Downtown Toronto Air Quality Prediction

Introduction

Urban air pollution and pollutant concentrations in atmosphere is a complex problem. The variability in atmospheric parameters such as humidity, wind speed, direction of wind and temperature combined with variability in sources generating the pollutants; such as traffic, heating and air conditioning systems, provide a challenge in predicting the levels of air pollution at any given point in time.

However, it is important to be able to improve the prediction performance for air pollutants. Ability to predict the levels of air pollutants will enable the government officials to invoke policies or regulations that controls the damage to human beings and reduce the impact to the climate.

The air pollutant prediction is a time series by nature. In this study, the objective is to compare the predictive performance of three different modelling techniques for time series and effectively predict the pollutant concentrations. The models that being compared are: ARIMA, Support Vector Machine (SVM) and Artificial Neural Network (ANN). Also, this study will attempt at investigating the effects of meteorological parameters such as wind speed, temperature, humidity and wind direction as dependent variables impacting the concentrations of pollutants and discuss the impact of addition of these factors. The models that being proposed for this purpose are: multivariate ARIMA, SVM and ANN.

Literature Review

The analysis of experimental data collected at specific intervals of time introduces a correlation between data collected in adjacent points in time which is called time series analysis (1). In the case of time series analysis, since the observed data are probabilistic in nature, the observed data are modelled as a sequence of random variables (2). Time series are classified into two categories of stationary and non-stationary. In the case of stationary time series, the mean and variance of data are constant while in non-stationary time series, the mean and/or variance of data are changing over time. In the case of non-stationary data, a differencing method is introduced to the data to convert them into stationary and therefore be able to use forecasting techniques. Other techniques such as smoothing is available which are not explored in this analysis (2). In addition, time series exhibit seasonality, noise and trending as well. Seasonality is referred to regular changes in the data patterns through time. Noise the added random component that makes the time series random and unpredictable. The trend is the moving average as time progressed and it demonstrates the changes in the data as increase or decreases.

One approach forecasting time series is Autoregressive, Integrated, Moving Average (ARIMA) model. One advantage of ARIMA model is that it is built on modelling the correlations in the time series data. Another advantage of ARIMA models is the reliance on the prior data in order to build future forecasts and availability of these models in many analytical tools such as R. Some disadvantages of ARIMA model can be expressed as the cost of computation, backward looking and inability to forecast long term (after a while the forecast tends to be a straight line) (3).

Another approach to the time series forecasting is leveraging Support Vector Machine (SVM) models. Traditionally SVM models are used for classification in pattern recognition problems. The SVM also demonstrated success in regression analysis. This application of SVM for solving regression problems have been demonstrated useful in modelling time series as well which has been explored in the research (4, 5, 6). It has been demonstrated that SVM can deal with the high dimensionality of time series and exhibit very good ability to modelling time series. High dimensionality is reference to the fact that time series data not only can be non-stationary, they often do have noise mixed in. The noise in the time series could lead to over-fitting or under-fitting problem which then results in poor performance of the model. A potential solution to this problem is to leverage the tuning parameters of SVM model such as Kernel, Regularization, Gamma and Margin. For the purposes of this study fine tuning kernel function is explored. Other approaches such as complex architecture using mixture of experts neural networks with SVM and Least Square Support Vector Machine (LSSVM) have been studied which provide improved prediction performance. However, these complex models are out of the scope of this study.

Artificial Neural Networks recently have been the focus of a large number of air pollution prediction studies (list of all ref. required here). The principle of neural networks is to mimic the characteristics of a human brain. A neural networks are structured by an input layer, a middle layer which is called the hidden layer and an output. The hidden layer is the layer in which the conversion of input to the desired output will happen (7). ANNs need a considerable amount of training data to be able to provide the output (8). The activation function in ANN is the

Different studies in the area of ANN and air pollutant prediction reviewed and three major approaches to increasing the performance of the models were found: 1) function selection, in some of the studies reviewed the effects of activation function on prediction performance is analyzed. By choosing different functions such as sigmoid, linear or tangent the reviewer studied the performance. 2) Back-propagation, in other studies demonstrated the effects of back-propagation in tuning ANN and increasing performance have and lastly 3) Perceptron architecture (multi-layered vs single layered), in another study the effects of Multi-layered perceptron (MLP) vs. single layer on performance is analyzed. For the purposes of this study, a single layered ANN with linear function is first considered. Since the intent of this study is to analyze the performance between ANN and other types of models (ARIMA vs SVM), this simplified approach is selected (9, 10, 11).

Dataset

The data used for the project are extracted from two sources:

1) Ontario Ministry of the Environment, Conservation and Parks for pollutant data (http://www.airqualityontario.com/history/index.php)

2) Government of Canada Environment and Natural Resources for the meteorological parameters (https://climate.weather.gc.ca/historical_data/search_historic_data_e.html)

The real time monitoring of pollutants data is carried out in a weather station located at Bay Street/Wellesley Street downtown Toronto. This location is selected specifically for it cyclical traffic during rush hours and being densely populated section of Toronto during the day time. The measured pollutants of interest are: CO, NO, NO2, NOX, O3 and PM2.5.

The meteorological data were selected from the nearest and best possible option available location source in downtown city of Toronto. The meteorological parameters obtained include: air temperature, relative humidity, wind speed and wind direction. Some additional parameters such as windchill and others are excluded from this study for simplicity and relevance purposes.

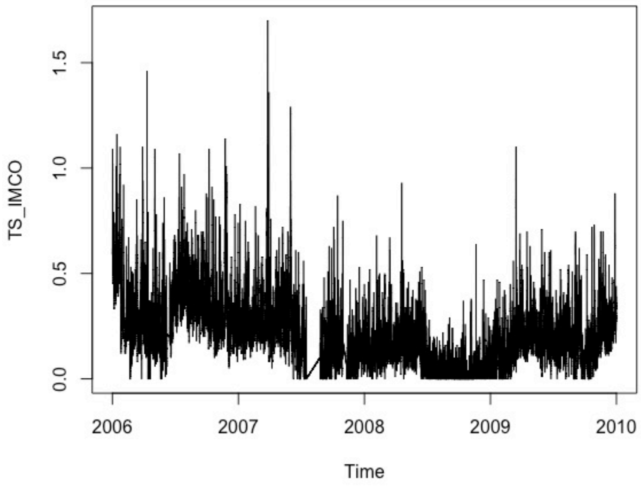
List of Pollutant Parameters:

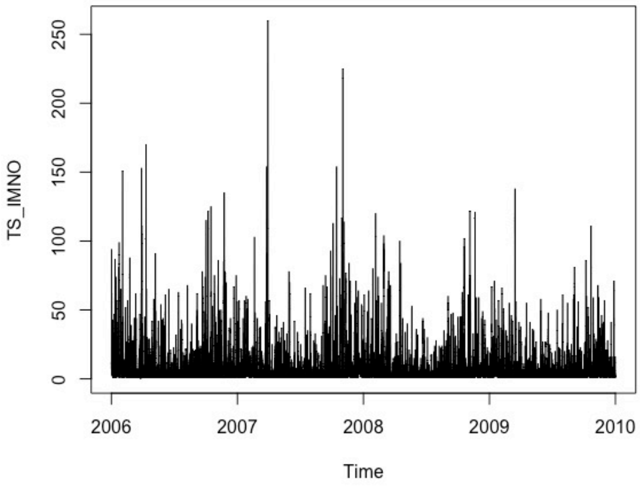
- Carbon Monoxide (CO) is a tasteless, colourless gas which is toxic in high concentration levels for animals and humans (source wikipedia). It is measured in parts per million (ppm), -999 for missing data and 9999 for invalid data is used in the dataset. Data is collected every day of the year on hourly basis (24 hours).
- Nitrogen Oxide (NO) is an irritant gas that in high concentrations causes air way irritations in humans. It is measured in parts per billion (ppb), -999 for missing data and 9999 for invalid data is used in the dataset. Data is collected every day of the year on hourly basis (24 hours).
- Nitrogen Dioxide (NO2) is a pollutant that is primarily produced due to burning fuels and it has harmful effects on human's respiratory system. It is measured in parts per billion (ppb), 999 for missing data and 9999 for invalid data is used in the dataset. Data is collected every day of the year on hourly basis (24 hours).
- Nitrogen Oxides (NOX) is a group of seven nitrogen based gases which are combined called NOX. NOX is the main source of acid rain and smog creation. It is measured in parts per billion (ppb), 999 for missing data and 9999 for invalid data is used in the dataset. Data is collected every day of the year on hourly basis (24 hours).
- Ozone (O3) is a pale blue gas with pungent smell and is found in low concentrations at lower levels of atmosphere. Ozone is a secondary pollutant that is produced when nitrogen oxides and volatile organic compounds react in the sunlight. It is measured in parts per billion (ppb), 999 for missing data and 9999 for invalid data is used in the dataset. Data is collected every day of the year on hourly basis (24 hours).
- Fine Particulate Matter (Fine Particulate Matter PM2.5) denoted as PM2.5 is respirable fine particles that are breathable. Its primary source of production is burning fuels from cars, power plants and airplanes. It is measured in micrograms per cubic metre (μg/m3), 999 for missing data and 9999 for invalid data is used in the dataset. Data is collected every day of the year on hourly basis (24 hours).

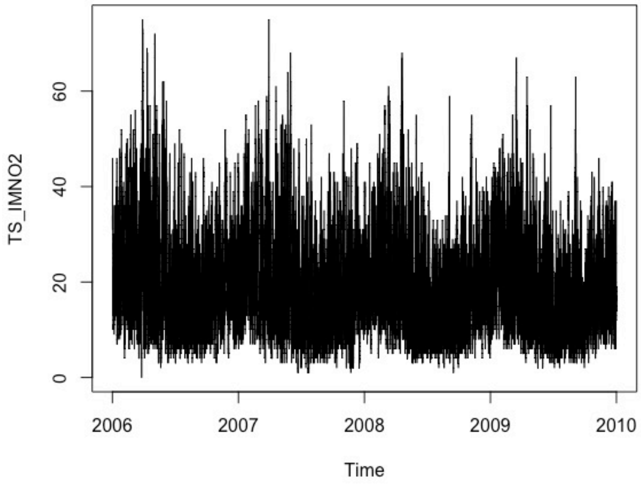
List of Meteorological Parameters

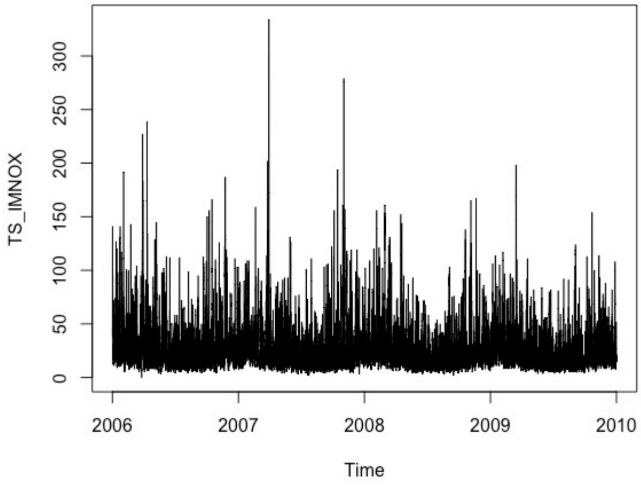
- Temperature, the temperature of the air in degrees Celsius (°C)
- Dew point, the dew point temperature in degrees Celsius (°C). This parameter is removed from the analysis for simplicity.
- Relative humidity, relative humidity in percent (%) is the ratio of the quantity of water vapour the air contains compared to the maximum amount it can hold at that particular temperature
- Wind direction (10's deg/tens of degrees), the direction (true or geographic, not magnetic) from which the wind blows.
- Wind speed, the speed of motion of air in kilometres per hour (km/h) usually observed at 10 metres above the ground.
- Visibility in kilometres (km) is the distance at which objects of suitable size can be seen and identified. This parameter is deemed unnecessary for the purposes of the analysis and therefore removed from the data set.
- Station pressure (kPa), the atmospheric pressure in kilopascals (kPa) at the station elevation. This parameter is removed from the analysis for simplicity.
- Humidex, is an index to indicate how hot or humid the weather feels to the average person. This parameter is removed from the analysis for simplicity.
- Wind Chill, is an index to indicate how cold the weather feels to the average person. This parameter is deemed unnecessary for the purposes of the analysis and therefore removed from the data set.
- Weather, observations of atmospheric phenomenon including the occurrence of weather and obstructions to vision have been taken at many hourly reporting stations. This parameter is removed from the analysis for simplicity.

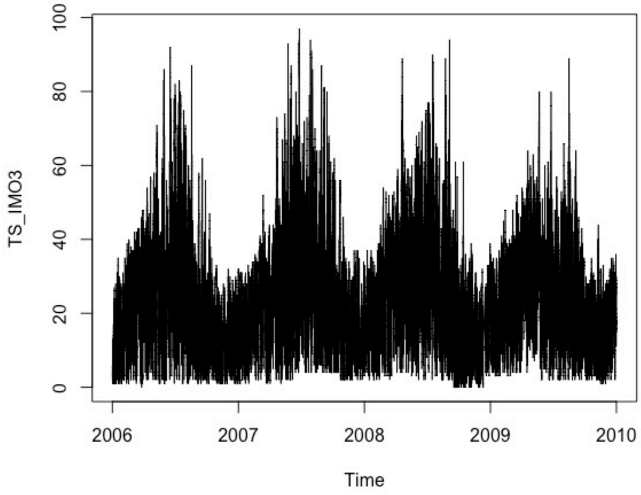
Below graphs are visualizations of the pollutants and meteorological parameters used in this analysis. All these data are in time series format. For details on how the data are obtained, assessed and cleaned please refer to: https://github.com/arshisal/Capstone-Project

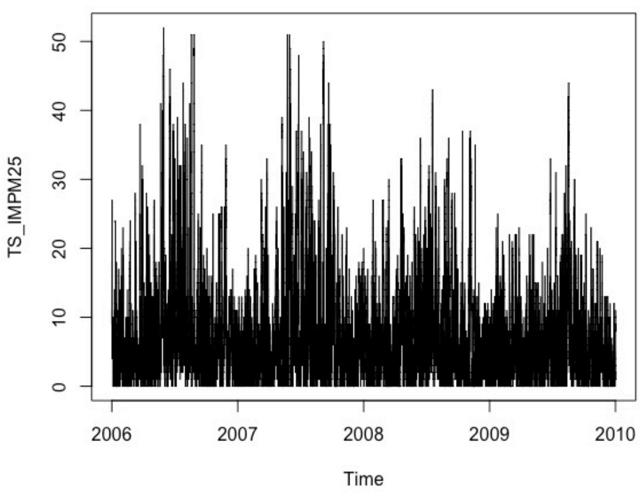


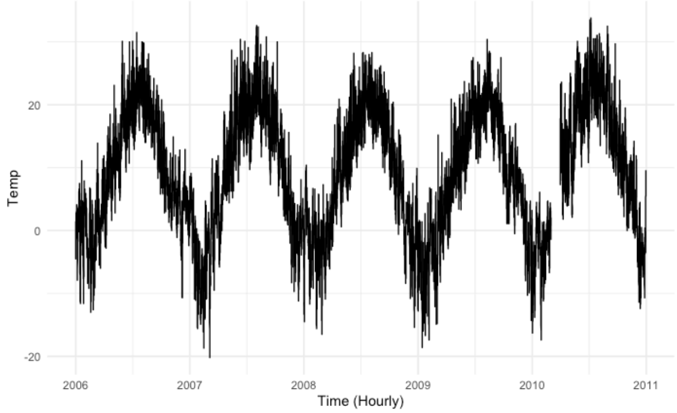


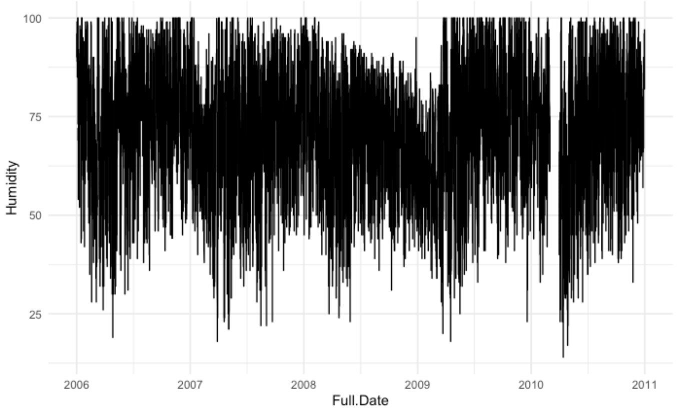


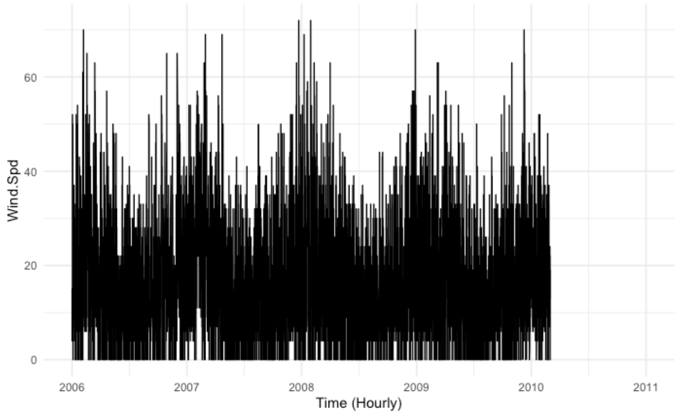


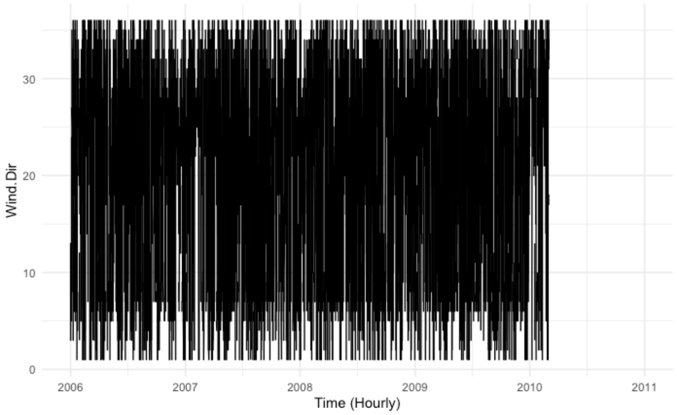












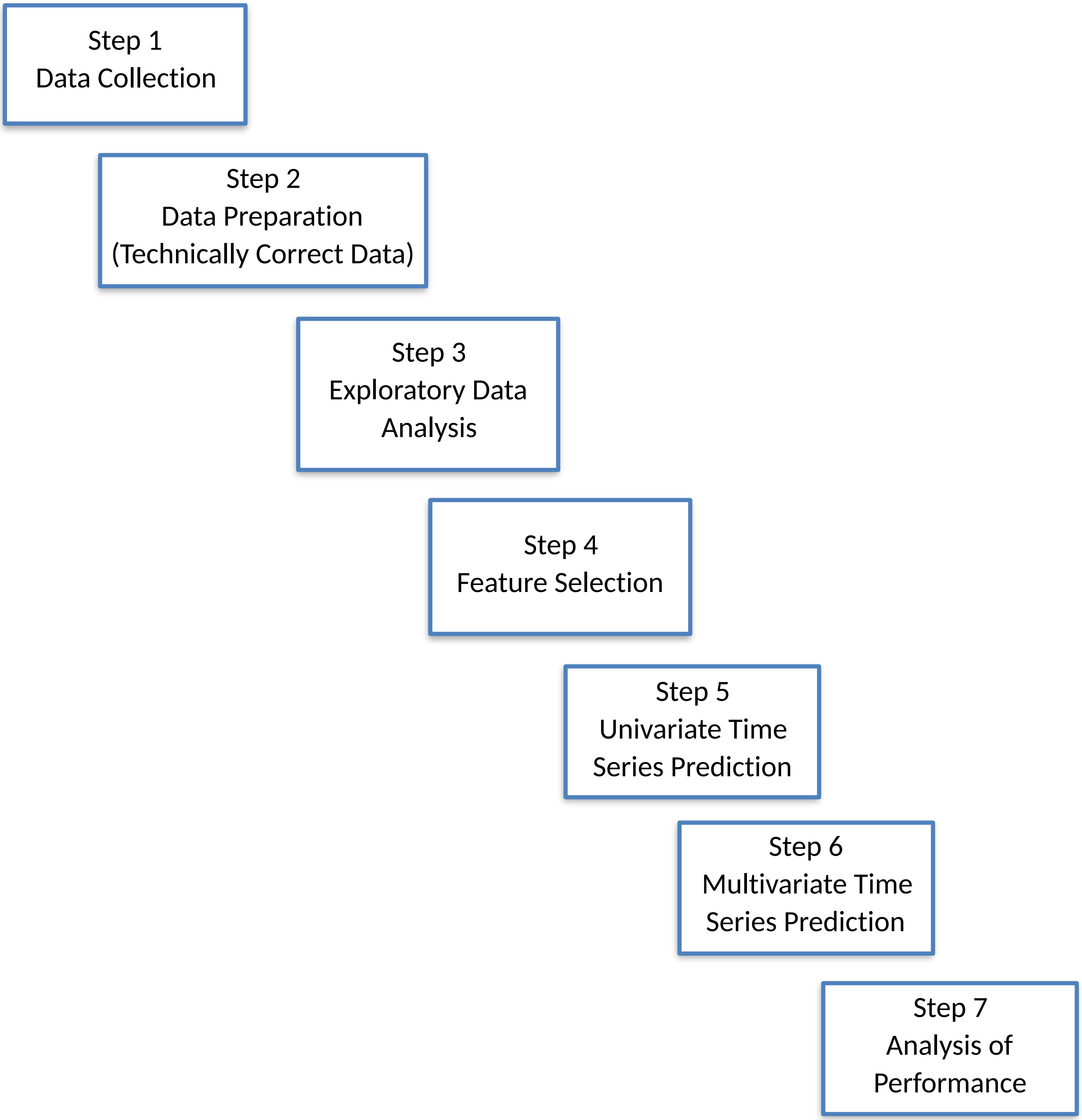
Assumptions

In this study certain assumptions have been made that are worth clarifying. The general assumption in this study is that all factors such as geography, chemical interactions, solar radiation effects on gaseous material are disregarded for simplicity. There are further more detailed models available that include such factors and are out of scope of this study.

Also any natural source of pollution should be taken into consideration when modelling air pollution such as wildfire, and volcanic activities. For example, forest fires and wind combined can increase the levels of CO and PM2.5. For the purposes of this study, neither of these factors are considered due to complexity and unavailability of data. Other concerns are the processing capabilities of machines used. For the purposes of this study, the analysis and modelling is organized in such a way that leveraging a laptop computer would be sufficient. Other complex models will require GPU and specialized processing power which are not considered for this study.

Approach

The approach for this study is as following:



Step 1: Data Collection

In this step, raw data is downloaded from their respective websites and combined together to create the data set. You can refer to the Github for details of how the raw data is converted to proper CSV files.

Step 2: Data Preparation

In this step, data is examined to verify the missing data, their distribution and impute missing data. In this steps also data is converted to time series for further analysis.

Step 3: Exploratory Data Analysis

In this step, and univariate and multi-variate correlations of time series are examined. Further data visualizations are added to enhance the understanding of the data and features. In this step also training and testing data sets are built.

Step 4: Feature Selection

In this step, PCA for time series is examined to assess the best feature for multivariate time series modelling.

Step 5: Univariate Time Series Prediction

In this step, each of the pollutants parameters sets are separately analyzed and predictive models are built.

Step 6: Multivariate Time Series Prediction

In this step, the multivariate prediction models for pollutant parameters are built and analyzed.

Step 7: Analysis of Performance

Upon completion of the predictions, performance and error for each of the univariate models and multivariate models are analyzed. The goal is to assess whether the inclusion of the atmospheric parameters as dependent variables improves the prediction.

References

1) Cryer, J.D. and Chan, K.S. (2008) Time Series Analysis: With Applications in R. Springer. Science + Business Media, London.

2) Davies, Robert, Coole, Tim, Osipyw David (2014) The Application of Time Series Modelling and Monte Carlo Simulation: Forecasting Volatile Inventory Requirements. Applied Mathematics, 2014, 5, 1152-1168

3) Gocheva-Ilieva, Snezhana Georgieva, Ivanov, Atanas Valen, Voynikova, Desislava Stoyanova, Boyadzhiev, Doychin Todorov (2013) Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. Stock Environ Res Risk Assess (2014) 28:1045-1060 DOI 10.1007/s00477-013-0800-4

4) Cao, L. (2003). Support vector machines experts for time series forecasting. Neurocomputing, 51, 321–339. doi: 10.1016/s0925-2312(02)00577-5

5) Sapanevych, Nicholas I., Sankar, Ravi (2009) Time Series Prediction Using Support Vector Machines: A Survey, IEEE Computational Intelligence Magazine, 1556-603X/09

6) Thissen et al. (2003) Using Support Vector Machines for Time Series Prediction. Chemometrics and Intelligence Laboratory Systems, vol. 69, no. 1-2, Nov 2003, pp. 35-49, 10.1016/s0169-7439(03)00111-4.

7) Ciaburro, Giuseppe, Venkateswaran, Balaji, (2017) Neural Networks with R, Smart models using CNN, RNN, deep learning, and artificial intelligence principles, Packt

8) Arhami, Mohammad, Kamali, Nima, Rajabi, Mohammad Mahdi, (2013) Prediction Hourly Air Pollutant Levels Using Aritificial Neural Networks Coupled with Uncertainty Analysis by Monte Carlo Simulations, Environ Sci Pollut Res (2013) 20:4777-4789, DOI 10.1007/s11356-012-1415-6

9) Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., & Shao, D. (2004, October 18). Progress in developing an ANN model for air pollution index forecast. Retrieved from https://www.sciencedirect.com/science/article/pii/S1352231004008118

10) Graupe, D. (2013). Principles of artificial neural networks. New Jersey: World Scientific.

11) Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007). Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. Atmospheric Environment, 41(14), 2967–2995. doi: 10.1016/j.atmosenv.2006.12.013