Predicting School Hour NO_2 Concentrations Using a Hybrid ARIMA and Multiple Linear Regression Ex Ante Forecasting Approach

By

Steven Futter

Thesis Project

Submitted in partial fulfillment of the

Requirements for the degree

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

September 2017

Dr. Alianna J. Maren, First Reader

Dr. Lynd Bacon, Second Reader

ABSTRACT

Predicting School Hour NO_2 Concentrations Using a Hybrid ARIMA and Multiple Linear Regression Ex Ante Forecasting Approach

Steven Futter

To minimize a pupil's exposure to harmful levels of nitrogen dioxide (NO_2) it is important that schools are able to understand the hourly variation in pollutant concentrations that can be expected each day. Informed decisions can then be made to minimize the amount of time that a child may be exposed to excessively high levels of pollution. In this paper, I use a hybrid ARIMA and multiple linear regression model approach to predict school hour NO_2 concentrations between 9 a.m. and 4 p.m. during a Monday to Friday school week at a site in Greenwich Eltham, London. A stepwise variable selection algorithm is used to select the best subset of predictors (pollutants and meteorological variables) as inputs into a multiple linear regression model, and an autoregressive integrated moving average (ARIMA) model is used to generate the forecasted hourly concentrations of each predictor. I demonstrate that the mean absolute error (MAE) of the response variable, hourly NO_2, can be reduced by over 50% on average for each hour of the school day; MAE reduced from 7.02 for the ex ante 'genuine forecast' model to 2.59 for the ex post hybrid ARIMA-stepwise variable selection model.

TABLE OF CONTENTS

LIST OF TABLES, ILLUSTRATIONS, FIGURES OR GRAPHS

INTRODUCTION

In December 1952, the combination of an anticyclonic (a period of high pressure and sinking air) weather pattern, windless conditions, coal emissions from citywide chimney use during a cold winter, vehicle exhaust emissions, and other pollutants formed a thick smog, covering the U.K.'s capital city of London for five straight days. As chronicled by Rieuwerts, J. (2016), as soon as weather conditions and wind patterns changed, the smoke and pollutants dispersed almost overnight, but not without the heavy smog wreaking havoc, ultimately causing the death of approximately 12,000 people.

More recently, many reports and newspaper articles have been published identifying the effects of poor air quality on human health. A report produced by Walton, H. et al. titled "Understanding the Health Impacts of Air Pollution in London", estimated that in 2010, there were 5,900 deaths in London that were associated with long-term exposure to nitrogen dioxide (NO_2), one of a group of gases called nitrogen oxides, of which U.K. road transport is estimated to be responsible for about 50% of the total emissions. In the same year, the U.K. government estimated that NO_2 emissions discharged by diesel-powered engines were responsible for 23,500 deaths (Walton, H. et al., 2015).

Children are more susceptible to ambient air pollution than adults. A study for the World Health Organization, "The Effects of Air Pollution on Children's Health and Development: A Review of the Evidence", note that the "special vulnerability of children to exposure to air pollution is related to several differences between children and adults. The ongoing process of lung growth and development, incomplete metabolic systems, immature host defences, high rates of

infection by respiratory pathogens and activity patterns specific to children can lead to higher exposure to air pollution and higher doses of pollutants reaching the lungs" (Binkova, B., Bobak, M., Chatterjee A., et al. (2004)). According to Asthma U.K., one in 11 children in the U.K. have been diagnosed with asthma, making U.K. childhood asthma prevalence rates one of the highest worldwide ("Asthma facts and statistics", n.d.). In fact, data provided by Asthma U.K. demonstrates that a child in the U.K. is admitted to hospital every 20 minutes because of an asthma attack.

Given the known effects of long-term exposure to NO_2, some local authority council governments across the U.K. have started to react. A report in The Times newspaper that two schools in South Yorkshire were closed due to air pollution caused by the busy roads surrounding the school in the article "Schools shut under a cloud of diesel" (Leake, 2015). The local council deemed the buildings a threat to children's health and had plans to move the schools to new sites that were a safer distance from the roads. Unfortunately, the option to relocate is not available for all schools who lack the funds or local authority support to do so themselves, or are otherwise limited by various factors. Aether, an air quality and climate change emissions consultancy, had earlier produced a report in highlighting the fact that many schools are located in high pollution areas in that year: 433 of the 1,777 primary schools in London are in locations where average concentrations of NO_2 exceeded the EU limit value (King, K., & Healy, S., 2013).

The emergence of air pollution in the U.K. as a significant problem to be prioritised is evidenced by the stance that the new London mayor, Sadiq Khan, is taking on the issue. In a press release by the London Assembly on January 24, 2017, "Mayor's new 'air quality' audits to protect

thousands of school kids" from the office of the Mayor of London, Mr. Khan plans to introduce new air quality audits into a number of London's most polluted schools. The Mayor states that every "...child deserves the right to breathe clean air in London and it is a shameful fact that more than 360 of our primary schools are in areas breaching legal pollution limits. Yesterday I was forced to issue the first 'very high' air pollution alert under my new comprehensive system, London's filthy air is a health crisis and our children are particularly vulnerable to the toxic effects of air pollution. This is why I'm doing everything in my power to safeguard Londoners' health and my new air quality audits are a strong step towards helping some of the most polluted schools in London identify effective solutions to protect pupils from toxic fumes." (*Mayor's new 'air quality' audits to protect thousands of school kids*, 2017).

According to the same press release, some possible audit recommendations could include moving school entrances and play areas to reduce exposure to busy roads, initiating 'no engine idling' schemes, restricting high emissions vehicles near school grounds, pedestrianizing school entrances, introducing green infrastructure such as 'barrier bushes', and encouraging walking and cycling by improving pathways to the school grounds. In addition to the potential recommendations outlined in the press release, through my research, I propose an additional step that could be taken by schools that may further reduce exposure to harmful levels of NO_2. Importantly, the proposal may not need as large of an investment from the city, so it may benefit schools who may not be part of the Mayor's audit. By predicting hourly variations in NO_2 across the school day, my research shows that it is possible for schools to adjust their activities ahead of the class schedule so that high physical exertion activities can be minimized, relocated, or brought inside the school during times where outside pollution levels may be excessively high. There is opportunity for schools to use predictive analytic tools to decrease the likelihood that

children may be exposed to excessively high concentrations of pollutants. It is upon these opportunities that I focus the attention of this paper. The U.K.'s Department for Environment, Food & Rural Affairs (DEFRA), recommends that "anyone experiencing discomfort such as sore eyes, cough or sore throat should consider reducing activity, particularly outdoors" ("Daily Air Quality Index - Defra, UK", n.d.) when levels of air pollution are high.

Figure 1 below shows the recommended actions and health advice that are provided by DEFRA.

Recommended Actions and Health Advice

Air Pollution Banding	Value	Accompanying health messages for at-risk individuals*	Accompanying health messages for the general population
Low	1-3	**Enjoy** your usual outdoor activities.	**Enjoy** your usual outdoor activities.
Moderate	4-6	Adults and children with lung problems, and adults with heart problems, **who experience symptoms**, should **consider reducing** strenuous physical activity, particularly outdoors.	**Enjoy** your usual outdoor activities.
High	7-9	Adults and children with lung problems, and adults with heart problems, should **reduce** strenuous physical exertion, particularly outdoors, and particularly if they experience symptoms. People with asthma may find they need to use their reliever inhaler more often. Older people should also **reduce** physical exertion.	Anyone experiencing discomfort such as sore eyes, cough or sore throat should **consider reducing** activity, particularly outdoors.
Very High	10	Adults and children with lung problems, adults with heart problems, and older people, should **avoid** strenuous physical activity. People with asthma may find they need to use their reliever inhaler more often.	**Reduce** physical exertion, particularly outdoors, especially if you experience symptoms such as cough or sore throat.

Figure 1: DEFRA Recommended Actions and Health Advice

For schools to avoid exposing their students to dangerous levels of air pollution, it is important for administrators to be given tools that forecast when pollution levels are most likely to be raised above the air quality index threshold from moderate to high. By knowing ahead of time

what the levels of pollutants such as NO_2 may be, schools can adapt their activities so that their students may “reduce physical exertion” or “consider reducing activity, particularly outdoors” as per the advice provided by the U.K. government ("Daily Air Quality Index - Defra, UK", n.d.). Currently, and at minimum, schools can rely upon the DEFRA website that provides the current and five-day forecast for pollution levels across the country.

Figure 2 shows the DEFRA website pollution forecast.

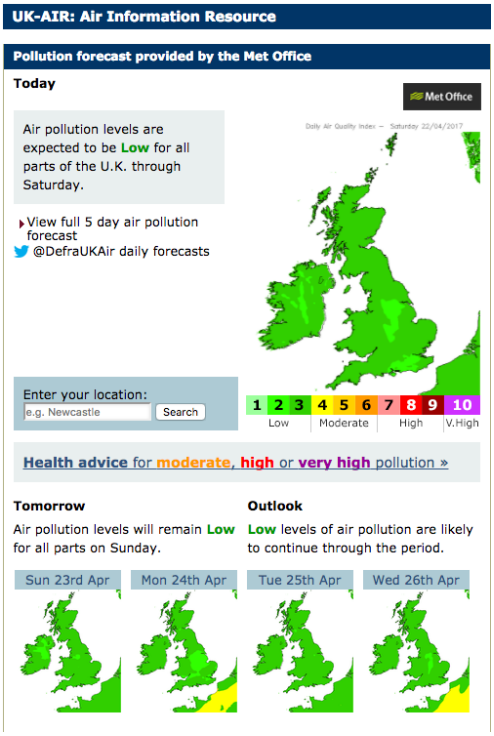

Figure 2: DEFRA Pollution Forecast Website

Schools based in London also have access to additional and more localized data that is provided by the Environmental Research Group, King's College London on the website London Air. The London Air website provides a map showing the current pollution levels as well as the one day ahead forecast for each local authority district across the city. Both of these tools are

excellent resources for schools that may be looking to minimize their pupils' exposure to pollution. However, as I demonstrate in this paper, hourly levels of pollution can fluctuate tremendously, thus requiring more detailed, sophisticated, and accessible predictive tools.

As highlighted in the hourly pollution concentration chart provided by The Economist magazine in Figure 3 below, cities such as Seoul and Hong Kong top global city rankings for NO_2 pollution. In Europe, London and Paris have daytime pollution levels that are consistently higher than the WHO's guidelines. New York and several other American cities have much cleaner air, which, as stated in the article, is partly due to the fact that diesel fuel,which emits more nitrogen dioxide, is less common in the United States. The article notes that residents may be able to reduce their exposure to pollution, "by changing their daily routines, such as commuting to work an hour earlier" ("Daily Chart: Comparing Urban Air Pollution", 2016).

Figure 3 shows that hourly variation of NO_2 varies differently across cities.

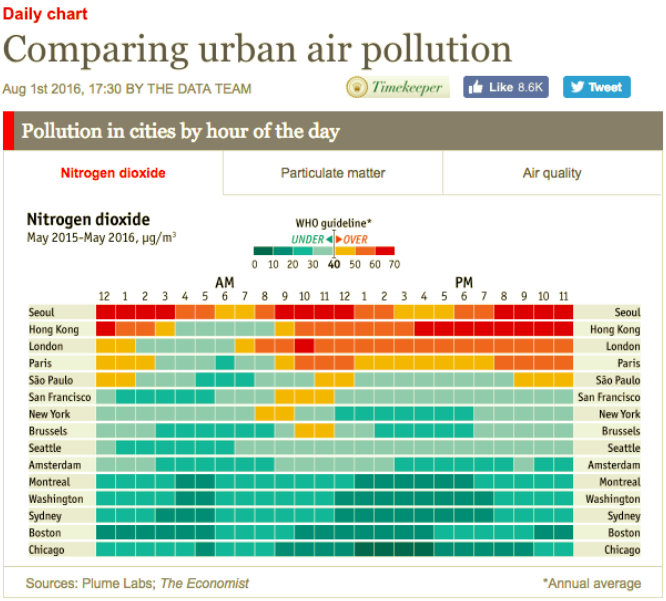

Figure 3: Comparing Urban Air Pollution

STATEMENT OF THE PROBLEM

To minimize a student's exposure to harmful levels of pollutants such as NO_2, schools need to understand not only the daily forecasted concentration levels provided by DEFRA and/or London Air, but also the daily hourly variation in pollutant concentrations that are expected each day before school begins. Without understanding the hourly variation in air quality, it is not possible for schools to provide their teachers with sufficient information to make informed decisions about the level of physical activity for each day, and whether or not physical activity would be best performed indoors or outdoors, so that student exposure to harmful pollutants can be minimized.

In this paper, I attempt to add additional clarity to the pollution prediction question by contextualizing it as a regression problem, using forecasted ARIMA model predictor variable inputs for each hour of the school day (9 a.m. to 4 p.m.). By evaluating the issue as a regression problem, prediction tools given to schools can provide an accurate hourly forecast of pollution concentration levels each day before school begins. By using a variable selection model such as the stepwise variable selection algorithm, the school is also able to understand what factor inputs are explaining the most variation in hourly NO_2 concentrations. Using a model that is easily comprehensible as well as a model that can be run quickly on a small computer is crucial to achieving accurate and useful results.

This paper intends to demonstrate the possibility, and benefits of, providing schools with a predictive tool that provides hourly air quality forecasts. Each day, school administrators and decision makers can run the predictive tool's model and review its output. The tool will

determine if levels of NO_2 at each hour of the school day are expected to exceed the European Union's Air Quality Standards limit, which is currently set at 40 μg/m3. In order for the predictive tool to be useful and accessible for schools, local car traffic and emissions data must be readily available: this means that each school must have access to a localized node that measures air quality in real time, along with access to local weather station data that measures wind speed and direction, barometric pressure, solar radiation, and temperature.

REVIEW OF THE LITERATURE

Literature Supporting the Data Modeling and Choice of Modeling Approach

Lalas, D. et al. (1982) analysed the meteorological parameters that affect sulphur dioxide (SO_2) pollution in Athens, demonstrating that wind speed, minimum temperature, and rain volume control SO_2 concentrations during the cold season while wind speed, wind direction, and relative humidity are key factor inputs in the warm season. Multiple regression models, both linear and non-linear were used to make prediction of the next day's SO_2 concentration with model outputs explaining more than 50% of the data variance. Additionally, Lalas, D. et al. (1982) showed that discriminant analysis classification functions, utilizing forecasted meteorological variables are capable of accurately predicting the occurrence of high pollution concentrations.

Discriminant functions are linear composites of the data chosen to maximize the variance between different groups relative to the variance within groups and to be linearly independent. By separating next day pollution into four levels of severity, this study showed that standardized canonical discriminant function coefficients can be used to determine which meteorological

indicators discriminate the levels. The research revealed that the two highest levels were dominated by the SO_2 concentrations of the previous day and the minimum temperature; the next two highest groups by the wind direction, the relative humidity, and the minimum temperature; the next two levels by the duration of rain, the wind speed, the wind direction and the relative humidity. This particular study applies to my research as it demonstrates the usefulness of the multiple regression model in predicting pollution levels while using a limited number of variables (in this case, precipitation, wind speed/direction, and relative humidity).

Niska, H. et al. (2004) demonstrated that a genetic algorithm can be used for selecting the inputs to design a multi-layer perceptron (MLP) model for forecasting hourly concentrations of nitrogen dioxide (NO_2) at a busy urban traffic station in Helsinki, Finland. “The modelling of real-world process such as air quality is generally a difficult task due to both their chaotic and non-linear phenomenon and high dimensional sample space”. The authors point out that neural networks (NN) have been used successfully in this domain, but that the selection of network architecture is still problematic and time consuming when developing a model for practical situations. “In the air quality forecasting, especially, the selection of optimal input subset (Jain and Zongker, 1997; John et al., 1994) becomes a tedious task due to high number of measurements from heterogeneous sources and their non-linear interactions. Moreover, due to a complex interconnection between the input patterns of NN and the architecture of NN (related to the complexity of the input and output mapping, the amount of noise and the amount of training data), the selection of NN architecture must be done simultaneously.” Niska, H. et al. highlight that evolutionary and genetic algorithms (GA) (Holland, 1975) have proven to be powerful techniques due to their ability to solve both linear and non-linear problems, but noted that using genetic algorithms for optimizing neural networks requires a high computational

requirement.

Additionally, Niska, H. et al. (2004) explain that the MLP model was chosen to be considered in the study due to the “extremely non-linear relationships” that “exist in the real world”. They argue that it is inappropriate to attempt to understand these problems using traditional regression. Additionally, another highlight of the MLP and reason why the MLP was chosen for their study was that the MLP “can be trained to approximate any smooth, measures (highly non-linear) function without prior assumptions concerning the data distribution”. Importantly, Niska, H. et al. demonstrated that a GA is capable of searching feasible high-level architectures and predictor variables so that the computational efforts by be reduced by eliminating irrelevant inputs. This leads to lower costs since a smaller amount of measures may be required. Using a simpler and reduced input model led to minimising the risk of noise by over-fitting.

In contradiction to Niska, H. et all (2004), Mckendry, I. G. (2002) compared the predictive accuracy of an multi-layer perceptron (MLP) artificial neural network (ANN) models with a traditional multiple regression (MLR) model daily maximum and average O3 and particulate matter (PM10 and PM2.5) forecasting. MLP particulate forecasting models show little if any improvement over MLR models and exhibit less skill than do O3 forecasting models. Meteorological variables (precipitation, wind, and temperature), persistence, and co-pollutant data are shown to be useful particulate matter predictors. If MLP approaches are adopted for particulate matter forecasting, training methods that improve extreme value prediction are recommended.

Zhang, G. (2003) used a hybrid artificial neural network (ANN) and autoregressive integrated moving average (ARIMA) approach to forecasting where the ARIMA model was used to predict the linear and the ANN the nonlinear.. “In this paper, we propose to take a combining approach to time series forecasting. The linear ARIMA model and the nonlinear ANN model are used jointly, aiming to capture different forms of relationship in the time series data. The hybrid model takes advantage of the unique strength of ARIMA and ANN in linear and nonlinear modeling. For complex problems that have both linear and nonlinear correlation structures, the combination method can be an effective way to improve forecasting performance. The empirical results with three real data sets clearly suggest that the hybrid model is able to outperform each component model used in isolation.”

In summary, much has been written about the prediction of various pollutants and how best to model the non-linearity presented in pollutant concentration levels. Past papers have used a variety of modeling techniques with varying degrees of predictive accuracy. There does not appear to be one model that performs better than all others and the choice of which model is best is to be debated. However, it is clear that modeling techniques, when applied in combination, can provide improvements in a model’s predictive accuracy.

METHOD

Data Collection

Hourly records of pollutant and meteorological variables were downloaded from the London Air Quality Network website that is maintained by King's College London. The data was obtained from one site in Greenwich Eltham, located in southeast London.

Figure 4 shows the London Air's air pollution monitoring portal.

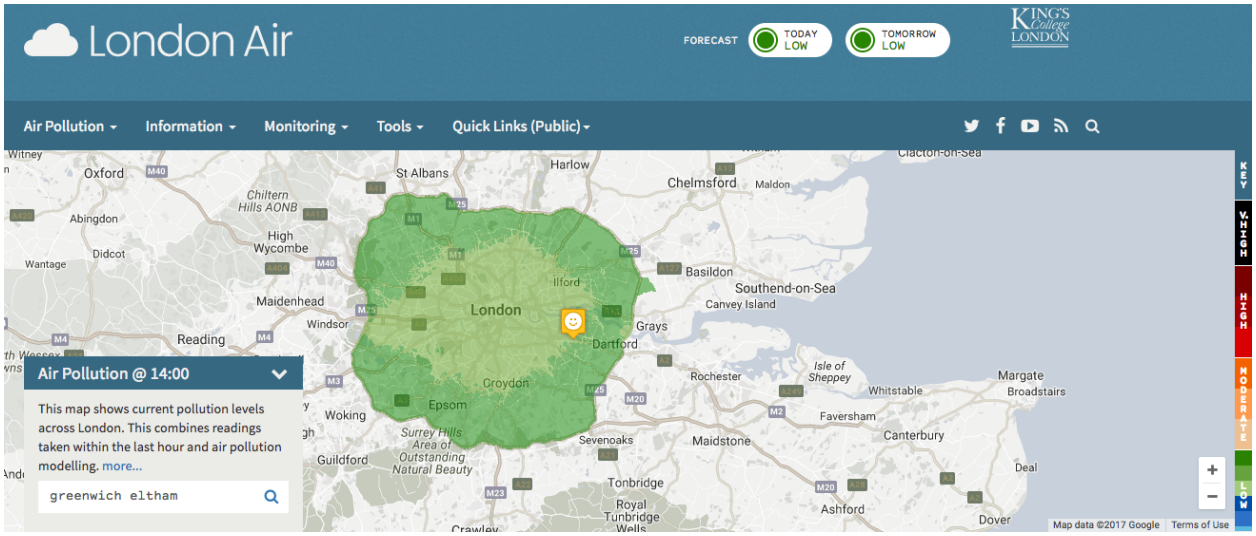

Figure 4: London Air Website Air Pollution Monitoring Portal

The Greenwich Eltham monitoring site utilizes three specific quality assurances and quality control (QA/QC) regulations set by the London Air Quality Network (LAQN), thus ensuring that all data collected is of a certain standard.

Data Preparation

Preparing the data for analysis and modeling is a critical step that has a substantial effect on the performance of any predictive model. Pollutant and meteorological data prepared for this paper was downloaded into .csv files from the London Air website, and the variables were inspected

for missing values. The raw data downloaded for Greenwich Eltham for date range January 1st 2008 to December 31st 2015 included the following variables: BP (barometric pressure), NO (nitric oxides), NO2 (nitrogen dioxide), NOX (oxides of nitrogen), O3 (ozone), PM10, PM2.5, RAIN (rain), RHUM (relative humidity), SOLR (solar radiation), SO2 (sulphur dioxide), TMP (temperature), WDIR (wind direction), and WSPD (wind speed); however, not all variables covered the same hours of day.

Missing Values

For the period January 1, 2008 to December 31, 2015 the Greenwich Eltham site had a large number of missing hourly observations, since not all variables were measured for the entire date range selected. Over 70% of the observations for RAIN, RHUM, BP, SOLR, and TEMP were missing, so to improve prediction accuracy, three data pre-processing steps were taken.

Step 1: Use nearby station data to replace the missing values

Values for RAIN, RHUM, BP, SOLR, and TEMP recorded at nearby stations Bexley Erith, Bexley Belvedere West, and Dagenham Rush Green were used as proxies for the Greenwich Eltham site. The Bexley Erith site was used for its barometric pressure (BP), Bexley Belvedere West for rain (RAIN), relative humidity (RHUM), and temperature (TEMP), and Dagenham Rush Green for solar radiation (SOLR). After using the imputed values for these four variables, the missing value percentages decreased from 80% to 34% for BP, from 96% to 32% for RAIN, from 92% to 31% for RHUM, from 71% to 14% for TEMP, and from 76% to 3% for SOLR.

Step 2: Use lagged and/or median values to replace missing values

One-hour, two-hour, and three-hour lagged values are used to replace any remaining missing values before removing rows that still include NA values. If a value is NA, I first use the one-hour lag; if the one-hour lag is also NA, I use the two-hour lag value. If the two-hour lag is NA, then the three-hour lag is used for all predictor variables, except for RAIN. Note that missing RAIN values are replaced by zero values, with zero being the median RAIN value. The median value for RAIN is used; since RAIN is not correlated to its lagged values, unlike the other variables in the data set. Having lived in London for a period of time, I can make this assumption without carrying out necessary correlation testing, but the results do stand with the data collected. RAIN has correlation coefficients of 0.42, 0.31, and 0.21 for the one, two, and three hour lagged RAIN values, respectively.

Step 3: Remove weekend and non-school hour observations (6am to 4pm)

Reduce data set to observations that represent the school day, which in England is generally from 9 a.m. to 4 p.m. from Monday to Friday. Additional hours are added prior to 9 a.m., since one- to three-hour lagged variables are used in the data modeling. This final step reduced the data set from 15,378 observations to 2,860.

Figure 5 shows the final reduction in missing values for the Greenwich Eltham site across the three data processing steps outlined above.

	Original Greenwich Eltham Data		Step 1: Use Nearby Station Data		Step 2: Replace NA with lag 1-3 hour and median values then remove remaining NA rows		Step 3: reduce data to school hours (6am-4pm), Monday-Friday	
Variable	Count Obs	% Missing	Count Obs	% Missing	Count Obs	% Missing	Count Obs	% Missing
BP	78912	0.8	78912	0.34	15378	0	2860	0
RAIN	78912	0.96	78912	0.32	15378	0	2860	0

RHUM	78912	0.92	78912	0.31	15378	0	2860	0
WDIR	78912	0.19	78912	0.18	15378	0	2860	0
PM2.5	78912	0.18	78912	0.17	15378	0	2860	0
WSPD	78912	0.16	78912	0.15	15378	0	2860	0
TEMP	78912	0.71	78912	0.14	15378	0	2860	0
PM10	78912	0.09	78912	0.09	15378	0	2860	0
NO	78912	0.09	78912	0.07	15378	0	2860	0
NO2	78912	0.09	78912	0.07	15378	0	2860	0
NOX	78912	0.09	78912	0.07	15378	0	2860	0
SO2	78912	0.07	78912	0.06	15378	0	2860	0
SOLR	78912	0.76	78912	0.03	15378	0	2860	0
O3	78912	0.02	78912	0.01	15378	0	2860	0

Figure 5: Missing Values in Raw Data

Training and Validation Data (70/30 Split)

70% of the pre-processed data was set aside as the training set, which is the largest group of all and is used to train the models. The testing set, or 30% of the data, is used to test the accuracy of the model by evaluating the prediction error of the model created, which is based upon the training data set. By retaining 30% of the data for testing, it is possible to prevent an overfitting of the model to the data, ensuring that the model can generalize well to new data provided.

Figure 6 shows how the observations are split between the training and testing data sets. Note that the size of the training and test data set combined is 2,082, rather than than 2,860. This is due to the fact the school day begins at 9 a.m. as opposed to 6 a.m.. However, the lagged

variables recorded between 6 a.m. and 8 a.m. are not lost; they are instead added as lagged variables to each hourly observation starting at 9 a.m. and ending at 4 p.m.

The total size of the data set is therefore 2,082 observations, as it is split between the training and testing data set, as outlined in Figure 6, below.

Data Set	Count Obs	Percent
Test	624	30
Train	1456	70

Figure 6: Training and Test Data Split

Exploratory Data Analysis

The 1st, 5th, 25th, 50th, 75th, 95th, and 99th percentiles, as well as minimum, maximum, mean, variance and percentage missing values are presented in Figure 7 below, providing an overview of the range of values found in the pre-processed data set. PM2.5, SO2, and SOLR each have negative values which may be due to setting the measuring instruments at the Greenwich Eltham air quality site to zero in a contaminated atmosphere. RAIN was particularly high on June 11, 2012 with a max value of 107 mm. Given the 99th percentile value for RAIN is 8 mm, this is a definite outlier; however, the measurement is a valid one that has been recorded by the Met Office ("Record rainfall - April to July 2012", 2012). As described previously, the percentage of missing observations was reduced to zero.

Figure 7 below provides an overview of the range of data values found in the data set.

	min	0.01	0.05	0.25	0.5	0.75	0.95	0.99	max	mean	variance	count	% missing
NO	0.0	0.70	1.2	2.50	4.30	9.12	40.31	94.87	268.0	10.27	353.01	2860	0
NO2	0.5	2.40	5.3	12.00	19.90	32.00	51.60	66.59	79.9	23.30	215.57	2860	0
NOX	2.9	4.80	8.6	16.70	26.80	45.90	107.02	207.18	489.4	39.04	1585.78	2860	0
O3	0.0	0.90	2.0	22.48	41.90	58.40	80.40	96.68	148.2	41.08	601.43	2860	0
PM10	0.5	5.10	7.7	12.47	17.30	25.10	48.23	66.20	96.4	21.09	170.06	2860	0
PM2.5	-4.0	1.96	4.0	7.00	10.50	17.30	37.50	53.68	90.4	14.25	122.86	2860	0
SO2	-7.5	-3.00	-1.3	1.60	4.30	7.90	13.50	16.80	20.4	5.09	20.63	2860	0
BP	959.0	975.00	989.0	1004.00	1011.00	1017.00	1028.00	1034.00	1039.0	1010.01	133.95	2860	0
RAIN	0.0	0.00	0.0	0.00	0.00	0.00	2.00	8.00	107.0	0.42	6.00	2860	0
RHUM	23.0	35.00	43.0	61.00	74.00	88.00	97.00	100.00	101.0	73.12	289.20	2860	0
SOLR	-3.0	0.00	2.0	27.00	62.00	132.00	354.05	612.05	767.0	103.53	14615.27	2860	0
TEMP	-3.0	1.00	3.0	8.00	12.00	15.00	21.00	25.00	29.0	11.84	29.33	2860	0
WDIR	0.0	4.00	27.0	102.00	209.00	254.00	337.00	355.00	359.0	189.66	9155.60	2860	0
WSPD	0.0	0.30	0.4	0.90	1.50	2.20	3.30	4.30	5.2	1.59	0.85	2860	0

Figure 7: Summary Table of Min, Max, Percentiles, Variance, Count, and Percent Missing

Range, Min, Max

A review of the distribution of NO_2 during the period 2012-2013 is represented below in Figures 8, 9, and 10. In Figure 8, we can see that the range of NO_2 concentration level recordings ranges from approximately zero to 80 µg/m3 with a median of approximately 20 µg/m3. Given that the yearly mean limit for NO_2 is 40 µg/m3, this places the Greenwich Eltham site within respectable limits of NO_2, in line with the European Union's standards.

Outliers

Figure 8 also provides evidence that outliers do exist in the data set over the years 2012-2013, where we have available data. The dots in Figure 8 represent these outliers, which represent data points that are one and a half times the interquartile range (IQR) higher than the third quartile (Q3) values. Although these values appear as outliers, given the fact that they are grouped together between 60 and 80 µg/m3, these values appear to be reasonable pollutant level limits that are in line with high concentration levels recorded in other major cities. The

outlier observations here appear to be valid and will be left in the data used in the predictive modeling section.

Figure 8 shows the range of NO_2 recordings as well as potential outliers.

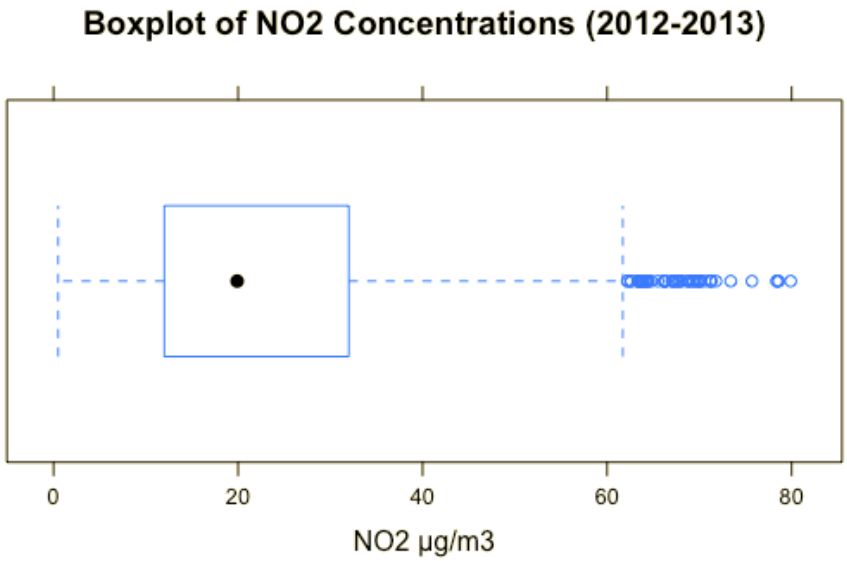

Figure 8: Boxplot of NO_2 Concentrations (2012-2013)

Further evidence that the outliers may be valid can be seen in the boxplot in Figure 9 below, which provides the hourly breakdown of NO_2 concentrations. The outliers falling between 60-80 µg/m3 are found in each hour of the school day, indicating that these observations were recording at different intervals. Only 8 a.m. and 4 p.m. have all data points within one and a half times the IQR.

Figure 9 shows the range of values and potential outliers for NO_2 recordings by hour of day.

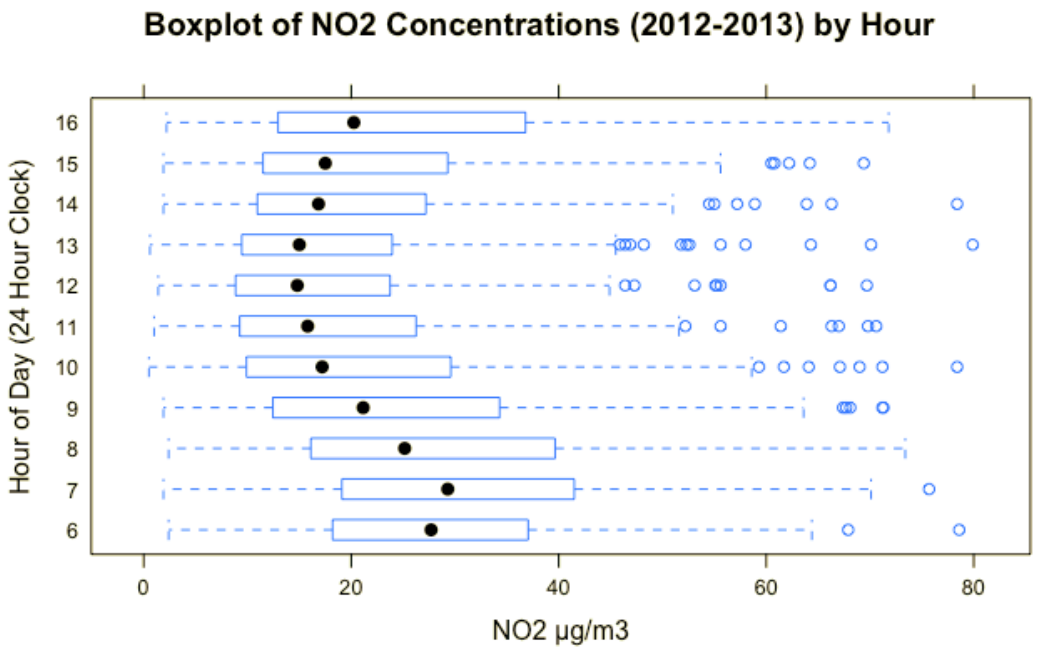

Figure 9: Boxplot of NO_2 Concentrations (2012-2013) by Hour of School Day

Daily Cycle of NO_2

The line chart in Figure 10 below plots the school week (Monday to Friday) data over a six month period in 2013; average hourly concentrations of NO_2 appear to be bi-modal. NO_2 concentrations rise during the early morning commute, peaking somewhere between 6-8 a.m., before declining steadily during 8 a.m. to 1 p.m., where concentrations are generally at a minimum level. Concentration levels then rise again between 1-4 p.m., as the evening commute begins and children return home from school. In addition to the daily cycle of NO_2 concentrations, it is also evident from Figure 10 that the colder months of year have higher average daily levels of NO_2 in comparison to the warmer months. The colder months, which in London is February, November, and December, each have higher average concentration levels of NO_2.

Figure 10 shows the average hourly NO_2 concentrations by hour of day across six months of the year in 2013. Colder months (February, November, and December) record, on average, higher concentrations of NO_2.

Average Hourly School Day NO_2 Concentrations (Six Months in 2013)

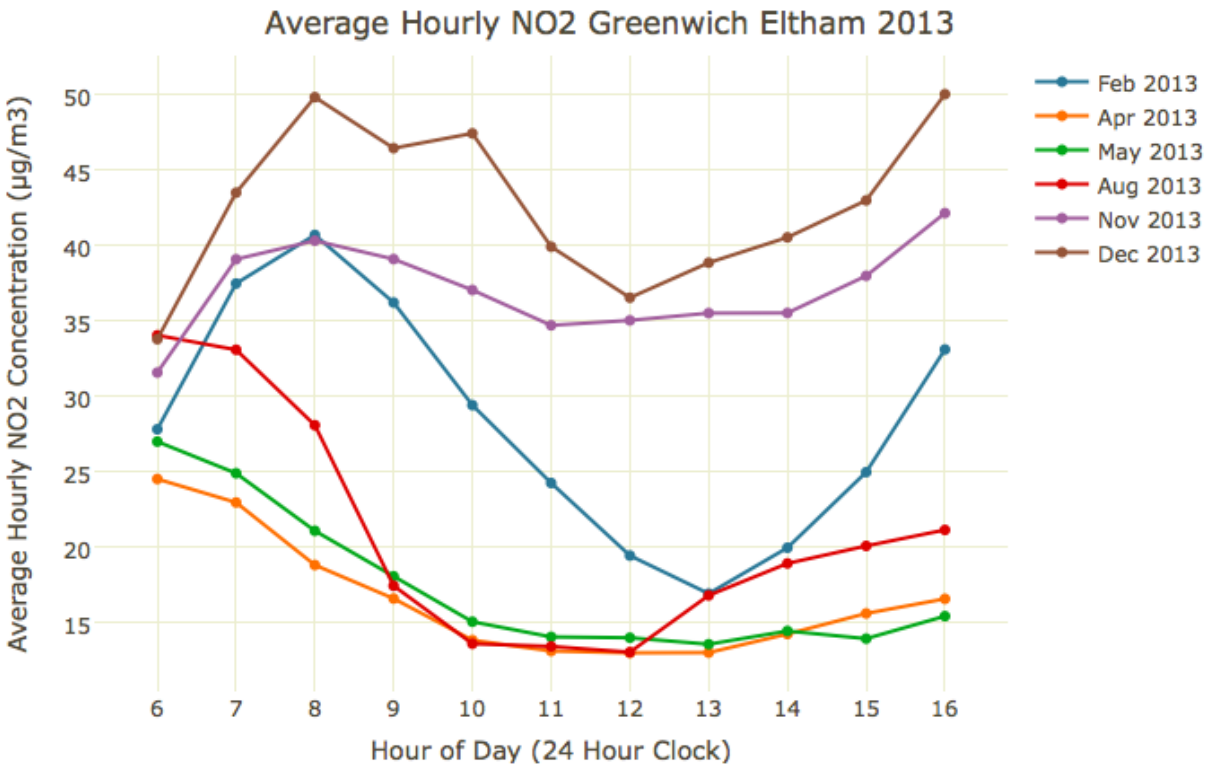

Figure 10: Average Hourly School Day NO_2 Concentrations (Six Months in 2013). Concentration of air pollutant NO_2 is given in micrograms(one-millionth of a gram) per cubic meter air, or µg/m3.

Correlation between NO_2 and the Meteorological Variables: BP, RAIN, RHUM,SOLR, TEMP, WDIR, and WSPD

As can be seen in the Figure 11 below, the NO2 variable observations are weakly positively correlated with barometric pressure (BP) and relative humidity (RHUM) where the Pearson correlation coefficient values are 0.17 and 0.41, respectively. This is indicative that higher levels of NO2 may be expected in higher pressure and humid weather patterns. Weak negative correlations are found between NO2 and solar radiation (SOLR), temperature (TEMP), and wind speed (WSPD) with Pearson correlation coefficients of -0.33, -0.35, and -0.45, respectively. As solar radiation, temperature and wind speed increase, concentrations of NO2 may be expected to decline.

Figure 11 shows the correlation of each meteorological variable with NO2. Values close to zero denote low or no correlation, whereas absolute values close to one denote a strong correlation.

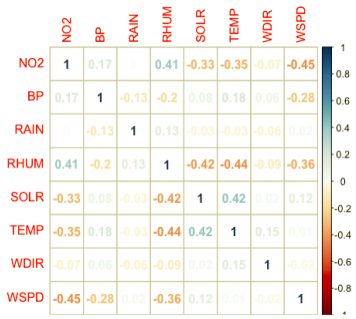

Figure 11: Correlation Matrix of NO_2 Against the Meteorological Variables

Correlation between NO_2 and the Pollutant Variables: NO, NOX, O3, PM10, PM2.5, SO2

In comparison to the correlation coefficients between NO_2 and the meteorological variables above, we can see in Figure 12 below that NO_2 is strongly positively correlated with NO (r = 0.64) and NOX (r = 0.83), weakly positively correlated with PM10 (r=0.42), PM2.5 (r=0.44), and SO2 (r=0.24), and strongly negatively correlated with O3 (r = -0.75). Given the strong correlation that exists between NO_2 and the other pollutants, it is valuable to include these variables in a

predictive model, although it is crucial to understand the increased variance that will be added to the coefficients when multicollinearity is present. Multicollinearity occurs when two or more predictor variables in a multiple linear regression model are highly correlated. Hyndman and Athanasopoulos note that the consequence is that the uncertainty associated with individual regression coefficients will be large and statistical tests (e.g. t-tests) on regression coefficients will be unreliable. However, Hyndman and Athanasopoulos also note that "in forecasting we are rarely interested in such tests...it will not be possible to make accurate statements about the contribution of each separate predictor to the forecast" (Hyndman, R. J., and Athanasopoulos, G., 2013).

Figure 12 shows the correlation of each pollutant variable with NO_2.

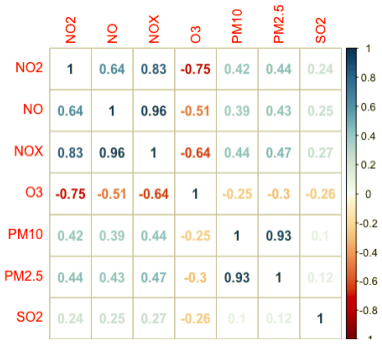

Figure 12: Correlation Matrix of NO_2 Against Other Pollutant Variables

Density Plot of NO_2 against Wind Direction

Southwesterly winds tend to bring lower levels of NO_2. From Figure 13 below, we can see that the green line (indicating a southwesterly wind coming from 180-degrees to 270-degrees in direction) is associated with lower NO_2 concentrations. This is due to the fact that southwesterly winds bring cleaner air from the Atlantic Ocean over London, then across the U.K.. The blue line

representing a northeasterly wind and the pink line representing a southeasterly wind may provide evidence that when winds primarily come from the European continent, NO_2 levels increase.

Figure 13 shows a density plot of NO_2 by wind direction.

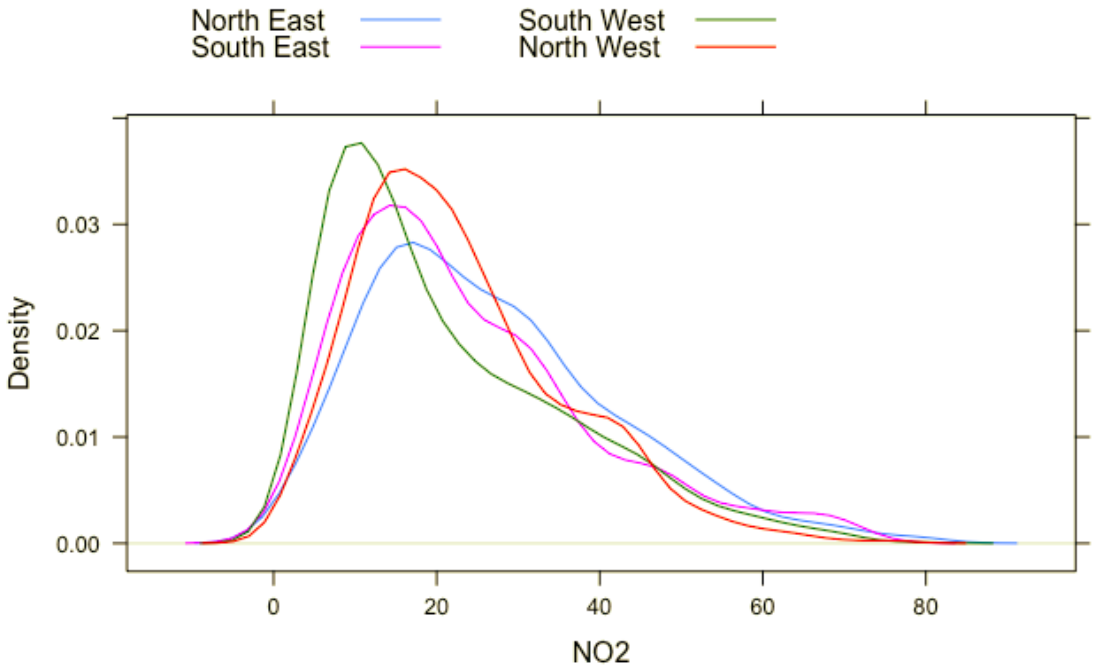

Figure 13: Density Plot of NO_2 Categorized by Wind Direction

Tree Plot

Before the modeling commences, variables were run through a tree plot to determine which predictor variables explain the most variation in NO_2. At the top of the tree, it is evident that NOX divides the data set, which means that this is considered the most important variable in predicting NO_2. Further down the tree, the first lag of NO_2 (lag.1.NO2) begins to emerge. According R's variable importance output summary, variables crucial to the determination of NO_2 levels are: NO, lag.1.NOX, O3, lag.1.O3, lag.2.NO2, and lag.1.NO.

Figure 14 shows a tree plot for determining NO_2 concentration levels.

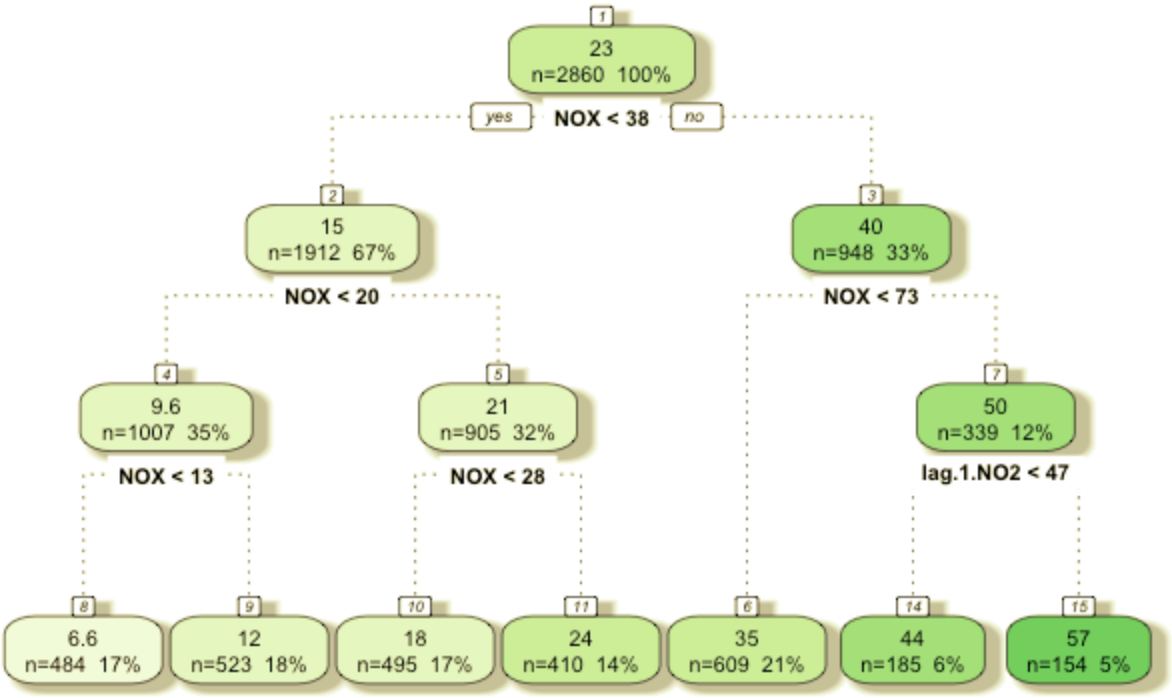

Figure 14: Tree Plot for NO_2 Concentration Prediction

Model Development

Using "a regression model to forecast time series data poses a challenge in that future values of the predictor variable are needed to be input into the estimated model, but these are not known in advance… When using regression models with time series data, we need to distinguish between two different types of forecasts that can be produced, depending on what is assumed to be known when the forecasts are computed" (Hyndman, R. J., and Athanasopoulos, G., 2013). As Hyndman and Athanasopoulos note, a comparative evaluation of *ex ante* and *ex post*

forecasts helps to separate any sources of prediction uncertainty.

Ex ante forecasts are those that are made using the information that is available in advance. Given that school starts at 9 a.m., the ex ante forecasts of NO_2 may be based upon data that is available before the start of school. In the pre-processed data set, this consists of pollutant and meteorological variable values between 6 a.m. and 8 a.m. Using this available data at the time of the forecast, the model produces what Hyndman and Athanasopoulos call a “genuine”, or e*x ante* forecast. *Ex post* forecasts are those that are made using later information on the predictors. The *ex post* forecasts of NO_2 use the actual observations of the pollutant and meteorological variable values for each hour in the school day (9 a.m. to 4 p.m.), once they have been observed; these kinds of forecasts also allow for observations of the performance of the predictive model. Therefore, the *ex post* forecast is not considered a legitimate forecast, since it uses current observations recorded for each hourly observation, which on new, future hourly observation data, are not known in advance.

The *ex ante forecast* can thus be expected to be less accurate than the *ex post* forecast. As future hour predictor variables change, an *ex ante forecast* cannot account for the future changes, which are of course unknown at the time of the forecast, whereas an *ex post* model will take the future changes into account since the *ex post* model includes observations that already exist. The difference between the two models therefore explains the uncertainty in the predictive model. If an *ex post* model performs well compared to the *ex ante* model, it is thus possible to conclude that the *ex post* model has been correctly built to maximize forecasting accuracy.

MODEL 1: Ex Ante Stepwise Variable Selection Model

Figure 15 below shows the *ex ante* stepwise variable selection model's mean absolute error (MAE) results on the training and test data sets. For instance, using lagged data collected at 6 a.m., 7 a.m, and 8 a.m., the forecast accessed by a school administrator at, say, 9 a.m., predicts air quality through to 4 p.m. within that same day. This means that the lag.1 variable represents observations from 8 a.m., the lag.2 variable represents observations from 7 a.m., and lag.3 variables are from 6 a.m..

Model	Formula	Training MAE	Testing MAE
NO2 Stepwise Variable Selection Time 9	lag.1.NO2 + lag.2.NO2 + lag.3.RAIN + lag.3.PM10 + lag.1.RHUM + lag.2.RHUM + lag.2.NOX + lag.3.SO2 + lag.1.O3	3.68	4.66
NO2 Stepwise Variable Selection Time 10	lag.1.NO2 + lag.1.NO + lag.2.NO2 + lag.3.PM10 + lag.1.RHUM + lag.2.RHUM + lag.3.BP + lag.3.SO2 + lag.2.NOX + lag.1.PM2.5 + lag.3.TEMP + lag.1.TEMP	4.95	5.95
NO2 Stepwise Variable Selection Time 11	lag.1.NO2 + lag.2.NO2 + lag.3.PM2.5 + lag.1.RHUM + lag.2.RHUM + lag.3.BP + lag.2.NOX + lag.3.SO2 + Day + lag.3.TEMP + lag.2.TEMP	5.55	6.1
NO2 Stepwise Variable Selection Time 12	lag.1.NO2 + lag.3.BP + lag.3.TEMP + lag.1.TEMP + lag.2.NOX + lag.3.SO2 + lag.3.O3	6.03	6.19
NO2 Stepwise Variable Selection Time 13	lag.1.NOX + lag.1.O3 + lag.3.SO2 + lag.3.BP + lag.1.RHUM + lag.2.RHUM + lag.2.NOX + lag.1.SO2 + lag.1.WDIR	6.95	7.5
NO2 Stepwise Variable Selection Time 14	lag.1.NOX + lag.1.NO2 + lag.2.WDIR + lag.3.RAIN + lag.3.SO2 + lag.3.BP + lag.1.RHUM + lag.2.RHUM	6.88	6.61
NO2 Stepwise Variable Selection Time 15	lag.1.NO2 + Year + lag.2.SO2 + lag.3.TEMP + lag.1.TEMP + lag.3.O3 + lag.1.RAIN + lag.3.WSPD + lag.2.NO + lag.1.SO2 + lag.1.PM10 + lag.2.WDIR	6.63	9.07
NO2 Stepwise Variable Selection Time 16	lag.1.NO2 + lag.1.RAIN + Year + lag.3.TEMP + lag.1.TEMP + lag.2.WDIR + lag.1.PM10	7.7	10.05

Figure 15: Ex Ante Stepwise Variable Selection Model Prediction Accuracy

The MAE increases the further ahead the forecast predicts in the school day: the 9 a.m. MAE is lowest, while the 4 p.m. MAE is highest. The difference between the training and testing MAE also increases similarly as the forecast predicts further ahead of the lagged data.

MODEL 2: *Ex Post* Stepwise Variable Selection Model

Figure 16 shows the *ex post* stepwise variable selection model MAE results on the training and test data sets. The hourly MAE results are based upon all current hour and one to three hour lagged variables. Lag.1 is one hour prior, lag.2 is two hours prior, and lag.3 is the three hour prior value of the pollutant and meteorological variable. Again, the *ex post* forecast is not to be considered a genuine forecast, since at each hour of the day the *ex post* forecast assumes that the current (future) hour values of pollutants and meteorological variables are known.

Model	Formula	Training MAE	Testing MAE
NO2 Stepwise Variable Selection (Ex Post) Time 9	NOX + NO + lag.2.WDIR	0.05	0.05
NO2 Stepwise Variable Selection (Ex Post) Time 10	NOX + NO + PM2.5 + lag.1.RHUM + lag.2.O3 + lag.1.SOLR + lag.3.SOLR + lag.2.SOLR + lag.2.PM10	0.05	0.06
NO2 Stepwise Variable Selection (Ex Post) Time 11	NOX + NO + lag.3.NO2	0.05	0.04
NO2 Stepwise Variable Selection (Ex Post) Time 12	NOX + NO + lag.3.O3 + TEMP	0.05	0.04
NO2 Stepwise Variable	NOX + NO + lag.1.RAIN + lag.1.NO	0.05	0.06

Selection (Ex Post) Time 13			
NO2 Stepwise Variable Selection (Ex Post) Time 14	NOX + NO + WSPD + RHUM + lag.2.SOLR	0.04	0.06
NO2 Stepwise Variable Selection (Ex Post) Time 15	O3 + NOX + NO + WSPD + lag.1.BP + lag.2.BP	0.05	0.05
NO2 Stepwise Variable Selection (Ex Post) Time 16	NOX + NO + lag.3.WDIR + lag.2.SOLR + lag.3.RHUM + lag.1.WDIR	0.05	0.04

Figure 16: Ex Post Stepwise Variable Selection Model Prediction Accuracy

As can be seen from Figure 16, the stepwise variable selection *ex post* model performs very well. The MAE values on both training and test data sets are very low (approx 0.04 to 0.06). Given that the *ex post* model's performance is very good, a genuine ex ante forecast's accuracy can be increased by improving forecasted hourly values of each predictor variable and using them as inputs into the ex post stepwise variable models fitted in Model 2. To forecast the predictors for each hour of the school day, an ARIMA model is used.

MODEL 3: *Ex Post* Stepwise Variable Selection on ARIMA predicted regressors

Figure 17 shows the hybrid *ex post* stepwise variable selection model. This can be considered a true forecast, since it can be performed using data that is available at 9 a.m. in each school day.

Model	Formula	Testing MAE
NO2 Stepwise Variable Selection (Ex Post Arima) Time 9	NOX + NO + lag.2.WDIR	2.59
NO2 Stepwise Variable Selection (Ex Post Arima) Time 10	NOX + NO + PM2.5 + lag.1.RHUM + lag.2.O3 + lag.1.SOLR + lag.3.SOLR + lag.2.SOLR + lag.2.PM10	2.59
NO2 Stepwise Variable Selection (Ex Post Arima) Time 11	NOX + NO + lag.3.NO2	2.58
NO2 Stepwise Variable Selection (Ex Post Arima) Time 12	NOX + NO + lag.3.O3 + TEMP	2.59
NO2 Stepwise Variable Selection (Ex Post Arima) Time 13	NOX + NO + lag.1.RAIN + lag.1.NO	2.59
NO2 Stepwise Variable Selection (Ex Post Arima) Time 14	NOX + NO + WSPD + RHUM + lag.2.SOLR	2.58
NO2 Stepwise Variable Selection (Ex Post Arima) Time 15	O3 + NOX + NO + WSPD + lag.1.BP + lag.2.BP	2.59
NO2 Stepwise Variable Selection (Ex Post Arima) Time 16	NOX + NO + lag.3.WDIR + lag.2.SOLR + lag.3.RHUM + lag.1.WDIR	2.59

Figure 17: Ex Post Forward Variable Selection Model Using ARIMA Forecasted Predictors

Note that the *ex post* forecast is made using ARIMA-predicted dependent variables, and the model used is the *ex post* stepwise variable selection in Model 2, above. By forecasting the school hour predictor variables and using the *ex post* stepwise variable selection model, the accuracy of the *ex ante* forecast (Model 1) can therefore be increased.

RESULTS

Because of its utilization of pre-existing data, the *ex post* forecasted hourly NO_2 concentrations are the most accurate and have the lowest test MAE (0.05) values for each hourly response variable. The *ex ante* model, which uses all available pollutant and meteorological variable observations from 6 a.m. to 8 a.m., is the less reliable model, with an average test MAE of 7.02. By comparison, Model 3, which uses the ARIMA-modeled predictor variable values, performs better than the *ex ante* model having an average test MAE of 2.59. Importantly, Model 3 can be considered a legitimate forecast, since it is taken with data that is available ahead of time. Using Model 3's hybrid approach as a key component of a predictive tool provided to schools would allow administrators and teachers to minimize the amount of time that schoolchildren are exposed to concentrations of NO_2 that surpass current E.U. and U.K. emissions standards.

	RESPONSE HOUR NO2 (MAE)								
MODEL	9:00 AM	10:00 AM	11:00 AM	12:00 AM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	Daily Average
Model 1: stepwise.ex.ante.train.MAE	3.683	4.953	5.554	6.031	6.946	6.88	6.629	7.704	6.05
Model 1: stepwise.ex.ante.test.MAE	4.657	5.954	6.101	6.193	7.499	6.607	9.071	10.045	7.02
Model 2: stepwise.ex.post.train.MAE	0.048	0.051	0.052	0.049	0.047	0.045	0.049	0.046	0.05
Model 2: stepwise.ex.post.test.MAE	0.051	0.064	0.045	0.041	0.055	0.059	0.05	0.043	0.05
Model 3: stepwise.ex.post.arima.test.MAE	2.587	2.59	2.584	2.588	2.588	2.584	2.586	2.59	2.59

Figure 18: Comparison of Model MAE Results

CONCLUSION

Through this research and modeling, schools can demonstrably benefit by being provided with tools that utilize hybrid ARIMA and multiple linear regression models to forecast NO_2 levels during the school day. By forecasting the hourly levels of NO_2 on an *ex ante* and *ex post* basis, it is evident that the challenge in forecasting hourly pollution concentrations each day is almost entirely related to the challenges of forecasting the intraday variation in the levels of other pollutants and meteorological variables. Evidence of this is provided by the fact that the average MAE recorded for the *ex ante* stepwise variable selection model (Model 1) hourly intervals is 7.02, whereas the *ex post* stepwise variable selection model (Model 2) had a much lower average MAE of 0.05 for the hourly responses. Since the *ex post* approach to modeling is not considered a forecast, the use of a third, ARIMA-based model was used to improve the intraday hourly concentration forecast of NO_2, which then gave this third model an MAE of 2.59. Given the fact that the accuracy of forecasting is almost entirely due to the accuracy with which one can forecast the predictor variable pollutant and meteorological variables, it can be concluded that further study may prove to be beneficial. The ARIMA-based model could also be improved by utilizing additional regressors. Rather than simply regressing the response variable on its own lagged values, it is likely that the ARIMA-based model forecasts could be improved by also factoring other variables inputs as regressors to provide a more dynamic output from the ARIMA forecast; the forecasted values for each variable would then be based upon the time trend, as well as other regression variables which may affect each predictor variable.

In this paper, the assumption is made that the predictive tools given to schools have access to pollutant and meteorological variable data points from a measuring station that is either on-site

or at a node site nearby. This will not always be the case. By using nearby stations to impute various levels of pollutants and/or meteorological conditions it will be more complicated to accurately predict response levels of pollutants. This must be an additional area for future study: what is the minimum data set required for a tool to accurately predict intraday pollution concentrations within a school's localized area, and how would such a minimum data set be determined? A thorough cost-benefit analysis is required to determine how prediction accuracy can improve for each variable that the school may need to pay to measure. Given the health risks associated with increased levels of air pollution, it must be recommended that schools adopt ARIMA model-based predictive tools in order to proactively protect the pulmonary health of schoolchildren.

REFERENCES

Mayor's new 'air quality' audits to protect thousands of school kids. (2017). Retrieved from https://www.london.gov.uk/press-releases/mayoral/air-quality-audits-to-protect-school-kids

Rieuwerts, J. (2016). An Air That Kills (1st ed.). CreateSpace Independent Publishing Platform.

Bodkin, H. (2017, April 21). Britain goes without coal for the first time 'since Industrial Revolution' Retrieved from http://www.telegraph.co.uk/news/2017/04/21/britain-set-historic-first-coal-free-day-since-industrial-revolution/

Leake, J. (2015). Schools shut under a cloud of diesel. Thetimes.co.uk. Retrieved 26 May 2017, from https://www.thetimes.co.uk/article/schools-shut-under-a-cloud-of-diesel-vtm27fr0pwx

King, K., & Healy, S. (2013, September 19). Analysing Air Pollution Exposure in London (Rep.). Retrieved from Greater London Authority website: https://www.london.gov.uk/sites/default/files/analysing_air_pollution_exposure_in_london_-_technical_report_-_2013.pdf

Daily Air Quality Index - Defra, UK. Uk-air.defra.gov.uk. Retrieved 26 May 2017, from https://uk-air.defra.gov.uk/air-pollution/daqi

Environmental Research Group, King's College London. (n.d.). Air Quality by Local Authority. Retrieved from https://www.londonair.org.uk/LondonAir/Default.aspx

Walton, H., Dajnak, D., Beevers, S., Williams, M., Watkiss, P., & Hunt, A. (2015, July 14). Understanding the Health Impacts of Air Pollution in London (Rep.). Retrieved from Transport for London and the Greater London Authority website: https://www.london.gov.uk/sites/default/files/hiainlondon_kingsreport_14072015_final.pdf

Binkova, B., Bobak, M., Chatterjee A., et al. (2004) The Effects of Air Pollution on Children’s Health and Development: A Review of the Evidence, World Health Organization, Geneva.

Paliatsos, A. G., & Nastos, P. T. (1999). RELATION BETWEEN AIR POLLUTION EPISODES AND DISCOMFORT INDEX IN THE GREATER ATHENS AREA, GREECE. 1(2), 91-97. Retrieved from https://journal.gnest.org/sites/default/files/Journal%20Papers/Paliatsos.pdf

Lalas, D., Veirs, V., Karras, G., & Kallos, G. (1982). An analysis of the SO2 concentration levels in Athens, Greece. Atmospheric Environment (1967),16(3), 531-544. doi:10.1016/0004-6981(82)90162-7

Perez, P., & Trier, A. (2001). Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile. Atmospheric Environment, 35(10), 1783-1789. doi:10.1016/s1352-2310(00)00288-0

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., & Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence,17(2), 159-167. doi:10.1016/j.engappai.2004.02.002

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2), 153–158.

John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In: Cohen, W., Hirsh, H. (Eds.), The 11th International Conference on Machine Learning. Morgan Kaufman Publishers, San Francisco, CA.

Mckendry, I. G. (2002). Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM10 and PM2.5) Forecasting. Journal of the Air & Waste Management Association,52(9), 1096-1101. doi:10.1080/10473289.2002.10470836

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing,50, 159-175. doi:10.1016/s0925-2312(01)00702-0

Air Quality Standards. (n.d.). Retrieved from http://ec.europa.eu/environment/air/quality/standards.htm

London Air Quality Network: Data Download. Retrieved from https://www.londonair.org.uk/london/asp/datadownload.asp

Daily Chart: Comparing Urban Air Pollution. (2016). Economist.com. Retrieved 26 May 2017, from http://www.economist.com/blogs/graphicdetail/2016/08/daily-chart

Record rainfall - April to July 2012. (2012). Met Office. Retrieved 26 May 2017, from http://www.metoffice.gov.uk/climate/uk/interesting/april-july2012

Hyndman, R. J., & Athanasopoulos, G. (2013). Regression with time series data. Retrieved from https://www.otexts.org/fpp/4/8

Hyndman, R. J., & Athanasopoulos, G. (2013). Selecting predictors. Retrieved from https://www.otexts.org/fpp/5/3

The Global Asthma Report (Rep.). (2014). Retrieved http://www.globalasthmareport.org/resources/Global_Asthma_Report_2014.pdf

Asthma facts and statistics. Asthma UK. Retrieved 26 May 2017, from https://www.asthma.org.uk/about/media/facts-and-statistics/

Air Quality in Lewisham: A Guide For Public Health Professionals. (2012). Retrieved 26 May 2017, from https://www.london.gov.uk/sites/default/files/air_quality_for_public_health_professionsal_-_lb_lewisham.pdf

Carrington, D. (2016). The truth about London's air pollution. The Guardian. Retrieved 26 May 2017, from https://www.theguardian.com/environment/2016/feb/05/the-truth-about-londons-air-pollution

EU Legislation on Passenger Car Type Approval and Emissions Standards. (2017). Retrieved from http://europa.eu/rapid/press-release_MEMO-16-4269_en.htm

Exhaust Emissions Testing. Dft.gov.uk. Retrieved 26 May 2017, from http://www.dft.gov.uk/vca/fcb/exhaust-emissions-testing.asp

How Can We Improve Childhood Asthma?. Kcl.ac.uk. Retrieved 26 May 2017, from http://www.kcl.ac.uk/kingsanswers/leadership/asthma-prevention.aspx

Non-road Mobile Machinery Exhaust Emissions directive. Conformance.co.uk. Retrieved 26 May 2017, from http://www.conformance.co.uk/adirectives/doku.php?id=mobilemachinery

Pollution and the respiratory health of London children. (2009). Qmul.ac.uk. Retrieved 26 May 2017, from http://www.qmul.ac.uk/media/news/items/smd/9833.html

The Future of Vehicle Emissions Testing and Compliance. (2015). Retrieved 26 May 2017, from http://www.theicct.org/sites/default/files/publications/ICCT_future-vehicle-testing_20151123.pdf

APPENDICES

Figure 19 below is a correlation matrix of each variable with its one to five hourly lagged values. Of all the variables, RAIN is the least correlated with its lagged values.

	Lag.1 <dbl>	Lag.2 <dbl>	Lag.3 <dbl>	Lag.4 <dbl>	Lag.5 <dbl>
NO.cor	0.8957486	0.7675851	0.6582850	0.5739301	0.5121494
NO2.cor	0.9350714	0.8360989	0.7370610	0.6513284	0.5836228
NOX.cor	0.9206813	0.8141738	0.7160171	0.6361530	0.5760974
O3.cor	0.9576038	0.8843299	0.8036708	0.7233739	0.6495869
PM10.cor	0.9395877	0.8887391	0.8381202	0.7902201	0.7450368
PM2.5.cor	0.9454446	0.8996108	0.8514144	0.8067862	0.7629247
SO2.cor	0.9452478	0.9151654	0.8891376	0.8678171	0.8460012
BP.cor	0.9983147	0.9954903	0.9911737	0.9855830	0.9788806
RAIN.cor	0.4207296	0.3119566	0.2072233	0.1619816	0.1199173
RHUM.cor	0.9557234	0.8641794	0.7520320	0.6301704	0.5071111
SOLR.cor	0.8521109	0.7291436	0.5782288	0.4224825	0.2772907
TEMP.cor	0.9845286	0.9541677	0.9130026	0.8654970	0.8162530
WDIR.cor	0.7979804	0.7185532	0.6660447	0.6263864	0.5880027
WSPD.cor	0.9311902	0.8575895	0.7892239	0.7239017	0.6616727

Figure 19: Correlation of each variable with one to five hour lagged values

Figure 20 below represents the count of observations from each month in 2010 to 2014 used in final data set. Note that there is no January data in the pre-processed data table due to missing data inputs across the predictor variables.

	2	3	4	5	6	7	8	9	10	11	12
2010	0	0	0	0	0	0	0	32	8	40	0
2012	0	0	128	64	88	0	40	160	184	136	40
2013	80	120	176	184	0	0	40	48	32	80	80
2014	80	80	64	56	0	40	0	0	0	0	0

Figure 20: Count of observations available in each Year and Month (Processed Data Set)