Short-term Predictions of Air Pollution Using Historical Data

Maria Liatsikou

Abstract In this report, I outline the methods I am using in an ongoing project for predicting the levels of air pollution in different areas of London. I am comparing the performance of different regression algorithms which are trained on historical data of air pollution. Based on the current results, I show that by looking further back into historical data, we can improve the performance of various machine learning algorithms for the task. However, the performance of the models gets worse when the pollution levels of the instances in the test set are different compared to those in the training set. Further weather, traffic and land-use-related variables will be added to my modelling to boost the performance and models that take into consideration the temporal dimension of the task will be incorporated.

1. Introduction

Air pollution is a major problem, especially in highly populated urban areas. The World Health Organization reports seven million deaths annually due to exposure to polluted air[1]. This project aims at predicting the values of air pollution throughout time and in different locations within London, which is a city where exceedances of air quality standards occur frequently.

Task Description: A data-centric approach is proposed in order to make short-term predictions of the values of pollutant concentrations, leveraging historical data and real-world data sources. The analysis is currently based on machine learning algorithms that leverage historical air pollution data. In specific, we aim at predicting the concentrations of various pollutants in different locations for the year 2018. The current analysis comprises two approaches (tasks):

- *Single Pollutant, Different Sites (SPDS):* predicting the values of one pollutant (NO2) in 10 sites across London.
- *Different Pollutants, Single Site (DPSS):* predicting the values of various pollutants (NO2, NOx, NO, PM10) in a single site (Islington).

In both approaches, we aim at predicting the concentration of the pollutant in the next time step. We achieve that by using the most recent concentration values of this pollutant.

2. Data

The source of the data is Londonair, the website of the London Air Quality Network (LAQN)[2]. Londonair offers information about air pollution in London and East England. More specifically, the

[1] https://www.who.int/airpollution
[2] https://www.londonair.org.uk

measurements concern the concentrations of eight pollutants in 120 active stations and meteorological data (e.g. rainfall and temperature) in a fine grained temporal resolution. For the purposes of this project, we downloaded data in a per-15-minute time resolution regarding the concentrations of NO2 in 2018 in 10 sites (see Table 1a) within Greater London (for the SPDS task) and the concentrations of four pollutants (see Table 1b) in the site of Islington (for the DPSS task).

Working on a per-site and per-pollutant basis, we converted the data to consecutive six-hour bins (i.e. daily bins: [00:00-06:00), [06:00-12:00), [12:00-18:00), [18:00-00:00]) by calculating the average value of the pollutant within each bin and replaced any missing values of the resulting time series of a certain pollutant in a single station, using linear interpolation. The mean and standard deviation of the time series are provided in Tables 1a, 1b. We also provide the respective histograms in Figures 1a, 1b. Moreover, after computing the Pearson correlation matrix in both cases, the corresponding heatmap is plotted in Figures 2a, 2b.

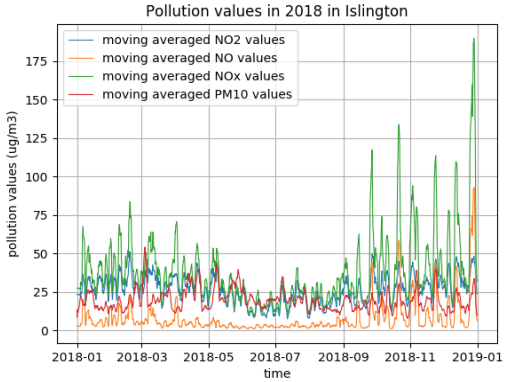
It is worth mentioning that NOx values in Islington vary a lot. To study this closer, we provide the Chart in Figure 1c of the moving-averaged (i.e. averaging over the last 10 binned values) time series of the four pollutants in this site, demonstrating a high rise of NOx in the final quarter of 2018.

Site	Average ± Std (NO2)
Bexley	21.2 ± 13.3
Brent	38.8 ± 18.1
Camden	50.3 ± 21.9
City of London	31.9 ± 13.6
Croydon	41.9 ± 18.8
Ealing	29.3 ± 16.1
Greenwich	34.6 ± 17.2
Islington	25.8 ± 14.2
Kensington	27.8 ± 16.3
Westminster	37.8 ± 16.0

(1a)

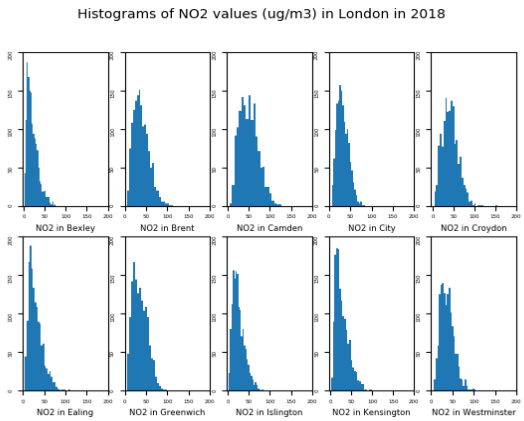
Pollutant	Average ± Std (Islington)
NO2	25.8 ± 14.2
NO	7.94 ± 19.1
NOx	37.9 ± 38.4
PM10	19.6 ± 9.7

(1b)

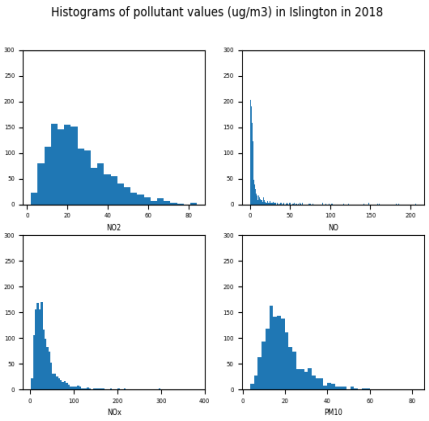


(1c)

Table 1: Mean and Standard Deviation of NO2 concentrations within London in 2018 (1a), Mean and Standard Deviation of various pollutants in Islington in 2018 (1b), Pollution values in Islington over time (Chart 1c).

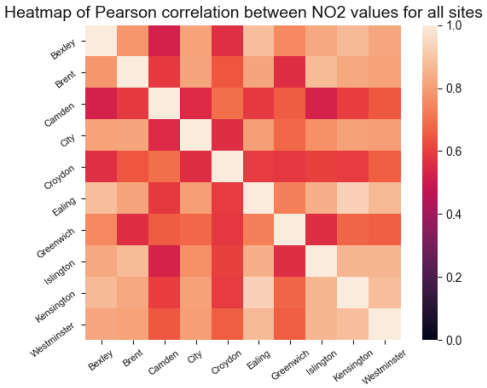


(1a)

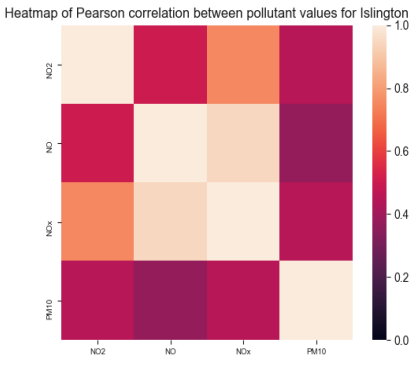


(1b)

Figure 1: Histograms of NO2 concentration values within Greater London in 2018 (1a), Histograms of pollution concentration values in Islington in 2018 (1b).



(2a)



(2b)

Figure 2: Heatmap of Pearson correlation between NO2 concentration values for 10 sites in London (2a), between four pollutants concentration values in Islington (2b).

3. Experiments

In this section we will describe the experiments that have been conducted thus far, including how we split the data (3.1), the number of the lagged variables that we have used (3.2), the machine learning algorithms that we have employed (3.3) and how we assess their performance (3.4). We perform the same steps in both tasks (SPDS, DPSS).

3.1 Train/Test Split

The dataset is split in training and test set using a percentage rate of 80/20. The test set contains the last values of our dataset. In the case of NN, we also used a development set for evaluation with a split of 70/10/20.

3.2 Features

The features of the prediction models are historical air pollution data. The number of features of each instance is the number of lagged variables used to predict the target value. We experiment with a range of such lagged variables (1-16). The values of the features of the training set are normalized to zero mean and unit variance. The same transformation is applied to the test set.

3.3 Models

We use several regression models for both tasks. Specifically, we use Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Support Vector Machine for regression (SVM), k Nearest Neighbours (kNN) and a Feed Forward Neural Network (NN). Each of the aforementioned ML algorithms is trained, while the models and results are kept in pickle and csv files respectively.

Parameter Selection: In both tasks, we use grid search in order to tune the parameters of each machine learning algorithm. In LASSO, we tune the learning rate (10^{-3}, …, 10^{3}); in RF, we tune the number of trees (50, 100, 200, 300, 400, 500); in SVM, we use an RBF kernel and we tune the C (10^{-3}, …, 10^{3}) and the γ parameters (10^{-3}, …, 10^{3}); in kNN, we tune the number of neighbours (2, 5, 10, 20, 50); and in NN, we tune the learning rate (10^{-4}, 10^{-3}, 10^{-2}), the number of neurons (10, 25, 50, 100, 200) and the mini-batch size (32, 64).

3.4 Evaluation

The metrics used for the evaluation of the models are the mean squared error (MSE) and the coefficient of determination (r^2 score) between the predicted and the target values. These metrics are plotted as a function of the number of features for each model.

4. Results

In this section we present the results of all algorithms in both tasks (SPDS, DPSS). We begin our analysis with the SPDS and then we proceed with DPSS.

4.1 Results on SPDS

In this task we try to predict the values of one pollutant (NO2) in 10 sites across London, using historical data, as described in Section 3.

The charts in Figures 3 and 4 show the Mean Squared Error (MSE) and the r^2 score of our models. Table 2 shows the best performing number of lagged features for each algorithm, averaged across the different sites. Despite the fact that NNs need a lot of data for training, and that our dataset is relatively small, in the vast majority NN achieves the highest r^2. However, LASSO achieves the lowest MSE. This inconsistency stems from the different nature of the two metrics.

MSE decreases as the temporal window rises. In fact, in most cases MSE and r^2 score improve rapidly until the number of features reaches the value of 5 and then small fluctuations are observed.

This signals that we can get a more accurate prediction of the air pollution levels by looking into the past. The best metric values are reached in the case of 5 features for kNN and 13 features for all the other models (see Table 2). Finally, the patterns in the variations for both metrics are similar across the different sites.

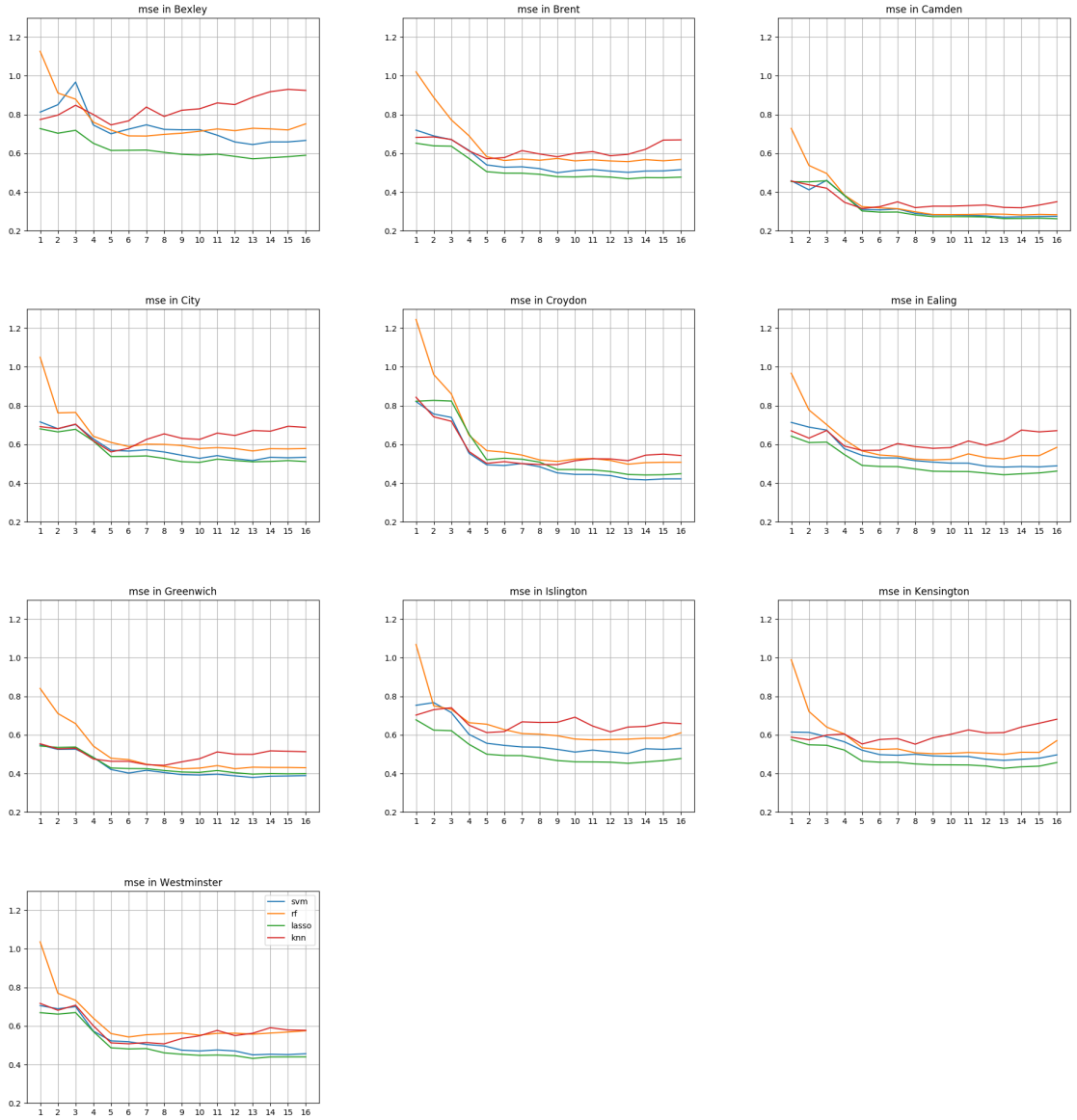


Figure 3: Mean Squared Error of the algorithms applied in SPDS task.

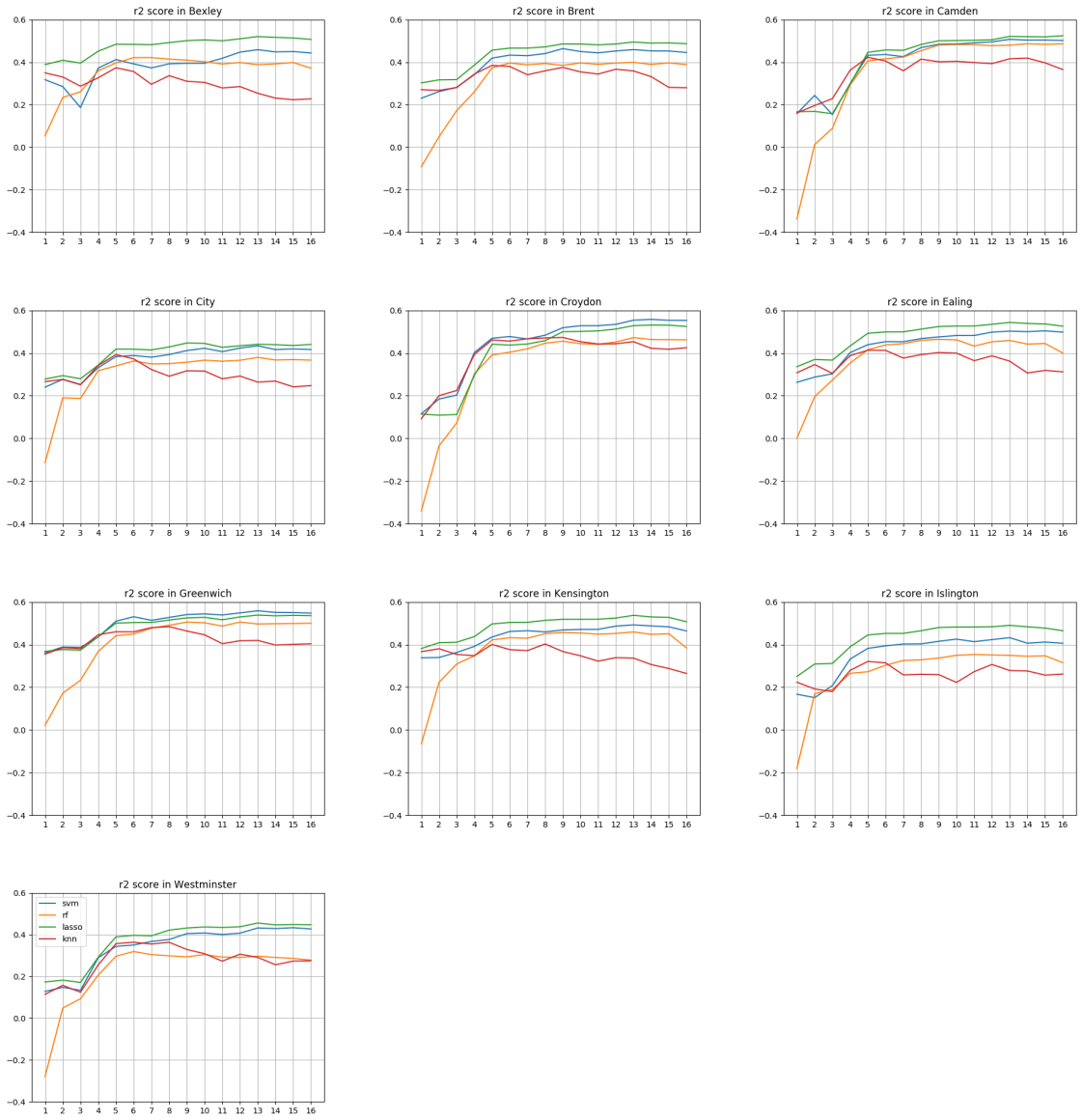


Figure 4: Coefficient of determination of the algorithms applied in SPDS task.

Model	Best r^2 (#features)	Best MSE (#features)
kNN	(5) 0.399	(5) 0.541
RF	(13) 0.418	(13) 0.522
SVM	(13) 0.483	(13) 0.464
LASSO	(13) 0.507	**(13) 0.441**
NN	**(13) 0.517**	(13) 0.508

Table 2: Best metric values for r^2 and MSE and the corresponding number of lagged features for each algorithm (SPDS task).

4.2 Results on DPSS

In this task, we focus on a single site (Islington) and we try to use historical data as our features (see Section 3) to predict the values of different pollutants.

The charts in Figure 5 show the Mean Squared Error (MSE) and the coefficient of determination of the models for the DPSS task. Table 3 shows the best performing number of lagged features for each algorithm, averaged for the different sites[3]. LASSO achieves the lowest errors and the highest r^2 for predicting all pollutants except for PM10, for which SVM is the best model. Contrary to SPDS task, the variations for both metrics are not really consistent for different pollutants. The values of MSE of the models predicting PM10 is independent of the number of features, while the error is rather high in the prediction of NO and NOx. Looking back in Figure 1c and taking into consideration that the test set contains the last values of the dataset, it is clear that the values of NO and NOx in the training set vary a lot compared to the test set, triggering the high error values of our models.

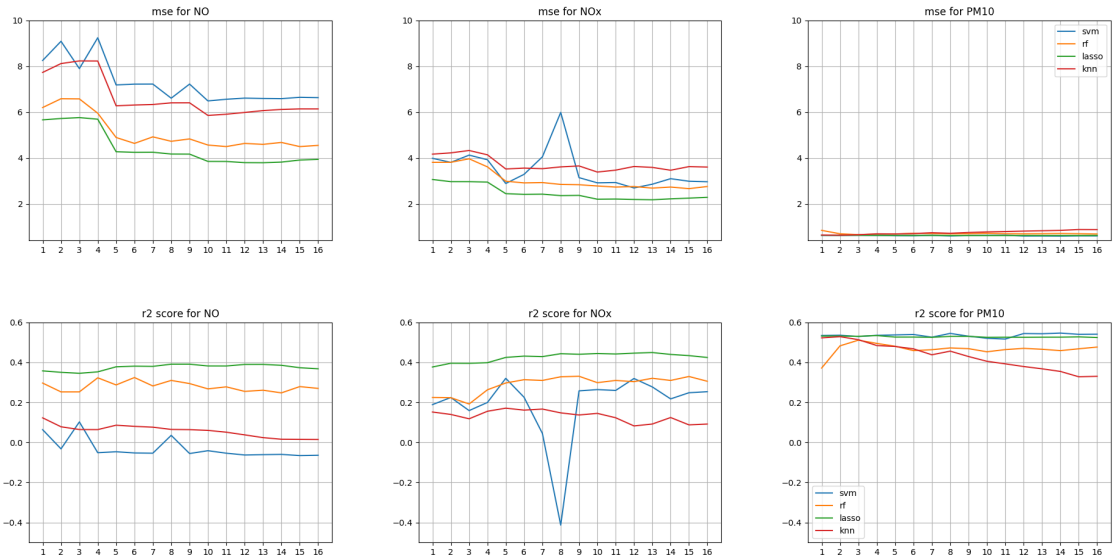


Figure 5: Mean Squared Error and coefficient of determination of the algorithms applied in DPSS task.

[3] The results of NN will be added once completed

Model	Best r^2 (#features)	Best MSE (#features)
kNN	(5) 0.264	(10) 2.677
RF	(8) 0.360	(15) 2.109
SVM	(12) 0.306	(12) 2.602
LASSO	**(13) 0.464**	**(13) 1.760**

Table 3: Best metric values for r^2 and MSE and the corresponding number of lagged features for each algorithm (DPSS task).

5. Future Steps

The analysis is currently based only on historical data. Following up, other real-world data sources will be leveraged and meteorological, traffic and land use data will be incorporated in various algorithms [1, 2]. In addition, ARIMA models and Convolutional Neural Networks will also be applied, in order to include the temporal dimension of this task. This way, we aim to build more robust and accurate models for predicting air pollution levels.

References

[1] Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S.N. and Weikum, G., 2016, May. Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. In 2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW) (pp. 54-59). IEEE.

[2] Zheng, Y., Liu, F. and Hsieh, H.P., 2013, August. U-air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1436-1444). ACM.