A 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 iii

List of Tables iv

List of Abbreviations v

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Research problem . . . 1
- 1.3 Research question . . . 2
- 1.4 Delimitations . . . 2

2 Extended Background 3
- 2.1 Ambient air pollution . . . 3
  - 2.1.1 Principal air pollutants . . . 3
  - 2.1.2 Ambient air pollution in Stockholm . . . 3
- 2.2 Forecasting air pollution . . . 4
  - 2.2.1 Forecasting as a regression problem . . . 4
  - 2.2.2 Linear regression models . . . 4
  - 2.2.3 Extensions of the linear model . . . 6
- 2.3 Summary and motivation for this work . . . 6

3 Methodology 7
- 3.1 Data retrieval and preprocessing . . . 7
  - 3.1.1 Data sources . . . 7
  - 3.1.2 Data preprocessing . . . 8
- 3.2 Model fitting and hyperparameter tuning . . . 10
  - 3.2.1 Multiple linear regression models . . . 10
  - 3.2.2 Deep learning models . . . 10

4 Results 11
- 4.1 Multiple linear regression models . . . 11

5 Discussion and Conclusions 12

6 Bibliography 13

Appendices **15**
A Monitoring stations 15
B Model diagnostics and summary statistics for the multiple linear regression models 16

List of Figures

3.1 NO_2 data for Torkel Knutssonsgatan. 8
3.2 Temporal variables for day as sine and cosine waves. 9
3.3 Rolling window approach for time-series data. 9

4.1 Predictions for the MLR model, and actual values, for (a) PM_{10} and (b) $PM_{2.5}$. 11

B.1 Residual plots for the OLS regression model. 16

List of Tables

A.1 Monitoring stations. 15
B.1 OLS Regression Results. 17
B.2 Robust linear Model Regression Results. 18

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 [1]. 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 $PM_{2.5}$ [2]. 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 [3].

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. [4–6]. With mechanistic models, the processes governing the evolution of air pollution is modeled mathematically, whereas statistical and machine learning models are more data-driven [6].

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 [6]. 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 [6, 7]. 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 [5]. Recent advances in machine learning however have shown promising results when it comes to air quality forecasts, especially deep neural networks [5, 6].

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 [6]. For statistical and/or machine learning methods, too simplistic models, lack of data, irrelevant input features, overfitting, etc., can limit prediction accuracy [6]. 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 [6]. 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 [1]. In addition to health risks, the global economic impacts are substantial due to lost labor productivity, increased health care costs, reduced crop yields, etc. [8]. 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 [1].

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) [2]. 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 [2]. 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 [2].

NO_2...

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) [9]. Since Stockholm has centralized district heating and few industries, the major source of local CO, NO_x, and PM pollution is vehicular traffic [3, 9]. 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 [9]. 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 [9].

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 [10]. 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" [3].

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 [4]. 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 [7, 11–13], 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 [4].

From a statistical 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 [13]. Learning the function f amounts to finding a set of parameters (or weights/coefficients) 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 [13]. 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 [13]. In Eq. (2.2.1) 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 Linear regression models

From the wealth of available regression techniques, multiple linear regression (MLR) has been extensively used to forecast and model air pollution [6]. Generally, if none of

[1] What is meant by cost and loss functions can vary slightly in the literature, but in this work, the same terminology as in Lindholm et al. [13] is adopted.

the basic model assumptions are violated, MLR is 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 appropriate [12].

Linear regression for time series data

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 [12]. To this end, adjustments to the MLR model can be made, some of which will require other parameter estimation techniques than the standard method of ordinary least squares (OLS). However, a simple and commonly used procedure to eliminate 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 [12]. The model in Eq. (2.2.2) can be fit with OLS, which in linear regression is the standard way of finding model parameters so that $J(\hat{\boldsymbol{\beta}})$ is minimized [14]. This is done by solving the so-called normal equations, and the least squares estimates of the model parameters are then (in matrix notation) given by

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

A commonly used test for detecting autocorrelation is the Durbin-Watson test, where the statistic will have a value of ~ 2 in the case if uncorrelated errors [12].

Robust regression

The errors of a MLR model should ideally be independent, have constant variance, and be approximately normally distributed [14]. If inference is to be made, the normality assumption is important. Deviations from normality can sometimes be reasonably ignored, however, when the error distribution has long (or heavy) tails, this can be a sign of frequent extreme values in the data, in which case so-called robust model fitting techniques are more appropriate than OLS [14].

One such technique is M-estimation, where a modified version of Eq. (2.2.1) is used to find the parameter estimates:

$$J(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \rho\big(\hat{y}_i - y_i\big)^2 \tag{2.2.4}$$

where ρ is...

Additional considerations for linear models

The extensive use of MLR for air pollution forecasts is many times motivated by its simplicity and straightforward implementation [6]. Another advantage is interpretability; for example, inference can be made on all input variables, allowing one to investigate their individual importance and relationship to the response variable [14]. However, the statistical properties of MLR make it rather restrictive, and not all violations of the assumptions can be remedied (e.g. non-linearity) [12].

Careful variable selection in MLR is also 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 [14]. To this end, variable selection techniques based on optimizing a criterion like the Akaike information criterion are common, and the condition number can be used as a diagnostic for multicollinearity (where a value below 100 is preferred) [14]. With many input variables, it is also common to use regularized versions of MLR such as ridge or lasso regression, in which an extra so-called "penalty" term is added to the cost function to shrink (and essentially stabilize) the model parameter estimates [13, 14].

2.2.3 Extensions of the linear model

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 [15]. 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 [16]. 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 [15].

Monitoring stations

In Stockholm County, there are 19 stationary sites for air pollution monitoring [10], and initially, data from each site was considered. For many stations the data series were irregular, 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 2.1.2, 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 [17]. 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) [15]. 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 A.1 in appendix A.

3.1.2 Data preprocessing

Initial preprocessig

The NO_2 data for Torkel Knutssonsgatan is shown in Fig. 3.1. (Similar plots for all stations are given in Fig. **??** in appendix A.) Some stations had short episodes with missing data, and linear interpolation was used to fill in the missing values. Missing weather data was also linearly interpolated, except for the variables atmospheric pressure and wind speed for which mean imputation was deemed more appropriate.

In Fig. 3.1, a notable reduction in NO_2 levels can be seen during 2020 and early 2021. This reduction is most likely due to the COVID-19 pandemic, and by late 2021, pre-pandemic NO_2 levels are again approached. Because of this, a train-test split (see below) was done to entirely avoid using the data for 2020.

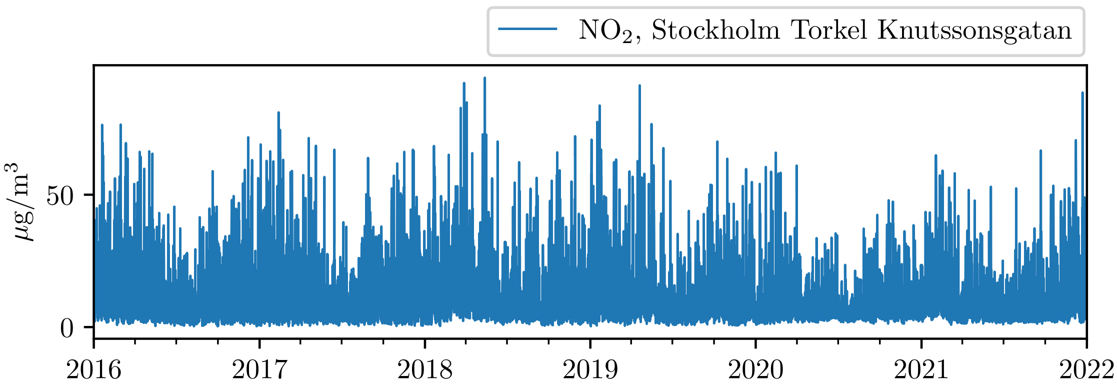

Figure 3.1: NO_2 data for Torkel Knutssonsgatan.

Creating temporal variables

In Fig. 3.1, yearly periodicity in the data can be seen, where levels tend to peak during winter months. 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 waves for day, week, and year. For example, the sine and cosine waves for day were calculated in the following way

$$\text{Sine day} = \frac{1}{2}\left(\sin\left(\text{timestamp} \cdot \frac{2\pi}{86,400}\right) + 1\right)$$

$$\text{Cosine day} = \frac{1}{2}\left(\cos\left(\text{timestamp} \cdot \frac{2\pi}{86,400}\right) + 1\right)$$

where timestamp is in UNIX epoch time[1] (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. Note that the sine and cosine waves were adjusted to oscillate between zero and one. The temporal variables for day in a 24 hour time window are shown in Fig. 3.2 on the following page.

[1]The UNIX epoch time for a given timestamp t is the number of seconds that has passed between January 1, 1970, and t.

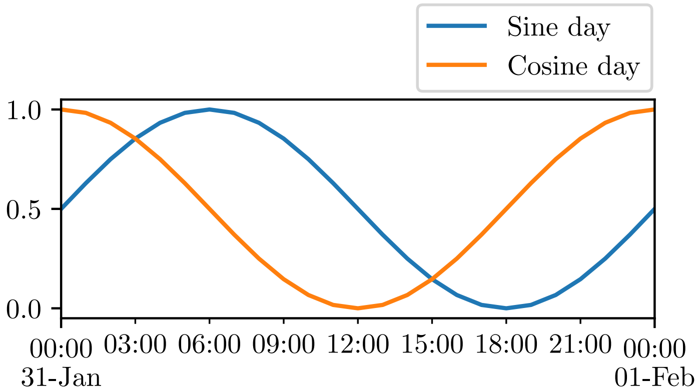

Figure 3.2: Temporal variables for day as sine and cosine waves.

Rolling windows

The rolling windows method extracts data sequences of certain lengths (the "windows") from the input data, and in each window is an "input window" and a "target window" [18]. For example, as shown in Figure 3.3 (where t indicate time steps), with a sequence of nine data points, the first eight observations would constitute the input window, and the ninth observation the target window. After extracting this sequence, a slide forward is made to extract the next sequence, and this is continued until observation n becomes the target window, at which point all the data have been processed.

In this work, rolling windows were used as input to the deep learning models, and different input window lengths were tested for making predictions of a target window one time step ahead (i.e., the next hour). More details are given in section 3.2.2 below.

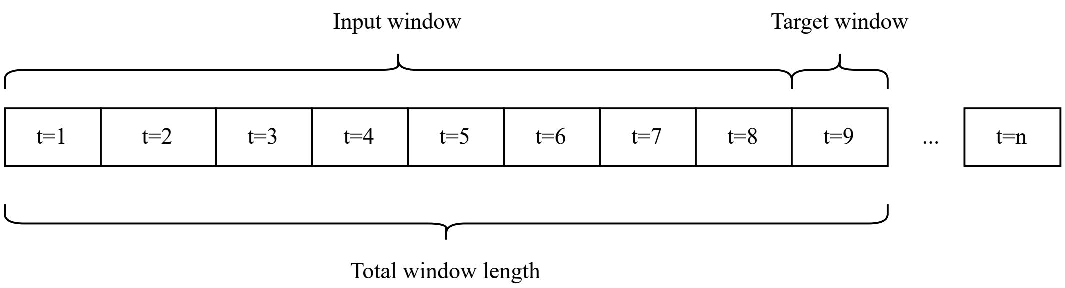

Figure 3.3: Rolling window approach for time-series data.

Train-test split

Lastly, the data was split into training, validation, and test sets, where the validation set was used for hyperparameter tuning for the deep learning models. The test set was taken as the most recent year of data (from 2021-01-01 to 2022-01-01). For the validation set, the data from 2019 was used (since 2020 was an unusual year with regards to air pollution levels). The remaining data was used for training (2016-01-01 to 2019-01-01). This ordered (as opposed to random) split is motivated by the time dependence in the data.

3.2 Model fitting and hyperparameter tuning

3.2.1 Multiple linear regression models

Initially, a simple linear regression model was fit with OLS where the response variable at lag one was used as predictor. No significant autocorrelation was seen with this model, but when also including the response variable at lag two as predictor, the Durbin-Watson statistic improved (i.e., was brought closer to 2). Including additional response variables after the first two lags did not lead to further improvements in terms of eliminating autocorrelation, however, the response variables at lag 24 and 25 turned out to also be highly significant.

It should be noted that a log transformation of the response as well as input variables were required to stabilize the variance of the errors, and also bring them closer to a normal distribution. Even so, deviation from normality was indicated, as can be seen in the residual plots in Fig. B.1 in Appendix B. More specifically, an examination of the residuals implied long-tailed errors, which is why a robust regression with M-estimates (and Huber's method) were used instead of OLS. Summary statistics for the OLS and robust regression are given in Table B.1 and B.2 in Appendix B.

Including data (also log transformed) from other stations improved the MLR model, as did the meteorological parameters and the temporal variables. Data from other stations were, similar as for the response variable, fit with the values at lag one (since future values of the predictors cannot be known).

3.2.2 Deep learning models

4. Results

4.1 Multiple linear regression models

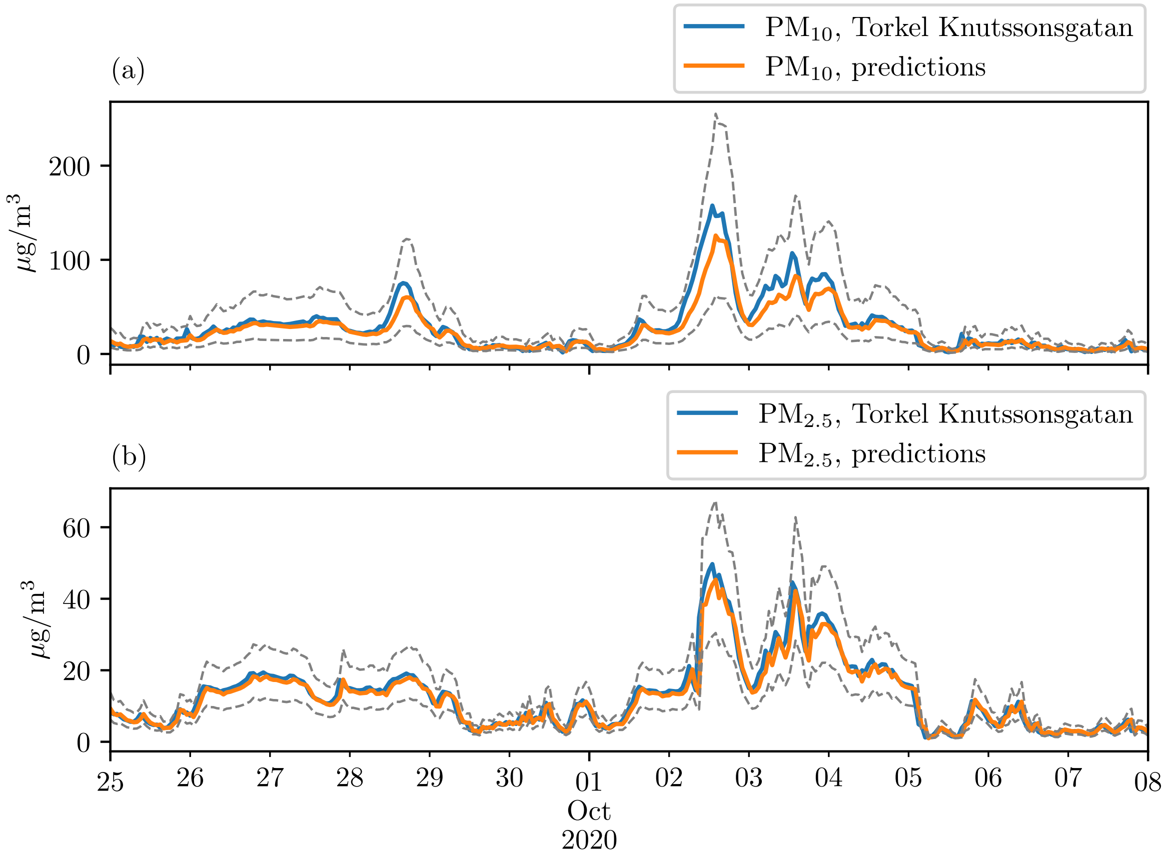

Figure 4.1: Predictions for the MLR model, and actual values, for (a) PM_{10} and (b) $PM_{2.5}$.

5. Discussion and Conclusions

6. Bibliography

[1] World Health Organization, "Ambient air pollution: a global assessment of exposure and burden of disease," tech. rep., World Health Organization, 2016.

[2] G. W. VanLoon and S. J. Duffy, *Environmental Chemistry.* London, England: Oxford University Press, 3 ed., Sept. 2010.

[3] SLB-analys, "Luften du andas - nu och de kommande dagarna: Utveckling av ett automatiskt prognossystem för luftföroreningar och pollen," tech. rep., SLB-analys vid miljöförvaltningen i Stockholm, 2021.

[4] M. El-Harbawi, "Air quality modelling, simulation, and computational methods: a review," *Environmental Reviews*, vol. 21, pp. 149–179, Sept. 2013.

[5] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep learning for air quality forecasts: a review," *Current Pollution Reports*, vol. 6, pp. 399–409, Sept. 2020.

[6] H. Taheri Shahraiyni and S. Sodoudi, "Statistical modeling approaches for PM10 prediction in urban areas; a review of 21st-century studies," *Atmosphere*, vol. 7, no. 2, 2016.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[8] OECD, *The Economic Consequences of Outdoor Air Pollution.* Paris: OECD Publishing, 2016.

[9] SLB-analys, "Luften i stockholm, Årsrapport 2021," Tech. Rep. 2022–5787, SLB-analys vid miljöförvaltningen i Stockholm, 2021.

[10] "Luftövervakning." `https://www.slb.nu/slbanalys/matningar/`. Accessed April 28, 2022.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer New York, 2009.

[12] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to time series analysis and forecasting.* Wiley Series in Probability and Statistics, Nashville, TN: John Wiley & Sons, 2 ed., Apr. 2015.

[13] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, *Machine Learning: A First Course for Engineers and Scientists.* Cambridge University Press, 2022.

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis.* Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Blackwell, 5 ed., Mar. 2012.

[15] SMHI, "Datavärdskap för luftkvalitet." `https://www.smhi.se/data/miljo/luftmiljodata`. Accessed May 3, 2022.

[16] "Environmental monitoring program area: Air." `https://www.naturvardsverket.se/en/environmental-work/environmental-monitoring/environmental-monitoring-program-areas/air/`. Accessed April 27, 2022.

[17] "Historiska data." `https://www.slb.nu/slbanalys/historiska-data-met/`. Accessed April 27, 2022.

[18] A. Gilik, A. S. Ogrenci, and A. Ozmen, "Air quality prediction using CNN+LSTM-based-based hybrid deep learning architecture," *Environmental Science and Pollution Research*, vol. 29, pp. 11920–11938, Sept. 2021.

Appendices

A Monitoring stations

Information about the monitoring stations from which data was used is summarized in Table A.1. In **??** on page ??, 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

B Model diagnostics and summary statistics for the multiple linear regression models

Residual plots from the OLS regression are shown in Fig. B.1 below. From plot (a) and (c), the long-tailed distribution of the errors can be seen, especially in plot (a) where the long tails are indicated by deviations from the straight line. Looking at plot (b), the variance appears stable, and the residuals are scattered in a reasonably random fashion and there is also no signs of non-linearity. The variance also appear stable over time, as indicated in plot (d). An additional numeric test for constant variance was performed where a regression line was fit to $\sqrt{|\hat{\varepsilon}|}$. This line had a non-significant slope (with $x = -0.0045$ and $p = 0.262$), giving further support to the homoscedasticity assumption [14]. In Table B.1 and B.2, summary statistics are shown for the OLS regression and robust regression, respectively.

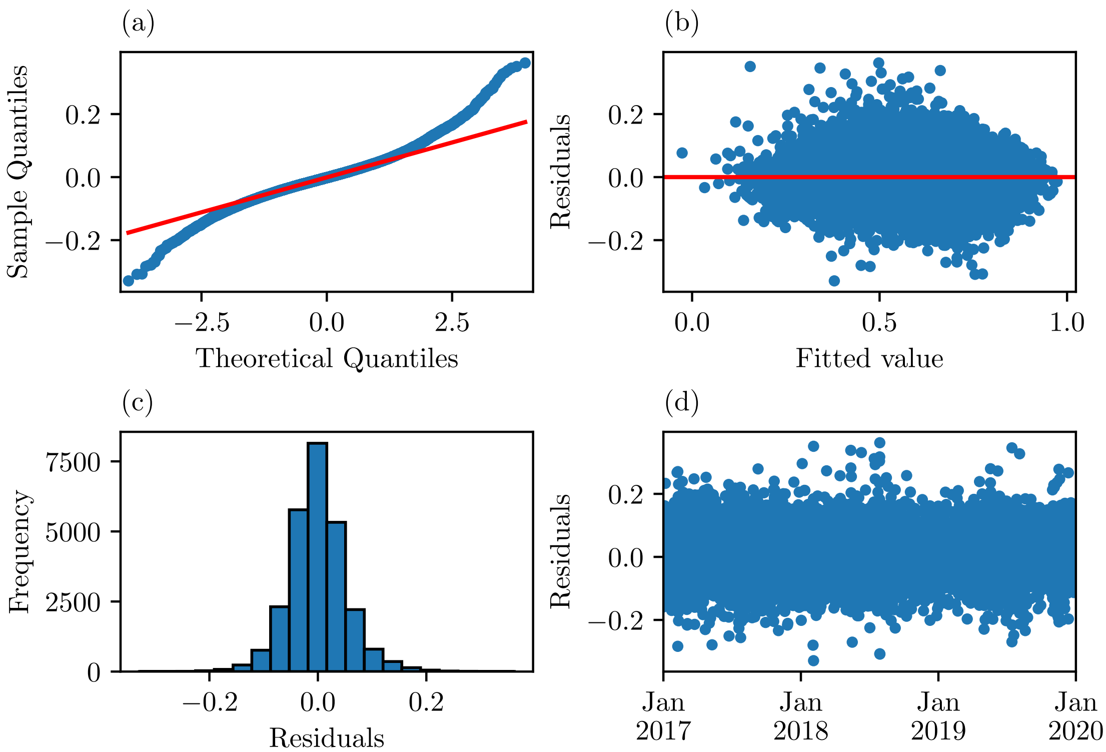

Figure B.1: Residual plots for the OLS regression model.

Dep. Variable:	NO$_2$, Torkel Knutssonsgatan	**R-squared:**	0.858
Model:	OLS	**Adj. R-squared:**	0.858
Method:	Least Squares	**F-statistic:**	1.136e+04
Date:	Fri, 19 Aug 2022	**Prob (F-statistic):**	0.00
Time:	12:42:52	**Log-Likelihood:**	39712.
No. Observations:	26281	**AIC:**	-7.939e+04
Df Residuals:	26266	**BIC:**	-7.927e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
intercept	0.1269	0.005	23.332	0.000	0.116	0.138
NO$_2$, Stockholm Torkel Knutssonsgatan, lag1	0.9222	0.006	142.835	0.000	0.910	0.935
NO$_2$, Stockholm Torkel Knutssonsgatan, lag2	-0.2021	0.006	-35.796	0.000	-0.213	-0.191
NO$_2$, Stockholm Torkel Knutssonsgatan, lag 24	0.0512	0.003	18.420	0.000	0.046	0.057
NO$_2$, Stockholm Hornsgatan 108 , lag1	0.0484	0.004	11.632	0.000	0.040	0.057
NO$_2$, Stockholm Sveavägen 59 , lag1	-0.0414	0.004	-10.076	0.000	-0.049	-0.033
NO$_2$, Stockholm E4/E20 Lilla Essingen, lag1	0.1133	0.005	23.308	0.000	0.104	0.123
Sine day	0.0017	0.001	1.370	0.171	-0.001	0.004
Cosine day	-0.0524	0.001	-37.946	0.000	-0.055	-0.050
Sine week	-0.0105	0.001	-10.986	0.000	-0.012	-0.009
Cosine week	0.0132	0.001	13.325	0.000	0.011	0.015
Temperature	-0.0038	0.003	-1.501	0.133	-0.009	0.001
Relative humidity	-0.0188	0.002	-8.710	0.000	-0.023	-0.015
Solar radiation	-0.0864	0.003	-32.090	0.000	-0.092	-0.081
Wind speed	-0.1473	0.004	-37.139	0.000	-0.155	-0.140

Omnibus:	1882.435	**Durbin-Watson:**	1.939
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6919.847
Skew:	0.298	**Prob(JB):**	0.00
Kurtosis:	5.442	**Cond. No.**	58.2

Table B.1: OLS Regression Results.

Dep. Variable:	NO$_2$, Torkel Knutssonsgatan	**No. Observations:**	26281
Model:	RLM	**Df Residuals:**	26266
Method:	IRLS	**Df Model:**	14
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Fri, 19 Aug 2022		
Time:	12:45:23		
No. Iterations:	23		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.1178	0.005	24.229	0.000	0.108	0.127
NO$_2$, Stockholm Torkel Knutssonsgatan, lag1	0.9596	0.006	166.213	0.000	0.948	0.971
NO$_2$, Stockholm Torkel Knutssonsgatan, lag2	-0.2101	0.005	-41.615	0.000	-0.220	-0.200
NO$_2$, Stockholm Torkel Knutssonsgatan, lag 24	0.0453	0.002	18.244	0.000	0.040	0.050
NO$_2$, Stockholm Hornsgatan 108 , lag1	0.0375	0.004	10.086	0.000	0.030	0.045
NO$_2$, Stockholm Sveavägen 59 , lag1	-0.0351	0.004	-9.571	0.000	-0.042	-0.028
NO$_2$, Stockholm E4/E20 Lilla Essingen, lag1	0.1014	0.004	23.344	0.000	0.093	0.110
Sine day	-0.0001	0.001	-0.130	0.896	-0.002	0.002
Cosine day	-0.0485	0.001	-39.317	0.000	-0.051	-0.046
Sine week	-0.0088	0.001	-10.269	0.000	-0.010	-0.007
Cosine week	0.0111	0.001	12.624	0.000	0.009	0.013
Temperature	-0.0027	0.002	-1.207	0.228	-0.007	0.002
Relative humidity	-0.0180	0.002	-9.309	0.000	-0.022	-0.014
Solar radiation	-0.0803	0.002	-33.360	0.000	-0.085	-0.076
Wind speed	-0.1320	0.004	-37.219	0.000	-0.139	-0.125

Table B.2: Robust linear Model Regression Results.