A Statistical and Machine Learning Approach to Air Pollution Forecasts

Simon Carlén

Degree project, 15 credits
Computer and Systems Sciences
Degree project at the master level
Spring term 2022
Supervisor: Sindri Magnússon
Co-supervisor: Ali Beikmohammadi

Stockholm University

Abstract

Synopsis

Background

Problem

Research Question

Method

Result

Discussion

Acknowledgements

Contents

List of Figures

List of Tables

List of Abbreviations

1. Introduction

1.1 Background

Outdoor air pollution is a major global environmental issue, linked to several serious health conditions, and causing millions of premature deaths every year [?]. Some principal air pollutants damaging to health include gaseous substances such as nitrogen oxides (NO_x), ground-level ozone (O_3), sulphur dioxide (SO_2), and carbon monoxide (CO), but also atmospheric aerosol particles such as PM_{10} and $PM_{2.5}$ [?]. In Stockholm, traffic is a major source of local air pollution, and though air quality is generally good, some streets experience short episodes with severe pollution levels, especially during winter and spring [?].

To protect public health, urban air is normally monitored. In addition to monitoring, forecasts of air quality (both hourly and daily) can be critical to regulatory authorities, and in general, there are two approaches to this; with mechanistic models or statistical and/or machine learning models. [?, ?, ?]. With mechanistic models, the processes governing the evolution of air pollution is modeled mathematically, whereas statistical and machine learning models are more data-driven [?].

From a statistical perspective, predicting air pollution is a time series regression problem, and there are many different regression techniques for forecasting and time series analysis [?]. These techniques can vary in complexity, from more simple linear models to deep neural networks capable of finding complex non-linear relationships in the data [?, ?]. Nonetheless, one of the main challenges with air pollution is that there are dependencies over both space and time (i.e., the data is spatio-temporal), and simpler models may not capture these dependencies [?]. Recent advances in machine learning however have shown promising results when it comes to air quality forecasts, especially deep neural networks [?, ?].

1.2 Research problem

Forecasts, be it for weather, stock returns, or future pandemics, are always associated with uncertainty and errors. Erroneous predictions made by existing air pollution forecasting systems, both mechanistic and statistical and/or machine learning-based, can be attributed to many causes. In the case of mechanistic models, there can be insufficient information in terms of the factors needed for simulation and modeling [?]. For statistical and/or machine learning methods, too simplistic models, lack of data, irrelevant input features, overfitting, etc., can limit prediction accuracy [?]. Nevertheless, atmospheric pollution is a very complex phenomenon depending on a multitude of factors across both space and time. Hence, the research problem addressed in this work is: *To capture and model the complex dynamics of air pollution with modern machine*

learning methods, with an emphasis on deep learning.

1.3 Research question

From a forecasting perspective, of special interest are episodes when pollution levels peak. Generally, this is also when existing forecasting systems tend to give the largest prediction errors [**?**]. Therefore, the research question this thesis tries to answer is: *How can machine learning, in particular deep learning, be used to forecast air pollution levels and pollution peaks?*

1.4 Delimitations

2. Extended Background

2.1 Ambient air pollution

Ambient air pollution is one of the greatest environmental and health concerns of the modern world. Worldwide, poor air quality causes millions of premature deaths every year and is linked to several adverse health effects such as respiratory problems, cardiovascular disease, and cancer [**?**]. In addition to health risks, the global economic impacts are substantial due to lost labor productivity, increased health care costs, reduced crop yields, etc. [**?**]. Outdoor air pollution has become a ubiquitous problem, affecting both cities and rural areas, and it is estimated that about 90% of the world's population are living in regions where air pollution levels exceed guidelines set by the World Health Organization [**?**].

2.1.1 Principal air pollutants

In densely populated urban areas, air pollution levels can periodically be severe, and with an accelerating urbanization, it has become imperative for regulatory authorities to closely monitor city air and try to mitigate the harmful effects of pollution. Commonly monitored substances include sulphur dioxide (SO_2), nitrogen oxides (NO_x, i.e., NO and NO_2), carbon monoxide (CO), ground-level ozone (O_3), volatile organic compounds (VOCs), and particulate matter (PM) [**?**]. Vehicular traffic is a major source of the gaseous pollutants NO_x, SO_2, CO, and VOCs, but certain industrial processes also contribute to emissions [**?**]. Ground-level O_3 (also a gas) is a so-called secondary pollutant that forms when NO_x and VOCs react on sunny days with little wind [**?**].

PM, the group of pollutants being the focus of this work, are atmospheric aerosol particles (i.e., particles suspended in the air). They have diverse origins, both natural and anthropogenic, and a complex chemical composition consisting of both solid and liquid species [**?**]. Some important sources of PM are forest fires, volcanic eruptions, sand/dust storms, sea spray, vehicular traffic, certain industrial processes, construction sites, and domestic combustion [**?**, **?**]. When entering the atmosphere directly by these routes, one denotes the PM as primary. However, PM can also be formed by the oxidation of gases such as SO_2, NO_x, and VOCs (followed by a complex chemical reaction process), in which case the PM is said to be secondary [**?**]. PM is also categorized by particle size (or more specifically, the aerodynamic diameter), and particles measuring smaller than 2.5 μm and 10 μm are denoted as $PM_{2.5}$ and PM_{10}, respectively [**?**].

Both PM_{10} and $PM_{2.5}$ can travel long distances from point sources (though $PM_{2.5}$ has a longer residence time in the atmosphere than PM_{10}), and local pollution can be affected by regional background levels [**?**, **?**]. PM levels are also dependent on weather conditions [**?**]. For example, temperature and solar radiation are related to the formation of secondary PM, and PM emission from roads, tires, brake wear, etc., can

be affected by precipitation and humidity [?, ?]. Both PM_{10} and $PM_{2.5}$ are hazardous and cause a wide range of health problems, though $PM_{2.5}$ can more easily penetrate the lungs [?]. In the European Union, annual mean limits are set to 40 $\mu g/m^3$ for PM_{10} and 20 $\mu g/m^3$ for $PM_{2.5}$ [?].

2.1.2 Ambient air pollution in Stockholm

In the city of Stockholm, environmental air quality standards are usually met, though some streets experience occasional episodes with severe pollution levels (e.g. Hornsgatan is one such street) [?]. Since Stockholm has centralized district heating and few industries, the major source of local CO, NO_x, and PM pollution is vehicular traffic [?, ?]. Mechanical wear by studded tires on asphalt and the wearing of brakes and tiers in motor vehicles contribute substantially to local levels of both PM_{10} and $PM_{2.5}$. For $PM_{2.5}$ however, contribution from sources outside of Stockholm is also significant [?]. Emission of SO_2 can come from the energy sector and waterborne transport, though local levels are also affected by outside sources. For O_3, long-range transport from mainland Europe is the single-most important factor contributing to locally measured levels [?].

The air in Stockholm County is monitored by Stockholms Luft- och Bulleranalys (SLB-analys), a unit in the Environment and Health Administration (EHA) of the city of Stockholm. SLB-analys are responsible for a number of monitoring stations measuring several air pollutants and some meteorological parameters in the Stockholm region, as well as a few stations outside of Stockholm [?]. In addition to monitoring the air, SLB-analys also model and forecast air pollution levels for the Stockholm metropolitan area, and their forecasts are available through a smartphone application, called "Luft i Stockholm" [?].

2.2 Forecasting air pollution

Having the possibility to forecast air pollution levels hours or days ahead can be extremely valuable to regulatory authorities in order to protect public health, and vulnerable groups in particular. In general, there are two broad categories of models for such forecasts; mechanistic models, and statistical and/or machine learning models [?]. This work is concerned with the latter type, and in the sections below a review follows. The mathematical and statistical theory behind many of the models is quite extensive [?, ?, ?, ?], but relevant theory will be covered briefly.

2.2.1 Forecasting as a regression problem

While mechanistic models are based on mathematical modelling of atmospheric processes along with other factors governing the distribution of air pollution (such as emission source characteristics, physico-chemical properties of pollutants, terrain and building design, etc.) statistical and/or machine learning models are entirely data-driven, being derived directly from measurements on the variables of interest [?]. From a statistical (or machine learning) perspective, forecasting air pollution can be viewed as a regression problem, in which a function f, mapping input data to a numerical output, is being approximated (or learned) from a training set of labeled input-output examples [?]. Learning the function f amounts to finding a set of parameters (or

weights) for the model, which in the case of a simpler regression technique can be only a handful, but possibly millions if a deep neural network is used [**?**]. Generally in regression, the weights are learned by minimizing a cost function

$$J(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{2.2.1}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated model parameters ($\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_n$), $\hat{y}_i$ is a prediction and y_i is a training data value [**?**]. In **??** the squared error loss is used as loss function, and the cost is simply the loss averaged over the training data.[1] Depending on the model, minimizing $J(\hat{\boldsymbol{\beta}})$ is approached differently, as explained further in the sections below.

2.2.2 Multiple linear regression models

From the wealth of available regression techniques, multiple linear regression (MLR) has been extensively used to forecast and model air pollution [**?**]. Generally, if none of the basic model assumptions are violated (i.e., linearity, independence, normality, and constant variance), MLR is often a straightforward method, especially for data with no temporal dependencies (so-called cross-sectional data). However, for time series data, the assumption of independent errors is often not approptiate [**?**].

If fitting a MLR model to time series data, successive errors will typically be correlated (often referred to as autocorrelation), and this will cause several problems with the model if the correlation is not accounted for [**?**]. To this end, adjustments to the MLR model can be made, some of which require other parameter estimation techniques than the usual least squares method (see below). However, a simple and commonly used procedure to get rid of the autocorrelation is to include one or more lagged values of the response variable as predictors. For example, if the value of the response variable at lag one (y_{t-1}) is included, the MLR model will have the form

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{2,t} + ... + \beta_k x_{k,t} + \varepsilon_t, \quad t = 1, 2, ..., T \tag{2.2.2}$$

where ε_t is the the error term, and t denotes time steps [**?**]. The model in **??** can be fit with the method of least squares, which in linear regression is the standard way of finding parameters so that $J(\hat{\boldsymbol{\beta}})$ is minimized [**?**]. This is done by solving the so-called normal equations, and the least squares estimates of the model parameters are then given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{2.2.3}$$

Careful variable selection in regression is crucial as it can influence the performance of a model, and one is often concerned with finding an optimal "subset" of predictors, where multicollinearity should also not be an issue [**?**]. To this end, variable selection techniques based on optimizing a criterion like the Akaike or Bayes information criterion are common, and typically multicollinearity is also tested for [**?**]. However, if one is reluctant to exclude variables, but multicollinearity still might be an issue, regularized versions of MLR can be used [**?**]. Two common techniques are L_1 and L_2 regularization,

[1] What is meant by cost and loss functions can vary slightly in the literature, but in this work, the terminology of Lindholm et al. [**?**] is adopted.

in which an extra so-called "penalty" term is added to the cost function to shrink (and essentially stabilize) the model parameter estimates [**?**]. In L_2 regularization (also called ridge regression), the parameters will be pushed towards small values, whereas in L_1 regularization (or lasso regression), some parameters will be driven to zero. Generally, L_1 and L_2 regularization can be used to prevent overfitting, and also for L_1 where some input variables can be eliminated, as a variable selection technique [**?**].

The extensive use of MLR for air pollution forecasts is many times motivated by its simplicity and straightforward implementation [**?**]. Another advantage is interpretability; for example, inference can be made on all input variables, allowing one to investigate their individual importance [**?**]. However, the assumption of linearity might not always hold, and large prediction errors have been observed at times of pollution peaks [**?**]. Moreover, with data from several (but nearby) monitoring stations, multicollinearity among the input variables can be an issue, which is why ridge or lasso regression are popular alternatives to the non-regularized MLR model [**?**].

2.2.3 Extensions of the linear model

More versatile and flexible regression models tend to give better forecasting results than linear models [**?**]. Some examples include regression trees, generalized additive models, and support vector machines (SVM) [**?**, **?**]. These models can handle more complex non-linear input-output relationships, and especially SVM has been successfully applied for PM_{10} prediction, sometimes with better results than artificial neural networks [**?**].

Artificial neural networks (ANNs), in particular the multilayer perceptron (MLP), have also been extensively used as a forecasting technique [**?**]. ANNs are flexible models able to handle non-linear input-output relationships, however, over-fitting can be an issue, especially with high-dimensional input and if training data is limited [**?**, **?**].

The MLP is a so-called feedforward neural network, in which a set of input data is taken and passed through several "hidden" layers made up of neurons (also called units), before an output is produced [**?**]. Deep neural networks can have many such layers (hence the term "deep" [**?**]), and each layer can have hundreds of units. Every layer produces a slightly more abstract representation of its input by non-linear transformations, and with several such transformations, complex relationships in the data can be learned [**?**].

Many other deep learning architectures than the MLP exist, such as convolutional neural networks (CNNs), or recurrent neural networks (RNNs). CNNs are commonly used for image recognition while RNNs (and variants thereof) normally are applied to sequential data.

2.3 Summary and motivation for this work

3. Methodology

The major steps of the implemented workflow were as follows;

Detailed descriptions of each step in the process are given in subsequent sections

3.1 Data retrieval and preprocessing

3.1.1 Data sources

Air pollution data was retrieved from the Swedish Meteorological and Hydrological Institute's (SMHI) centralized database for air quality measurements [**?**]. This data is part of the national and regional environmental monitoring of Sweden, a program coordinated and funded by the Swedish Environmental Protection Agency (Swedish EPA) and the Swedish Agency for Marine and Water Management. There are in total ten different program areas, of which air is one, and all data are licensed under CC0 and therefore freely accessible to the public [**?**]. For the national air monitoring (under Swedish EPA's responsibility), SMHI acts as a national data host and stores (quality checked) historical data reported yearly from municipalities in Sweden [**?**].

In Stockholm County, there are 19 stationary sites for air pollution monitoring [**?**], and initially, data from each site was considered. However, many stations have irregular data series, and not all stations measure the same set of pollutants. Due to this, data from three sites with hourly measurements of PM_{10} and $PM_{2.5}$ (in $\mu g/m^3$) for the time period 2016-01-01 to 2022-01-01 was chosen, giving a total of 52,609 data points. For the station at which PM predictions subsequently were to be made (Torkel Knutssonsgatan), hourly data of NO_2 was also included. As described in section **??**, SLB-analys also monitor several weather parameters, and hourly measurements of temperature (in °C), precipitation (mm), atmospheric pressure (hPa), relative humidity (as %), solar radiation (W/m^2), and wind speed (m/s) were also included from the station at Torkel Knutssonsgatan. The meteorological data were downloaded from SLB-analys' webpage [**?**].

In general, air pollution monitoring can be classified by the surrounding area (rural, rural-regional, rural-remote, suburban, and urban), and by the predominant emission sources (background, industrial, or traffic) [**?**]. The chosen stations included data from both traffic and background monitoring, in urban as well as rural-regional areas. More information about the stations are given in Table **??** in appendix **??**.

3.1.2 Data preprocessing

Time series plots for PM_{10} and $PM_{2.5}$ at Torkel Knutssonsgatan are shown in **??**. (Similar plots for all stations are given in Fig. **??** in appendix **??**.) Some stations had short episodes with missing data, and linear interpolation was used to fill in the missing

values. However, for the $PM_{2.5}$ data at Torkel Knutssonsgatan (plot (b) in **??**), due to the rather big gap at the beginning of 2019, a train-test split was done to entirely avoid this period (see below). Missing weather data was also linearly interpolated, except for the variables atmospheric pressure and wind speed for which mean imputation was deemed more appropriate. Moreover, before use in any of the models, all data were min-max normalized (i.e., scaled to the interval $[0, 1]$).

It should be noted that some PM values were negative. However, negative values are expected since automated measuring instruments for PM (due to noise) may produce values between zero and the negative detection limit, especially when there are rapid changes in humidity (SMHI, personal communication, April 11, 2022). These values are therefore not to be considered any more "incorrect" than positive values, though it may at first seem contradictory to include them in an analysis.

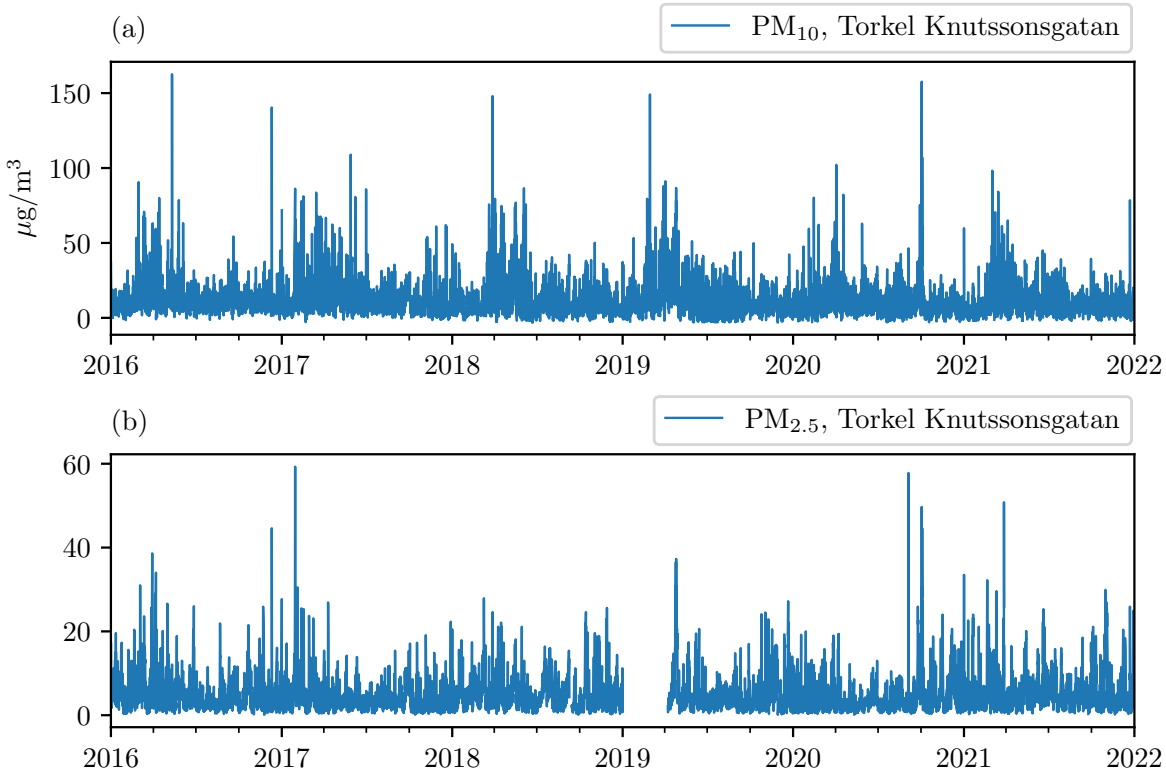

Figure 3.1: Time series plots for (a) PM_{10} and (b) $PM_{2.5}$ at Torkel Knutssonsgatan.

Feature engineering In **??**, yearly periodicity in the data can be seen, especially for PM_{10} where levels tend to be higher during spring. Daily and weekly periodicity is also expected since traffic intensities vary throughout the day and week. To account for this, timestamps were converted to temporal variables as sine and cosine signals for day, week, and year. For example, the sine and cosine signals for day were calculated in the following way

$$\text{Sine day} = \sin\left(\text{timestamp} \cdot \frac{2\pi}{86,400}\right)$$

$$\text{Cosine day} = \cos\left(\text{timestamp} \cdot \frac{2\pi}{86,400}\right)$$

where timestamp is in seconds (and with 86,400 seconds in 24 hours, dividing by this term is necessary). The calculations were done similarly for week and year, except for the term in the denominator which instead was set to seconds per week and seconds per year, respectively. The temporal variables for day in a 24 hour time window are shown in **??**.

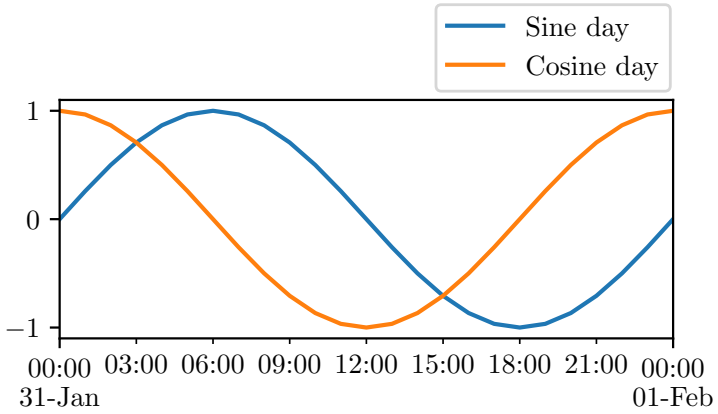

Figure 3.2: Temporal variables for day as sine and cosine signals.

Sliding windows from the data were also created. The sliding window approach is used for time-series forecasting where windows (or sequences of certain lengths, also called frames) are extracted from the input data [**?**, **?**]. In each window, there are two "sub-windows"; the input window and the target window, and the target window is offset by some amount of time from the input window. For example, as shown in Figure **??**, the total window length is nine time steps, and the first eight time steps is the input window used to predict the target window (in this case having a length of one) one time step in the future. After extracting a data sequence, the window slides to the right one (or more) steps and extracts the next sequence. This is continued until time step n at which point all the data have been processed. In this work, input windows of different lengths were tested to make short-term predictions for a target window with a length of one (more details are given in section **??** below).

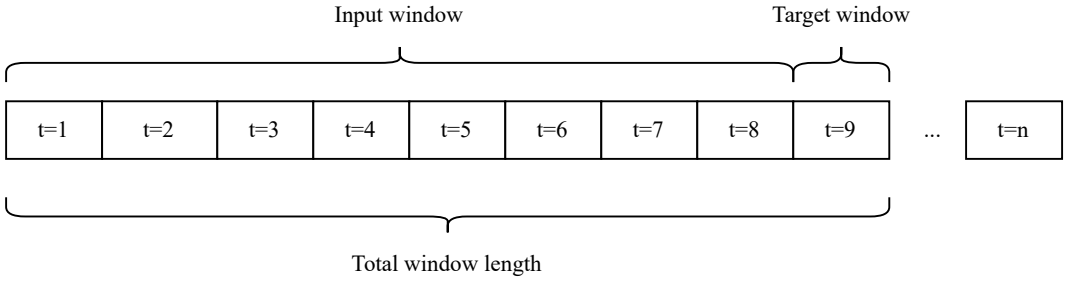

Figure 3.3: Sliding window approach for time-series data.

Train-test split Lastly, the data was split into training, validation, and test sets, where the validation set was used for hyperparameter optimization. The test set was taken as the most recent year of data (from 2019-09-16 to 2020-09-16), the validation

set was taken as the year prior to the test data (2018-09-16 to 2019-09-16), and the remaining data was used for training (2016-01-01 to 2018-09-16). This split is motivated by the fact that the data is in the form of time-series, where each observation has a specific time-stamp and where successive observations are (in this case) positively autocorrelated.

3.2 Hyperparameter tuning and model fitting

3.2.1 Multiple linear regression models

With data from three stations in the Stockholm area, some collinearity was expected, and regularized MLR models were initially tried. However,

3.2.2 Deep learning models

4. Results

4.1 Multiple linear regression models

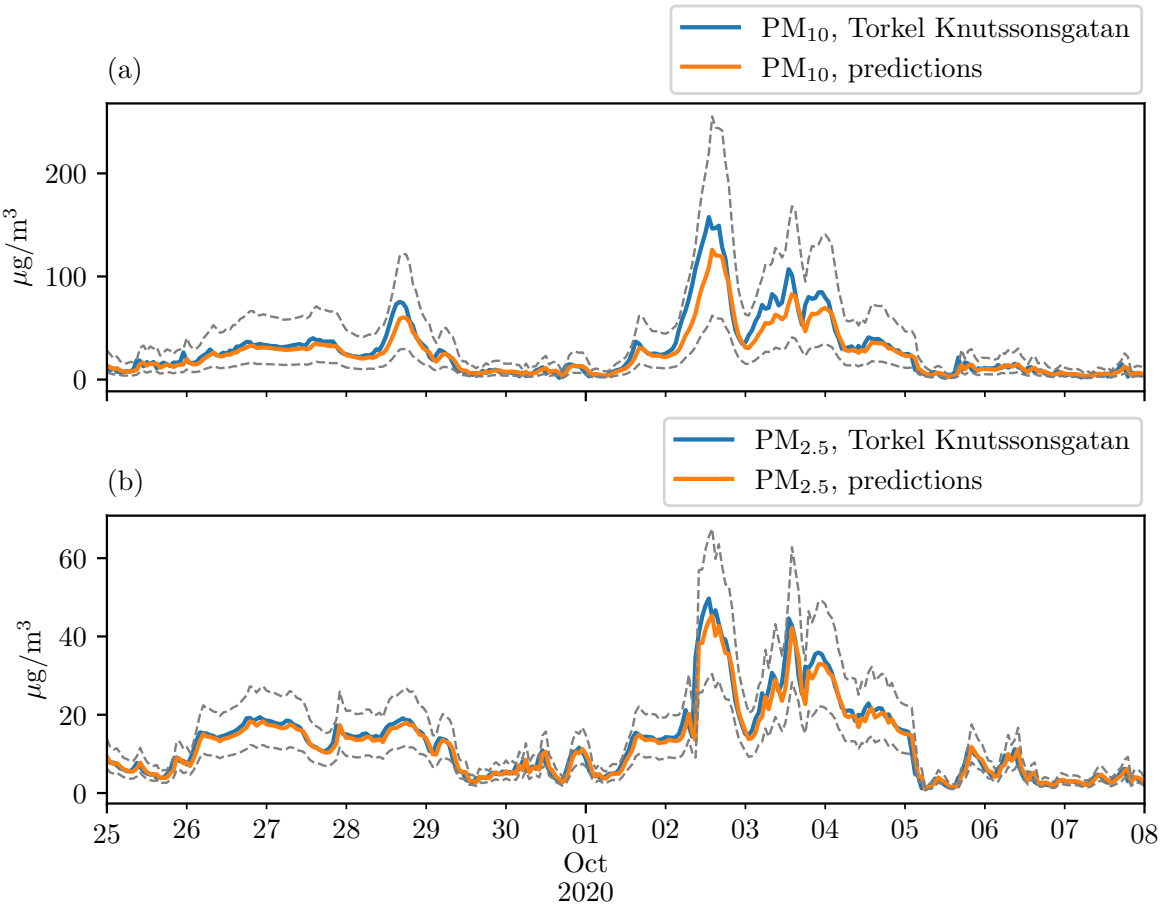

Figure 4.1: Predictions for the MLR model, and actual values, for (a) PM_{10} and (b) $PM_{2.5}$.

5. Discussion and Conclusions

Appendices

A Monitoring stations

Information about the monitoring stations from which data was used is summarized in **??**. In **??**, time series plots of PM_{10} and $PM_{2.5}$ at each station are shown.

Table A.1: Monitoring stations.

Station	Station code	Longitude	Latitude	Type of monitoring	Parameters
Norrtälje, Norr Malma	18643	18.631313	59.832382	Rural-Regional Background	PM_{10}, $PM_{2.5}$
Stockholm, Hornsgatan 108	8780	18.04866	59.317223	Urban Traffic	PM_{10}, $PM_{2.5}$
Stockholm, Torkel Knutssonsgatan	8781	18.057808	59.316006	Urban background	PM_{10}, $PM_{2.5}$, NO_2, meteorological parameters

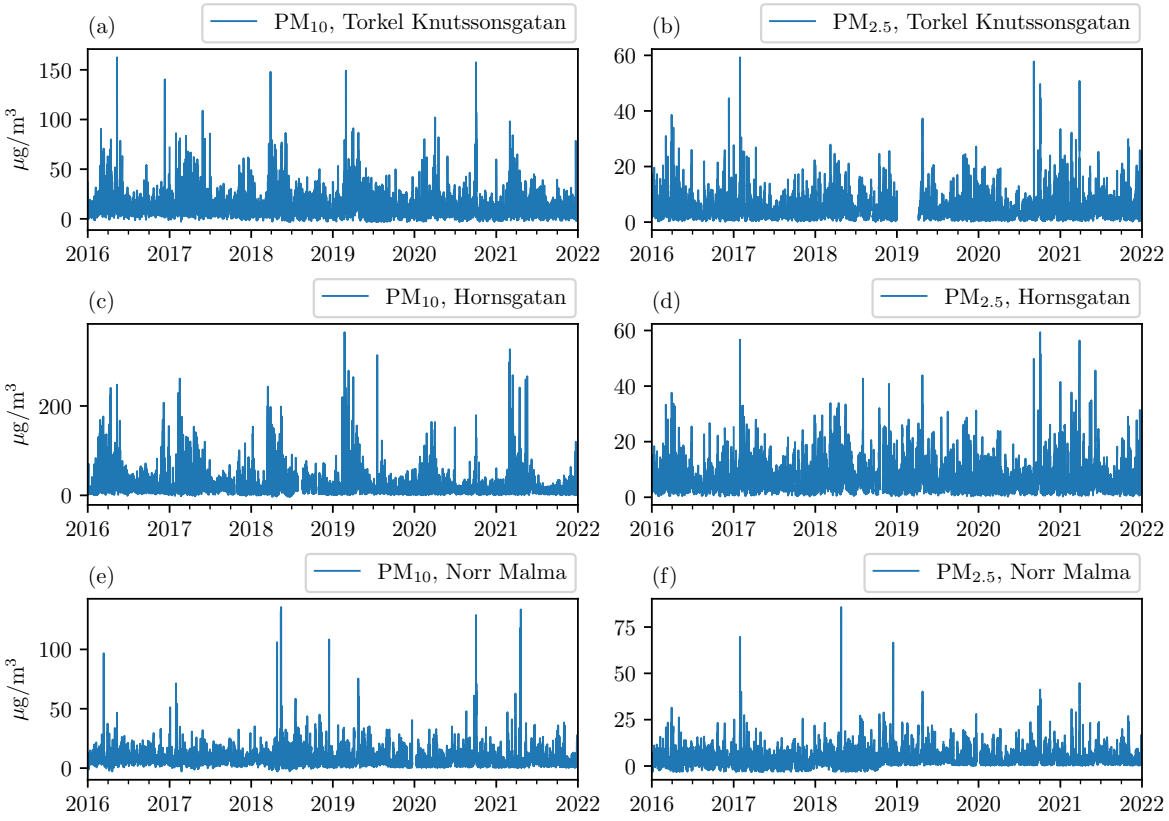

Figure A.1: Time series plots for PM_{10} and $PM_{2.5}$ at all stations.

B Model diagnostics for the multiple linear regression models

Diagnostics plots are shown in **??** and **??** below and models statistics are shown in tables...

Dep. Variable:	PM_{10}, Torkel Knutssonsgatan	**R-squared:**	0.755
Model:	OLS	**Adj. R-squared:**	0.755
Method:	Least Squares	**F-statistic:**	2.026e+04
Date:	Thu, 11 Aug 2022	**Prob (F-statistic):**	0.00
Time:	23:46:52	**Log-Likelihood:**	-10505.
No. Observations:	26328	**AIC:**	2.102e+04
Df Residuals:	26323	**BIC:**	2.106e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
intercept	0.0656	0.009	6.974	0.000	0.047	0.084
PM$_10$, Torkel Knutssonsgatan lag1	0.6335	0.007	96.861	0.000	0.621	0.646
PM$_10$, Torkel Knutssonsgatan lag2	0.1072	0.006	17.494	0.000	0.095	0.119
PM$_10$, Hornsgatan lag1	0.1234	0.004	30.778	0.000	0.116	0.131
PM$_10$, Norr Malma lag1	0.0808	0.004	19.644	0.000	0.073	0.089

Omnibus:	7362.227	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	122434.478
Skew:	-0.906	**Prob(JB):**	0.00
Kurtosis:	13.408	**Cond. No.**	21.8

Table B.1: OLS Regression Results for PM_{10}

Dep. Variable:	$PM_{2.5}$, Torkel Knutssonsgatan	**R-squared:**	0.927
Model:	OLS	**Adj. R-squared:**	0.927
Method:	Least Squares	**F-statistic:**	8.370e+04
Date:	Thu, 11 Aug 2022	**Prob (F-statistic):**	0.00
Time:	23:20:09	**Log-Likelihood:**	4594.4
No. Observations:	26328	**AIC:**	-9179.
Df Residuals:	26323	**BIC:**	-9138.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
intercept	0.0386	0.003	11.092	0.000	0.032	0.045
PM$_2.5$, Torkel Knutssonsgatan lag1	1.1199	0.007	167.681	0.000	1.107	1.133
PM$_2.5$, Torkel Knutssonsgatan lag2	-0.2116	0.006	-35.111	0.000	-0.223	-0.200
PM$_2.5$, Hornsgatan, lag1	0.0340	0.004	8.867	0.000	0.026	0.041
PM$_2.5$, Norr Malma, lag1	0.0218	0.001	15.576	0.000	0.019	0.025

Omnibus:	5704.682	**Durbin-Watson:**	1.987
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	161104.434
Skew:	-0.393	**Prob(JB):**	0.00
Kurtosis:	15.093	**Cond. No.**	21.3

Table B.2: OLS Regression Results for $PM_{2.5}$