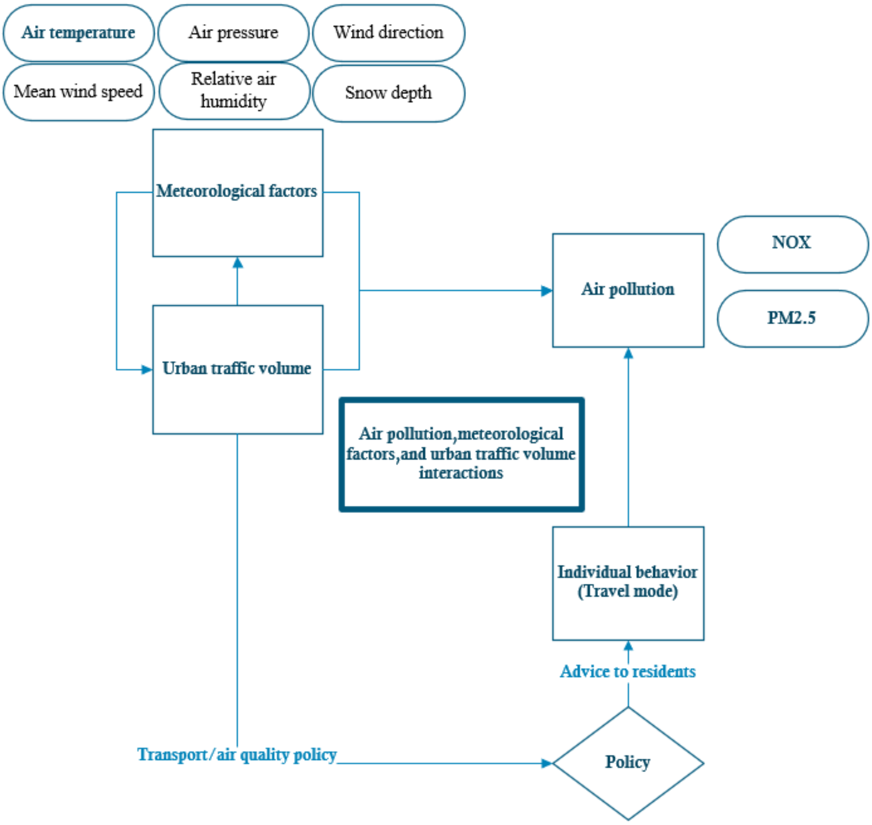
Therefore, I propose two research questions: (1) Under the interaction of traffic, weather, and air pollution, what is the impact of traffic and weather on air pollution? (2) Which approach can better predict traffic-related air pollution, machine learning or traditional statistical approaches? Figure 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 Figure 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 $\mathrm{PM}_{2.5}$ 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.

Figure 1

Flow Chart 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:

$$P_t = f\,(T_t, M_t) \quad (1)$$

Here P_t is a pollutant, NOx or $PM_{2.5}$, T_t is traffic volumes. $M_{t,j}$ are meteorological variables, subscript j is the metrological variables number, from 1, 2…J. It includes j_1 = air temperature, j_2 = wind direction, j_3 = mean wind speed, j_4 = snow depth, j_5 = relative air humidity, j_6= air pressure. Subscript t is time, from 1,2, 3… T, 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 an ARMAX model as follows:

$$P_t = f\,(T_t, M_{t,j}, \mathrm{lag}) \ (2)$$

And an ARDL model is as follows:

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

Here I also 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.

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

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.

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

Table 1

Equation (5)'s Equation Term Explanation

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

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

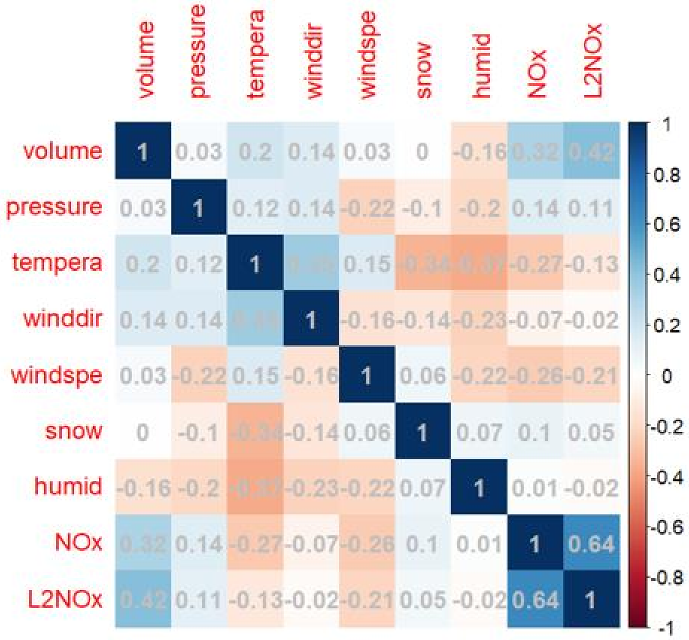
There exist several complications to estimating the model. First, because the inclusion of more weather variables complicates the impact of traffic on air pollution, there will be higher demands on the model when estimating air pollution. Second, vehicle emissions undergo chemical reactions in the air. 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 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.

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



	volume	pressure	tempera	winddir	windspe	snow	humid	NOx	L2NOx
volume	1	0.03	0.2	0.14	0.03	0	-0.16	0.32	0.42
pressure	0.03	1	0.12	0.14	-0.22	-0.1	-0.2	0.14	0.11
tempera	0.2	0.12	1	[illegible]	0.15	-0.34	-0.37	-0.27	-0.13
winddir	0.14	0.14	[illegible]	1	-0.16	-0.14	-0.23	-0.07	-0.02
windspe	0.03	-0.22	0.15	-0.16	1	0.06	-0.22	-0.26	-0.21
snow	0	-0.1	-0.34	-0.14	0.06	1	0.07	0.1	0.05
humid	-0.16	-0.2	-0.37	-0.23	-0.22	0.07	1	0.01	-0.02
NOx	0.32	0.14	-0.27	-0.07	-0.26	0.1	0.01	1	0.64
L2NOx	0.42	0.11	-0.13	-0.02	-0.21	0.05	-0.02	0.64	1

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.

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

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.

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.

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 two specific approaches, SVM and DT. In theory, there are other alternative machine learning approaches, such as deep neural networks, 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 two of 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 two machine-learning methods can be found in Appendix 4.

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. 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. Since SVM is sensitive to missing values, so I performed missing value imputation at the very beginning.

Traffic Data

The traffic volume data from SVV is measured as the number of approved vehicle registrations during the relevant hour. Figure 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 nearest weather station to obtain the meteorological data.

Figure 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.

Figure 2

Traffic registration map with data monitoring stations

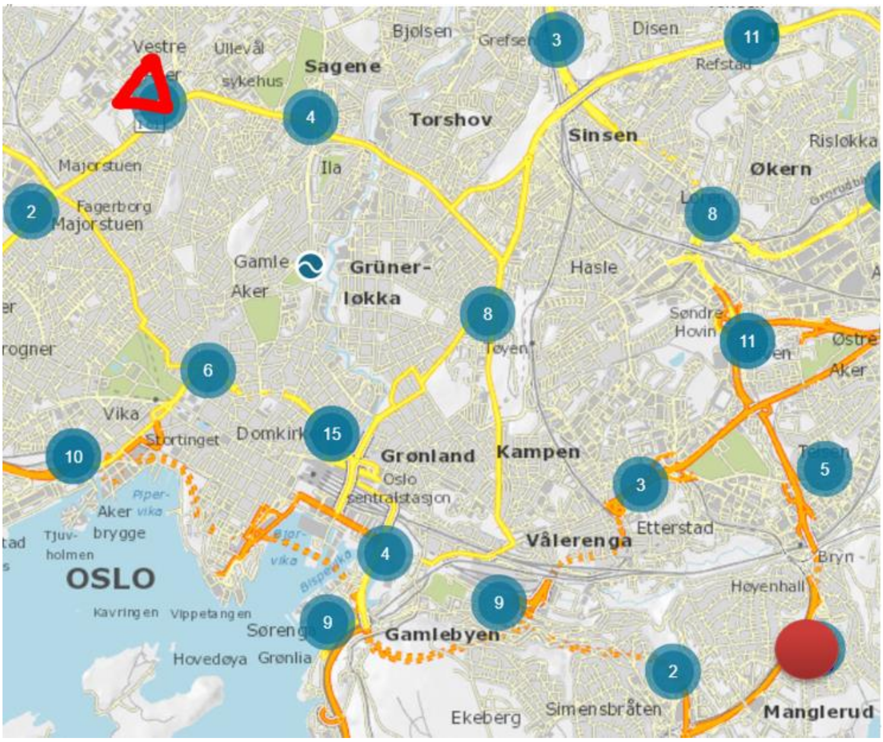
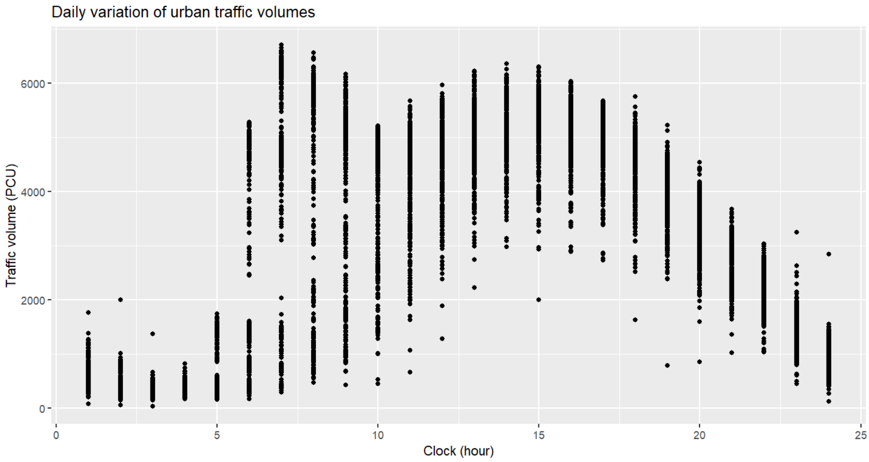


Figure 3 shows the daily variation in traffic volumes. These increase from 6:00, with the first peak at 7:00. The traffic volumes also increase from 10:00, and the second peak is at 14:00. The rush hour is from 7:00 to 9:00, and from 13:00 to 16:00.

Figure 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.

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. Figure 4 and Appendix 2 show the monthly variation of the metrological factors and traffic volumes. Figure 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 Figure 4 Panel B, it can be seen that 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.

As shown in Panel C in Figure 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.

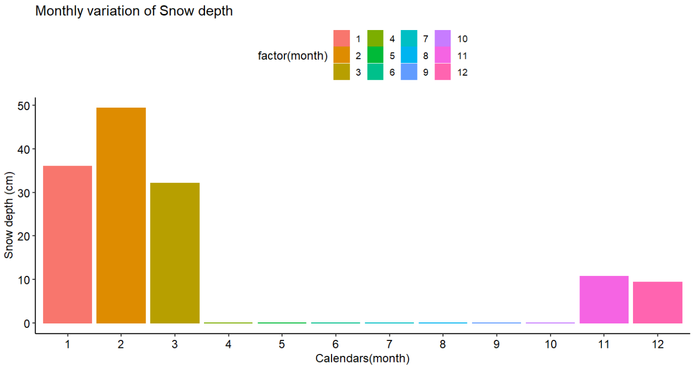
Ten datasets

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

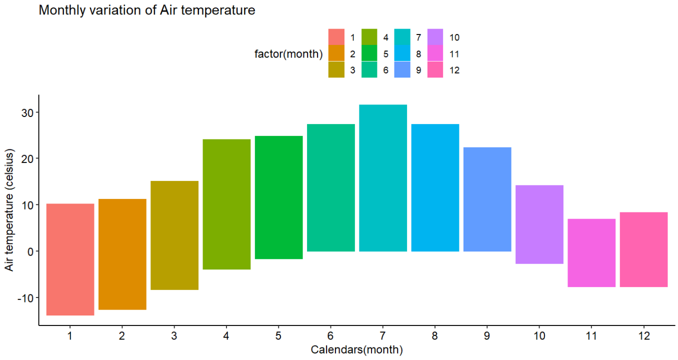
Figure 4

The Meteorological Variables with Seasonality

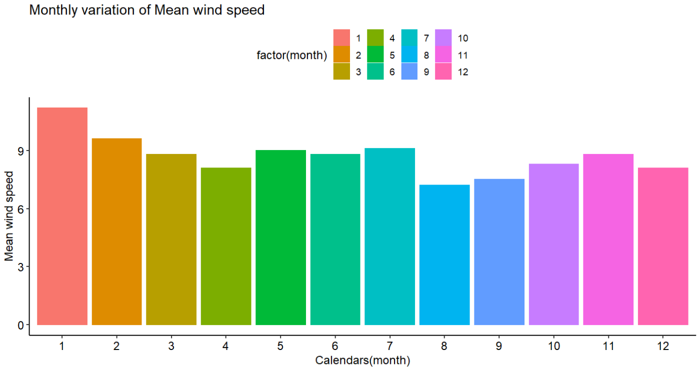
Panel A



Panel B



Panel C



Note: These demonstrate that the snow depth and air temperature have seasonality, while mean wind speed doesn’t have seasonal differences.

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}, $\mathrm{PM}_{2.5}$, 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), $\mathrm{PM}_{2.5}$ 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. Initially, all data are standardized using max-min normalization, $\mathrm{x}^* = \frac{\mathrm{x}-\mathrm{x}_{\min}}{\mathrm{x}_{\max}-\mathrm{x}_{\min}}$, which converts the original data to the range [0 1].

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.

	Variables	**NOx**	**$PM_{2.5}$**
Observations	**8760**		
Adjusted R^2	**0.74 (NOx as the dependent variable)**		
	0.81 ($PM_{2.5}$ as the dependent variable)		
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)

	Variables	**NOx**	**PM$_{2.5}$**
Observations	**8760**		
Adjusted R^2	**0.74 (NOx as the dependent variable)** **0.81 (PM$_{2.5}$ as the dependent variable)**		
	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)**
	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. This table has two pages.

	Variables	**NOx**	**$PM_{2.5}$**
Observations	**8760**		
Adjusted R^2	**0.73 (NOx as the dependent variable)** **0.78 ($PM_{2.5}$ as the dependent variable)**		
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***
		(0.007)	(0.008)
	L.ma	-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)
	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

	Variables	**NOx**	**$PM_{2.5}$**
Observations	**8760**		
Adjusted R^2	**0.73 (NOx as the dependent variable)**		
	0.78 ($PM_{2.5}$ as the dependent variable)		
		(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 two machine learning algorithms to predict air pollution concentrations, and then compare their prediction accuracy.

Figure 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 two machine learning algorithms. In the ten 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. At the same time, I compare the performance of ARMAX in these ten models. I find ARMAX has the smallest MAE, MSE, and RMSE when using the summer subset to predict NOx, the autumn subset to predict $PM_{2.5}$. This indicates that the prediction results of ARMAX depend greatly on seasonal factors, and in summer and autumn, the air quality is more dependent on weather, traffic, and time variables.

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 called a loss function in the field of machine learning, so it represents the predictive power.

Together these show that when predicting air pollution, the prediction power of the ARMAX model and ARDL model is nearly the same.

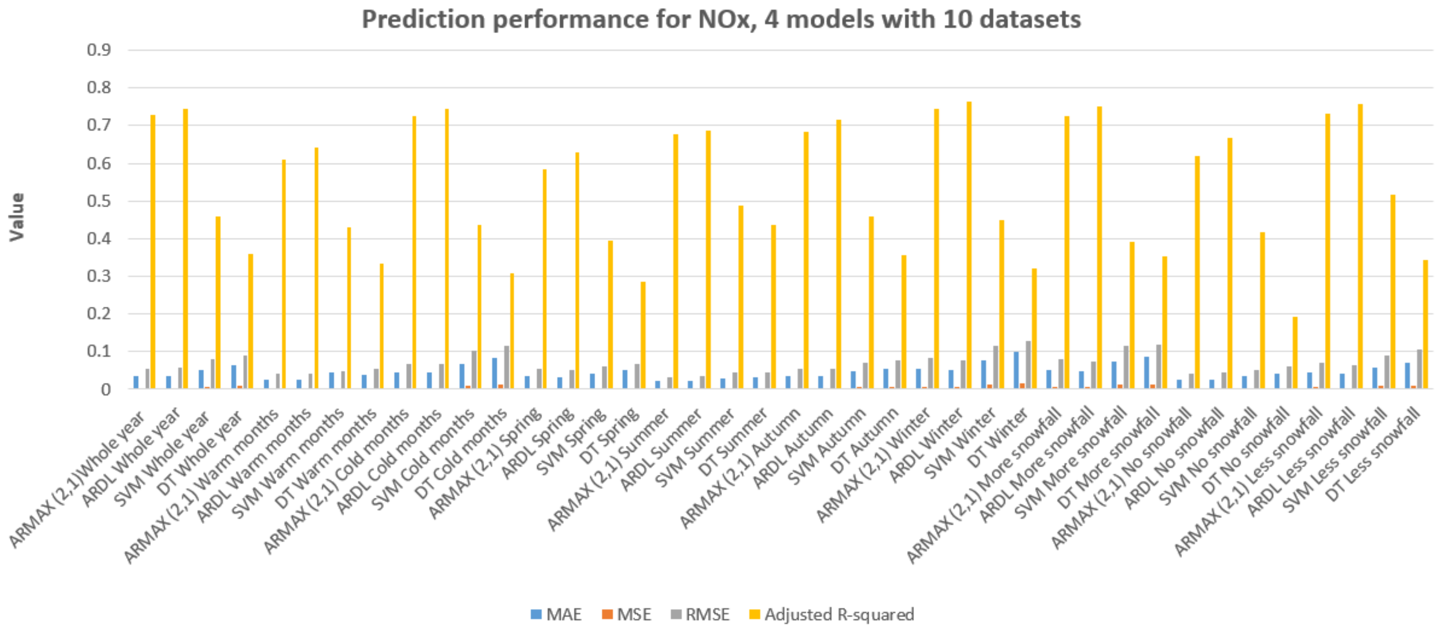
In Figure 5 Report A, for the NO_x concentration prediction, I find that in all the datasets, compared with the ARMAX model, the ARDL model has a larger adjusted R^2. In Figure 5 Report B, for the prediction of $PM_{2.5}$ concentration, I find that in the warm months, spring, no snowfall, and less snowfall, 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.

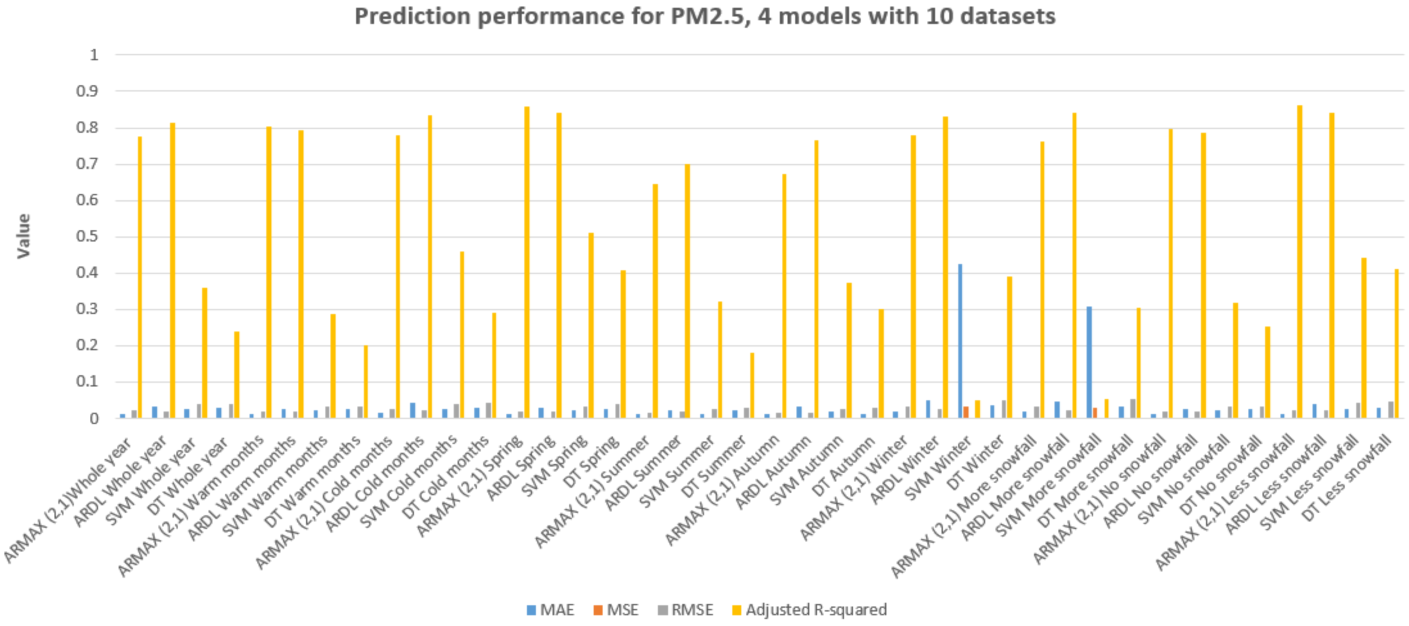
Figure 5 shows the prediction results of the four models, while a detailed prediction accuracy comparison table of the four models can be found in Appendix 5.

Figure 5

Prediction Accuracy for NOx and $PM_{2.5}$

Report A





Note: The x-axis in the figure is the four models in ten datasets, and the y-axis is the value of the model evaluation methods. Among them, blue represents the value of MAE, gray is RMSE, yellow is adjusted R^2, and orange is MSE.

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 divide the data into nine data subsets according to Norwegian meteorological and temporal variables, so the four models are evaluated ten 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 ten 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 recommend reducing going out.

References

Aldrin, M., & Haff, I. H. (2005). Generalized additive modeling of air pollution, traffic volume, and meteorology. Atmospheric Environment, 39(11), 2145–2155. https://doi.org/10.1016/j.atmosenv.2004.12.020

Amos, J.. Polluted air cause 5.5 million deaths a year new research says[EB/OL]. (2016-02-13)[2021-08-17].https://www.bbc.com/news/science-environment-35568249

Baena-Cagnani, C. E., Patiño, C. M., Cuello, M. N., Minervini, M. C., Fernández, A. M., Garip, E. A., … Gómez, R. M. (1999). Prevalence and severity of asthma and wheezing in an adolescent population. *International Archives of Allergy and Immunology*, 118(2–4), 245–246. https://doi.org/10.1159/000024087

Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, *15*(4), 780. doi 10.3390/ijerph15040780. PMID: 29673227; PMCID: PMC5923822

Bathmanabhan, S., & Saragur Madanayak, S. N. (2010). Analysis and interpretation of particulate matter - PM10, PM2.5, and PM1 emissions from the heterogeneous traffic near an urban roadway. *Atmospheric Pollution Research*, 1(3), 184–194. https://doi.org/10.5094/APR.2010.024

Bjørgen, A., & Ryghaug, M. (2022). Integration of urban freight transport in city planning: Lesson learned.*Transportation Research Part D: Transport* and *Environment*, *107*, 103310. ISSN 1361-9209,https://doi.org/10.1016/j.trd.2022.103310

Briggs, D. J., Collins, S., Elliott, P., Fischer, P. H., Kingham, S., Lebret, E., Pryl, K., Reeuwijk, H. V., Smallbone, K., & Veen, A. V. (1997). Mapping urban air pollution using GIS: A regression-based approach. *International Journal of Geographical Information Science, 11*, 699–718.

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. *Environmental Health: A Global Access Science Source*, 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 mortality in Nanjing, China: Evidence highlighting the roles of cumulative exposure and mortality displacement. *Chemosphere*, 265, 129035. https://doi.org/10.1016/j.chemosphere.2020.129035

Conte, M., & Contini, D. (2019). Size-resolved particle emission factors of vehicular traffic derived from urban eddy covariance measurements. *Environmental Pollution*, *251*, 830–838. doi: 10.1016/j.envpol.2019.05.029. Epub 2019 May 11. PMID: 31125813

Conte, M., Donateo, A., & Contini, D. (2018). Characterization of particle size distributions and corresponding size-segregated turbulent fluxes simultaneously with CO2 exchange in an urban area. *Science of the Total Environment*, 622–623, 1067–1078. https://doi.org/10.1016/j.scitotenv.2017.12.040

Currie, J., & Neidell, M. (2005). Air Pollution and Infant Health: What Can We Learn from California's Recent Experience*? The Quarterly Journal of Economics*, 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 New Jersey. *Journal of Health Economics*, 28(3), 688–703. https://doi.org/10.1016/j.jhealeco.2009.02.001

European Environmental Agency. (2020). Air quality in Europe 2020 report. *EEA Report.* Retrieved from https://www.eea.europa.eu//publications/air-quality-in-europe-2020-report

Folgerø, I. K., Harding, T., & Westby, B. S. (2020). Going fast or going green? Evidence from environmental speed limits in Norway. *Transportation Research Part D: Transport and Environment*, 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 positive effect in London? *Environmental Pollution*, 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 development from 10 to 18 years of age: A cohort study. *Lancet*, *369*(9561), 571–577. doi: 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 meteorological normalization models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics*, *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 charge. *Journal of Public Economics*, *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 Norway. *Health Economics*, 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 quality management of changes in air quality during lockdown in Auckland (New Zealand) in response to the 2020 SARS-CoV-2 epidemic. *Science of the Total Environment*, *746*,141129. doi: 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. *Journal of Ambient Intelligence and Smart Environments*, 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, and Soil Pollution*, *226*, 394 https://doi.org/10.1007/s11270-015-2663-4

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. *Journal of Environmental Management*, *217*, 164–174. ISSN 0301-4797, https://doi.org/10.1016/j.jenvman.2018.03.094

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. *Atmospheric Environment*, 122, 133–141. https://doi.org/10.1016/j.atmosenv.2015.09.019

Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on DWT Decomposition. *Procedia Computer Science*, 48(C), 173–179. https://doi.org/10.1016/j.procs.2015.04.167

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *Journal of Electrical and Computer Engineering*, 2017. 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 Climate*, 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 modeling simulations for hazardous air pollutants. *Atmospheric Environment, 40(26)*, 5087–5096. https://doi.org/10.1016/j.atmosenv.2005.12.044

Moazami, S., Noori, R., Amiri, B. J., Yeganeh, B., Partani, S., & Safavi, S. (2016). Reliable prediction of carbon monoxide using a developed support vector machine. *Atmospheric Pollution Research*, 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 ports of Los Angeles. *Journal of Human Resources*, 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. *Journal - of Economic Literature*, 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 assessment of air pollution from traffic. *Transportation Research Procedia*, 25(June), 3354–3369. https://doi.org/10.1016/j.trpro.2017.05.210

Pothirat, C., Chaiwong, W., Liwsrisakun, C., Bumroongkit, C., Deesomchok, A., Theerakittikul, T., … Phetsuk, N. (2019). Acute effects of air pollutants on daily mortality and hospitalizations due to cardiovascular and respiratory diseases. *Journal of Thoracic Disease*, 11(7), 3070–3083. https://doi.org/10.21037/jtd.2019.07.37

Qu, H., Lu, X., Liu, L., & Ye, Y. (2019). Effects of traffic and urban parks on PM10 and PM2.5 mass concentrations. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*, 45(2), 0–5647. https://doi.org/10.1080/15567036.2019.1672833

Shaban, B. K., Kadri, A., & Rezk, E. (2016). Urban air pollution monitoring system with forecasting models. In IEEE Sensors Journal, 16(8), 2598–2606. doi: 10.1109/JSEN.2016.2514378

Smallbone, K. (2000). A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. Science of The Total Environment, 253(1–3), 151–167. ISSN 0048-9697,https://doi.org/10.1016/S0048-9697(00)00429-0

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. Transport Policy, 99(August), 251–261. https://doi.org/10.1016/j.tranpol.2020.08.025

Wærsted, E. G., Sundvor, I., Denby, B. R., & Mu, Q. (2022). Quantification of the temperature dependence of NOx emissions from road traffic in Norway using air quality modeling and monitoring data. *Atmospheric Environment: X*, 13(x), 100160. https://doi.org/10.1016/j.aeaoa.2022.100160

Zhang, P. G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neu -rocomputing*, 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

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.
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.
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)[2]	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.

[2] 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
Pollutant	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

In the explanation of a pollutant’s negative values, I find that the percentage of missing values for the whole year dataset is 4.24％. I have conducted missing value imputation. 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.

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.

The Ten Datasets Used 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.

I have four subsets according to the season:

Dataset 4, spring, from March to May, includes 2208 observations.

Dataset 5, summer, from June to August, includes 2208 observations.

Dataset 6, autumn, from September to November, includes 2184 observations.

Dataset 7, winter, from December to February, includes 2160 observations.

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

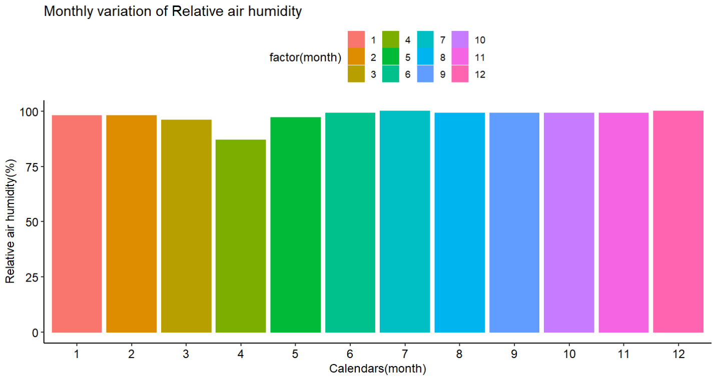
Dataset 8, months with more snowfall, when the snow depth is greater than 30cm, from January to March, includes 2159 observations.

Dataset 9, months without snowfall, when the snow depth is equal to 30cm, from April to October, includes 5136 observations.

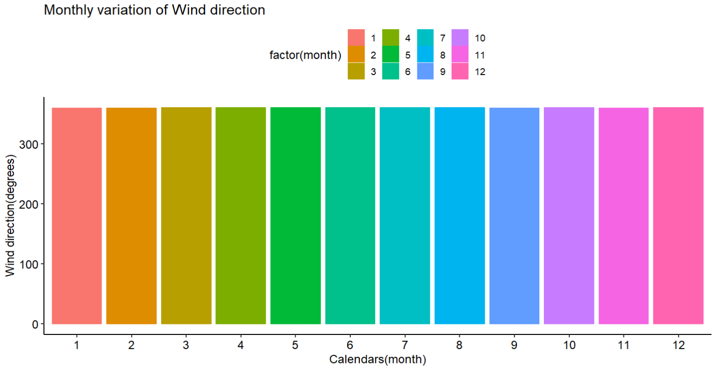
Dataset 10, months with less snowfall, when the snow depth is less than 15cm, in November and December, includes 1465 observations.

Appendix 2 Monthly Variation of Meteorological Factors

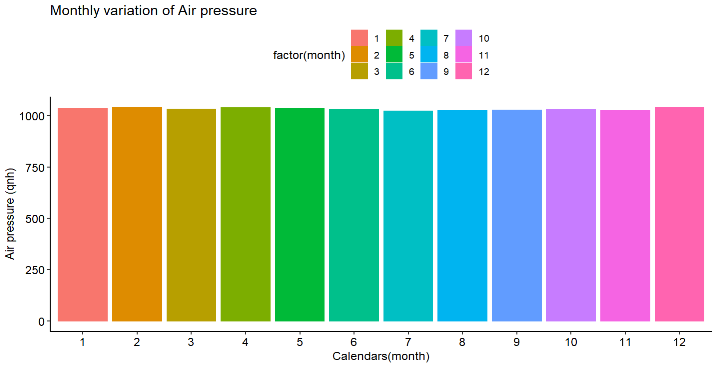
Panel A



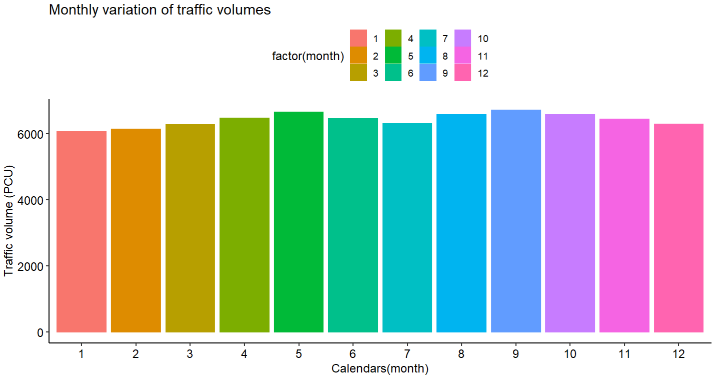
Panel B



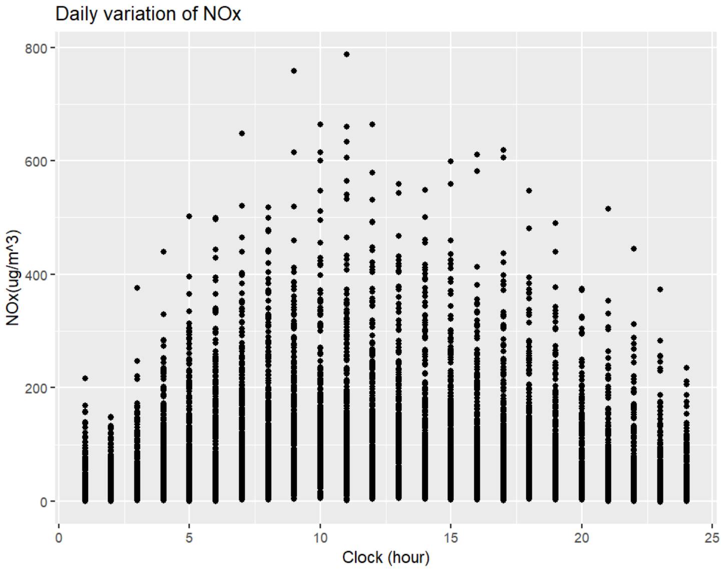
Panel C

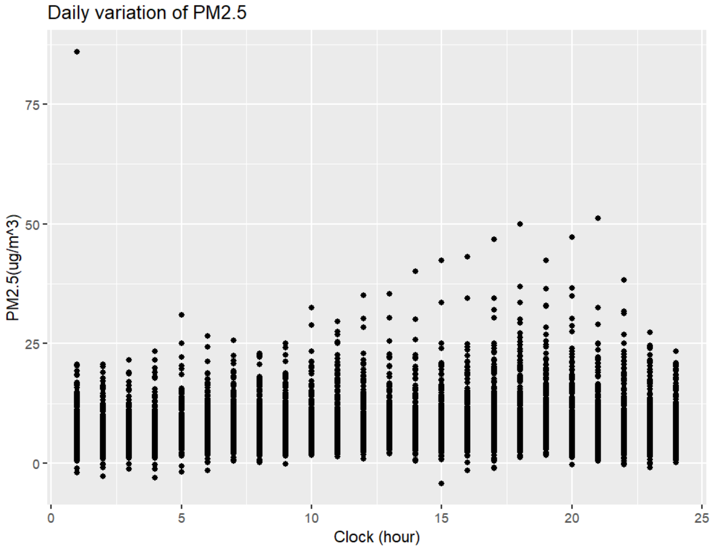


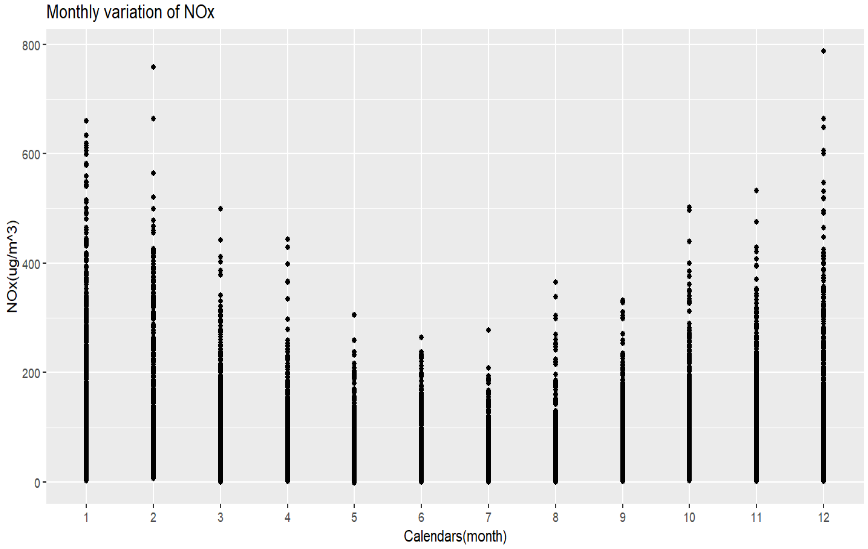
Panel D

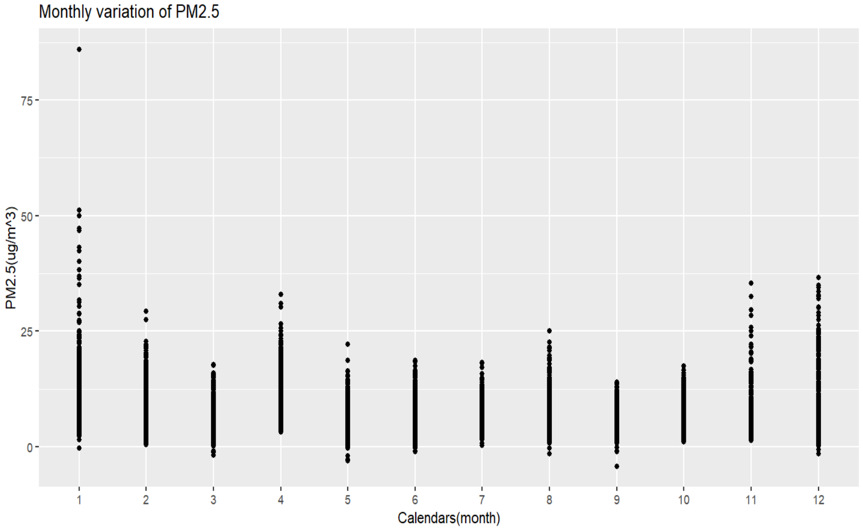

Monthly variation of traffic volumes
factor(month)
1
2
3
4
5
6
7
8
9
10
11
12
Traffic volume (PCU)
0
2000
4000
6000
1
2
3
4
5
6
7
8
9
10
11
12
Calendars(month)

Daily and Monthly Variations of Pollutants


Daily variation of NOx


Daily variation of PM2.5


Monthly variation of NOx
NOx(ug/m^3)
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
Calendars(month)


Monthly variation of PM2.5
PM2.5(ug/m^3)
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
Calendars(month)

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, and the computational complexity does not depend on the dimension of the eigenspace. Thus, it can avoid the "dimension disaster", so it is good at solving high-dimensional problems, and large eigenspaces and obtains a lower error rate.

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:

$$a^{*} = (a_1^{*}, \dots, a_n^{*})^T$$

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

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

（4）Constructing the decision boundary：

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

From this, I will have the decision function：

$$f(x) = \mathrm{sgn}\left(g(x)\right)$$

After constructing the decision boundary：

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

I have the decision equation

$$f(x_i) = \mathrm{sgn}\left(g(x)\right)$$

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 take air pollution as the dependent variable and traffic and weather as the independent variables. First, I sort the characteristics of each independent variable, and the cut point s is selected, and then I get:

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

Where $C_1$and $C_2$are:

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

Among them, $N_1$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:

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

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

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.

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.

Figure A

Time Series Diagram of NOx

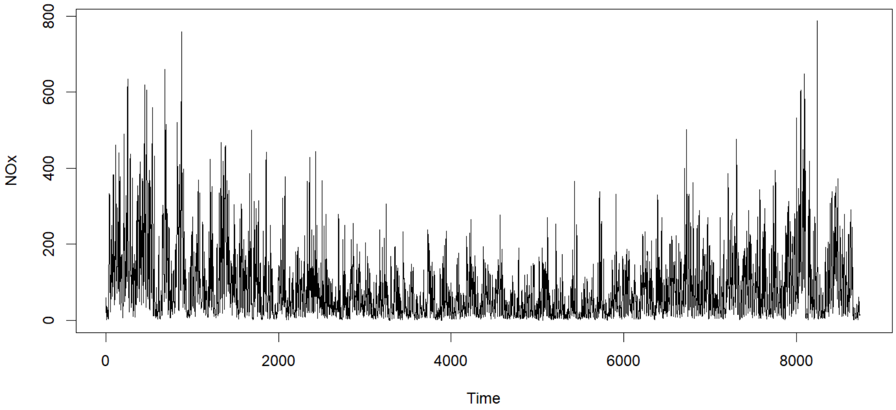
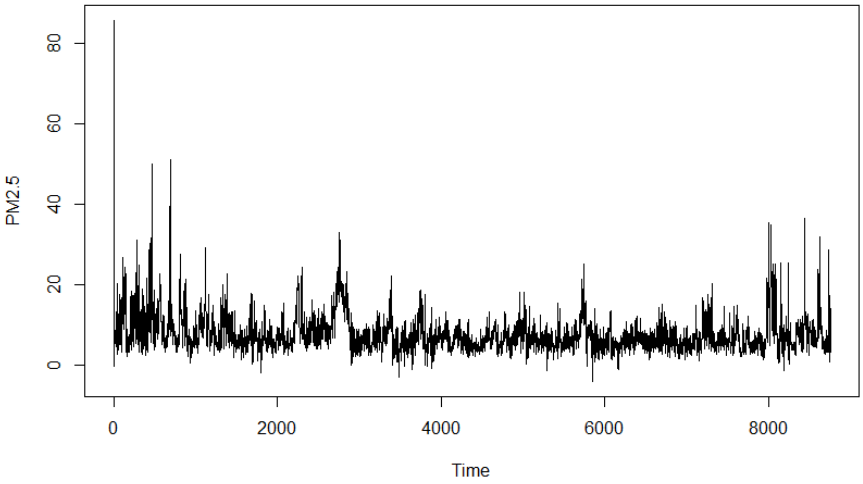


Figure B

Time Series Diagram of $PM_{2.5}$



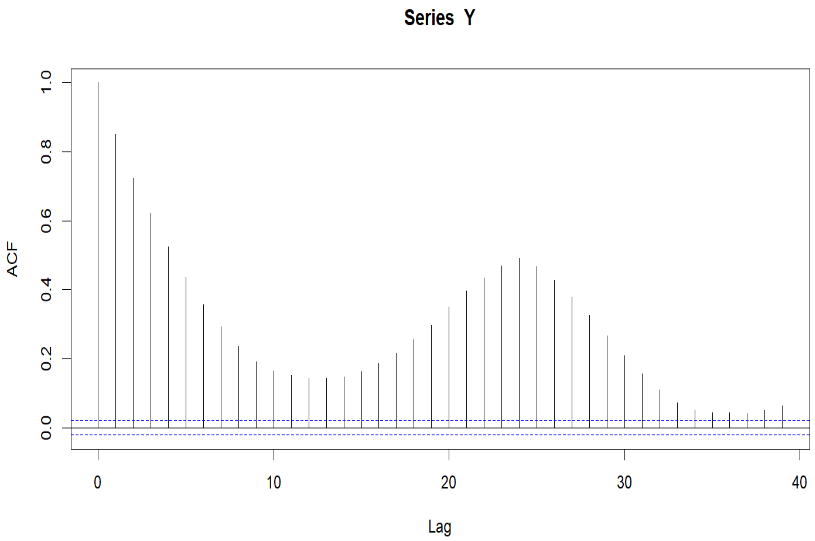
For further verification, I use the Augmented Dickey-Fuller Test to test for the stationarity of the time series. The Augmented Dickey-Fuller Test (ADF) is a modified version of the Dickey-Fuller Test, that excludes the influence of autocorrelation. The null hypothesis is that the data are nonstationary. 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.

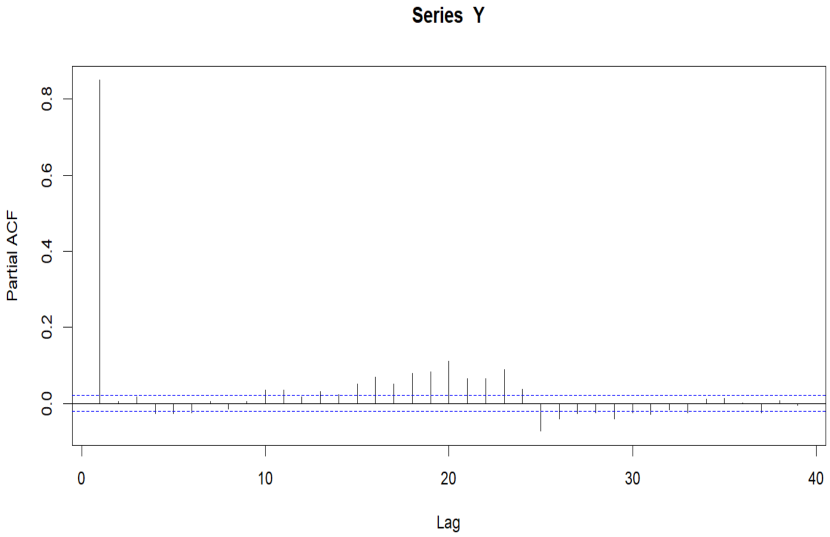
Autocorrelation Function (ACF) refers to the linear relationship between the sequence value and the lag value at any time t $(\mathrm{t} = 1,2,3, \ldots, \mathrm{n})$. An ACF plot, also known as a "correlogram," refers to a plot with the lag value i as the x-axis and the autocorrelation coefficient as the y-axis. A correlation coefficient value between $\mathrm{X_t}$ and $\mathrm{X_{t-i}}$ 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 2^{nd}-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.

Figure C

ACF and PACF Graph





The best model with the smallest AIC[3] 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, we selected the four ARMAX models most likely to have the smallest AIC values, using NO_x as the dependent variable, and we compared their AIC. We found that the AIC of ARMAX (2,0,1) = 72946.03, which is the smallest AIC value, meaning this ARMAX model is the best.

[3] $AIC = （2k - 2L）/n$.
$L = -(n/2) * \ln(2 * pi) - (n/2) * \ln(sse/n) - n/2$, 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
$\boldsymbol{Sigma^2}$	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, we selected the four ARMAX models most likely to have the smallest AIC values, using $PM_{2.5}$ as the dependent variable, and we compared their AIC. We 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 Four Models in Ten Datasets

I use two machine learning algorithms, SVM 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^2**
	Whole year	ARMAX (2,1) Whole year	**0.034**	**0.0031**	**0.056**	0.7285
		ARDL Whole year	0.0351	0.0032	0.0566	**0.7438**
		SVM Whole year	0.0500	0.0064	0.0800	0.4573
		DT Whole year	0.0626	0.0082	0.091	0.3596
	Warm months	ARMAX (2,1) Warm months	**0.0248**	**0.0017**	**0.0408**	0.6097
		ARDL Warm months	0.0254	0.0017	0.0409	**0.6353**
		SVM Warm months	0.0445	0.0025	0.0496	0.429
		DT Warm months	0.0380	0.0029	0.0543	0.3348
	Cold months	ARMAX (2,1) Cold months	0.0447	0.0047	0.0689	0.7240
		ARDL Cold months	**0.0441**	**0.0046**	**0.0681**	**0.7409**
		SVM Cold months	0.0676	0.0107	0.1036	0.4379
		DT Cold months	0.0827	0.0132	0.115	0.3079
	Spring	ARMAX (2,1) Spring	0.0339	0.003	0.055	0.5850
		ARDL Spring	**0.0312**	**0.0026**	**0.0506**	**0.6189**
		SVM Spring	0.0416	0.0038	0.0613	0.3948
		DT Spring	0.0500	0.0044	0.0667	0.2856
	Summer	ARMAX (2,1) Summer	**0.0217**	**0.0011**	**0.0336**	0.6762
		ARDL Summer	0.0218	0.0012	0.0348	**0.6781**
		SVM Summer	0.0277	0.0019	0.0434	0.4866
		DT Summer	0.0314	0.0021	0.0455	0.4359
	Autumn	ARMAX (2,1) Autumn	**0.0345**	**0.0028**	**0.053**	0.6851
		ARDL Autumn	0.035	0.0028	0.05301	**0.7067**
		SVM Autumn	0.0491	0.005	0.0704	0.4589
		DT Autumn	0.0543	0.0059	0.0767	0.3569
	Winter	ARMAX (2,1) Winter	0.054	0.0067	0.0822	0.7448
		ARDL Winter	**0.0501**	**0.0058**	**0.0767**	**0.7559**
		SVM Winter	0.0780	0.0136	0.1167	0.4480
		DT Winter	0.0981	0.0167	0.1294	0.3217
	More snowfall	ARMAX (2,1) More snowfall	0.0520	0.0062	0.0789	0.7251
		ARDL More snowfall	**0.0476**	**0.0053**	**0.073**	**0.7435**
		SVM More snowfall	0.0745	0.0131	0.1146	0.3929
		DT More snowfall	0.0856	0.014	0.1185	0.3517
	No snowfall	ARMAX (2,1) No snowfall	**0.0252**	**0.0017**	**0.0411**	0.6178
		ARDL No snowfall	0.0272	0.0019	0.0436	**0.6635**
		SVM No snowfall	0.0344	0.0027	0.0517	0.4167
		DT No snowfall	0.0428	0.0037	0.0608	0.1928
	Less snowfall	ARMAX (2,1) Less snowfall	0.0444	0.0048	0.0693	0.7323
		ARDL Less snowfall	**0.0416**	**0.0043**	**0.0653**	**0.7472**
		SVM Less snowfall	0.0583	0.0083	0.091	0.5160
		DT Less snowfall	0.0708	0.0112	0.106	0.3434

PM$_{2.5}$		**Model**	**MAE**	**MSE**	**RMSE**	**Adjusted R^2**
	Whole year	ARMAX (2,1) Whole year	**0.0144**	**0.0005**	**0.0232**	0.7756
		ARDL Whole year	0.0351	0.0005	0.0214	**0.8128**
		SVM Whole year	0.0260	0.0015	0.0392	0.3601
		DT Whole year	0.0293	0.0016	0.0400	0.2407
	Warm months	ARMAX (2,1) Warm months	**0.0129**	**0.0003**	**0.0187**	**0.8024**
		ARDL Warm months	0.0254	0.0004	0.0190	0.7894
		SVM Warm months	0.0236	0.0012	0.0340	0.2883
		DT Warm months	0.0261	0.0013	0.0340	0.2013
	Cold months	ARMAX (2,1) Cold months	**0.0146**	0.0007	0.0255	0.7811
		ARDL Cold months	0.0441	**0.0005**	**0.0229**	**0.8333**
		SVM Cold months	0.0251	0.0015	0.0389	0.4615
		DT Cold months	0.0309	0.0020	0.0447	0.2910
	Spring	ARMAX (2,1) Spring	**0.0131**	**0.0003**	**0.0187**	**0.8599**
		ARDL Spring	0.0312	0.0004	0.0193	0.8379
		SVM Spring	0.0244	0.0012	0.0352	0.5110
		DT Spring	0.0278	0.0015	0.0387	0.4093
	Summer	ARMAX (2,1) Summer	**0.0121**	**0.0003**	**0.0177**	0.6467
		ARDL Summer	0.0218	0.0003	0.0182	**0.6923**
		SVM Summer	0.0120	0.0007	0.0273	0.3206
		DT Summer	0.0228	0.0009	0.0300	0.1829
	Autumn	ARMAX (2,1) Autumn	**0.0119**	0.0003	0.0172	0.6734
		ARDL Autumn	0.0350	**0.0003**	**0.0168**	**0.7604**
		SVM Autumn	0.0183	0.0007	0.0271	0.3740
		DT Autumn	0.0120	0.0008	0.0286	0.3033
	Winter	ARMAX (2,1) Winter	**0.0195**	0.0012	0.0340	0.7780
		ARDL Winter	0.0501	**0.0008**	**0.0280**	**0.8252**
		SVM Winter	0.4240	0.0324	0.0025	0.0500
		DT Winter	0.0376	0.0026	0.0515	0.3897
	More snowfall	ARMAX (2,1) More snowfall	**0.0184**	0.0011	0.0325	0.7639
		ARDL More snowfall	0.0476	**0.0006**	**0.0241**	**0.8364**
		SVM More snowfall	0.3087	0.0288	0.0030	0.0550
		DT More snowfall	0.0325	0.0030	0.0552	0.3042
	No snowfall	ARMAX (2,1) No snowfall	**0.0131**	0.0004	0.0190	**0.7970**
		ARDL No snowfall	0.0272	**0.0003**	**0.0187**	0.7838
		SVM No snowfall	0.0231	0.0011	0.0332	0.3199
		DT No snowfall	0.0253	0.0012	0.035	0.2524
	Less snowfall	ARMAX (2,1) Less snowfall	**0.0135**	**0.0005**	**0.0219**	**0.8621**
		ARDL Less snowfall	0.0416	0.0006	0.0237	0.8340
		SVM Less snowfall	0.0271	0.0020	0.0446	0.4436
		DT Less snowfall	0.0294	0.0021	0.0458	0.4113

Chapter 2

Using Support Vector Machine and Decision Tree to Predict Mortality Related to Traffic, Air Pollution, and Meteorological Exposure in Norway*

Cong Cao [a], Jan Morten Dyrstad [a], Shilpa Rao [b], Colin P. Green [a]

[a] Department of Economics, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway

[b] Division for Infectious Diseases and Environmental Health, Norwegian Institute of Public Health, PO Box 222 Skøyen, 0213 Oslo, Norway

Abstract

Cardiovascular and respiratory disease (CPD) is a leading cause of death worldwide. There is increasing evidence that air pollution and exposure to extreme weather conditions have important contributory roles. In practice, understanding the interaction of these factors is difficult due to the complexity of the relationship between CPD, air pollution, and environmental factors in general. This paper returns to this point and uses machine learning approaches to explore these relationships focusing on four cities in Norway, as well as investigating whether meteorological factors and air pollution have a synergistic effect on CPD. We demonstrate that machine learning outperforms regression models in terms of the accuracy of predicting CPD mortality, as regression models are prone to overfitting with the increase in variables. We show the importance of the interaction between weather and air pollution. We demonstrate that extreme weather is associated with higher CPD mortality, as is exposure to air pollution in the form of NOx and particulate matter. These effects are most pronounced for older 75-year-old individuals. Our results suggest policy responses for mitigating the negative health impacts, especially for vulnerable age subgroups.

* Acknowledgment: Cong Cao is grateful for the research funding from Statens vegvesen (the Norwegian public roads administration) and also acknowledges data received from the European Union's Horizon 2020 Project EXHAUSTION (Grant JO: 820655). She is also grateful for the comments that were made by participants at the Norwegian Public Health Institution Method Development and Analytics Department’s meetings, the Nordic Annual Environmental and Resource Economics (NAERE) Conference (Uppsala, Sweden), the Urban Transitions Conference (Sitges, Spain) in 2022, and the Statistics Norway Research Department (Oslo, Norway) in 2023.

E-mail: cong.cao@ntnu.no

Keywords: Machine Learning, Air Pollution, Mortality, Health Policy

JEL Classification: C53, C52, I18, Q5, R4

1 Background

Every year approximately 17.9 million people die from cardiovascular disease (CVD); which accounts for 32.8% of all deaths worldwide (Folkehelseinstituttet, 2021). There is growing evidence showing that environmental exposure to extreme air temperatures and air pollution contributes to the risk of cardiovascular disease and respiratory disease (CVD and RD), which together constitute cardiopulmonary diseases (CPDs) (Phung et al., 2016; Zhao et al., 2017). In addition, living with symptoms of CPD can reduce the quality of life and well-being. Early detection and expanded targeting of health programs are important for preventing heart disease (Vogel et al., 2021). In line with the above, medical resource demands for patients with CPD have increased (Ahmad et al., 2018; Thompson & Dulin, 2019). Selecting accurate models to better predict CPD health outcomes from exposure to air pollution and meteorological conditions has the potential to lower CPD mortality. A better understanding of this link can lead to a more efficient allocation of scarce healthcare resources and reduce mortality (Huntink et al., 2015).

A key aspect is the development of approaches that accurately predict CPD outcomes while including the role of exposure to air pollution. The complication is the potential role of meteorological conditions in influencing how air pollution affects health outcomes. There is little evidence on how the interaction between air pollution and weather affects health outcomes (Areal et al., 2021). At the same time, there has been a focus on the direct impact of extreme air temperatures on CPD, but not the effects of other extreme meteorological factors (Weilnhammer et al., 2021). In general, forecasting models in this area typically focus on single measures of air pollution and/or air temperature. Together, these shortcomings suggest a potential role for machine learning approaches in predicting the effect on health outcomes of air pollution, interactions with meteorological factors, and direct effects of weather conditions.

Related to this, a broader literature has developed aimed at testing the relative performance of machine learning and "traditional" statistical approaches in predicting health outcomes. This literature is inconclusive. For instance, the performance of long short-term memory (LSTM) networks in predicting daily hospital admissions due to CPD has been compared with traditional regression models, and their prediction accuracy is significantly better than traditional stepwise regression approaches (Navares & Aznarte, 2020). In contrast, other research has shown that traditional statistics models exhibit superior prediction accuracy to machine learning approaches in certain settings. For example, Kamińska et al. (2018) demonstrate that traditional regression models outperform several machine learning approaches in predicting hospital admission rates, while Rajula et al. (2020) show that traditional statistical methods outperform machine learning algorithms when the number of patients is larger than the number of variables studied. Côté et al (2022) compare the prediction accuracy of nine machine learning algorithms and two traditional algorithms concerning fruit and vegetable consumption, and they demonstrate that they have similar accuracy.

This paper aims to identify models that are effective in predicting CPD mortality from meteorological and air pollution exposure. Apart from this, we investigate whether there is an interactive effect of meteorological factors and air pollution on CPD. We study these questions in the context of Norway, which provides an interesting setting for several reasons. First, the major cities in Norway vary substantially in their geographical setting and climatic conditions. In addition, experience large variations in meteorological conditions across the year. Second, we have access to high-quality and detailed records of daily CPD mortality at the municipal level for up to ten years. Together this provides an advantageous setting for our purposes. While Machine Learning has been used in other countries to predict the relationship we are studying, to the best of our knowledge we are the first to do so using Norwegian data.

Our paper makes several contributions to the existing literature. We add to the literature that explores predictive models of CPD mortality, comparing machine learning and traditional statistical models. We do so in a setting where we endeavor to capture the effect of a wide range of air pollution and meteorological factors on CPD mortality, where we allow these to have interactive and lagged effects (Grigorieva & Lukyanets, 2021). In addition, we allow for these effects to vary across different age groups (Næss et.al., 2007). As we demonstrate, this is important as it highlights different risk factors across age groups. Together, all these provide more evidence likely to help develop appropriate policy responses.

The rest of the paper is organized as follows. Section 2 outlines the conceptual framework, and Section 3 describes the methods and materials we use in this paper. Section 4 presents the results and discussion. The conclusion and discussion are in the fifth section.

2 Conceptual Frameworks

2.1 Recent Review Papers

There are several recent review papers on the effects of weather, traffic, and air pollution exposure on CPD. Ai et al. (2020) provide a systematic review of the literature on changes in mortality and hospital admissions due to air pollution and weather exposure. They suggest that future research on health outcomes of environmental exposure should focus on predicting mortality associated with these environmental factors. Lan and Wu (2022) review the evidence on the influence of air pollution and other factors on 'CPD aging'.[4] The consensus from this literature is that exposure to air pollution and extreme air temperatures increases CVD. They additionally point out that research on the most important influencing factors of CPD aging remains challenging. There are also many inconsistent conclusions about the length of time from exposure to death. Pothirat et al. (2019) find that the association between exposure to air pollution such as $PM_{2.5}$ and PM_{10} , and mortality is mainly reflected in early exposure, that is, within 7 days. Chen et al. (2021) also noted that the effect of air temperature on cardiorespiratory mortality was higher with a lagged cumulative effect of 0 to 7 days.

There exists an area of literature that examines the effect of extreme weather on CPD that focuses primarily on the role of air temperatures. There is, however, little research on the wider effects of weather variations. For example, Weilnhammer et al. (2021) summarize research on the health consequences of extreme weather in Europe. They show that while both extremes (high and low) increase various mortality rates including CPD mortality, the effects on CPD mortality are concentrated in extremely high air temperatures. They conclude that there is a lack of research on other extreme weather conditions, which is important. For instance, for understanding the effects of climate change on health outcomes. Zafeiratou et al. (2021) make similar points in surveying the literature on the relationship between long-term environmental exposure and health. Their paper also highlights the fact that extreme cold weather may be associated with CPD outcomes due to sympathetic physiological responses leading to increased blood pressure. Finally, Grigorieva and Lukyanets (2021) summarize the literature on the effects of heat and air pollution exposure on cardiorespiratory health outcomes. They point out that the interaction effects between air pollution and weather characteristics on health are understudied. They suggest that these effects are likely to

[4] CPD aging refers to the aging of the body system of the elderly, such as arteriosclerosis or high blood pressure, to develop CPD.

be synergistic, i.e., interaction effects will lead to higher overall variations in health outcomes than air pollution or weather effects alone.

2.2 Approaches to Predicting CPD

There exists a body of research that explores different approaches to estimating the effect of weather and air pollution on CPD outcomes. For example, Tian et al. (2019) use Poisson regression models to explore the relationship between air temperature variations and hospital admissions due to CPD. Zhu et al. (2017) use three time series models to predict discharge from hospitals and demonstrate that traditional linear time series models combined with a weighted Markov Chain model have improved forecasting performance.

The rapid development of artificial intelligence has led to a wide application of machine learning methods in the medical field (Awan et al., 2019), for instance, Rumsfeld et al. (2016) use the Naive Bayesian algorithm to predict CPD disease risk. After that, Suri et al. (2019) conduct a comparison of LSTM and other machine learning algorithms to predict the risk of CPD disease, and the results showed that LSTM has a higher prediction accuracy than other machine learning algorithms. In the application of machine learning in the CPD disease prediction field (LeCun et al., 2015), decision tree (DT) and support vector machine (SVM) have the advantage of being able to solve nonlinear relationships in high-dimensional data (Golas et al., 2018; Futoma et al., 2015; Krittanawong et al., 2017–2019). This together motivates the use of machine learning approaches in this paper.

There is a small body of meta-research for Norway on the links between air pollution, weather conditions, and health. Nafstad et al. (2004) look at the relationship between male mortality and air pollution in a 26-year long-term cohort study of Norwegian men that includes both all-cause and cause-specific mortality. They find that exposure to N0x is associated with increased male mortality, but also note that no single epidemiological study could establish a causal relationship between them. Madsen et al. (2012) assess the effect of air pollution on hourly respiratory-related mortality in Oslo. They find that increases in $PM_{2.5}$ and NO_2 concentrations increase CPD mortality. In a study of respiratory hospitalization rates due to traffic-related air pollution, Oftedal et al. (2003) uses daily hospitalization rate data, which from 1995 to 2000. They find that benzene was the pollutant most associated with respiratory morbidity, but PM_{10} is the most important pollutant for overall health. However, the effect is very weak. In Norway, the main sources of benzene include vehicle exhaust and wood burning for heating.

The literature on the relationship between weather and health in Norway has produced mixed results. Stene et al. (2001) study the relationship between daily mortality and air temperature in Oslo, including several weather variables and air pollution factors. They find that air temperatures below 10 degrees Celsius increase mortality. Carter et al. (2016) describe the health effects of climate change on the local elderly. In the study across Norway, Sweden, and Finland, they conclude that heatwaves, icing, and freezing weather all contribute to mortality in the elderly.

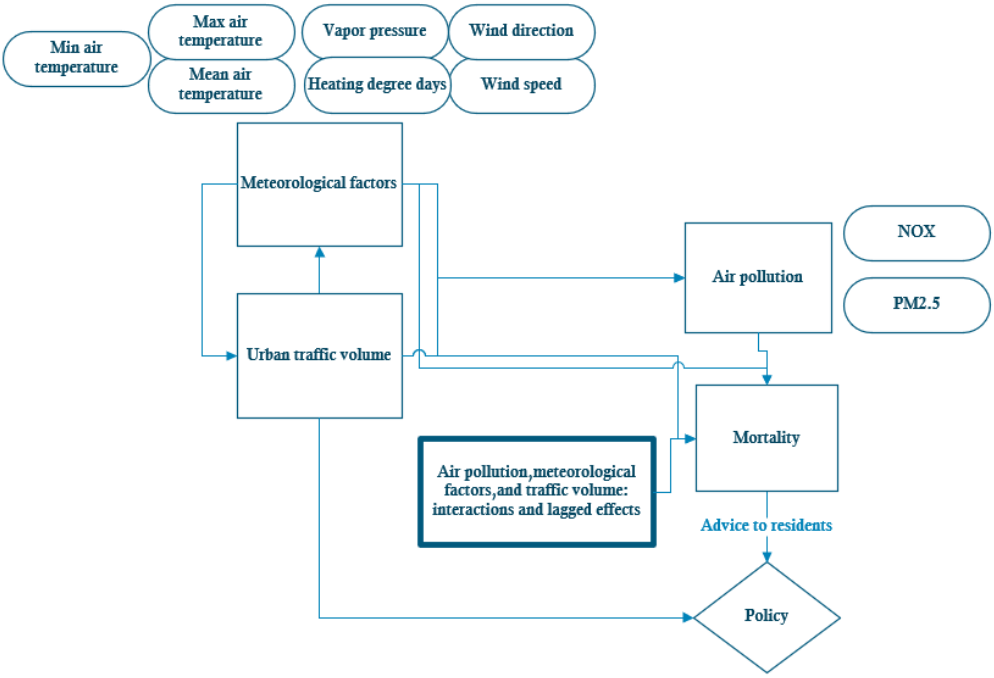
There is little Norwegian evidence on air pollution effects on CPD across different population subgroups. In a study on nitrogen oxides and particulate matter for cause-specific death, Næss et al. (2007) studied the mortality in people aged from 51 to 70, and from 71 to 90, using daily data covering four years. They find that patients with underlying chronic diseases and the elderly are more susceptible to the effects of air pollution. In addition, they also suggest that different subgroups of the population, such as younger populations, need to be considered when formulating policies to improve air pollution.

Research Questions

Figure 1 provides an overview of the research questions: Are there interactions between these environmental factors? Is the synergistic effect greater than the effect of single factors? What are the most critical environmental impact factors? Is the predictive performance of machine learning better than that of traditional statistical models? Are there other extreme weather factors that increase CPD, and do other extreme weather factors have a more substantial impact on CPD than extreme air temperatures, for example, mean wind speed, wind gust, vapor pressure, and heating degree days?

Figure 1

Research questions visualization



3 Methods

We use Support Vector Machine (SVM) and Decision Tree (DT) for CPD mortality prediction, K Nearest Neighbor (KNN) for missing value imputation, and Random Forest (RF) to find the most critical prediction factors. These methods have been widely used to explore the relationship between health outcomes and environmental exposure.

SVM is a classification method based on dimension theory and structural risk minimization. The advantage of SVM is that it can balance model fitting and model complexity to select the optimal model, which can better adapt and identify new samples. The basic principle of SVM is to assume that the hyperplane used is strictly linearly separable. Its approach is to separate the two types of points in space, let the distance between the hyperplane and the two classes be the widest, and finally obtain the optimal split hyperplane equation.

DT uses entropy to judge the internal nodes of a tree, where each node represents a judgment on an attribute, and each leaf node represents a classification result. The generation of the DT involves several steps. The principle of the decision tree is a "like this if-else process." That is, if a certain condition is met, it will be divided into a category and forms a branch, otherwise it will be divided into other categories and forms other tree branches. Therefore, the iterative, decision-making process of the decision tree is drawn as an algorithm tree.

KNN is a supervised learning algorithm in machine learning. KNN finds the K nearest neighbors on the training set for a new prediction instance; then, by a voting method, it divides into the class with the one that has the closest distance to the new prediction instance, here class means group. The principle of KNN is that in the feature space, samples of the same category should be clustered together. In a dataset, when there is a new input observation, this observation will find the K closest observations within the dataset. Most of these K observations belong to a certain category, and the newly inputted observations are classified into this category. The selection of the K value in the KNN algorithm is important. The usual approach is to select a smaller K value first according to the distribution of the sample. However, there are trade-offs. If the K value is too small. it will be prone to overfitting. If a larger K value is selected, the overall model will be simpler, leading to larger training errors which creates prediction errors and results in low prediction accuracy.

Both KNN and RF have been shown to exhibit superior performance in missing value interpolation, pattern recognition and classification, and prediction when compared to traditional statistical approaches. RF can be used both for classification and regression analysis. Here we use it for classification purposes. The principle is that when there is a new prediction sample, each decision tree in the random forest is judged separately and determines which class the sample belongs to; and the class that has been selected the most is often chosen as the class of the prediction sample. RF is an algorithm composed of multiple decision trees, that belongs to the ensemble learning branch of machine learning. It uses the method of bagging, that is, first randomizing the observations and column variables of the data, to obtain multiple classification trees. All these classification trees are aggregated to obtain the final random forest. Since the final model is the average of all classification tree results, it reduces model variance and makes the model more stable. At the same time, RF approaches are not sensitive to missing data and robust to imbalanced data.

Logistic regression models represent a standard approach in the field of cardiovascular prediction. For instance, it is a commonly used approach to estimating cardiology risk (Goldstein et.al., 2017). Its advantage is the simplicity of specification and estimation. However, such approaches face a range of known difficulties with complex data structures, or as the volume of data becomes large (Côté et al., 2022). For validation, we contrast logistic regression models with machine learning approaches, regarding the predictive performance.

We evaluate model performance using several standard approaches. These are Area Under the Curve (AUC),[5] which is the curve under the Receiver Operating Characteristic (ROC) curve, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and adjusted R-squared (R^2). AUC represents one of the most widely used evaluation methods in machine learning, and existing studies find that it has several advantages compared to other evaluation methods. For example, AUC does not depend on the size of the threshold for the selected decision. The AUC is a measure of model performance. The horizontal of the ROC curve represents the false positive rate (FPR), $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$, TP = true positive, TN = true negative, FP = false positive, and FN = false negative. FPR indicates that the one that has been singled out (the prediction

[5] $AUC = \frac{\sum_{postive\ Class} rank_i - \frac{M(1+M)}{2}}{M"N}$; explanation: First, sort by score. Score represents the probability that the test sample belongs to the positive sample. The rank with the largest score is n, the rank with the second largest score is N-1, and so on; the smallest score is 1, and then the ranks of all positive samples are accumulated, minus the M-1 combination of two positive samples, divided by M×N.

is "positive"), and which is correct (the predicted value is equal to the true value), accounted for the percentage of the total predicted positive. The vertical represents the true positive rate (TPR), $\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{TN})$, which represents the one that has been singled out (the prediction is "positive"), but which is wrong (the predicted value is not equal to the true value), accounted for the percentage of the total predicted negative. The larger the TPR, or the larger the FPR, the higher the accuracy. AUC represents a probability value between 0.1 and 1. When used as a judgment criterion for model evaluation, a larger value means that the current algorithm has a higher probability of ranking positive samples ahead of negative samples for better classification.

We use root mean squared error (RMSE), Mean Absolute Error (MAE), and mean absolute percentage error (MAPE) as alternative approaches to measuring prediction error. In all cases, smaller values indicate less prediction error. In addition, we use adjusted R^2 and AIC[6] to evaluate the model, and the assumption is that the error of the model follows an independent normal distribution. The number of parameters in the model is given by k, L is the log-likelihood value, and the related formula is in a footnote.[7]

[6] $\mathrm{AIC} = (2k - 2L)/n$. Likelihood value formula: $L = -(n/2) * \ln(2 * pi) - (n/2) * \ln(sse/n) - n/2$, where n is the sample size, SSE is the sum of squared residuals, and L is usually a negative number.

[7] $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2}, \in [0, +\infty)$

$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^* - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0, 1]$, Ajusted $R^2 = \{1-[(1-R^2)(n-1)/((n-k-1))]\}$

where y_i^* represents the predicted value, and y_i is the real value, so assuming that y_i has a variance of 1 unit, then the R^2 indicates how much the variance of the residual is reduced after using this model. n represents the amount of data in the data set, and k represents the number of independent variables.

3.1 Raw Data

We use daily data from 2009 to 2018 from Oslo, Bergen, Trondheim, and Tromsø, which are distributed across different geographical regions, with resulting variations in climatic conditions. A critical reason for choosing these cities was data availability over the entire period. Variables include mortality (CPD in different age groups and total natural mortality), pollutants ($\mathrm{NO_x}$, $\mathrm{PM_{2.5}}$), meteorological variables, and traffic volume. The data preprocessing process is shown in Appendix 1, and we use K nearest neighbors for imputation. Table 1 provides summary statistics of the data.

We focus on two pollutants, $\mathrm{PM_{2.5}}$ and NOx. These are known to be linked to poorer health outcomes and differ in the main sources of emissions that generate these pollutants. For instance, burning wood for the fire is a primary source of $\mathrm{PM_{2.5}}$, while NOx is primarily generated by vehicle exhaust and industrial production.

The mortality data, including natural mortality and CPD mortality, are obtained from the Cause of Death Registry of Norway. Cardiovascular disease and respiratory disease (CVD and RD) together constitute cardiopulmonary diseases (CPDs) (Basu, 2009; Turner et al., 2012). The data are daily at the municipal level. Daily average air temperature data are taken from the seNorge2 dataset (Lussana et al., 2016). In mainland Norway, seNorge2 provides high-resolution daily air temperature data. The final data include 2,990,756 observations and 20 variables. The air pollution data come from the Norwegian Air Research Institute (NILU), and the weather data come from the Norwegian Meteorological Institute (MET).

Table 1

Statistical Description of the Original Data

Abbreviation	Variable explanations	Min	Max	Mean	Std.Dev.
CPDall	Total cardiovascular disease plus respiratory disease mortality	0	15	2.94	1.98
CPD0_64	Cardiovascular disease plus respiratory disease mortality between the ages of 0 and 64 years	0	5	1.71	0.42
CPD65_74	Cardiovascular disease plus respiratory disease mortality between the ages of 65 and 74 years	0	5	1.24	0.52
CPD75_84	Cardiovascular disease plus respiratory disease mortality between the ages of 75 and 84 years	0	7	1.50	0.79
CPD85plus	Cardiovascular disease plus respiratory disease mortality for people over the age of 85 years	0	11	2.02	1.29
TV (PCU)	Traffic volume (Passenger Car Unit)	0	8270	3489.17	2843.37
NOx (µg/m^3)	Nitrogen oxide, which is a gas mixture, composed of nitrogen and oxygen (µg/m^3)	0	514	65.83	73.54
$PM_{2.5}$ (µg/m^3)	Particulate matter with aerodynamic equivalent diameter less than or equal to 2.5 microns (µg/m^3)[8]	-1	65	11.41	9.48
Tmax (°C)	Maximum air temperature (°C)	-16	34.6	9.75	8.12
Tmean (°C)	Mean air temperature (°C)	-20.40	26.1	6.28	7.20
Tmin (°C)	Minimum air temperature (°C)	-24.30	21.70	3.29	6.86
HDD (°C)	Heating degree days, 17 °C (°C)，represents the energy demand for indoor heating, and the value represents how much the hourly mean temperature is lower than 17 degrees; if it is equal to, or greater than, 17 degrees, the value is 0	0	37.40	10.86	6.97
VP (kPa)	Daily mean vapor pressure 24h (kPa)	0.90	21.50	7.60	3.48
WS(m/s)	Average of mean wind speed from main obs 24 h (m/s)	0	14.80	2.97	1.68
WG(m/s)	Mean wind gust 24 h (m/s)	1	135	6.03	2.85
Nall	Total natural mortality (per person)	1	26	5.84	4.12
	Number of observations			14,606	

Note: Here is the statistical description of the raw data. The data have not been cleaned and organized, so there are outliers. For example, the daily maximum concentration of NOx is 514 µg/m^3, but the mean value is 65.83 µg/m^3, so this maximum value we consider an outlier. The same as wind gust (WG). Here is the statistical description of the raw data. Mortality is the number of deaths per 1000 people per year.

[8] Due to the measurement error of the data collected at the air monitoring station, the value sometimes changes between 0 and -5.

3.2 Variable Importance

As an initial step, we seek to explore the relative importance of variables as predictors of total mortality from CPD. To identify the environmental factors most critical to CPD mortality. We do so use random forest approaches to generate the variable importance map reported in Table 2. The mean squared error (%IncMSE) provides the effect of removing a given variable on prediction accuracy. Hence, higher values indicate a larger effect of a given variable on prediction accuracy.

Table 2

The Variable Importance Results from Random Regression Forest

Variables	%IncMSE
Traffic volume (TV) (PCU)	2.57
NOx (µg/m³)	32.45
$PM_{2.5}$ (µg/m³)	20.84
Temp max (°C)	17.63
Temp mean (°C)	24.93
Temp min (°C)	17.31
HDD (°C)	9.44
VP (kPa)	15.11
WS(m/s)	15.39
WG(m/s)	16.83

Note: The second column of values represents the %IncMSE value, and the first column is the variables.

Table 2 reveals that NOx has the largest %MSE, followed by mean air temperature, $PM_{2.5}$, and mean relative air humidity. This provides an initial indication of the importance of both air pollution and meteorological conditions for health outcomes. It should also be noted that also that many additional weather factors appear to have a role in prediction accuracy. This provides a further indication that there are effects beyond the average air temperature. Note, however, that caution must be exercised in interpreting the effect of the traffic volume. This appears to be relatively unimportant in predicting CPD mortality but reminds us that traffic volume is itself a major cause of NOx concentrations.

3.3 Empirical Approaches

We next explore the role of weather conditions and air pollution, and their interactions, in explaining CPD mortality. Our initial approach is to include traffic volume as an independent variable, recognizing that this is the main effect of the two air pollutants explored in this paper. We focus on the effect of the interaction between air pollution and meteorological factors on CPD mortality, as well as on the impact of a wide set of meteorological factors on CPD mortality.

We initially estimate a panel model. Along with the direct effects of traffic volume, air pollution, and meteorological conditions, we also include interactions between them.

We use four cities in our study and include control variables for the month of the year. We additionally include city-fixed effects. Hence, we estimate the effects of changes in meteorology, air pollution, and traffic conditions (and their interactions) within cities on city changes in CPD. We observe both CPD and natural mortality for different age groups. As a result, we estimate variants of equation (1) for these different groups along with total mortality. The above can be summarized as:

$$\mathrm{Y} = \mathrm{f}\,(\mathrm{T}, \mathrm{P}_{NO_x}, \mathrm{P}_{PM_{2.5}}, \mathrm{M}, \mathrm{X}) \qquad (1)$$

Where Y is mortality, T is traffic volume, and P_{NO_x} and $\mathrm{P}_{PM_{2.5}}$ are NOx and $\mathrm{PM}_{2.5}$, respectively. M are meteorological factors, and X represents control variables. Besides single factors, we also include lagged effects, interaction terms, and lagged effects of the interaction terms. We consider interaction terms of traffic volume and air pollution, the interaction of each weather factor, and air pollution.

The expected findings are that traffic, weather, air pollutants alone, and the interaction between these environmental factors will have an effect on mortality and that mortality will rise with the increase of traffic volume and air pollutant concentration.

We also expect nonimmediate death after exposure to these environments, so there is a lagged effect. Based on previous literature, the first seven days after exposure to the environment have a high effect on mortality; we choose two, four, and six days as lag effects to explore (Chen et al., 2021; Pothirat et al., 2019). Finally, we also expect that some age groups are more sensitive to different environmental factors.

We consider the initial model we use, we call it Regression Model (RM) (1), as a full model and build two simpler models as a point of comparison, respectively (RM) (2) and (RM) (3). For regression model (RM) (2), as the simplest model, we consider the single environmental factors, without considering the interaction terms between air pollution and weather, or lagged effects. The input variables are only traffic volume and all the individual weather factors. For the regression model (RM) (3), as an intermediate model, we only consider the single-factor effect and the interaction terms between air pollution and weather, without considering the lagged effect (see details in Table 3).

Table 3

Independent Variables Included in the Three RM Models

	RM (1)	**RM (2)**	**RM (3)**
	Full model	Simple model	Intermediate model
Single factors	Yes	Yes	Yes
Interaction terms	Yes	No	Yes
Lagged effects	Yes	No	No
Interaction terms' lagged effects	Yes	No	No

4 What Is the Role of Environmental Factors in CPD Mortality?

4.1 Use RM (1) Model to Estimate CPD Total Mortality

We first explore the role of environmental factors in CPD mortality using the full RM (1) model from Section 3.3. RM (1) is a systematic model; it traverses all variables, variable interaction items, and lagged effects. The R^2 is 0.64, which shows that our model has good explanatory power.

Appendix 2 reports the associations between a range of effects on CPD mortality. We will focus on those that are statistically significant at the ** $P < 0.05$ and *** $P < 0.01$ levels. Often these associations are largest in lagged exposure, highlighting the need to incorporate these effects in the model. For instance, exposure to NOx two days ago has the largest association with total CPD mortality, followed by the interaction of $PM_{2.5}$ and minimum air temperature from six days ago, as well as the traffic volume. Only the $PM_{2.5}$ does not have a statistically significant effect on the total mortality of CPD. While the minimum air temperature has a positive effect on CPD; that is, as the minimum air temperature increases, CPD mortality increases. However, with exposure six days before, under the interactive effect of $PM_{2.5}$ and the minimum air temperature, the impact on CPD mortality is negative, and this interaction effect reduces CPD mortality.

As a point of comparison, Appendix 2 reports estimates for natural mortality. The relationship between our environmental risk factors and natural mortality is mainly due to single environmental factors, such as the effect of NO_x and the mean air temperature on natural mortality. This provides supportive evidence that our main approach isolates the role of risk factors primarily linked to CPD mortality.

At the same time, in Appendix 2, the group "the month of the year", there are clear seasonal and yearly patterns in mortality. We can see that in both the CPD mortality and the natural mortality, there is a statistically significant change with the month, showing that January has the highest mortality. From the yearly patterns, we find that from 2009 to 2018, CPD mortality has had a statistically significant change from 2013, showing that CPD mortality decreases over time.

Our results fit with a report by the Norwegian Public Health Administration (FHI) (Raknes et al., 2022). The report shows that, from 1971 until 2020, cardiovascular disease mortality has been on a downward trend, and compared with 1987, the mortality rate has been reduced by half. This is consistent with our results. The report from the FHI also shows that the reasons for the decline, include reductions in factors such as smoking and cholesterol, as well as better medical treatment, meaning that people pay more attention to their health.

4.2 Compare the Estimated Results of RM Models (1), (2), and (3)

We use the three RM models to estimate the total CPD mortality from environmental factors and try to compare the estimated results of these three models to see if they are different. We focus on estimating whether they are at statistically significant levels at ** $P < 0.05$ and *** $P < 0.01$. As shown in Appendix 3, RM (1) emphasizes the statistically significant effect of traffic volume on mortality. The simplest model, RM (2), shows that NOx, mean air temperature, and maximum air temperature all have a statistically significant effect on CPD mortality. Model RM (3) reflects the statistically significant effect of the interaction term of NOx and traffic volume on CPD mortality. The time variable is consistent in all three models, indicating that January each year has the highest mortality and that the CPD mortality has been decreasing this decade. By comparing the R^2 of the three models, we find that RM (1) has the largest R^2, considering, that environmental factors such as traffic, weather, and air pollution can hardly explain 60% or 40% of CPD mortality. Because regarding the impact of CPD mortality, in addition to the environment, there are many other factors, for instance, the confounding factors caused by seasonal influenza, diet, exercise, and so on. Thus, we consider that RM (1) and RM (3) may be overfitting. However, if compare the adjusted R^2, the explanatory power of the model is more in line with the actual phenomenon, and the adjusted R^2 values of the three models are nearly the same.

4.3 Estimates of CPD Mortality by Age Group Using RM (1)

We use the model RM (1), which estimates the CPD mortality split by broad age groups in Table 4. Because the RM (1) traverses all interactions and lagged effects, it can explore the most comprehensive information. We focus on estimating statistically significant at ** $P < 0.01$ level. The focus is on older age groups who face a higher risk of CPD. This reveals heterogeneous associations across age groups. Among environmental factors, we find that meteorology and air pollution have different effects on different age groups, and with lagged effects.

Among the effects of single environmental factors: exposure to same-day $PM_{2.5}$ is associated with higher CPD mortality in the 65-74 years age group, but not in other age groups. Exposure to maximum air temperature six days earlier affected CPD mortality in the 75- to 84-year-old group.

The impact of the interaction of air pollution and weather on CPD mortality is as follows, exposure to the interaction of NOx and mean air temperature six days earlier have a statistically significant effect only for 65- to 74-year-old people, but this is a negative effect. Exposure to the interactive effect between the maximum air temperature and NOx two days earlier will decrease the CPD mortality of the over 85 years age group.

The interaction between the mean wind speed and NOx on the same day, the CPD mortality of the 0- to 64-year-old subgroup decreased. Although the only mean wind speed, and only NOx on the same day, have no statistical significance for the CPD mortality, the interaction term can reduce mortality of this age group. Four days of earlier exposure to the interactive effect of mean wind gusts and NOx will lead to an increase in CPD mortality in this young age group.

The interactive effect of $PM_{2.5}$ and mean air temperature exposure four days earlier has a statistically significant effect only on the 65-74 years age group. We also find that exposure to the interaction effect of $PM_{2.5}$ and minimum air temperature six days earlier will lead to a decrease in the CPD mortality of the 75- to 84-year-old group.

Our results show that exposure to only NOx, and only maximum air temperature increases CPD mortality, except for the interactive effect of mean wind gust and NOx, all other interaction results in less CPD mortality.

Considering the differences in patterns between age groups, traffic volume has a statistically positive effect on mortality mainly in older age subgroups (more than 75+), and as traffic increases, so does the CPD mortality. CPD mortality has fallen significantly over the decade in the 0–64 and 75–84 age groups, and for the 85+ age group it has dropped since 2016.

Table 4

Contributors to CPD Mortality by Age Group, Oslo, Bergen, Trondheim, and Tromsø, 2009–2018, Daily. This table has three pages.

	Contributors	0~64	65~74	75~84	85 plus
Observations		14606	14606	14606	14606
$\boldsymbol{R^2}$		0.680	0.157	0.478	0.747
Adjusted $\boldsymbol{R^2}$		0.005	0.002	0.016	0.022
Single environment contributors and their lagged effects, and the magnitude are expanded by 10 to the 5th power.	**Traffic**	0.33	0.15	1.52*	2.13*
	Traffic $_{t-2}$	-0.03	0.61*	-0.74	-0.03
	Traffic $_{t-4}$	-0.08	0.47	0.22	-0.01
	Traffic $_{t-6}$	-0.06	-0.20	0.36	0.31
	NOx	5.59	-3.65	6.55	13.71
	NOx $_{t-2}$	6.06	2.01	11.71*	14.29
	NOx $_{t-4}$	-7.50*	2.59	-7.65	7.77
	NOx $_{t-6}$	2.45	2.86	-2.16	16.02
	PM$_{2.5}$	17.27	**90.60****	7.84	-83.89
	PM$_{2.5t-2}$	14.47	3.48	58.56	12.92
	PM$_{2.5t-4}$	29.81	9.49	98.27*	-115.90
	PM$_{2.5t-6}$	-13.30	-53.19	7.57	44.33
	Temp (Mean)	0.16	0.34	-0.24	0.18
	Temp (Mean) $_{t-2}$	-0.19	-0.62	0.48	-0.25
	Temp (Mean) $_{t-4}$	-0.12	3.86	-0.77	-0.15
	Temp (Mean) $_{t-6}$	0.02	-0.12	0.06	-1.46*
	Temp (Max)	-6.15	4.87	9.69	0.21
	Temp (Max) $_{t-2}$	2.23	0.22	3.13	2.77
	Temp (Max) $_{t-4}$	0.83	3.86	-3.90	9.30
	Temp (Max) $_{t-6}$	0.28	-6.66	**17.98****	7.78
	Temp (Min)	4.18	2.17	25.53	33.22
	Temp (Min) $_{t-2}$	-4.72	7.21	2.13	0.63
	Temp (Min) $_{t-4}$	5.40	-5.41	-2.98	-7.04
	Temp (Min) $_{t-6}$	-3.18	3.60	-3.43	-127.00
	HDD	-0.08	8.57	-7.10	-4.16
	HDD $_{t-2}$	5.88	-2.27	5.00	20.15
	HDD $_{t-4}$	-3.22	-2.01	-3.16	15.12
	HDD $_{t-6}$	4.15	-1.89	16.14*	14.62
	VP (kPa)	19.31	12.02	27.59	41.89
	VP (kPa) $_{t-2}$	5.11	-8.58	18.8	4.85
	VP (kPa) $_{t-4}$	-5.49	2.11	-15.14	23.66
	VP (kPa) $_{t-6}$	0.29	-21.00*	-1.16	-5.38
	WS (m/s)	4.97	2.47	67.73	-156.46
	WS (m/s) $_{t-2}$	28.21	12.00	-8.11	-34.25
	WS (m/s) $_{t-4}$	0.31	-50.38	-26.29	38.74
	WS (m/s) $_{t-6}$	30.33	7.09	9.64	-50.16
	WG (m/s)	-5,83	-21.58	1.68	-56.90
	WG (m/s) $_{t-2}$	6.19	2.81	6.30	2.36
	WG (m/s) $_{t-4}$	6.59	-20.79	2.61	-28.04
	WG (m/s) $_{t-6}$	-4.41	-3.41	32.58	1.29
Interaction of NOx and weather	**NOx *Temp (Mean)**	8.99e-04	-1.91e-03	-6.66e-03	-5.86e-03
	NOx *Temp (Mean) $_{t-2}$	-1.86e-03	-1.09e-03	-6.64e-04	-1.43e-03
	NOx *Temp (Mean) $_{t-4}$	8.31e-04	5.83e-03	-4.68e-03	-3.79e-04

	Contributors	0~64	65~74	75~84	85 plus
contributors and the magnitude is expanded by 10 to the 5th power.	**NOx *Temp (Mean)$_{t-6}$**	1.77e-03	-8.40e-03**	3.69e-03	1-34e-03
	NOx * Temp (Max)	-6.14e-03	0.16*	0.09	0.13
	NOx * Temp (Max)$_{t-2}$	-7.66e-03	0.12	-0.02	-0.40**
	NOx * Temp (Max)$_{t-4}$	0.07	-0.11	0.13	0.06
	NOx * Temp (Max)$_{t-6}$	0.04	-0.11	-0.24*	-0.13
	NOx * Temp (Min)	-3.63e-03	0.06	0.10	0.17
	NOx * Temp (Min)$_{t-2}$	-0.08	0.03	-0.17	-0.12
	NOx * Temp (Min)$_{t-4}$	0.02	-0.16*	-0.01	-0.26
	NOx * Temp (Min)$_{t-6}$	-0.15*	-0.13	0.11	-0.05
	NOx * HDD	0.04	-1.52e-04	-0.10	0.18
	NOx * HDD$_{t-2}$	-0.08	0.06	-0.13	-0.20
	NOx * HDD$_{t-4}$	-0.02	-0.15	0.30*	0.21
	NOx * HDD$_{t-6}$	-9.66-03	-0.02	-0.17	-0.26
	NOx * VP	-0.02	-0.22	0.06	0.57
	NOx * VP$_{t-2}$	-0.05	-0.07	-0.30	-0.28
	NOx * VP$_{t-4}$	-0.08	0.32	0.06	-0.27
	NOx * VP$_{t-6}$	0.06	-0.17	0.24	-0.56
	NOx * WS (Mean)	-0.09**	0.25	-1.10	0.10
	NOx * WS (Mean)$_{t-2}$	-0.06	-0.20	0.60	0.76
	NOx * WS (Mean)$_{t-4}$	0.35	-0.55	-0.34	0.26
	NOx * WS (Mean)$_{t-6}$	0.28	0.54	0.71	0.75
	NOx * WG (Mean)	-0.14	0.09	0.22	0.13
	NOx * WG (Mean)$_{t-2}$	-0.14	-0.18	-0.23	0.50
	NOx * WG (Mean)$_{t-4}$	0.41**	0.09	0.09	-0.17
	NOx * WG (Mean)$_{t-6}$	-0.51	0.04	-0.05	-0.08
Interaction terms of NOx with traffic volume, and their lagged effects.	NOx * Traffic	3.70e-04	-1.81e-03	-0.01	6.45e-04
	NOx * Traffic$_{t-2}$	1.02e-03	-2.71e-03	1.34e-04	2.86e-03
	NOx * Traffic$_{t-4}$	8.59e-5	-8.28e-04	7.90e-03	-5.57e-03
	NOx * Traffic$_{t-6}$	-3.55e-03	4.06e-03	1.97e-04	-0.01
Interaction of $PM_{2.5}$ and weather contributors and the magnitude is expanded by 10 to the 5th power.	$PM_{2.5}$ * Temp (Mean)	-0.01	0.01	3.66e-03	0.03
	$PM_{2.5}$ * Temp (Mean)$_{t-2}$	8.12e-03	5.24e-08	0.02	-0.05
	$PM_{2.5}$ * Temp (Mean)$_{t-4}$	0.01	-0.06**	-0.03	-9.35e-5
	$PM_{2.5}$ * Temp (Mean)$_{t-6}$	7.4e-03	0.03	0.02	0.06
	$PM_{2.5}$ * Temp (Max)	0.13	-1.30*	0.09	-1.01
	$PM_{2.5}$ * Temp (Max)$_{t-2}$	-0.50	-0.07	0.54	1.10
	$PM_{2.5}$ * Temp (Max)$_{t-4}$	-0.77	-0.21	-1.15	1.00
	$PM_{2.5}$ * Temp (Max)$_{t-6}$	0.83	-0.06	-0.50	-0.13
	$PM_{2.5}$ * Temp (Min)	0.82	-0.60	-0.90	0.78
	$PM_{2.5}$ * Temp (Min)$_{t-2}$	-0.30	0.30	-1.15	2.05
	$PM_{2.5}$ * Temp (Min)$_{t-4}$	-0.50	1.31	1.81	0.13
	$PM_{2.5}$ * Temp (Min)$_{t-6}$	-0.09	-1.33	-1.95*	-2.55
	$PM_{2.5}$ * HDD	-0.04	-0.41	0.44	-1.00
	$PM_{2.5}$ * HDD$_{t-2}$	-0.20	0.13	0.77	-1.65
	$PM_{2.5}$ * HDD$_{t-4}$	-0.13	-0.17	-1.23	1.29
	$PM_{2.5}$ * HDD$_{t-6}$	0.34	0.13	-0.68	-0.76
	$PM_{2.5}$ * VP	-1.51	0.14	0.04	-1.60
	$PM_{2.5}$ * VP$_{t-2}$	1.36	-1.28	0.37	-0.93
	$PM_{2.5}$* VP$_{t-4}$	-0.44	-0.38	-0.48	1.47
	$PM_{2.5}$ * VP$_{t-6}$	0.10	0.84	1.39	-0.09
	$PM_{2.5}$ * WS (Mean)	1.71	-7.99	-3.09	4.60
	$PM_{2.5}$ * WS (Mean)$_{t-2}$	-2.14	2.45	6.31	-4.21

	Contributors	**0~64**	**65~74**	**75~84**	**85 plus**
	$PM_{2.5}$ * WS(Mean) $_{t-4}$	-1.58	-1.99	-5.43	3.64
	$PM_{2.5}$ * WS (Mean) $_{t-6}$	-2.04	5.87	-0.54	-0.84
	$PM_{2.5}$ * WG (Mean)	-1.72	-0.55	-1.18	4.23
	$PM_{2.5}$ * WG (Mean) $_{t-2}$	-0.60	0.68	1.85	-1.76
	$PM_{2.5}$ * WG (Mean) $_{t-4}$	0.17	0.25	-2.12	6.39
	$PM_{2.5}$ * WG (Mean) $_{t-6}$	-1.28	1.43	1.43	-1.00
Interaction terms of $PM_{2.5}$ *with* traffic volume, and their lagged effects.	$PM_{2.5}$ * Traffic	2.43e-03	-0.18	-0.01	0.02
	$PM_{2.5}$ * Traffic $_{t-2}$	-0.02	-0.03	6.42e-03	-0.08
	$PM_{2.5}$ * Traffic $_{t-4}$	0.02	-0.02	-0.64	0.01
	$PM_{2.5}$ * Traffic $_{t-6}$	-0.01	-0.01	0.02	0.02
Time variables	The month of the year				
	February	0.01	-0.02	-0.01	0.01
	March	0.02	-0.02	-0.03	-0.06
	April	**0.02****	-0.06	-0.06	**-0.15****
	May	0.001	-0.05	-0.08	-0.12*
	June	-0.01	-0.02	-0.07	-0.13*
	July	0.02*	-0.08	-0.03	-0.02
	August	0.02*	-0.08	-0.02	-0.06
	September	-0.01*	-0,07	-0.05	-0.09
	October	-0.005	-0.03	**-0.12****	**-0.19*****
	November	-0.02*	-0.05	**-0.12*****	**-0.25*****
	December	-0.01*	-0.05	**-0.08****	-0.11*
	The year 2009				
	2010	0.02	-0.008	-0.04	-0.03
	2011	-0.006	-0.004	**-0.08****	0.02
	2012	-0.02	0.03	**-0.12*****	0.004
	2013	**-0.05****	0.02	**-0.12*****	-0.06
	2014	-0.04*	0.02	**-0.20*****	-0.06
	2015	**-0.05****	0.03	**-0.16*****	-0.08*
	2016	**-0.05****	0.005	**-0.18*****	-0.09*
	2017	**-0.05****	0.04*	**-0.15*****	**-0.21*****
	2018	**-0.05****	0.06	**-0.19*****	**-0.20*****
	_cons	1.14***	1.31***	1.40***	2.29***
	i.holiday	-0.02	-0.03	0.05	-0.12

Notes: Here t-2, t-4, and t-6 represent lags of 2 days, 4 days, and 6 days, respectively. Two digits after the decimal point are taken. *, **, and *** indicate statistical significance at the 0.1, 0.05, and 0.01 levels, respectively; $PM_{2.5}$ * Windspeed (Mean) $_{t-6}$ represents exposure six days ago to the interaction of $PM_{2.5}$ and mean wind speed. We use a two-way fixed model using fixed time and fixed cities.

4.4 Prediction Performance Comparison

We seek to compare the predictive performance of machine learning algorithms and regression models. To do so, we split the data into two parts: a 75% training dataset, used to estimate all the models in the period from 1st January 2009 to 21st December 2015, and a 25% testing dataset for prediction, with the period being from 22nd December 2015 to 31st December 2018. As discussed earlier, we use a variety of approaches, including SVM, three regression models (RM), and DT, to train the model in the training set, and use the test set for evaluating the model's predictive performance out of the sample. Table 5 and Table 6 report model evaluation results.

Table 5

Performance Comparison Results of ML and Regression Models

	SVM	**DT**	**RM (1) Full model**	**RM (2) Simple model**	**RM (3) Intermediate model**
MAE	**1.209**	1.332	1.434	1.435	1.442
MSE	**2.943**	3.036	3.447	3.414	3.436
RMSE	**1.715**	1.742	1.857	1.848	1.854
$\boldsymbol{R^2}$	0.238	0.214	**0.642**	**0.067**	**0.431**
AUC	0.715	0.**733**	0.680	0.683	0.682

Table 6

Performance comparison results of three RM models.

	RM (1) Full model	**RM (2) Simple model**	**RM (3) Intermediate model**
R^2	**0.642**	**0.067**	**0.431**
Adjusted R^2	**0.0337**	**0.0327**	**0.0333**

In our case, the MAE, MSE, and RMSE of SVM and DT are smaller than the three RM models; SVM and DT also have a larger AUC than the three RM models. This means that machine learning does exhibit better prediction performance than traditional RM models in predicting CPD mortality. The RM models are relatively simple and prone to overfitting, which means that the models can fit the data well on the training set, but they cannot fit the data well on the new data outside the training set. Therefore, the RM model results in poor prediction performance.

Focusing solely on the prediction accuracy of SVM and DT, SVM performs better than DT in most cases. One possible reason for this is that although DT can handle nonlinear features, it is easy to overfit the model as the tree depth increases. SVM performs well in processing large-scale feature space datasets because it only relies on samples at the classification boundary compared to all data when constructing a classification surface.

Comparing the three RM models' results from Tables 5 and 6, the interpretation of the results from a model design perspective is that we derive the importance of including lagged effects and interaction terms. In our RM (1), we not only include a single environmental factor, but we also include interactions between environmental factors, as well as the lagged effect on the dependent variable and the interaction terms' lagged effects, so this more comprehensively considers the link between traffic, weather, and air pollution. RM (3) includes the single environment factors and interaction terms of environmental factors but not the lagged effects, while RM (2) only includes a single environmental factor, without considering the interaction terms and lagged effects. Our results show that the MAE, MSE, and RMSE values of the three RM models are not very different, but the R^2 is very different. This indicates that the prediction accuracy of these three models is close. Considering the R^2 of RM (2) is 0.642, and R^2 of RM (3) is 0.431, in real life, air pollution, weather, and traffic, those environmental factors cannot explain 64.2% or 43.1% of the cause of CPD mortality. In addition to the environment, seasonal flu, diet, and other living habits also have a big impact on CPD mortality. It is indicated that RM (1) and RM (3) are overfitting. Regarding adjusted R^2 , the explanatory power of the three RM models is similar, and they are closer to the actual phenomenon.

5 Conclusion and Discussion

This paper explores the effect of meteorological conditions, air pollution, and their interaction on CPD mortality in Norway. We use different machine learning algorithms to analyze and predict the effects. At the same time, we include the influence of other extreme weather factors besides extreme air temperature on CPD. This is, to the best of our knowledge, the first study to do so with Norwegian data.

To investigate whether meteorological factors and air pollution have interaction effects, as well as whether extreme meteorological conditions other than extreme air temperatures increase CPD mortality. We first used the K nearest neighbor algorithm to impute the missing values of the raw data. We then use random forests to perform the variable importance analysis to identify key variables affecting CPD. We further build a panel model for analysis. Finally, we seek to examine whether machine learning exhibits better predictive performance than traditional statistical regression models.

Our main finding is that weather and air pollution have different effects on different age subgroups, and there are lagged effects. The most critical environmental factors affecting CPD mortality are NO_x, $PM_{2.5}$ and air temperature. In addition to extreme air temperature, other weather factors also contributed to increased CPD mortality. We find the importance of splitting by age group because we identify the environmental factors with the biggest risk of CPD mortality for specific age subgroups. When predicting CPD mortality, we find that the predictive performance of machine learning is better than that of traditional regression models; when the number of variables increases, the RM model tends to overfit, which further reflects the predictive advantages of machine learning.

A report from the Norwegian Institute of Public Health (Raknes, 2022) shows that the Norwegian government has already instigated preventive treatments for people under 75 years old, such as primary and specialist health services for CPD patients. The results of this paper indicate that the focus should be more specifically on the two critical holidays of Christmas and Easter. The government could consider providing additional medical services during the two holidays in communities with people aged over 75. Due to the increase in CPD mortality in people aged over 75 in January and given the usual cold weather in Norway, the government should consider additional grants for people aged over 75 to better handle the cold weather. Our policy recommendations could also be extended to other countries beyond Norway.

Different age subgroups are affected differently by the environment; therefore, we recommend that policymakers, according to the distribution characteristics of the medical needs of patients of different age groups, allocate the current medical resources, thereby promoting the transfer of economic risks among patients of different age groups.

Ethical Approval

Ethical Approval from the Regional Committees for Medical and Health Research Ethics (REK)

References

Ahmad, T., Lund, L. H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., Dahlström, U., O'Connor, C.M., Felker, G. M., & Desai, N. R. (2018). Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association*, *7*(8), e008081. doi: 10.1161/JAHA.117.008081. PMID: 29650709; PMCID: PMC6015420

Ai Ahad, M. A., Sullivan, F., Demšar, U., Melhem, M., & Kulu, H. (2020). The effect of air-pollution and weather exposure on mortality and hospital admission and implications for further research: A systematic scoping review. *PLoS ONE*, 15(10 October), 1–24. https://doi.org/10.1371/journal.pone.0241415

Areal AT, Zhao Q, Wigmann C, Schneider A, Schikowski T. The effect of air pollution when modified by temperature on respiratory health outcomes: A systematic review and meta-analysis. *Science of the Total Environment*. 2022 Mar 10;811:152336. doi: 10.1016/j.scitotenv.2021.152336. Epub 2021 Dec 14. PMID: 34914983.

Baena-Cagnani, C. E., Patiño, C. M., Cuello, M. N., Minervini, M. C., Fernández, A. M., Garip, E. A., … Gómez, R. M. (1999). Prevalence and severity of asthma and wheezing in an adolescent population. *International Archives of Allergy and Immunology*, 118(2–4), 245–246. https://doi.org/10.1159/000024087

Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, 15(4), 1–44. https://doi.org/10.3390/ijerph15040780

Basu, R. (2009). High ambient temperature and mortality: A review of epidemiologic studies from 2001 to 2008. *Environmental Health*, *8*, 40 https://doi.org/10.1186/1476-069X-8-40

Bathmanabhan, S., & Saragur Madanayak, S. N. (2010). Analysis and interpretation of particulate matter - PM10, PM2.5, and PM1 emissions from the heterogeneous traffic near an urban roadway. *Atmospheric Pollution Research*, 1(3), 184–194. https://doi.org/10.5094/APR.2010.024

Bjørgen, A., & Ryghaug, M. (2022). Integration of urban freight transport in city planning: Lesson learned. *Transportation Research Part D: Transport and Environment*, 107(May). https://doi.org/10.1016/j.trd.2022.103310

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., … Van Der Veen, A. (1997). Mapping urban air pollution using gis: A regression-based approach. *International Journal of Geographical Information Science*, 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. *Environmental Health: A Global Access Science Source*, 6, 1–12. https://doi.org/10.1186/1476-069X-6-23

Carter, T. R., Fronzek, S., Inkinen, A., Lahtinen, I., Lahtinen, M., Mela, H., … Terama, E. (2016). Characterizing vulnerability of the elderly to climate change in the Nordic region. *Regional Environmental Change*, 16(1), 43–58. https://doi.org/10.1007/s10113-014-0688-7

Chay, K. Y., & Greenstone, M. (2003). Mortality: Evidence From Geographic. *Quarterly Journal of Economics*, 118(3), 1121–1167.

Chen, Q., Wang, Q., Xu, B., Xu, Y., Ding, Z., & Sun, H. (2021). Air pollution and cardiovascular mortality in Nanjing, China: Evidence highlighting the roles of cumulative exposure and mortality displacement. *Chemosphere*, 265, 129035. https://doi.org/10.1016/j.chemosphere.2020.129035

Conte, M., & Contini, D. (2019). Size-resolved particle emission factors of vehicular traffic derived from urban eddy covariance measurements. *Environmental Pollution*, 251(2019), 830–838. https://doi.org/10.1016/j.envpol.2019.05.029

Conte, M., Donateo, A., & Contini, D. (2018). Characterization of particle size distributions and corresponding size-segregated turbulent fluxes simultaneously with CO2 exchange in an urban area. *Science of the Total Environment*, 622–623, 1067–1078. https://doi.org/10.1016/j.scitotenv.2017.12.040

Côté, M., Osseni, M. A., Brassard, D., Carbonneau, É., Robitaille, J., Vohl, M., Lemieux, S., Laviolette, F., & Lamarche, B. (2022). Are machine learning algorithms more accurate in predicting vegetable and fruit consumption than traditional statistical models? An exploratory analysis. *Frontiers in Nutrition*, *9*. https://doi.org/10.3389/fnut.2022.740898

Folgerø, I. K., Harding, T., & Westby, B. S. (2020). Going fast or going green? Evidence from environmental speed limits in Norway. *Transportation Research Part D: Transport and Environment*, 82, 102261. https://doi.org/10.1016/j.trd.2020.102261

Folkehelseinstituttet. (2021). Hjerte - og karsykdommer i Norge, 1–35. Retrieved from https://www.fhi.no/nettpub/hin/ikke-smittsomme/Hjerte-kar/#om-hjerte-og-karsykdommer

Font, A., & Fuller, G. W. (2016). Did policies to abate atmospheric emissions from traffic have a positive effect in London? *Environmental Pollution*, 218, 463–474. https://doi.org/10.1016/j.envpol.2016.07.026

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, *56*, 229–238. doi: 10.1016/j.jbi.2015.05.016. Epub 2015 Jun 1. PMID: 26044081

Gauderman, W. J., Vora, H., McConnell, R., Berhane, K., Gilliland, F., Thomas, D., … Peters, J. (2007). Effect of exposure to traffic on lung development 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

Goals, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., Kvedar, J., & Jethwani, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*, *18*(1), 44. doi: 10.1186/s12911-018-0620-z. PMID: 29929496; PMCID: PMC6013959

Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart Journal*, *38*(23), 1805–1814. doi: 10.1093/eurheartj/ehw302. PMID: 27436868; PMCID: PMC5837244

Green, C. P., Heywood, J. S., & Navarro, M. (2016). Traffic accidents and the London congestion charge. *Journal of Public Economics*, 133, 11–22. https://doi.org/10.1016/j.jpubeco.2015.10.005

Green, C., & Krehic, L. (2022). An extra hour was wasted. Bar closing hours and traffic accidents in Norway. *Health Economics (United Kingdom),* 31(8), 1752–1769. https://doi.org/10.1002/hec.4550|

Grigorieva, E., & Lukyanets, A. (2021). The combined effect of hot weather and outdoor air pollution on respiratory health: A literature review. *Atmosphere*, *12*(6), 790. https://doi.org/10.3390/atmos12060790

Gryech, I., Ghogho, M., Elhammouti, H., Sbihi, N., & Kobbane, A. (2020). Machine learning for air quality prediction using meteorological and traffic-related features. *Journal of Ambient Intelligence and Smart Environments*, 12(5), 379–391. https://doi.org/10.3233/AIS-200572

Gualtieri, G., Crisci, A., Tartaglia, M., Toscano, P., & Gioli, B. (2015). A statistical model to assess air quality levels at urban sites. *Water, Air, and Soil Pollution*, 226(12). https://doi.org/10.1007/s11270-015-2663-4

Huntink, E., Wensing, M., Klomp, M. A., & Van Lieshout, J. (2015). Perceived determinants of cardiovascular risk management in primary care: Disconnections between patient behaviors, practice organization, and the healthcare system. *BMC Family Practice*, 16(1), 1–13. https://doi.org/10.1186/s12875-015-0390-y

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. *Journal of Environmental Management*, 217, 164–174. https://doi.org/10.1016/j.jenvman.2018.03.094

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. *Atmospheric Environment*, 122, 133–141. https://doi.org/10.1016/j.atmosenv.2015.09.019

Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on DWT Decomposition. *Procedia Computer Science*, 48(C), 173–179. https://doi.org/10.1016/j.procs.2015.04.167

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *Journal of Electrical and Computer Engineering*, 2017. https://doi.org/10.1155/2017/5106045

Krittanawong, C., Bomback, A. S., Baber, U., Bangalore, S., Messerli, F. H., & Wilson Tang, W. H. (2018). Future direction for using artificial intelligence to predict and manage hypertension. *Current Hypertension Reports*, *20*(9), 75. doi: 10.1007/s11906-018-0875-x. PMID: 29980865

Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. H. W., Halperin, J. L., & Narayan, S. M. (2019). Deep learning for cardiovascular medicine: A practical primer. *European Heart Journal 40*(25), 2058–2073. doi: 10.1093/eurheartj/ehz056. PMID: 30815669; PMCID: PMC6600129

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, *69*(21), 2657–2664. doi: 10.1016/j.jacc.2017.03.571. PMID: 28545640

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 Climate*, 10(2014), 703–719. https://doi.org/10.1016/j.uclim.2014.03.008

Lan, Y., & Wu, S. Impacts of environmental insults on cardiovascular aging. (2022). *Current Environmental Health Reports*, *9*(1), 11–28. doi: 10.1007/s40572-022-00335-x. Epub 2022 Feb 1. PMID: 35103958

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi 10.10 38/nature14539. PMID: 26017442

Luecken, D. J., Hutzell, W. T., & Gipson, G. L. (2006). Development and analysis of air quality mo -deling simulations for hazardous air pollutants. *Atmospheric Environment,* 40(26), 5087–50 96. https://doi.org/10.1016/j.atmosenv.2005.12.044

Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O. E., & Andersen, J. (2018). seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the prese -nt day. *Earth System Science Data*, *10*, 235–2249. https://doi.org/10.5194/essd-10-235-201 8

Madsen, C., Rosland, P., Hoff, D. A., Nystad, W., Nafstad, P., & Naess, O. E. (2012). The short-term effect of 24-h average and peak air pollution on mortality in Oslo, Norway. *European Jou -rnal of Epidemiology*, *27*(9), 717–727. doi 10.1007/s10654-012-9719-1. Epub 2012 Jul 27. PMID: 22836233

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 Pollution Research*, 7(3), 412–418. https://doi.org/10.1016/j.apr.2015.10.022

Moholdt, T., Afoakwah, C., Scuffham, P., McDonald, C. F., Burrell, L. M., & Stewart, S. (2021). Excess mortality at Christmas due to cardiovascular disease in the HUNT study prospective population-based cohort in Norway. *BMC Public Health*, *21*(1), 549. doi 10.1186/s12889-02 1-10503-7. PMID: 33743642; PMCID: PMC7980726

Næss, Ø., Nafstad, P., Aamodt, G., Claussen, B., & Rosland, P. (2007). Relation between the concentration of air pollution and cause-specific mortality: Four-year exposures to nitrogen dioxid -e and particulate matter pollutants in 470 neighborhoods in Oslo, Norway. *American Journa l of Epidemiology*, *165*(4), 435–443. https://doi.org/10.1093/aje/kwk016

Nafstad, P., Håheim, L. L., Wisløff, T., Gram, F., Oftedal, B., Holme, I., Hjermann, I., & Leren, P. (2004). Urban air pollution and mortality in a cohort of Norwegian men. *Environmental Heal -th Perspectives*, *112*(5), 610–615. doi: 10.1289/ehp.6684. PMID: 15064169; PMCID: PMC 1241929

Navares, R., & Aznarte, J. L. (2020). Deep learning architecture to predict daily hospital admissions. *Neural Computing and Applications*, *32*, 16235–16244 https://doi.org/10.1007/s00521-020-04840-8

Oftedal, B., Nafstad, P., Magnus, P., Bjørkly, S., & Skrondal, A. (2003). Traffic-related air pollution and acute hospital admission for respiratory diseases in Drammen, Norway 1995–2000. *European Journal of Epidemiology*,*18*(7), 671–675. doi 10.1023/a:1024884502114. PMID: 12952141

Parry, I., W. H., Walls, M., & Harrington, W. (2007). Automobile externalities and policies. *Journal of Economic Literature*, *45*(2), 373–399. doi 10.1257/jel.45.2.373

Pasquier, A., & André, M. (2017). Considering criteria related to spatial variabilities for the assessment of air pollution from traffic. *Transportation Research Procedia,* 25(June), 3354–3369. https://doi.org/10.1016/j.trpro.2017.05.210

Phung, D., Thai, P. K., Guo, Y., Morawska, L., Rutherford, S., & Chu, C. (2016). Ambient temperature and risk of cardiovascular hospitalization: An updated systematic review and meta-analysis. *Science of the Total Environment*, *550*, 1084–1102. doi: 10.1016/j.scitotenv.2016.01.154. Epub 2016 Feb 9. PMID: 26871555

Pothirat, C., Chaiwong, W., Liwsrisakun, C., Bumroongkit, C., Deesomchok, A., Theerakittikul, T., … Phetsuk, N. (2019). Acute effects of air pollutants on daily mortality and hospitalizations due to cardiovascular and respiratory diseases. *Journal of Thoracic Disease*, 11(7), 3070–3083. https://doi.org/10.21037/jtd.2019.07.37

Qu, H., Lu, X., Liu, L., & Ye, Y. (2019). Effects of traffic and urban parks on PM10 and PM2.5 mass concentrations. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects,* 45(2), 0–5647. https://doi.org/10.1080/15567036.2019.1672833

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina* (Kaunas), *56*(9), 455. doi: 10.3390/medicina56090455. PMID: 32911665; PMCID: PMC7560135

Raknes, G. (2022). Hjerte- og karsykdommer bidro til økt døde - lighet i 2021, 1 – 6.

Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology*, *13*(6), 350–359. doi 10.1038/nrcardio.2016.42. Epub 2016 Mar 24. PMID: 27009423. *Science of the total environment*, Volume 772,2021,145383, ISSN 0048-9697,https://doi.org/10.1016/j.scitotenv.2021.145383

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 po

-licies in Norway. *Transport Policy*, 99(August), 251–261. https://doi.org/10.1016/j.tranpol.2020.08.025

Stene, L. C., & Nafstad, P. (2001). Relation between the occurrence of type 1 diabetes and asthma. *Lancet*, *357*(9256), 607–608. doi: 10.1016/S0140-6736(00)04067-8. PMID: 11558491

Suri, J. S., Maindarkar, M. A., Paul, S., Ahluwalia, P., Bhagawati, M., Saba, L., Faa, G., Saxena, S., Singh, I. M., Chadha, P. S., Turk, M., Johri, A., Khanna, N. N., Viskovic, K., Mavrogeni, S., Laird, J. R., Miner, M., Sobel, D. W., Balestrieri, A., …, Fouda, M. M. (2022). Deep learning paradigm for cardiovascular disease/stroke risk stratification in Parkinson's disease affected by COVID-19: A narrative review. *Diagnostics* (Basel), *12*(7), 1543. doi 10.3390/diagnostics12071543. PMID: 35885449; PMCID: PMC9324237

Thompson, M. E., & Dulin, M. F. (2019). Leveraging data analytics to advance personal, population, and system health: Moving beyond merely capturing services provided. *North Carolina Medical Journal*, *80*(4), 214–218. doi: 10.18043/ncm.80.4.214. PMID: 31278180

Tian, Y., Liu, H., Si, Y., Cao, Y., Song, J., Li, M., Wu, Y., Wang, X., Xiang, X., Juan, J., Chen, L., Wei, C., Gao, P., & Hu, Y. (2019). Association between temperature variability and daily hospital admissions for cause-specific cardiovascular disease in urban China: A national time-series study. *PLoS Med*, *16*(1), e1002738. doi 10.1371/journal.pmed.1002738. PMID: 30689640; PMCID: PMC6349307

Turner, L. R., Barnett, A. G., Connell, D., & Tong, S. (2012). Ambient temperature and cardiorespiratory morbidity: A systematic review and meta-analysis. *Epidemiology*, *23*(4),594-606. doi: 10.1097/EDE.0b013e3182572795. PMID: 22531668

Varapongpisan, T., Frank, T. D., & Ingsrisawang, L. (2022). Association between out-patient visits and air pollution in Chiang Mai, Thailand: Lessons from a unique situation involving a large data set showing high seasonal levels of air pollution. *PLoS ONE*, 17(8 August), 1–14. https://doi.org/10.1371/journal.pone.0272995

Vogel, B., Acevedo, M., Appelman, Y., Bairey Merz, C. N., Chieffo, A., Figtree, G. A., Guerrero, M., Kunadian, V., Lam, C. S. P., Maas, A. H. E. M., Mihailidou, A. S., Olszanecka, A., Poole, J. E., Saldarriaga, C., Saw, J., Zühlke, L., & Mehran, R. (2021). The *Lancet* women and cardiovascular disease commission: Reducing the global burden by 2030. *Lancet*, 397(10292), 2385–2438. doi: 10.1016/S0140-6736(21)00684-X. Epub 2021 May 16. PMID: 34010613

Wærsted, E. G., Sundvor, I., Denby, B. R., & Mu, Q. (2022). Quantification of the temperature dependence of NOx emissions from road traffic in Norway using air quality modeling and monitoring data. *Atmospheric Environment: X*, 13(x), 100160. https://doi.org/10.1016/j.aeaoa.2022.100160

Weilnhammer, V., Schmid, J., Mittermeier, I., Schreiber, F., Jiang, L., Pastuhovic, V., Herr, C., & H einze, S. (2021). Extreme weather events in Europe and their health consequences: A system -atic review. *International Journal of Hygiene and Environmental Health*, 233, 113688. doi: 10.1016/j.ijheh.2021.113688. Epub 2021 Jan 30. PMID: 33530011

Zafeiratou, S., Samoli, E., Dimakopoulou, K., Rodopoulou, S., Analitis, A., Gasparrini, A., Stafoggi -a, M., De' Donato, F., Rao, S., Monteiro, A., Rai, M., Zhang, S., Breitner, S., Aunan, K., Sc hneider, A., Katsouyanni, K., & EXHAUSTION project team. (2021). A systematic review on the association between total and cardiopulmonary mortality/morbidity or cardiovascular risk factors with long-term exposure to increased or decreased ambient temperature. *Science of the Total Environment,* 772, 145383. doi: 10.1016/j.scitotenv.2021.145383. Epub 2021 Ja -n 27. PMID: 33578152

Zhang, P. G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neu -rocomputing*, 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

Zhao, Q., Zhao, Y., Li, S., Zhang, Y., Wang, Q., Zhang, H., Qiao, H., Li, W., Huxley, R., Williams, G., Zhang, Y., & Guo, Y. (2018). Impact of ambient temperature on clinical visits for cardio --respiratory diseases in rural villages in northwest China. *Science of the Total Environment*, *612*, 379–385. doi: 10.1016/j.scitotenv.2017.08.244. Epub 2017 Sep 1. PMID: 28858748

Zhu, T., Luo, L., Zhang, X., Shi, Y., & Shen, W. (2017). Time-series approaches for forecasting the number of hospitals' daily discharged inpatients. In *IEEE Journal of Biomedical and Health Informatics*, *21*(2), 515–526. doi: 10.1109/JBHI.2015.2511820

Appendix 1 The Data Pre-processing

The raw data usually are not clean and well formed; we must remove noise, missing values, and outliers, to suit the application of machine learning approaches, so data preprocessing improves the accuracy and efficiency of the machine learning approaches. Missing data will lead to incomplete datasets, which will affect the quality of data analysis and the accuracy of the results, especially for high-dimensional data or compound data, and even cause errors in the algorithms. Therefore, the preprocessing of missing values is important. Commonly used methods include the deletion method, the statistical imputation method, and the machine learning imputation method. It is important to choose an appropriate imputation method according to the data.

Appendix 2 Associations Between a Range of Exposure Effects to CPD Mortality and Natural Mortality

Contributors to CPD Mortality and Natural Mortality from Environmental Causes, Oslo, Bergen, Trondheim, and Tromsø, 2009–2018, Daily. This table has three pages.

	Environmental contributors	**CPD total mortality**	**Natural total mortality**
Observations		14,606	14606
R^2		0.64	0.61
Adjusted R^2		0.037	0.033
Single contributors and their lagged effects, and the magnitude is expanded by 10 to the 5th power.	**Traffic**	**4.13****	-0.69
	Traffic $_{t-2}$	-0.19	-0.35
	Traffic $_{t-4}$	0.52	1.98
	Traffic $_{t-6}$	0.40	-1.41
	NOx	5.21	**25.25****
	NOx $_{t-2}$	**34.06****	32.98
	NOx $_{t-4}$	4.783	-34.16
	NOx $_{t-6}$	19.17	40.83
	$PM_{2.5}$	31.82	82.91
	$PM_{2.5t-2}$	-27.70	25.94
	$PM_{2.5t-4}$	-19.56	360.78
	$PM_{2.5t-6}$	14.60	122.00
	Temp (Mean)	0.44	**22.46****
	Temp (Mean) $_{t-2}$	-0.58	2.01
	Temp (Mean) $_{t-4}$	-0.65	8.36
	Temp (Mean) $_{t-6}$	-1.50	9,42
	Temp (Max)	8.62	-102.43
	Temp (Max) $_{t-2}$	83.50	-79.18
	Temp (Max) $_{t-4}$	10.11	-135.31
	Temp (Max) $_{t-6}$	19.39	-125.64
	Temp (Min)	65.10*	-43.67
	Temp (Min) $_{t-2}$	52.60	132.80
	Temp (Min) $_{t-4}$	-10.03	18.92
	Temp (Min) $_{t-6}$	-15.77	26.54
	HDD	-2.76	-188.83
	HDD $_{t-2}$	28.77*	13.72
	HDD $_{t-4}$	6.72	-453.16
	HDD $_{t-6}$	33.03*	-38.37
	VP	100.81	163.44*
	VP $_{t-2}$	20.20	-0.23
	VP $_{t-4}$	5.14	11.67
	VP $_{t-6}$	27.24	153.69*
	WS(m/s)	-81.29	8.99
	WS(m/s) $_{t-2}$	-21.40	59.31
	WS(m/s) $_{t-4}$	-37.61	-120.06
	WS(m/s) $_{t-6}$	-3.25	82.62
	WG(m/s)	-82.63	46.54
	WG(m/s) $_{t-2}$	17.67	39.33
	WG(m/s) $_{t-4}$	-37.63	-59.08
	WG(m/s) $_{t-6}$	26.04	116.74
	NOx * Temp (Mean)	-0.01*	-0.01

	Environmental contributors	**CPD total mortality**	**Natural total mortality**
Interaction of NOx and environmental contributors, and their lagged effects, and the magnitude is expanded by 10 to the 5th power.	NOx * Temp (Mean) $_{t-2}$	-5.05e-03*	-4.54e-03
	NOx * Temp (Mean) $_{t-4}$	1.60e-03	-4.07e-03
	NOx * Temp (Mean) $_{t-6}$	-1.61e-03	-1.40e-07
	NOx * Temp (Max)	0.37	0.27
	NOx * Temp (Max)$_{t-2}$	-0.34	-0.36
	NOx * Temp (Max)$_{t-4}$	0.13	0.12
	NOx * Temp (Max) $_{t-6}$	-0.44*	-0.60
	NOx * Temp (Min)	0.33	0.26
	NOx * Temp (Min) $_{t-2}$	-0.33	-0.63
	NOx * Temp (Min) $_{t-4}$	-0.42	0.33
	NOx * Temp (Min) $_{t-6}$	-0.04	0.19
	NOx * HDD	0.11	0.03
	NOx * HDD $_{t-2}$	-0.36	0.29
	NOx * HDD $_{t-4}$	0.33	0.27
	NOx * HDD $_{t-6}$	-0.46	-0.87*
	NOx * VP	0.39	1.17
	NOx * VP $_{t-2}$	-0.69	-1.18
	NOx * VP $_{t-4}$	0.04	0.45
	NOx * VP $_{t-6}$	-0.44	-1.38
	NOx * Windspeed (Mean)	-1.66	-0.92
	NOx * Windspeed (Mean) $_{t-2}$	11.00	0.44
	NOx * Windspeed (Mean) $_{t-4}$	-0.28	0.58
	NOx * Windspeed (Mean) $_{t-6}$	2.30*	3.8*
	NOx * Wind Gust (Mean)	0.30	-0.71
	NOx * Wind Gust (Mean) $_{t-2}$	-0.04	0.90
	NOx * Wind Gust (Mean) $_{t-4}$	0.41	0.63
	Nox * Wind Gust (Mean) $_{t-6}$	-0.15	-0.10
Interaction terms of NOx with traffic volume, and their lagged effects.	NOx * Traffic	-0.01	-1.62e-03
	NOx * Traffic $_{t-2}$	1.30e-03	0.02
	NOx * Traffic $_{t-4}$	-1.59e-03	-3.81e-03
	NOx * Traffic $_{t-6}$	-8.19e-03	-0.02
Interaction of PM$_{2.5}$ and weather factors, and their lagged effects , and the magnitude is expanded by 10 to the 5th power.	PM$_{2.5}$ * Temp (Mean)	0.03	0.01
	PM$_{2.5}$ * Temp (Mean) $_{t-2}$	-0.02	0.03
	PM$_{2.5}$ * Temp (Mean) $_{t-4}$	-0.07	-0.06
	PM$_{2.5}$ * Temp (Mean) $_{t-6}$	0.11	0.11
	PM$_{2.5}$ * Temp (Max)	-1.26	-2.81
	PM$_{2.5}$ * Temp (Max) $_{t-2}$	1.06	3.39
	PM$_{2.5}$ * Temp (Max) $_{t-4}$	-1.14	-3.52
	PM$_{2.5}$ * Temp (Max) $_{t-6}$	0.03	0.07
	PM$_{2.5}$ * Temp (Min)	0.10	2.11
	PM$_{2.5}$ * Temp (Min) $_{t-2}$	0.89	1.48
	PM$_{2.5}$ * Temp (Min) $_{t-4}$	2.75	2.13
	PM$_{2.5}$ * Temp (Min) $_{t-6}$	**-5.92****	-5.77
	PM$_{2.5}$ * HDD	-1.66	-2.99
	PM$_{2.5}$ * HDD $_{t-2}$	1.98	1.22
	PM$_{2.5}$ * HDD $_{t-4}$	-2.30	-1.10
	PM$_{2.5}$ * HDD $_{t-6}$	0.44	-1.64
	PM$_{2.5}$ * VP	-2.93	-1.31
	PM$_{2.5}$ * VP $_{t-2}$	-0.49	-2.54
	PM$_{2.5}$ * VP $_{t-4}$	0.17	1.68
	PM$_{2.5}$ * VP $_{t-6}$	-2.25	3.86
	PM$_{2.5}$ * Windspeed (Mean)	2.42	-7.50

	Environmental contributors	**CPD total mortality**	**Natural total mortality**
	PM$_{2.5}$ * Windspeed (Mean)$_{t-2}$	-5.36	9.12
	PM$_{2.5}$ * Windspeed (Mean)$_{t-4}$	6.53	8.71
	PM$_{2.5}$ * Windspeed (Mean)$_{t-6}$	1.51	3.55
	PM$_{2.5}$ * Wind Gust (Mean)	0.79	1.48
	PM$_{2.5}$ * Wind Gust (Mean)$_{t-2}$	0.16	-2.52
	PM$_{2.5}$ * Wind Gust (Mean)$_{t-4}$	4.68	0.08
	PM$_{2.5}$ * Wind Gust (Mean)$_{t-6}$	0.06	-0.43
Interaction terms of PM$_{2.5}$ with traffic volume, and their lagged effects.	PM$_{2.5}$ * Traffic	-0.01	0.13
	PM$_{2.5}$ * Traffic$_{t-2}$	-0.12	-0.03
	PM$_{2.5}$ * Traffic$_{t-4}$	-0.01	0.09
	PM$_{2.5}$ * Traffic$_{t-6}$	0.04	8.80e-03
Time variables	**The month of the year**		
	February	-0.01	-0.11
	March	-0.10	**-0.28****
	April	**-0.25*****	**-0.40*****
	May	**-0.24****	**-0.49*****
	June	**-0.24***	**-0.48****
	July	-0.11	-0.33
	August	-0.14	-0.36
	September	**-0.22***	**-0.39***
	October	**-0.36*****	**-0.55*****
	November	**-0.43*****	**-0.59*****
	December	**-0.25*****	**-0.36*****
	The year 2009		
	2010	-0.05	-0.03
	2011	-0.07	0.12
	2012	**-0.11***	0.05
	2013	**-0.22*****	-0.13
	2014	**-0.27*****	**-0.19***
	2015	**-0.27*****	-0.14
	2016	**-0.31*****	**-0.21***
	2017	**-0.36*****	**-0.25***
	2018	**-0.43*****	**-0.33****
	_cons	3.15***	134.97***
	i.holiday	-0.12	-0.07

Notes: Here t-2, t-4, and t-6 represent lags of 2 days, 4 days, and 6 days, respectively. Two digits after the decimal point are taken. *, **, and *** indicate statistical significance at the $P < 0.1$, $P < 0.05$, and $P < 0.01$ levels, respectively; PM$_{2.5}$ * Windspeed (Mean)$_{t-6}$ represents exposure six days ago to the interaction of PM$_{2.5}$ and average mean wind speed. Mortality is the number of deaths per 1000 people per year. This is a two-way fixed effects model with fixed cities and fixed time.

Appendix 3 Estimating the Total CPD Mortality from Environmental Factors with Three RM Models.

Using Three RM Models to Estimate the Contributors to CPD Mortality from Environmental Causes, Oslo, Bergen, Trondheim, and Tromsø, 2009–2018, Daily. This table has two pages.

	Environmental contributors	**RM (1)** **Complex**	**RM (2)** **Simple**	**RM (3)** **Middle**
Observations		14,606	14606	14606
R^2		0.64	0.07	0.43
Adjusted R^2		0.0337	0.0327	0.0333
Single contributors and their lagged effects, and the magnitude is expanded by 10 to the 5th power.	**Traffic**[9]	**4.13****	0.48	2.06*
	NOx	5.21	**5.99****	18.37
	$PM_{2.5}$	31.82	-23.44	22.76
	Temp (Mean)	0.44	**-2.21****	-1.19
	Temp (Max)	8.62	**31.23****	23.49
	Temp (Min)	65.10*	19.49	12.60
	HDD	-2.76	23.83	29.74
	VP (kPa)	100.81	7.98	57.32
	WS (m/s)	-81.29	-58.33	57.85
	WG (m/s)	-82.63	-37.22	-68.72
Interaction of NOx and environmental contributors, and their lagged effects, and the magnitude is expanded by 10 to the 5th power.	NOx * Temp (Mean)	-0.01*	N/A[10]	-0.01
	NOx * Temp (Max)	0.37	N/A	-0.03
	NOx * Temp (Min)	0.33	N/A	0.03
	NOx * HDD	0.11	N/A	-0.09
	NOx * VP	0.39	N/A	0.29
	NOx * WS (Mean)	-1.66	N/A	-0.45
	NOx * WG (Mean)	0.30	N/A	-0.18
Interaction terms of NOx with traffic volume, and their lagged effects.	**NOx * Traffic**	-0.01	N/A	**-0.02****
Interaction of $PM_{2.5}$ and weather factors, and their lagged effects, the magnitude is expanded by 10 to the 5th power	$PM_{2.5}$* Temp (Mean)	0.03	N/A	-0.01
	$PM_{2.5}$ * Temp (Max)	-1.26	N/A	0.68
	$PM_{2.5}$ * Temp (Min)	0.10	N/A	0.40
	$PM_{2.5}$ * HDD	-1.66	N/A	-0.43
	$PM_{2.5}$ * VP	-2.93	N/A	-5.1*
	$PM_{2.5}$ * WS (Mean)	2.42	N/A	-6.97
	$PM_{2.5}$ * WG (Mean)	0.79	N/A	2.16
Interaction terms of $PM_{2.5}$ with traffic volume, and their lagged effects.	$PM_{2.5}$ * Traffic	-0.01	N/A	-0.04
Time variables.	**The month of the year**			
	February	-0.01	-0.04	-0.03

[9] The estimated value for traffic here is 4.13e-5, which is the number of deaths per 1000 people per year, and Norway has about 5 million inhabitants, so the effect on 5 million is 0.2065, therefore each increase in traffic in Norway results in about 0.2 deaths from CPD per year.

[10] N/A represents no estimated value.

March	-0.10	**-0.16****	**-0.15***
April	**-0.25*****	**-0.35*****	**-0.33*****
May	**-0.24****	**-0.36*****	**-0.35*****
June	**-0.24***	**-0.42*****	**-0.41*****
July	-0.11	**-0.33*****	**-0.34*****
August	-0.14	**-0.38*****	**-0.38*****
September	**-0.22***	**-0.42*****	**-0.40*****
October	**-0.36*****	**-0.48*****	**-0.45*****
November	**-0.43*****	**-0.51*****	**-0.50*****
December	**-0.25*****	**-0.28*****	**-0.28*****
The year 2009			
2010	-0.05	-0.03	-0.03
2011	-0.07	-0.07	-0.08
2012	**-0.11***	**-0.11***	**-0.11***
2013	**-0.22*****	**-0.21*****	**-0.22*****
2014	**-0.27*****	**-0.29*****	**-0.30*****
2015	**-0.27*****	**-0.26*****	**-0.27*****
2016	**-0.31*****	**-0.32*****	**-0.33*****
2017	**-0.36*****	**-0.40*****	**-0.40*****
2018	**-0.43*****	**-0.47*****	**-0.47*****
_cons	3.15***	**3.55*****	3.40***
holiday	-0.12	-0.12	-0.12

Notes: Here t-2, t-4, and t-6 represent lags of 2 days, 4 days, and 6 days, respectively. Two digits after the decimal point are taken. *, **, and *** indicate statistical significance at the 0.1, 0.05, and 0.01 levels, respectively. Mortality is the number of deaths per 1000 people per year. This is a two-way fixed effects model with fixed cities and fixed time. The difference between RM (1) and RM (3) is whether it includes lagged effects. Only RM (1) includes lagged effects, so there is no comparison of the results of lagged effects. This table does not include the results of lagged effects. The results of RM (1) lagged effects are in Appendix 2.

Chapter 3

Off-Premises Alcohol Availability and Traffic Accidents: Evidence from the Extension of the Norwegian Wine Monopoly[*]

Cong Cao [a], Colin P. Green [a]

[a] Department of Economics, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway

Abstract

Alcohol availability has been demonstrated to influence a range of social outcomes, including traffic accidents, injuries, and fatalities. Existing literature has focused primarily on on-premises availability, yet there are marked variations in off-premises availability with the potential for large effects on traffic safety. We return to this issue in Norway and exploit large changes in the off-premises availability of high-strength alcohol availability through the expansion of government wine monopoly stores over the last 20 years. We demonstrate variations in the effect of off-premises availability according to the margin of adjustment. Municipalities that open their first store experience an increase in accidents that are concentrated in those involving light injuries. These are concentrated in the period directly following the opening. Conversely, expansions of stores beyond the initial store appear to reduce accidents, including those involving serious accidents or fatalities. Our results have implications for alcohol and road safety policy.

JEL Classification: I18, R41

Keywords: Alcohol Availability, Alcohol Policy, Traffic Accidents

[*] Acknowledgment: Cong Cao is grateful for the research funding from Statens vegvesen (the Norwegian public roads administration).

1 Introduction

The effects of alcohol across a range of health outcomes are a focus of numerous policy interventions. A key aspect is traffic accidents and fatalities. At the forefront of policy interventions in this area has been criminalizing drunk driving and lowering permissible blood alcohol levels while driving. At the same time, a growing literature shows that the availability of alcohol has the potential to substantially influence road accidents. A focus of the literature has been on timing-related availability in on-premises sales. This literature primarily uses changes in permissible on-premises drinking hours as a source of variation. The evidence from this resulting literature is mixed. For instance, while Vingilis et al. (2005) find no impact of extensions of on-premises sales on traffic fatalities in Ontario, Canada, in a second paper they show increases in Windsor, Canada, and a reduction in bordering, later-closing, Detroit from the harmonization of opening hours across the border (Vingilis et al, 2006). Green et al. (2014) demonstrate reductions in traffic accidents from a large liberalization in bar closing hours in England and Wales. While Biderman et al. (2010) examine the effect of a restriction of bar closing hours in the Sao Paulo Metropolitan Area and demonstrate a reduction in fatal traffic accidents. Recently, Green and Krehic (2022) in a study of Norway, the focus of this paper, demonstrate an overall zero effect of opening hours on accidents, that hides marked variations which may reflect differences in access to public transport at bar closing hours.

A less understood factor is the underlying level of alcohol availability in general. This, to an extent, reflects difficulties in measuring. Yet, the availability of alcohol off-premises, both in terms of the time of the day and locations, is marked in many jurisdictions. At the same time, it is an area of active government intervention, and off-premises drinking may often dwarf on-premises drinking. Along these lines, Marcus et al. (2015) demonstrate that changes in alcohol availability have large effects on alcohol consumption, and through this, large effects on alcohol-related harms. In a similar vein, Heaton (2012) demonstrates the marked effects of liberalizing Sunday off-premises alcohol availability on crime.

We return to this issue focusing on Norway, which presents an interesting point of study for a range of reasons. First, and foremost, drinks with over 4.7% alcohol content are only available at government-owned and run stores ('Vinmonopolet' – herein the wine monopoly).[11] While the first of these stores was established in Oslo as early as 1922, there were still only 54 stores across all of

[11] As the name suggests the original wine monopoly sold only wine.

Norway by the end of 1965. However, recent years, and the 20 years covered by our data, have witnessed a dramatic increase in both the number of stores and geographic coverage of these stores. As we argue later, this is a substantial increase in the availability of medium to high-strength alcohol. This has potential implications for a range of important social outcomes, including traffic accidents and fatalities which form the focus of our paper.

Our main approach uses the roll-out of these stores as a source of variation in alcohol availability. We combine this at a municipal level (approx. 415 in Norway in our period of analysis) with administrative data on traffic incidents to estimate the effect of off-premises availability on traffic accidents. We explore two different margins of availability. First, we examine the opening of monopoly stores in municipalities without a store (extensive margin). Second, we examine the effects of increased availability within a given municipality in the form of increases in stores in municipalities with existing stores (intensive margin). Doing so demonstrates contrasting results. Opening a store in a municipality that previously had no store leads to a small but statistically significant increase in accidents in the order of 1.5 accidents a year. At the same time, increases in the intensive margin of availability lead to reductions in accidents. These patterns are robust to a range of standard concerns, along with alternative context-specific variations in availability, such as proximity to the Swedish border (a source of cheaper alcohol).

There are a number of advantages of our setting in addition to the wide range of changes in off-premises availability we observe. Off-premises availability, through monopoly stores, varies at the municipal level while other policy changes likely to confound estimates are set at a higher level. Most notably, broader policies regarding drink-driving penalties, limits, and other off-premises alcohol laws are all nationally set. These laws and policies simply do not vary in our period of analysis. Similarly, policing decisions are not made at the same level, or by the same authorities, as those who determine monopoly store openings and locations.

2 Literature Review

There is existing, largely descriptive, literature on the relationship between off-premises alcohol availability and traffic safety. These are often cross-sectional studies, see, for instance, Kelleher et al. (1996) and Tonkin et al. (1997). Many of these studies demonstrate a positive association between alcohol availability and traffic accidents. For instance, Jewell et al. (1995) demonstrate a positive correlation between alcohol supply and alcohol-related traffic accidents across Texas counties, where travel costs of obtaining alcohol appeared as a primary factor. Smith (1988) examines Sunday trading in Brisbane and demonstrates a significant increase in Sunday traffic casualties, and this appears to be related to the hours of Sunday opening. Treno et al. (2007) study the relationship between the number of alcohol stores and traffic accidents (including non-alcoholic accidents) in California, they find that locations with more bars and off-premises stores had more traffic accidents. Wang et al. (2020) study traffic accidents in Tianjin, China, from 2011 to 2013, and shows a relationship between the density of alcohol retail stores and traffic accidents. Similarly, Morrison et al. (2016) demonstrate a positive relationship between the density of alcohol sales points, including on-premises locations, and traffic accidents in Melbourne, Australia. Gruenewald et al. (2010) find a positive relationship between the density of off-premises alcohol stores and the incidence of traffic injuries among young people aged 21 to 29, and they also observe that the incidence is higher in densely populated areas. Lipton et al. (2021) discuss the relationship between road characteristics, alcohol-related traffic accidents, and alcohol sales stores by using data from 2006 to 2009 in 50 cities in California and find that off-premises stores contribute to all alcohol- and non-alcohol-related accidents, and road characteristics do not change this relationship, meaning both are positively related regardless of road characteristics.

Other research uses before and after analysis of changes in alcohol availability. For instance, Han et al. (2014) examine the changes in accidents in one Texas city following the implementation of a new alcohol licensing policy that increased off-premises alcohol availability. They find that traffic accidents decreased after the policy was implemented. Trolldal (2005) examines the impact of Canadian alcohol retail privatization on traffic accidents. Across the period of 1950 to 2000, he finds no effect of privatization on fatal traffic accidents.

There is a range of other studies that find zero or even a reduction in traffic accidents when off-premises alcohol availability is higher. For instance, Ponicki et al. (2013) examine data from 58 counties in California from 1999 to 2008. They find that off-premises alcohol stores' density is statistically negatively correlated with traffic accidents, but statistically positively correlated with bar density. They note that the effects are very small. Tang (2013) reaches a similar conclusion in a study of 254 counties in Texas from 1975 to 1996. They emphasize that this may reflect that off-premises stores in this setting sell relatively low-alcohol alcohol products, such as beer and wine. Avdic et al. (2021) conducts a study on the increase in the business hours of alcohol retail stores and the related impacts in 290 cities in Sweden from 2008 to 2015 and they find that there is no effect on alcohol-related adverse effects, including traffic accidents.

There exists a small, related literature for Norway. This literature often also studies drugs or psychotropic drugs together with alcohol, and most are medical studies (Valen et al., 2019). For example, Gjerde et al. (2011) study fatal accidents in south-eastern Norway from 2003 to 2008 and find that together alcohol and drugs, as well as alcohol alone, increase the risk of fatal road traffic accidents. Pasnin et al. (2021) study the alcohol and drug use status of drivers in fatal traffic accidents from 2016 to 2018 for all of Norway. They find that high blood alcohol or drug concentrations above the limit are still the main contributing factors to fatal traffic accidents. Gjerde et al. (2023) explore fatal accidents and alcohol and drug use in Norway, from 2011 to 2020. They find that alcohol use is one of the main causes of fatal traffic accidents.

Although alcohol control can help reduce drunk driving and improve road safety, policy implementation is challenging. Norström et al. (2013) study 32 years of data on drinking and driving under the influence in Norway and Sweden. They find that the control of total alcohol consumption can help reduce the incidence of drinking and driving, but differs across the two countries, and is closely related to public opinion, drinking culture, and population segmentation. Middleton et al. (2010) study the impact of changes in the number of days of alcohol sales, including off-premises alcohol sales, they compare the United States and many countries including Norway, and find that increasing alcohol sales on weekends, will increase the risk of negative harms, such as drinking-related traffic accidents. Reducing the number of days of sales or the overall alcohol consumption will reduce this risk, but this faces opposition from the alcohol industry and a range of economic factors.

3 Institutional Framework and the Data

Norway has a national alcohol policy concerning off-premises alcohol sales. Alcohol that is 4.7% or under can be bought from stores such as supermarkets, but only up to 8 pm on weekdays, 6 pm on Saturdays, and not on Sundays. Stronger alcohol is only available for purchase from government-run specialty stores, namely the wine monopoly (Vinmonopolet). The wine monopoly was originally established in 1922 as an effort to regulate alcohol sales, and reduce, amongst others, home production and consumption of alcohol. The original store was in Oslo, and inhabitants of Norway outside of Oslo could purchase alcohol from this store via mail order. All other sales of alcohol were illegal in

Norway, and at this time low-alcohol drinks were not available from other stores. In the 1960s, there was an expansion of these stores to other major cities in Norway such as Bergen, Trondheim, and Stavanger. Slowly, starting from the 1970s, there was expansion in the number of openings across Norway such that by 1999, the start of our analysis period, there existed 131 stores in 100 municipalities in Norway. Yet, there remained 315 municipalities, covering a population of 2,086,035, without local access to a wine monopoly store. At the same time, while mid-size to large city municipalities had monopoly stores by this point, there were typically few of them and they were concentrated in the city center and/or a handful of locations.[12]

Importantly, for our purposes, the following twenty years have witnessed a quite dramatic increase in the number of wine monopoly stores in Norway, such that by 2019 there were 326 stores, but still 121 municipalities without a store. Figure 1 provides an overview of both the growth in the number of monopoly stores in all municipalities and the growth in those municipalities that initially already have at least one store coverage. These figures reflect that there are now substantially fewer municipalities without stores. There has been quite a marked expansion in-store availability in the largest cities. As an example of the latter point, Oslo has increased from 17 stores in 1999 to 30 stores by the end of 2019. Despite these increases over the period, there remains quite limited geographic and temporal availability. While opening hours are limited at wine monopoly stores. Standard openings are typically weekdays 10 am until 6 pm; 10 am-3 pm on Saturdays; and closed on Sundays.[13]

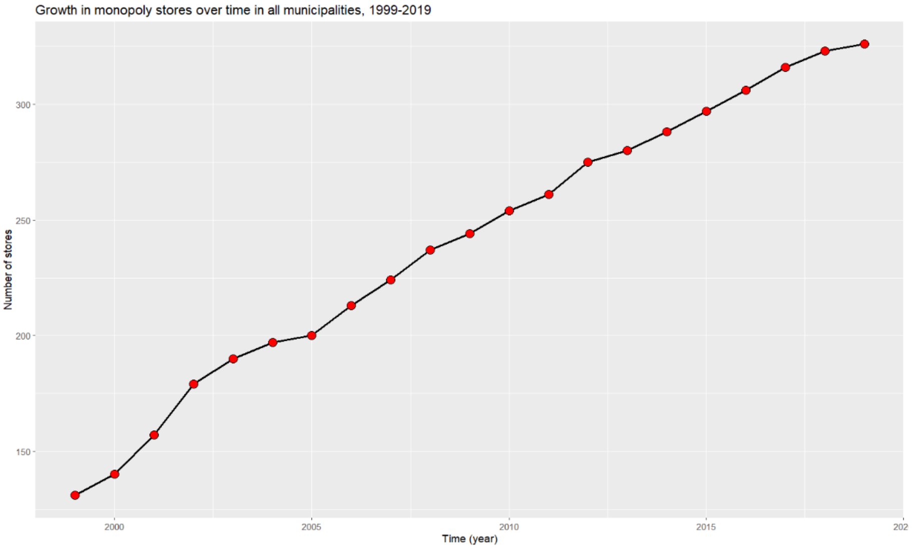
[12] For instance, at the start of 1999 Oslo municipality had only 17, Bergen 6, and Trondheim 4 stores. These numbers, as we discuss further, increased dramatically over the following 20 years.

[13] Some smaller stores have shorter opening hours. Notably the timing opening hours remained unchanged in our period of analysis. The extension of opening hours to 6pm, from 5pm, occurred during the covid period which, as we discuss later, we omit from our analysis.

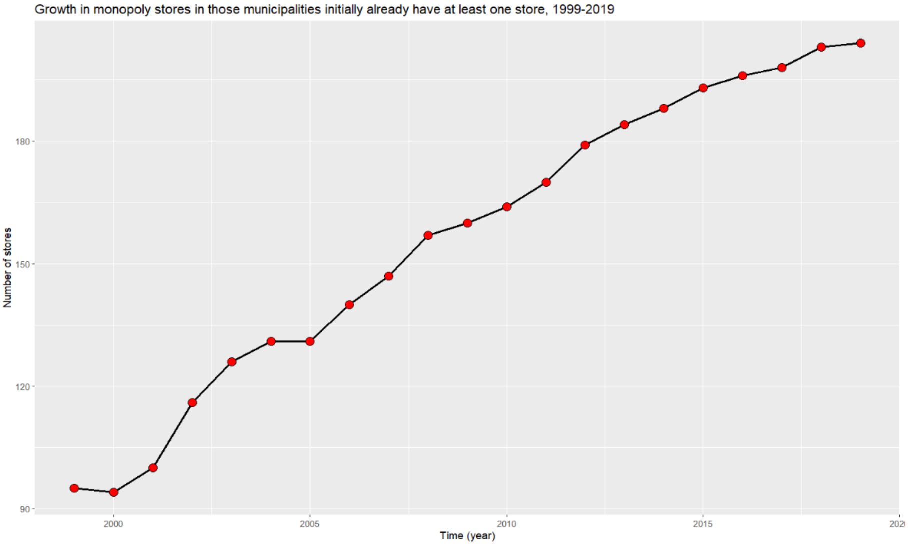
Figure 1

Growth in Monopoly Stores Over Time in Norway, 1999-2019

Panel A Growth in Monopoly Stores Over Time in All Municipalities



Panel B Growth in Monopoly Stores in those Municipalities Initially already Have At Least One Store



Data

Data on monopoly stores comes from administrative data available from the Vinmonopolet, and publicly available data from the Bronnøysund company register. From these data, we can determine the location of all stores in Norway, both in terms of their address and the municipality in which they are located, along with the first time they opened. It is this information that forms the basis of our measure of off-premises alcohol availability. While these stores vary in size and range, they typically do not differ in terms of the general types of alcohol available or available during opening hours. We treat each store as having a homogeneous effect on alcohol availability.

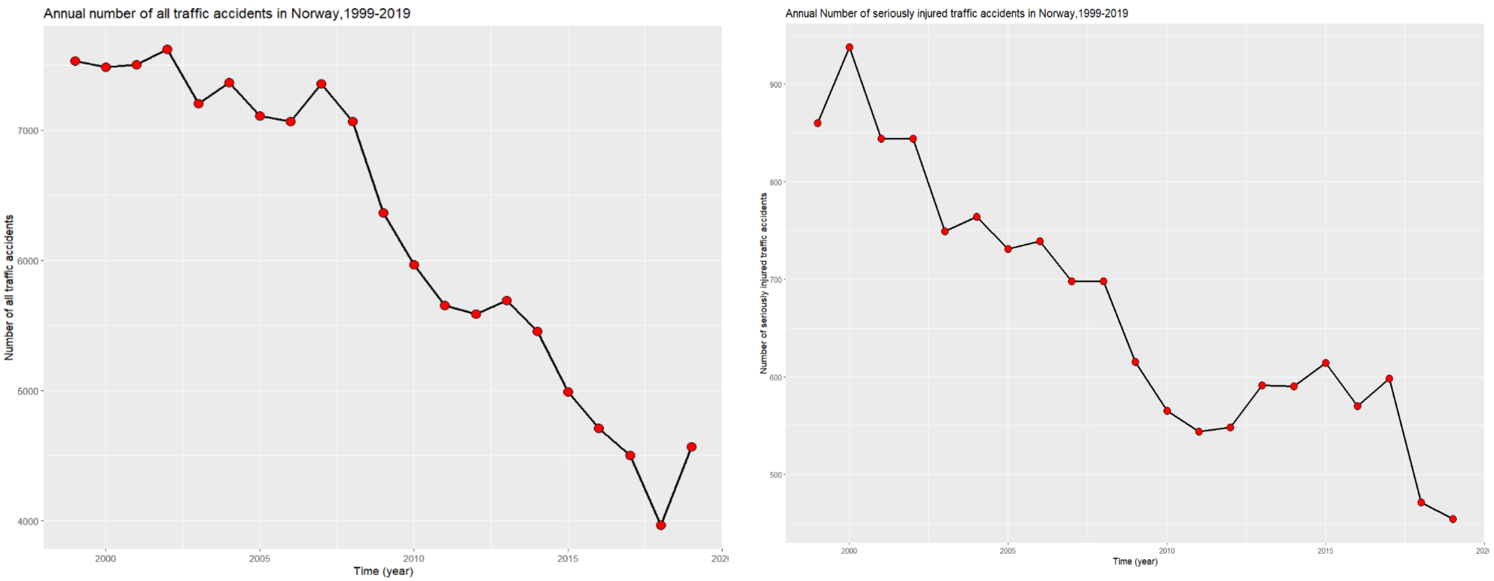
Road accident data are obtained from the Norwegian Public Roads Administration (NPRA) for the period 1999 to 2019, which included all motor vehicle accidents reported to the police in 415 municipalities. We stop our analysis period before the Covid period, which was characterized at times by changes in both alcohol availability and traffic behavior. These data allow us to match an accident location to whether the corresponding municipality had a monopoly store. We conduct all analyses at a monthly level. This leads to some measurement errors concerning stores that are open during the month. We explore the robustness of our results to alternative treatments for this.[14]

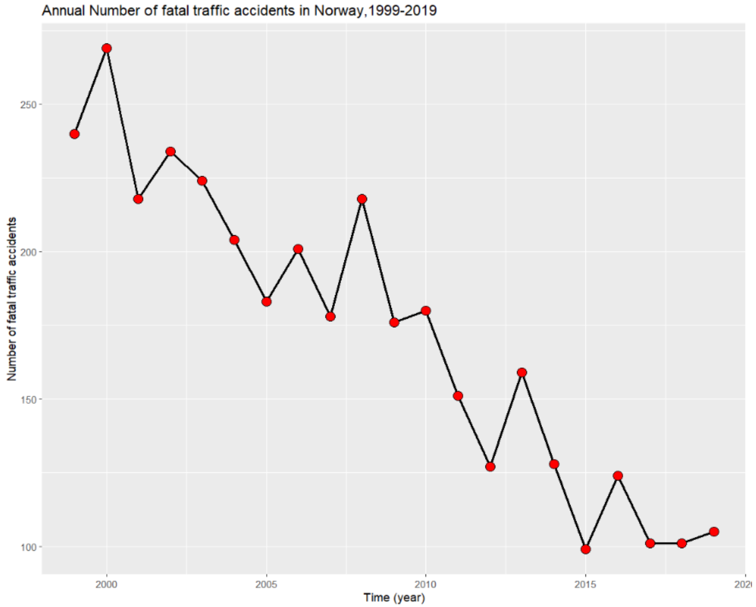
Figure 2 provides an overview of traffic accidents over this period. In general, and in common with several other countries, traffic accidents have been decreasing over this period. As can be seen, this is true for all accidents, and serious and/or fatal accidents alone. From Figure 2, we can see that the number of traffic accidents has been decreasing year by year, with a slight increase in 2002, 2005, and 2008, showing the lowest value in 2018, and rising again in 2019. According to a report by the Norwegian public health administration (Folkehelseinstituttet, 2021), traffic accident-related deaths have been declining since 1973, as traffic accidents account for a large proportion of deaths from external causes.

[14] A complicating factor is that there were substantial municipal mergers in 2020. Our approach is to use the municipalities that existed prior to these changes.

Figure 2

Annual Accident Numbers in Norway, 1999–2019





Note: The x-axis in the figure represents the year from 1999 to 2019, the y-axis represents the number of traffic accidents per year, and each point in the figure represents the number of accidents in that year.

Table 1 provides summary statistics of our data. We report statistics for all municipalities, but also those municipalities who began the period without a monopoly store, those who started with at least one, and those who change from zero to at least one in the period. Naturally, a key difference across these types of municipalities is that those that initially had at least one store been more populous. They also experienced higher numbers of accidents across all types.

Additionally, we report differences between those municipalities that were without a store at the beginning of our data period and gained at least one in the period and did not. In 1999, 76.1 % of all municipalities had no wine monopoly store; over the two-decade period, 32.6 % of municipalities went from having no store to having at least one wine monopoly store. The difference between the two is that the from 0 to 1 municipal group has more traffic accident rates, regardless of what type of accident or when it happened.

Table 1

Descriptive Statistics of the Data (1999–2019)

	(1) All	(2) Zero Stores (In 1999)	(3) At Least One	(4) Changed from 0 to 1
Accidents per month	1.256	0.729	2.49	0.898
Serious Accidents per month	0.135	0.091	0.274	0.106
Fatal Accidents per month	0.035	0.026	0.062	0.030
Light Accidents per month	1.023	0.564	2.494	0.698
Accidents that happen on the weekend per month	0.315	0.201	0.683	0.242
Single-vehicle accidents per month	0.431	0.317	0.795	0.373
Multiple-vehicle accidents per month	0.106	0.052	0.278	0.068
Population	11,492	5,971	29,047	7853
Young adults-men	1,251	591	3,351	779
Young adults-women	1,204	556	3,263	733
Number of Municipalities	415	315	100	135

Note: All values are mean values.

4 Empirical Approach

Our main approach is to estimate variants of:

$$Acc_{it} = \beta_0 + \beta_1 Monopoly_{it} + \gamma X_{it} + \alpha_i + \delta_t + \varepsilon_{it} \quad (1)$$

Where Acc_{it} is the number of accidents in municipality i in month t. Vector X includes the population level and the number of young adults (aged 18 to 34) in municipality i in year t. α_i and δ_t captures municipal and time-fixed effects, respectively, and ε_{it} is a random idiosyncratic error. Our parameter of interest β_1 is identified by within municipality variation in wine monopoly openings. We focus on two margins. First, an extensive margin where municipalities that previously had no wine monopoly store experience an opening. Second, an intensive margin where municipalities increase the number of monopoly stores in the period. This approach reflects a view that these might generate heterogeneous effects due to both their different implications for alcohol availability, and the fact that new openings largely occur in more rural settings, and this has been previously demonstrated to have implications for drunk driving and traffic accidents (Green and Krehic, 2022). As no municipality experiences closures during this period, in practice, this means that the estimates come from the opening of the first wine monopoly in a municipality, and increases in the number of stores, respectively.

A standard concern is that there exist unobserved variables that influence both off-premises alcohol availability and traffic safety. Most obviously, more densely populated places are also more likely to have monopoly stores. At the same, areas with fewer / no stores may be in more remote areas with more difficult driving conditions. Together, this suggests that OLS estimates of (1) may lead to biased results, and the direction of this bias is unclear. Our main strategy is to include municipal fixed effects. This removes levels in both accidents and stores such that the estimate of interest in (1) is identified by municipal changes in both off-premises availability and accidents. In addition, we include year-fixed effects to account for secular changes over the whole of Norway, for instance, the general reductions in accident levels witnessed earlier. We further include month-of-year fixed effects to account for seasonal variations in traffic accidents which reflect the marked variations in driving conditions over the year in Norway. Naturally, these approaches do not mitigate all potential problems, most notably time-varying unobservables influence both accident levels and the number of stores at the municipality level. We adopt approaches to examine this later.

5 Results

Table 2 reports estimates of the relationship between the opening of the first monopoly store and accidents in a municipality. All standard errors are clustered at the municipal level. The first column reports OLS estimates of this relationship without any controls. This demonstrates a large positive correlation between store openings and traffic accidents, just less than 2 more accidents per month. Column 2 introduces municipal fixed effects leading to a dramatic reduction in this estimate. This likely reflects the fact that more populous municipalities are more likely to have at least one monopoly show. The resultant FE estimates suggest an increase of 0.16 accidents per month. This compares to an average monthly number of 0.73 for those municipalities that initially began without a store (Table 1), and hence this is a substantial increase in percentage terms (approx. 22% increase). The final column reports estimate after the introduction for overall population numbers, and numbers of young men and young women, respectively. Their introduction does not dramatically affect the coefficient of interest.

Table 2

The Influence of Opening an Off-Premises Wine Monopoly Store on Traffic Accidents (1999–2019)

	(1)	(2)	(3)
	OLS	FE	FE + Controls
Store [0,1]	1.873***	0.158***	0.176***
	(0.488)	(0.0248)	(0.0248)
log (population)			0.231*
			(0.130)
log (young men 18-34)			-0.457***
			(0.0920)
log (young women 18-34)			-0.610***
			(0.0996)
Constant	0.872***	1.348***	6.014***
	(0.0804)	(0.0329)	(0.583)
Observations	104,124	104,124	104,124
R^2	0.053	0.026	0.030
Number of Municipalities	415	415	415

Note: Robust standard errors are clustered at the municipality level in brackets. ***, **, and * indicate statistical significance at 1%, 5% and 10%, respectively.

Table 3 reports analogous estimates but where the key independent variable is instead the number of monopoly stores. Simple OLS estimation suggests a positive association between the number of stores and traffic accidents. Again, this most likely reflects the correlation between municipal population and store location. Introducing municipal fixed effects (Column 2) changes the estimate, and flips its sign, to a reduction of approximately three-quarters of an accident per month in a municipality. In addition, controlling for changes in municipal demographic characteristics does not substantively change this estimate. The third column's results are close to the second column. This suggests that for municipalities that already have wine monopoly stores, adding more wine monopoly stores will reduce the local traffic accident rate.

Table 3

The Influence of the Number of Wine Monopoly Stores on Traffic Accidents (1999–2019)

	(1)	(2)	(3)
	OLS	FE	FE + Controls
Stores #	2.200***	-0.779***	-0.763***
	(0.304)	(0.0120)	(0.0120)
log (population)			0.245*
			(0.128)
log (young men 18-34)			-0.479***
			(0.0903)
log (young women 18-34)			-0.351***
			(0.0978)
Constant	0.439***	1.730***	4.821***
	(0.107)	(0.0320)	(0.573)
Observations	104,122	104,122	104,122
R^2	0.583	0.064	0.066
Number of Municipalities	415	415	415

Note: Robust standard errors are clustered at the municipality level in brackets. ***, **, and * indicate statistical significance at 1%, 5% and 10%, respectively.

6 Heterogeneity and Robustness

Concerns about the effects of alcohol availability on traffic accidents are naturally concentrated in those incidents that lead to injury or death. In addition, as our data comes from reported accidents with likely underreporting, these accident types are more reliably reported. Both factors lead us to explore the effects of store openings on accidents of differing severity. We adopt the approach of using the injury of the greatest severity to classify a given accident.

Table 4 reports variants of our preferred specification but where the dependent variable is the number of accidents where the injury levels are light, serious, and fatal accidents, respectively. This reveals several points. First, the effect of opening the first monopoly store is concentrated entirely on accidents resulting in light injuries. Expanding the availability of alcohol in this manner increases these types of accidents by a little over 2 accident per year in each municipality. There are quite precisely zero effects for accidents leading to serious or fatal injuries. The pattern for the number of stores is again distinct. There are substantial reductions in light accidents because of increases in the number of stores in each municipality. These are sizeable, with just under one accident reduction per municipality per month with an extra store opening. In addition, there are small but statistically significant reductions in serious and fatal injuries.

Table 4

Accident Severity and Monopoly Stores, 1999–2019

	(1) Light	(2) Serious	(3) Fatal	(4) Light	(5) Serious	(6) Fatal
Store [0,1]	0.191***	0.0004	0.003			
	(0.022)	(0.005)	(0.003)			
Stores #				-0.842***	-0.009***	-0.010***
				(0.011)	(0.003)	(0.001)
log (population)	0.063	-0.023	-0.009	0.195*	-0.021	-0.008
	(0.117)	(0.028)	(0.013)	(0.114)	(0.028)	(0.013)
log (young men 18-34)	-0.441***	-0.040**	-0.007	-0.482***	-0.041**	-0.007
	(0.083)	(0.020)	(0.010)	(0.080)	(0.020)	(0.009)
log (young women 18-34)	-0.598***	-0.004	-0.005	-0.328***	-0.0009	-0.002
	(0.089)	(0.021)	(0.010)	(0.087)	(0.021)	(0.010)
Constant	7.063***	0.611***	0.197***	4.951***	0.589***	0.171***
	(0.523)	(0.125)	(0.059)	(0.509)	(0.125)	(0.060)
Observations	104,124	104,124	104,124	104,122	104,122	104,122
R^2	0.038	0.011	0.004	0.092	0.011	0.005
Number of Municipalities	415	415	415	415	415	415

Note: Robust standard errors are clustered at the municipality level in brackets. ***, **, and * indicate statistical significance at 1%, 5% and 10%, respectively.

Table 5

Accident Type and Monopoly Stores, 1999–2019

	(1) Single	(2) Multi	(3) Night	(4) Weekend	(5) Single	(6) Multi	(7) Night	(8) Weekend
Store [0,1]	0.027***	0.153***	0.0482***	0.042***				
	(0.0104)	(0.0196)	(0.00909)	(0.00886)				
Stores #					-0.120***	-0.642***	-0.192***	-0.179***
					(0.005)	(0.009)	(0.004)	(0.004)
Constant	1.472***	5.405***	1.946***	1.813***	1.152***	3.695***	1.434***	1.337***
	(0.246)	(0.462)	(0.214)	(0.209)	(0.246)	(0.453)	(0.213)	(0.208)
Observations	104,124	104,124	104,124	104,124	104,122	104,122	104,122	104,122
R^2	0.025	0.024	0.024	0.020	0.030	0.065	0.041	0.035
Number of Municipalities	415	415	415	415	415	415	415	415

Note: Robust standard errors are clustered at the municipality level in brackets, ***, **, and * indicate statistical significance at 1%, 5%, and 10%, respectively.

In Table 5 we further explore heterogeneity. We do this according to whether the accidents were single or multiple vehicles, occurred at night (18:00 to 6:00), or over the weekend. The former two outcomes are motivated again by a view that multiple-vehicle accidents may be both more severe and less likely to suffer from under-reporting (a particular concern is under-reporting of single-vehicle accidents if the driver is intoxicated). The focus on nights and weekends reflects a desire to focus more clearly on times of day and periods of the week when drinking, and drunk driving, are likely to be more concentrated. In terms of the number of vehicles involved, these results suggest that the effects of opening a first store are concentrated almost entirely in multiple-vehicle accidents. Similarly, although not as striking, the effects of increasing numbers of stores are also concentrated in multiple vehicle accidents. As suggested above, one implication of these results is that it is less likely that our results are generated by reporting variations with respect to monopoly store openings. The columns which report estimates for nights and weekends are less dramatic. Both types of changes in store openings have effects on accidents over these two, are not mutually exclusive. Yet, when compared to the main estimates in Tables 2 and 3, it suggests that comprise only approximately half of the changes in overall accidents.

As mentioned before, a concern is that the openings of monopoly stores may reflect time-varying changes in municipal accident rates and traffic safety. For instance, consider our estimates that suggest expanding the number of stores in a municipality reduces traffic accidents. This could reflect the fact that improving traffic safety is a factor determining the likelihood of an application for an additional monopoly store being granted. Alternatively, it could simply reflect the efforts of larger towns and cities in Norway to reduce traffic accidents in this period (and it is these cities that have experienced the largest expansion in store numbers). More generally, the key identifying assumption of these models is that, in the absence of changes in monopoly stores, accidents in municipalities where alcohol availability changed would have followed a similar evolution to municipalities without these changes. Exploring this motivates an event study analysis which provides an approach to considering anticipatory effects where there should be none. In addition, it allows for analysis of the time pattern of changes in traffic accidents following changes in stores.

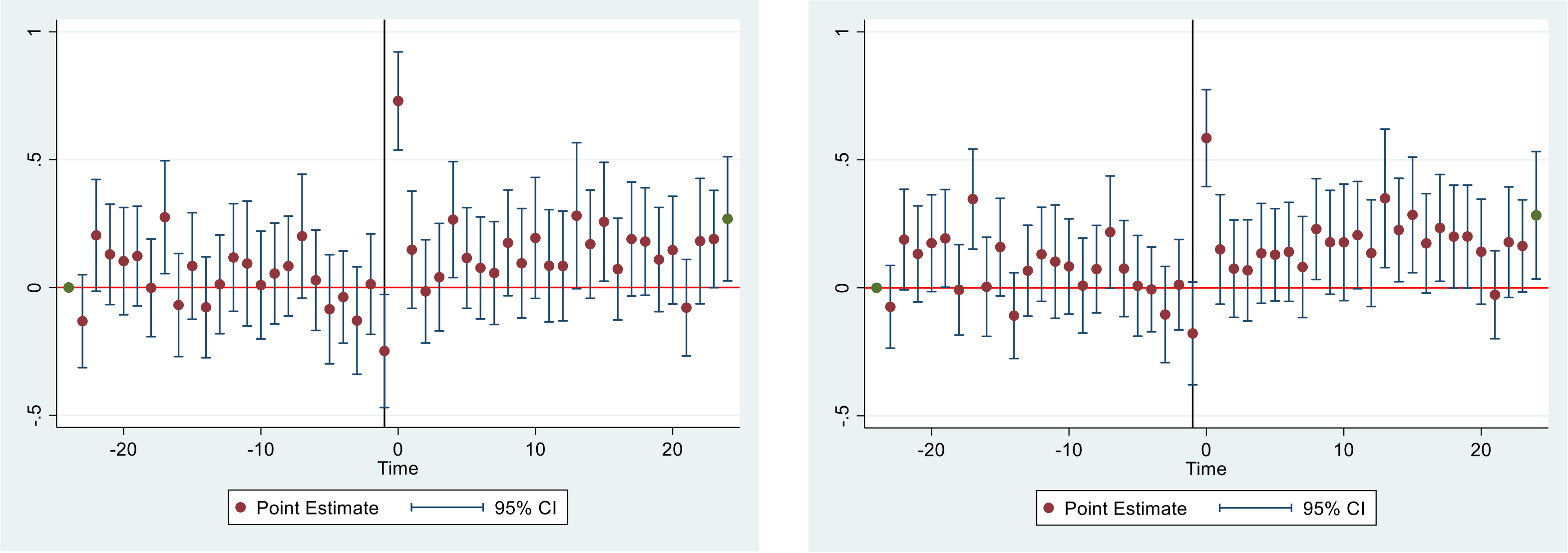
Here we focus on the extensive margin, i.e., movements from zero to one. This approach more readily lends itself to event study analogues of (1), especially as store openings are in an absorbing state.

Figure 3 presents event study estimates for all accidents, and those leading to light, serious and fatal injuries, respectively. These are monthly event study estimates of the impact of moving from 0 to 1 monopoly store on accidents. Note, however, that for serious and fatal injuries this pushes the data hard so some caution in interpreting these results. A few points are worth noting. In general, these results are supportive of the parallel trends assumption. With few exceptions, the pre-change confidence intervals overlap with zero. This suggests that openings of the first store do not reflect downward trends in traffic accidents in (soon to be) treated municipalities. However, there is some evidence of an anticipation effect right before opening. The month before the store opening has off-trend lower accidents, this is too short a time to be itself part of the store opening decision process. This is followed by the largest increase in accidents occurring in the month of opening. These patterns could reveal intertemporal smoothing of drinking behaviour, and related driving, over the periods before and after store opening.

These results do, additionally, lead to a concern that our prior estimates solely reflect these two periods. However, in unreported estimates, we re-estimated our main models excluding both the period before and the period of opening, and in general, our main results are unaffected. This fits with panels A and B, which demonstrate evidence that store opening is associated with a small, but higher, accident level following store opening.

Figure 3

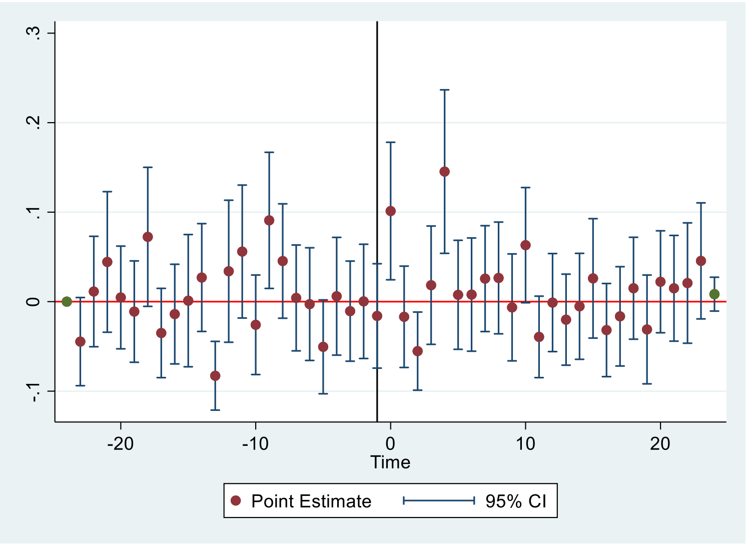
Effect of Monopoly Store Opening in Municipality on Traffic Accidents, Event Study Estimates



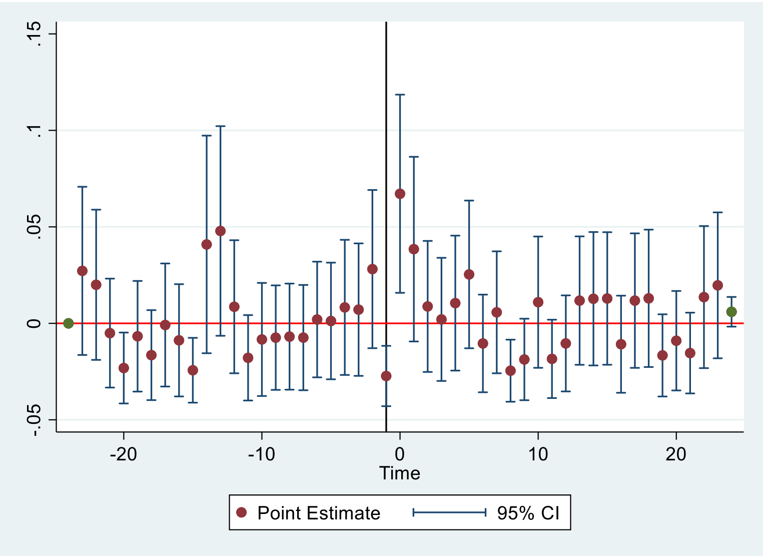
Panel A: All Accidents

Panel B: Light Injuries

Panel C: Serious Injuries



Panel D: Fatal Injuries



A more general point is the use of appropriate control groups, and this, in turn, relates to recent developments in the literature on differences in differences and two-way fixed effects models (Roth et al., 2023). For instance, one reading of the results from Table 2 and Table 3 is that municipalities, where the number of stores has increased from 1 to more, are unlikely to be a good control group for those who shift from zero to one store. This reflects the fact that many of these municipality's experience increases in the number of stores and this, in and of itself, influences traffic accidents. The opposite is true for estimates of increases in the number of stores. For the former case, we highlight that in the event study estimates above we adopt a standard approach and only include in the estimating sample those municipalities who had no monopoly stores in 1999.

For the extensive margin, as a first step, we use an alternative control group that consists of those municipalities that have at least one store. We do this for two reasons. The first is simply a concern that municipalities with no monopoly stores are typically smaller and may present a poor comparison group for larger municipalities that experience growth in store numbers over our period of analysis. The second is that currently we are, in our number of stores estimates, averaging the effects over new openings (zero to 1) which appear to increase accidents, and expansions of the number of stores which appear to reduce accidents.

These estimates are reported in Table 6 and fit with expectations. Focusing on columns 2 and 3, excluding those municipal-month observations without stores, and hence identifying based on changes from 1 and upwards, leads to an increase in the absolute size of the negative effects of wine monopoly stores. In the (3) column, they are in the order of a reduction in 1.5 accidents per month in the municipality.

An alternative approach to examine the same issue is to use the full sample but split the number of stores into groups. We replace a number of stores with dummy indicators for 1 to 2 stores, 3 to 4 stores, 5 to 6 stores, and 7 or more. This allows us to examine the non-linear effects of the store opening, but at the same time separate the extensive and intensive margins within one regression. These results are reported in the final column of Table 6. They fit with previous estimates. Recalling that no municipality experiences reductions in the number of stores within the period, then the first estimate reveals a positive effect of moving from zero to 1 or 2 stores in the municipality. This fits with, for instance, Table 2. Beyond this point, there are reductions in accidents as the number of stores increases. This suggests again that pooling all municipalities and including stores as a linear term underestimates the intensive margin effect.

Table 6

The Influence of the Number of Wine Monopoly Stores on Traffic Accidents (1999–2019) Excluding Observations with Zero Stores

	(1) OLS	(2) FE	(3) FE + Controls	(4) FE + Controls Non-Linear
Monopoly Stores	2.447***	-1.511***	-1.490***	
	(0.224)	(0.024)	(0.024)	
Stores 1-2				0.211***
				(0.024)
Stores 3-4				-0.893***
				(0.098)
Stores 5-6				-1.409***
				(0.180)
Stores 7+				-2.822***
				(0.188)
Constant	-0.978***	4.736***	2.764*	6.415
	(0.318)	(0.091)	(1.564)	(0.600)
Observations	38,829	38,829	38,829	104,124
R^2	0.639	0.151	0.154	0.290
Number of Municipalities	270	270	270	415

Note: Robust standard errors are clustered at the municipality level in brackets, ***, **, and * indicate statistical significance at 1%, 5%, and 10%, respectively.

7 Conclusion

How alcohol availability influences traffic safety is an important issue. We argue that while much is known about the influence of on-premises availability, less is known about off-premises availability and that this is an important element of alcohol consumption. We explore this using the expansion of the Norwegian wine monopoly over the past 20 years. This leads to substantial changes in off-premises availability.

In doing so, we focus on two margins of adjustment. The first is that group of municipalities that experienced the opening of their first store. This represents a one-time large increase in the local availability of medium to heavy-strength alcohol. We demonstrate that this leads to small increases in traffic accidents, concentrated amongst accidents causing light injuries, and to some extent in the period directly following the opening. We contrast this with expansions in the number of stores in municipalities with some pre-existing availability. These experience larger reductions in traffic accidents, with some suggestions of reductions in more serious injuries and fatalities.

Similar to the literature on bar opening hours, our results suggest that the effect of off-premises availability on traffic accidents is context and margin specific. To explore this further, future research should focus on the underlying mechanisms generating these different effects. A more general point is that in formulating regulations regarding off-premises alcohol availability, policymakers should be aware that this has implications for traffic safety.

References

Avdic, D., & von Hinke, S. (2021). Extending alcohol retailers' opening hours: Evidence from Sweden. *European Economic Review*, 138(July), 103830. https://doi.org/10.1016/j.euroecorev.2021.103830

Biderman, C., De Mello, J. M. P., & Schneider, A. (2010). Dry laws and homicides: Evidence from the Sao Paulo metropolitan area. *The Economic Journal*, 120(543), 157–182. doi: 10.1111/j.1468-0297.2009.02299.x

Folkehelseinstituttet. (2021). Hjerte - og karsykdommer i Norge, 1–35. Retrieved from https://www.fhi.no/nettpub/hin/ikke-smittsomme/Hjerte-kar/#om-hjerte-og-karsykdommer

Giacopassi, D., & Winn, R. (1995). Alcohol availability and alcohol-related crashes: Does distance make a difference? *American Journal of Drug and Alcohol Abuse*, 21(3), 407–416. https://doi.org/10.3109/00952999509002706

Gjerde, H., & Frost, J. (2023). Prevalence of alcohol and drugs among drivers killed in road traffic crashes in Norway during 2011–2020. *Traffic Injury Prevention*, 24(3), 256–261. https://doi.org/10.1080/15389588.2023.2174801

Gjerde, H., Normann, P. T., Christophersen, A. S., Samuelsen, S. O., & Mørland, J. (2011). Alcohol , psychoactive drugs, and fatal road traffic accidents in Norway: A case-control study. *Accident Analysis and Prevention*, 43(3), 1197–1203. https://doi.org/10.1016/j.aap.2010.12.034

Green, C. P., Heywood, J. S., & Navarro, M. (2014). Did liberalizing bar hours decrease traffic accidents? *Journal of Health Economics*, 35, 189–198. doi 10.1016/J.JHEALECO.2014 .03.007

Green, C., & Krehic, L. (2022). An extra hour wasted? Bar closing hours and traffic accidents in No -rway. *Health Economics (United Kingdom),* 31(8), 1752–1769. https://doi.org/10.1002/hec.4550

Gruenewald, P. J., Freisthler, B., Remer, L., Lascala, E. A., Treno, A. J., & Ponicki, W. R. (2010). E -cological associations of alcohol outlets with underage and young adult injuries. *Alcoholism*

: Clinical and Experimental Research, 34(3), 519–527. https://doi.org/10.1111/j.1530-0277.2009.01117.x

Han, D., Shipp, E. M., & Gorman, D. M. (2015). Evaluating the effects of a large increase in off-sale alcohol outlets on motor vehicle crashes a time-series analysis. *International Journal of Injury Control and Safety Promotion*, 22(4), 320–327. https://doi.org/10.1080/17457300.2014.908223

Heaton P. Sunday Liquor Laws and Crime. *Journal of Public Economics*. 2012 Feb;96(1-2):42-52. doi: 10.1016/j.jpubeco.2011.08.002. PMID: 22125348; PMCID: PMC3224020.

Kelleher, K.J., Pope, S.K., Kirby, R.S., & Rickert, V.I. (1996). Alcohol availability and motor vehicle fatalities. *Journal of Adolescent Health*, 19, 325–330.

Lipton, R., Banerjee, A., Ponicki, W. R., Gruenewald, P. J., & Morrison, C. (2021). Impacts of confounding roadway characteristics on estimates of associations between alcohol outlet densities and alcohol-related motor vehicle crashes. *Drug and Alcohol Review*, 40(2), 239–246. https://doi.org/10.1111/dar.13156

Marcus, J., & Siedler, T. (2015). Reducing binge drinking? The effect of a ban on late-night off-premise alcohol sales on alcohol-related hospital stays in Germany. *Journal of Public Economics,* 123, 55–77. https://doi.org/10.1016/j.jpubeco.2014.12.010

Middleton, J. C., Hahn, R. A., Kuzara, J. L., Elder, R., Brewer, R., Chattopadhyay, S., Lawrence, B. (2010). Effectiveness of policies maintaining or restricting days of alcohol sales on excessive alcohol consumption and related harms. *American Journal of Preventive Medicine*, 39(6), 575–589. https://doi.org/10.1016/j.amepre.2010.09.015

Morrison, C., Ponicki, W. R., Gruenewald, P. J., Wiebe, D. J., & Smith, K. (2016). Spatial relationships between alcohol-related road crashes and retail alcohol availability. *Drug and Alcohol Dependence,* 162, 241–244. https://doi.org/10.1016/j.drugalcdep.2016.02.033

Norström, T., & Rossow, I. (2013). Population drinking and drunk driving in Norway and Sweden: An analysis of historical data 1957-89. *Addiction*, 108(6), 1051–1058. https://doi.org/10.1111/add.12126

Pasnin, L. T., & Gjerde, H. (2021). Alcohol and drug use among road users involved in fatal crashes in Norway. *Traffic Injury Prevention*, 22(4), 267–271. https://doi.org/10.1080/15389588.2021.1887854

Ponicki, W. R., Gruenewald, P. J., & Remer, L. G. (2013). Spatial panel analyses of alcohol outlets and motor vehicle crashes in California: 1999-2008. *Accident Analysis and Prevention*, 55, 135–143. https://doi.org/10.1016/j.aap.2013.03.001

Smith, D. I. (1988). Effect on traffic accidents of introducing Sunday alcohol sales in Brisbane, Australia. *Substance Use and Misuse*, 23(10), 1091–1099. https://doi.org/10.3109/10826088809056188

Tang, M. C. (2013). The multitude of alehouses: The effects of alcohol outlet density on highway safety. *B.E. Journal of Economic Analysis and Policy*, 13(2), 1023–1050. https://doi.org/10.1515/bejeap-2012-0051

Tonkin, R. S., & Kelleher, K. J. (1997). Alcohol availability and motor vehicle fatalities (multiple letters). *Journal of Adolescent Health*, 21(2), 74–75. https://doi.org/10.1016/S1054-139X(97)00131-6

Treno, A. J., Johnson, F. W., Remer, L. G., & Gruenewald, P. J. (2007). The impact of outlet densities on alcohol-related crashes: A spatial panel approach. *Accident Analysis and Prevention*, 39(5), 894–901. https://doi.org/10.1016/j.aap.2006.12.011

Trolldal, B. (2005). An investigation of the effect of privatization of retail sales of alcohol on consumption and traffic accidents in Alberta, Canada. *Addiction*, 100(5), 662–671. https://doi.org/10.1111/j.1360-0443.2005.01049.x

Valen, A., Bogstrand, S. T., Vindenes, V., Frost, J., Larsson, M., Holtan, A., & Gjerde, H. (2019). Driver-related risk factors of fatal road traffic crashes associated with alcohol or drug impairment. *Accident Analysis and Prevention*, 131(0424), 191–199. https://doi.org/10.1016/j.aap.2019.06.014

Vingilis, E., McLeod, A., Seeley, J., Mann, R., Beirness, D., & Compton, C. (2005). Road safety impact of extended drinking hours in Ontario. *Accident Analysis & Prevention*, 37(3), 549–556. doi 10.1016/J.AAP.2004.05.006

Wang, S., Chen, Y., Huang, J., Liu, Z., Li, J., & Ma, J. (2020). Spatial relationships between alcohol outlet densities and drunk driving crash: An empirical study of Tianjin in China. *Journal of Safety Research,* 74, 17–25. https://doi.org/10.1016/j.jsr.2020.04.011

ISBN 978-82-326-7200-4 (printed ver.)
ISBN 978-82-326-7199-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology