National College of Ireland

Forecasting of Delhi Air Pollution With The Help of Performance Evaluation of Advanced Time Series Models

MSc Research Project
Data Analytics

MOHAMMAD SHAHID MEMON
X16150775

School of Computing
National College of Ireland

Supervisor: Mr. Vikas Tomer

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	MOHAMMAD SHAHID MEMON
Student ID:	X16150775
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Mr. Vikas Tomer
Submission Due Date:	13/08/2018
Project Title:	Forecasting of Delhi Air Pollution With The Help of Performance Evaluation of Advanced Time Series Models
Word Count:	7232

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting of Delhi Air Pollution With The Help of Performance Evaluation of Advanced Time Series Models

MOHAMMAD SHAHID MEMON
X16150775
MSc Research Project in Data Analytics

12th August 2018

Abstract

Air quality prediction is a hot topic for many researchers due to its effect on health. Researchers from every corner of the world are trying to develop the better model to predict the air quality. Delhi, the capital of India has been listed the most polluted cities on the earth by WHO,2014 which is a serious concern for the nation. This research evaluates the forecasting accuracy of the PM2.5 and PM10 fine particulate matter concentration in air by advanced time-series models. However, In Delhi air pollution, very limited work had been done in past. This research is accomplished using advanced time series models such as TBATS, ARIMA (Auto Regression Integrated Moving Averages), Simple Exponential Smoothing, Holt model and Neural Network. While this is the first research on forecasting the Delhi air pollutants such as PM10 and PM2.5 using TBATS, and Feedforward Neural Network. Throughout the research, it was found that neural network with feedforward single layer performing better than any other applied model but its execution time is very large, so it is only appropriate for short-term data whereas for long-term data ARIMA is the best model.

1 Introduction

In recent years, air pollution has become an increasingly prominent problem in the world. According to the World Health Organization (WHO) on 12th May 2016, More than 80 % of the worlds people are suffering from air pollution especially the people who live in urban areas where air quality levels exceed the WHO limits(Wang and Zhao; 2018). According to World Economic Forum, in 2018, Indian cities such as New Delhi, Varanasi and Patna listed in most polluted cities in the world based on the amount of particulate matter less than 2.5 micrograms found in every cubic meter of air, for this research data has been collected from WHOs Database of more than 4300 cities in the world (*Pollutedcities*; 2018). Due to rapid economic development, industrialization, and urbanization, as well as the energy consumption and exponentially increased use of vehicles, air pollution has become a serious environmental issue in India over the few decades(Yang et al.; 2018). The Governments and citizens have expressed increasing concern regarding air pollution because of it affects human health and sustainable development worldwide. The situation

in developed countries is better, as they have already paid a heavy price to control their pollution level, and they are continuously investing money to maintain the same(Wang and Zhao; 2018). Also, the world together is working towards saving the environment and the best example is **"The Paris Agreement"** (new global climate change agreement).

So, for this research, we have considered Delhi as a research area and data is sourced from Indian government website [1]. This study is not only used for forecasting the air pollution but also compare the different time series models like TBATS, ARIMA (Auto Regression Integrated Moving Averages), Simple Exponential Smoothing, Holt model and Feedforward Neural Network. While this is the first research on using TBATS, and Feedforward Neural Network models for forecasting the Delhi air pollution.

This paper is outlined in the following way. First, we discuss the Motivation and Domain Overview followed by the project specification which includes research question,and the purpose of the study. Section 2 covers related work that cover the research that has been done so far in air pollution and why our research is relevant and original. Literature covers both time series analysis and other prediction models used in this area. Section 3 covers the proposed methodology used for the project. In section 4 implementation of different time series models on Delhi air pollution. Section 5 covers evaluation in which model comparison and validation is covered and section 6 concludes the paper with reference to future work in air pollution prediction.

1.1 Motivation

Air pollution has become a serious concern in many developing countries, especially in India, and could generate adverse effects on human beings. The Delhi, capital of India, has been suffering from air pollution issues over the last decades. In Delhi, concentration of PM2.5 and PM10[2] air pollutants, has fifteen timers higher than defined guidelines of the World Health Organization (WHO). In Delhi, asthma cases have been increased by 900 percent during December 1998 to December 1999 according to All India Institute of Medical Sciences (AIIMS)(Kumar and Goyal; 2011). We could avoid 7490 deaths in Delhi by a 141.6 g m3(Using standard US metric) reduction in PM10 according to "Brandon and Hommann (1995)". The situation in Delhi is becoming worse day by day, one out of ten schools in Delhi is suffering from asthma, and this situation will become more dangerous, if we couldn't come up with an alternative way. And we can only find out the alternative way when we have Air quality early-warning systems. This is the motivation behind this research topic and provides better prediction mechanism of air pollution that helps to mitigate health-related issues. So, this topic is well worth to investigate.

1.2 Domain Overview

1.2.1 Air Quality Forecasting

Air quality forecasting is one of the most challenging real-world application domains for machine learning algorithms(Tzima et al.; 2011). Many scientists focus on this hot topic. The nature of the problem with its non-regular, non-linearity and good quality of data. This research effort deals with the particulate matter PM10, and PM2.5 air pollutants, and data has been sourced from Indian government website[3] and we have decided to

[1]Data Source- Government of India website: `http://cpcb.nic.in/`

[2]PM- Particulate Matter

[3]Data Source- Government of India website: `http://cpcb.nic.in/`

choose Anand Vihar, New Delhi DPCC station and data that is used for this research is every hour from 01/01/2016 to 25/02/2018.

1.2.2 Stationary of Data

According to (Wei; 2006), Purpose of time series analysis can be defined as observation over time. Time series data can be decomposed in four patterns such as Trends (long-term direction),Cyclic/Irregular (unsystematic, short-term fluctuation), Seasonal (calendar based movements) and Regular (systematic). Furthermore, Data is said to be stationary if it's statistical properties such as mean, variance, and autocorrelation etc are constant(*Stationarity and differencing*; 2018) and it should be said non- stationary if statistical properties of data are non-constant. Non-stationary data affects the forecasting ability of the model, so data should be checked whether it is stationary or non-stationary. If data is non- stationary then the series will need to be differenced until the data are stationary. The Air quality data was basically found as non-stationary data as p-value from Box test is below 0.05 which shows it is not a white noise. Therefore, in forecasting of Air quality models like TBATS, ARIMA (Auto Regression Integrated Moving Averages), Simple Exponential Smoothing, Holt model and Single Layer Feedforward Neural Network would be a better choice.

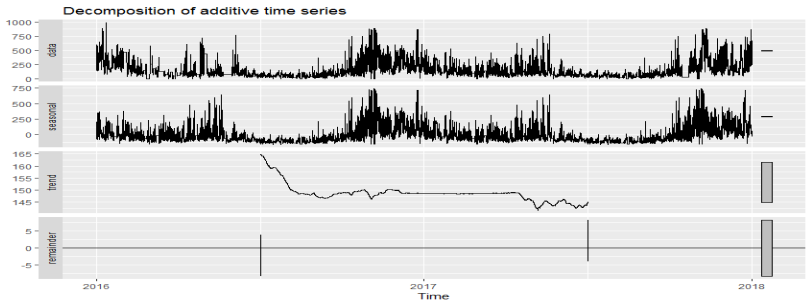

Figure 1: Decomposition of PM2.5

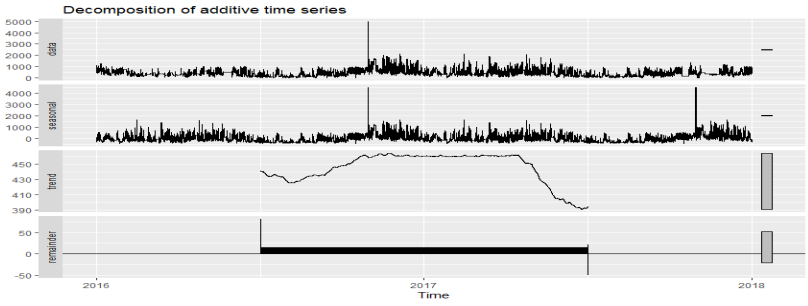

Figure 2: Decomposition of PM10

1.3 Project Requirement / Specification

1.3.1 Research Question

This research project addressed the Business Question as follows:
Research Objective 1: To what extent, the performance of PM2.5 and PM10 air pollutants

in Delhi can be predicted using advanced time series models?
Research Objective 2: To evaluate the performance of advanced forecasting models.

1.3.2 Purpose

Purpose of this research is to predict the PM2.5 and PM10 air pollutants in Delhi using advanced time series models. Today, Delhi is facing a serious issue of air pollution due to its effect on the human health as well as the environment. According to WHO, air pollution is responsible for many health-related issues such as skin and eye infection, irritation of the nose, throat, and eyes. So, it is crucial to continuous monitoring and analyzes of air pollution levels. Appropriate plans can be devised to control the air pollution when they are exceeding the levels of pollutants(Gore and Deshpande; 2017). So, predicting air pollution is an important prerequisite for monitoring, estimating, and mapping unknown pollution values which can be done by with our research.

2 Related Work

Air pollution issues have attracted worldwide researchers attention because of it globally effects on human health, especially in developing countries like China and India. According to the WHO report, Air pollution is one of the single largest health risks(Wendel; 2014). A lot of academic work and many great pieces of research in the past have been done on the prediction of air pollution. So, the prediction is implying to forecast the future based on historical data. Mainly statistical method and machine learning algorithms have been used for forecasting. In this section, we will critically review the past related work and would provide the brief comparison between them along with the providing necessary opinion on the same. To provide a better understanding of past work, we have divided this section into followings:

- Literature Based on Health Concern by Air Pollution.
- Literature Based on Machine Learning.

2.1 Literature Based on Health Concern by Air Pollution

Air pollution is a major contribution to several health-related issues. In 2014, in the United Nations Environment Program (UNEP) it stated that there is an urgent need to reduce the air pollution level globally as in most of the cities in the world air pollution level do not meet the WHO guidelines for acceptable pollutant levels especially in developing countries like India(Fotopoulou et al.; 2016). According to OECD (Organization for Economic Co-Operation and Development), in 2012, globally 3.7 million people died because of air pollution, that has been increased from 2010 by 3 million (Fotopoulou et al.; 2016). To show the relationship between air pollution and lung cancer (Yoon et al.; 2016) have used the time series method, they have focused on PM10 and PM2.5 air pollutant in the US. Their result confirmed that the high exposure of PM10 and PM2.5 adversely influences lung cancer risk. In India, especially in Delhi, the air pollution is a major concern for people. In Delhi, since the early 1990s air pollution levels have exceeded most other cities in developing countries(Nagpure et al.; 2014). To understand the air quality impacts on heath, in 2014,(Nagpure et al.; 2014) have done their research in Delhi

Air Quality; they have divided Delhi into different regions, their studied covered data from year 1991- 2000. Throughout their research, they have found that air pollution and because of that health risk has been increased by 100 % in year 2000 compared to 1991. A limitation of their study is the lack of analysis on air pollutants and other environmental factors which could cause responsible for poor air quality (Nagpure et al.; 2014). That has been proven by(Sindhwani and Goyal; 2014), In their study they found that there has been a tremendous increases number of vehicles in Delhi during year 2000-2010 which are causing a high level of air pollution. Their study concluded that implementation of CNG and phasing out of old vehicles from Delhi road helped to reduce the air pollution. Delhi government continuously trying to adopt new ideas, for instance, Odd- even traffic rule implementation during winter in year 2016 but somehow this implementation doesn't work in Delhi and didn't help to reduce the traffic emission in Delhi (Chandra et al.; 2018).

2.2 Literature Based on Machine Learning

Machine learning is an effective way to analyze and predicts the output based on historical data. As it is very popular among researchers so, many researchers have their own version of it's definition. Developing countries like China and India are more concern by many researchers. In 2008,(Jiang et al.; 2008), have used BP neural network (BPNN) to predict the air quality based on a rough set theory for that they obtained data from "The Urban air quality and the key air pollution emission source monitoring system of Jinan (city of China)", they utilized a rough set theory to reduce the monitoring data of the pollution. Top of that they created a topological structure of the multilayer to defined rules that helped to eliminate the redundancy. Finally, they trained their model using the BP arithmetic and compare their model to general neural network prediction in which their model performs better than a general neural network. Their study was appreciable especially the extraction rules that they defined for their model, but (Jiang et al.; 2008) have not talked about execution time, which is usually large for training data. To minimize this problem (Wang et al.; 2009) have used multilayer back propagation neural network (BPNN), and they have trained their model using particle swarm optimization (PSO) that helps BP to achieve the high accuracy for forecasting. They had tried to be optimized the long training time that (Jiang et al.; 2008) has not considered in their study. For that, they combined BP with rough- set. Also, they achieved rule extraction and decision table reduction using a different method such as weighted set method (basal method).

Similarly,(Chan and Jian; 2013)have used neural network based knowledge discovery system that helps to overcome the limitation of ANN and help to predict the PM10 and PM2.5 air pollutant level. Their system consists of two layers, a) prediction of air pollution based on relevant air pollution factors such as traffic, air temperature, wind speed and relative humidity b)To extract the explicit knowledge from ANN. They have compared their model performance to linear regression, where an applied model is better than regression but they have not used the feature engineering process that raised a question of model accuracy.(Kurt and Oktay; 2010)have also used a neural network to forecast the air pollutants such as SO2, CO, PM10 levels three days in advance for Besiktas district in Turkey. They also considered geographical factors as well in their study that helps to enhance the accuracy of forecasting.

In 2016,(Li et al.; 2016), have analysed air quality level in Beijing, using spatiotem-

poral deep learning (STDL) method in which they used the stacked auto encoder to extract inherent air quality features. SAE is used greedy layer approach to train the data that is bit time- consuming process. They have used RMSE, MAE and MAPE statistical method to evaluate their model performance. They have also implemented STANN (spatiotemporal artificial neural network), ARMA (regression moving average) and SVR(support vector regression) model on the same data set to compare their performance with the STDL model. Overall, the performance of STDL is better than other applied models. But in their work also they have not mentioned about the execution time of each model similar to (Jiang et al.; 2008).

As stated by the WHO, PM2.5 air pollutant is more harmful than other pollutant that causes many healths related issues that motivates (Jiang et al.; 2017) to analyze and predict the PM2.5 in China. They have considered Jing-Jin-Ji and Pearl River area for their research, they have built a hybrid model name called HML- AFNN. It is a combination of High Dimension Association Rule (HDAR), Learning Vector Quantization (LVQ) and Adaptive Fuzzy Neural network (AFNN). They have compared their model with plain AFNN through MAE, MAPE and Band Error, and they found that HML-AFNN model is providing more accurate results than plain AFNN, their study is intense and their motivation also well defined behind their research. In 2017, a grateful effort has been shown by(Liu et al.; 2017), they studied 3 areas of China name as Beijing-Tianjin-Shijiazhuang to analyse the AQI, they have also considered input feature as the weather. They have used Multi-dimensional collaborative SVR model, through 4– fold cross validation they have estimate the error of the model. To check the performance of their model they have also used MSE, RMSE and MA like (Gorai and Mitra; 2017) but to make their model more accurate they tested their model on different -different data sets based on different- different weather conditions. The MAPE for all the cases falls between 0.05 to 0.09 which is quite good, also the RMSE value is almost same for training and testing (75 % training and 25% testing) data set i.e. less than 12 that shows their model is not face the problem of overfitting the data.

Some researchers focused on the other air pollutants as well like Ozone, NOx etc.(Hasham et al.; 2004) have used ANN to analyse and forecast the NOx concentration hourly basis using 4– layer back propagation network. They have considered the industry stack emission rates, meteorological data and traffic count as input variables. Throughout their research they received resulting model fit (Coefficient of determination R2 = 0.63) and precision of model prediction (RMSE= 1.8 * 10-3 ppm as NO2), in their research they have not used any other models so it is not trustable that their model is better than any other models.(Castro et al.; 2008) have developed Interval Type-2 Fuzzy Neural network (IT2FNN) hybrid method that helps to predict the impact of meteorological pollutants on ozone (O3). For meteorological they have considered Wind speed (WSS), wind direction (WDS), temperature (T), humidity, rainfall and solar radiation. In their hybrid model for time series prediction they have used Mackey-Glass differential delay equation. Overall, they found that IT2FNN helped to improve the time series forecasting compared to ANFIS, NN and autoregressive forecasting methods with least RMSE of 5.6. They have also found that in case of non- linear effects IT2FNN doesn't perform better than NN and ANFIS. The only prediction of ozone does not give the overall result of air pollution and ozone is not that much harmful than PM pollutants stated by(Jiang et al.; 2017).

Hybrid systems that combine Artificial Neural Networks with other forecasters have been widely employed for time series forecasting. A great effort done by (de Mattos Neto et al.; 2017), have developed a new forecast method called Nonlinear Com-

bination (NoLiC) for forecasting the PM pollutant data(PM2.5 and PM10) that is the combination of time series forecaster and error series forecaster. In time series, is typically be difficult to ensure linear and non-linear patterns. Their method is mainly for univariate data. To evaluate the performance of model they have used six statistical metrics such as Mean Squared Error (MSE), Prediction of Change in Direction (POCID), U of Theil Statistics (THEIL), Mean Absolute Percentage Error (MAPE), Average Relative Variance (ARV), and Index of Agreement (IA). They have found that their accuracy of the applied model is great, but can be improved further to use Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model in top of ARIMA in future.

In the year 1999, (Salcedo et al.; 1999) have also developed a novel time series method to analyze air pollution of Oporto area ((PR, MT, and ML Sampling sites). They have identified the long-term trend and of cyclical or periodic components. They have used a stepwise approach called SATSA Model to analyze the daily concentration of strong acidity (SA) and black smoke (BS). Before applying the model, they have decomposed the time series to identify if any trends and seasonality available in data or not. In each step, their model is performing a correlation analysis of residuals. That helps to identify the white noise which is a crucial step in time series method. To check the trand of SA and BS, they have performed their model in a different-different time periodic that helps to evaluate their model as well.(Wang and Niu; 2010)have used Seasonal ARIMA (autoregressive integrated moving average) model and MCMC(Markov chain Monte Carlo) method to analyzed the Los Angeles long beach air pollution PM 2.5 for that they have used traffic data from 1997 to 2008 that seems to be a very huge volume of data. They concluded that the applied model Seasonal ARIMA is helpful to predict the air pollution and then compared the ARMA and ARIMA model where ARIMA performed better than ARMA. Their study is not much significant as they have not provided proper justification to choose only seasonal ARIMA over other time series method. (Bhardwaj and Pruthi; 2016) and (Taneja et al.; 2016) have studied and analysed the Delhi air pollution. (Bhardwaj and Pruthi; 2016) have more focused on major air pollutants NO, NO2, CO, SO2, NOx and PM2.5 trends before and after odd-even schema for that they have used statistical approach and time series method whereas (Taneja et al.; 2016) have used linear regression and multilayer perception to predict the air pollution in Delhi. Nonetheless, the significance of their study is questionable as they have not used appropriate time series method and proper justification also not provided for applied method.

2.3 Identified the Gaps

(Bhardwaj and Pruthi; 2016),(Taneja et al.; 2016),(Nagpure et al.; 2014) and (Sindhwani and Goyal; 2014) have done a great amount of work by analyzing and predicting the Delhi air pollution. As most of the authors have used visualization or simple time series model to analyze Delhi air pollution, they should also use advanced time series methods to predict the air pollutants more accurately. To overcome this problem every 60-minute frequency data has considered for the analysis to see the daily change in PM2.5 and PM10 concentration. Advanced Time series models are implemented which are discussed in the implementation section.

2.4 Summary of Related Work

After scrutinizing the related work, the first task is to identify the research variable. Inspired from (Jiang et al.; 2017),(Nagpure et al.; 2014), and (Yoon et al.; 2016), we have chosen PM10 and PM2.5 as research variable. PM2.5 and PM10 play an important role to calculate Indian AQI as well.[4], According to CPCB,
AQI=Max (I1,I2,I3,...,In)
Ip = Concentration of pollutant.
In addition, the literature reviewed in the previous section is the backbone of our project that helped in identifying the appropriate models to predict the air pollution and statistical measure to evaluate them. While this is the first research on using TBATS, and Feedforward Neural Network models for forecasting the Delhi air pollution. However, TBATS has been used in Gold price prediction by (Hassani et al.; 2015). The performance of the model is appreciable which inspired us to use this model for the research.

3 Methodology

For this research, we had used the CRISP-DM [5] Methodology. CRISP- DM is one of the most preferable methodology among professional because it can be altered and modified according to business requirements. For this research, we had followed six hierarchical process of CRISP -DM according to (Brockwell and Davis; 2016). Looking at the business requirement (i.e. short-term air quality prediction) we had come up with our modified version of the CRISP-DM as shown in the figure 3.

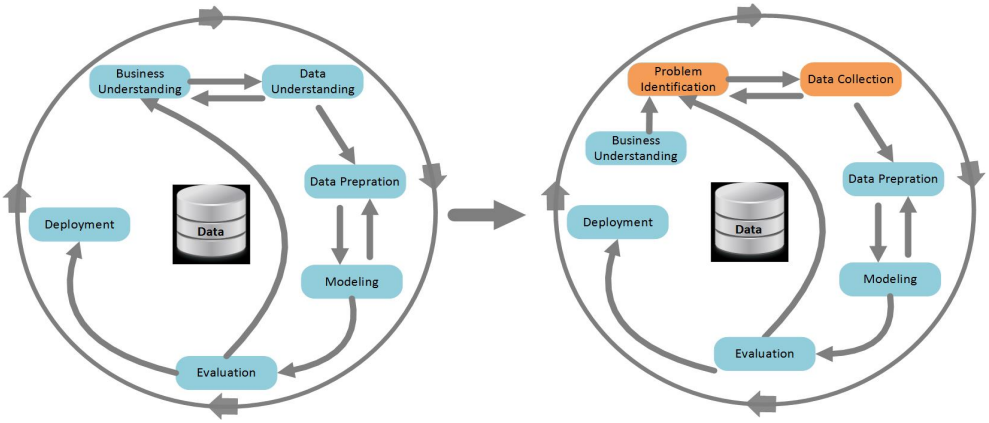

Figure 3: Modified CRISP-DM

3.0.1 Business Understanding

The Research objective, at first, we looked at the research scope. It is important to be aware of the overall goal of the research. As we are predicting the Delhi air pollution (PM2.5 and PM10 air pollutants), the underlying knowledge of methodology that used for air pollution was researched which eventually resulted in the formulation of Data Understanding plan.

[4]AQI- Air Quality Index
[5]CRISP-DM- Cross Industry Standard Process Data Mining

3.0.2 Data Understanding

Data understanding and Business understanding are naturally correlated for that we followed an iterative process. For this research, data is sourced from CPCB[6] India, it is an official website from the government of India. We had chosen Anand Vihar, New Delhi DPCC station data, we had extracted PM2.5 and PM10 hourly basis data from 01-01-2016 to 25-02-2018 to keeping the research objective and problem definition in mind. We had missing values in our data that could risk of our model accuracy, and we could fix this with the method that we will talk about in later steps.

3.0.3 Data Preparation

Data Preparation is an important and longest phase of any research in Machine learning. Data needs to be cleaned from any impurities like missing values or strange characters. Proper data preparation helps to improve the performance of models. For this research, we had used Python under Anaconda Jupyter environment to do data preparation. The reason to use Anaconda environment over standard python environment because:

- Anaconda is a combination of packages and tools which provides the python in it.
- It is providing package management, easy installation of add-ons and other required packages.
- Jupyter notebook provides the functionality of writing code and viewing the output at the same time, also it provides visualization within notebook itself.

Firstly, we had selected only essential variables for the time series analysis, then with the help of forward filling (ffill) function we had deal with the missing data. As in the dataset there were not high fluctuation of PM2.5 and PM10 concentration within a day, thats the reason to choose ffill() function. ffil() functions fills the missing values with previous values, Whereas bfill() function (back word fill) dealing with the missing data based on future values, that is not relevant with our case, so we decided to choose ffill() function over bfill() function. After cleaning all the data, we had checked again if there is any null value present or not using isnull() function. As stated by (Buzzi-Ferraris and Manenti; 2011) outlier detection is one of the major phase in data cleaning to get more the accurate result, so for that, we used box plot to check the outliers and handled. Once data has been clean we had applied various time series model for that we will talk about in later steps.

3.0.4 Data Modeling

After the data preparation, we had applied the time series based forecasting models on a clean dataset. Time series forecasting is not easy, unlike the simpler problem of classification and regression. In time series, forecasting is based on previously observed values. To handling the missing data with the appropriate approach is required when fitting and evaluating models. It also supports in modeling, providing the additional structure like seasonality and trends that can be helped to improve the model performance. As inspired by literature (Wang and Niu; 2010), we had checked that GARCH model is applicable in this research or not, but the implementation of this will be the part of future work.

[6]CPCB: CENTRAL POLLUTION CONTROL BOARD

However, TBATS and Feedforward Neural Network model have not been used earlier in Delhi Air pollution as no literature have been found on Delhi air pollution prediction. So, we had implemented these models, also basic models such as Nave method,and mean method used for forecasting before implementing Exponential smoothing, holt linear trend, ARIMA and Feedforward Neural Network. ARMA method was not in scope due to non-stationary nature of data. TBATS model is also used because

- It has capability to handle long seasonality as we had in our dataset.
- It has capability to handle non-linear features effectively.

Unlikely Naive method, that forecast the future values based on recent observations which is not appropriate in our case and mean method that forecast the values based on average of all the observation which is also not appropriate, so we had considered simple exponential smoothing (SES) which forecast the value based on all the observation and provided heavily weighted to recent observation than earlier observation. It deals with data having no trends and no seasonality. Holts linear trend method is a just extended version of SES that has deals with a trend and we can see in the decomposition of data in fig:1 and 2, that we had a trend, so that's the reason to consider this method for this research. ARIMA model also implemented as we had univariate data, and ARIMA is one of the most efficient and famous algorithm for additive time series data. A single layer feed forward neural network also has been implemented as many pieces of literature stated that neural network in time series is providing good accuracy. It has the capability to robust the noise in input data. Also, we had univariate data so that's why we had chosen single layer feed forward neural network.

3.0.5 Evaluation

It is important to evaluate the model performance using the appropriate parameters. In this section, we had compared the performance of all the applied time series models based on RMSE (Root mean squared error), MAE (mean absolute error) and execution time of models, inspired from literature. ACF and PACF plots are also used to check residuals values with BOX test and AICC value also considered where required.

RMSE (Root mean squared error): RMSE is measures of the difference between sample or population values and the values observed.

$$RMSE : \sqrt{mean(e_i^2)}$$

$$ForecastError : e_i = y_i - y_j$$

MAE (mean absolute error): It is difference of two continuous variables. It also helps to predict the average magnitude of the prediction error. (*otexts*; 2018)

$$MAE : mean(\|e_i\|)$$

$$ForecastError : e_i = y_i - y_j$$

And for the execution time of models, we had used proc.time() command, as output we had received the user, system and elapsed time. Where user time is related to the execution of codes, system times related to system process time such as opening and closing time that normally depends on system configuration and the elapsed time is the difference in times since we started the stopwatch. Based on these performance parameters, models having the least error in predicting the next 15 days air pollution is regarded as the best model.

3.0.6 Deployment

There is no real deployment now as this analysis was merely done for research. Nonetheless, producing this report could be the deployment in this case.

3.1 Design Models Used in Research

This research provided the comparatively analysis between different time series models.

3.1.1 Simple Exponential Smoothing:

In simple exponential smoothing, forecast is based on weighted average where the recent observations have been weighted heavily than an earlier observation that means forecasting depends more on last observation (*otexts*; 2018). In this model point, forecast values are the average of all the predicted values and this model is good with data that having no trend and no seasonality. So, the forecast at time t+1 is:

$$S_t = aY_t + (1 - a).S_{t-1}$$

,
Yt= most recent observation,
St = most recent forecast.

3.1.2 Holts linear trend model:

Holts linear trend method is an extended version of SES that considered trend in the model. This model can map the trend accurately without any assumption. It is just a combination of average values in the series and trend or we can say this method involves a forecast equation and two smoothing equation as mentioned below:

Forecast equation:

$$Y_{t+h,t} = L_t + hb_t$$

Level Equation:

$$L_t = aY_t + (1 - a)(l_{t-1} + b_{t-1})$$

Trend equation:

$$b_t = B * (L_t - L_{t-1}) + (1 - B*)b_{t-1}$$

Where Lt denotes an estimate level of the time series at time t, a is the smoothing parameter (between 0 to 1), and bt denotes and estimate of the trend in time tt (*otexts*; 2018).

3.1.3 ARIMA

ARIMA stands for Autoregressive Integrated Moving average. It is the most widely used in time series problem, often for non- stationary and univariate data. ARIMA model represented as ARIMA (p,d,q) where

- P is the number of autoregressive terms
- d is the number of non-seasonal differences needed for stationarity, and
- q is the number of lagged forecast errors in the prediction equation (*otexts*; 2018)

The moving average represents the lags of the prediction errors and the autoregressive represents the lag of differenced series. If d=0, then model becomes ARMA, If q=0 and d=0, than model becomes an AR and if p=0 and d=0 than model becomes MA.

3.1.4 ARIMA/GARCH

ARIMA model doesn't reflect the recent changes; it is only focuses on analyzing the time series. To deal with this ARIMA- GARCH can been used. ARIMA-GARCH encountered new information and analyses the series based on conditional variance. It is also deal with the volatility and non- linearity of data (Pham; 2018). In this research, we had checked ARIMA-GARCH is appropriate for our data or not, implementation will be the part of future work.

3.1.5 TBATS

TBATS models are applied where multiple seasonal data are present or it has capability to handle complex seasonal data. This model is a combination of various models. In TBATS T refer to trigonometry used for seasonality, B refer to Box Cox transformation for heterogeneity, A refer to ARMA error for short-term dynamic, T refer to trend and S refer to Seasonality includes multiple and non- integer periods (De Livera et al.; 2011). In TBATS parameter are selected as follows:

TBATS {W,{P,Q} @ {M,K}}
Where,
W = Box-Cox transformation parameter
P,Q = ARMA errors
@ = Damping parameter
M,K = seasonal period and Fourier term

There are some advantages of TBATS modeling:

- It handles nonlinear features.
- It allows for any autocorrelation in the residuals.
- It can effectively handle long seasonality data.

Thats the reason for selecting this model for analysis as in our dataset we had nonlinear features and long seasonality.

3.1.6 Feed forward Neural Network

Feed forward neural network is the first type of artificial neural network, in which we used a single layer. As shown in figure 4. It consists of a single layer of output nodes, the input layer is directly connected to the output via weights, it means it is fully connected. In the input layer there is no computation has been performed. Typically, it takes a larger time to train the data(*wikipedia*; 2017).

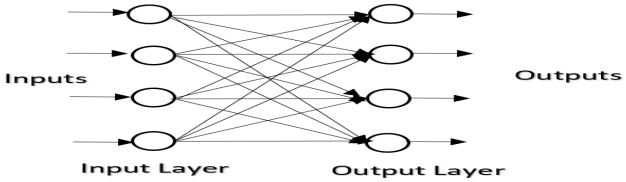

Figure 4: Single Layer Feedforward Neural Network Architecture

4 Implementation

The implementation has been done using R programming language. The outline of the implementation has been explained below in Fig:5.

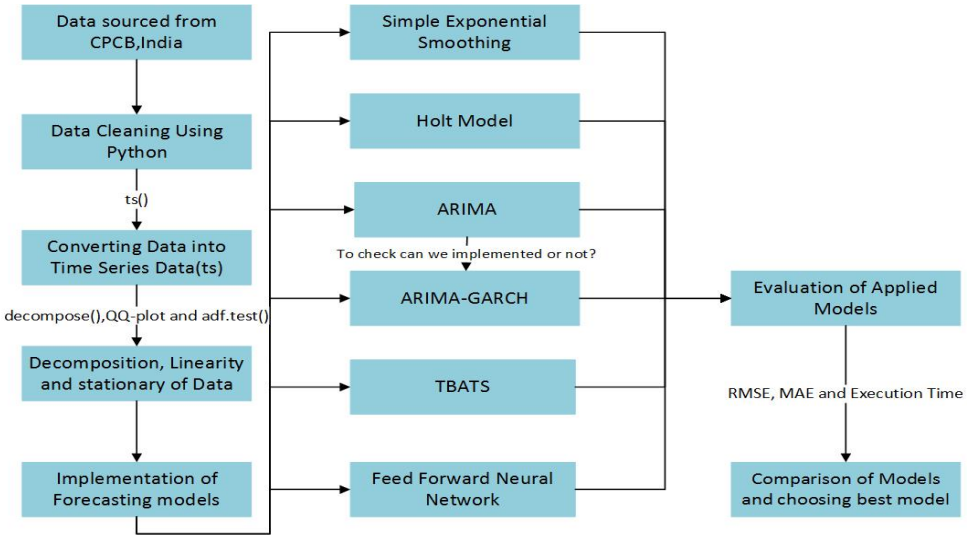

Figure 5: Implementation Process Flow

For this research, data is collected from CPCB India, it is an official website from the government of India. We had chosen Anand Vihar station to collect the air pollutant PM2.5 and PM 10 data which is hourly based. After that we had cleaned the data using Python and excel, cleaned data has been loaded in R language using read.csv() function. To process the data sets, all the required packages has been installed like readr','zoo','forecast','gplot2','lubridate','xts','fpp2' and 'tseries' using install.packages() function. Once these packages have been installed, then with the help of library() function all these packages are loaded in memory. Before implementation any model data is converted into time series with the help of ts() function in R. After that Delhi air pollution data has been decomposed in seasonal, trend and randomness (Salcedo et al.; 1999)(Fig: 1

and Fig: 2). After decomposition non- linearity of data have been checked using qqnorm() and qqline() function (fig:6).

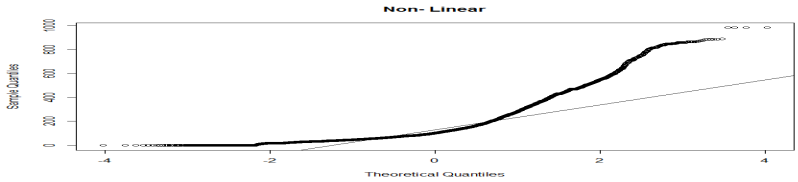

Figure 6: Q-Q Plot

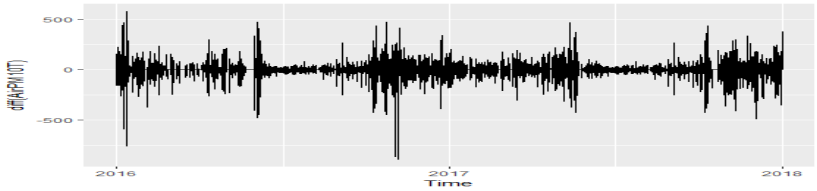

Figure 7: Differencing of Time Series

Then we had conducted Dickey- Fuller test to see the time series is stationary or not. As p-value for the test is below 0.05, that means it is a non- stationary series. To checking the white noise Box.test() and ACF (Auto- correlation function) has been conducted. These two-test showed that data is not a white noise as p-value is below 0.5 for box-test and there are some trends and seasonality available in data. Differencing has been performed and plot to make time series data to stationary as it is an important part of the ARIMA model(fig:7). Now we had implemented Simple Exponential Smoothing model(SES), but before that, we had also implemented basic model such as Nave method, and mean method. For SSE model we had used ses() function which is an inbuilt forecast package in R. Therefore, PM2.5 and PM10 pollutants have been forecasted by SES model for the next 15 days. Autoplot has been used to visualize the forecast values and forecasted values is extract in excel using write.table(). Then we had applied Holts model on the data set using holt() function. It is also an inbuilt function in R where code for holt model is present in the forecast package. So, through holt() function, we had forecasted the PM10 and PM 2.5 air pollutants value for next 15 days, and with the help of autoplot() forecasted values have been visualized. Then forecasted value is extracted in csv format to local memory from R studio using write.table() function.

Now, we had implemented the ARIMA model in a data set. To select the best fit model, auto.arima() has been the best option as it selects the value of p,d,q automatically with less execution time. We had tried to select (p,q,d=1) using grid search as well but its execution time is very large compared to auto.arima(). In our research best fit model for PM2.5 pollutant is (1,1,1) and for PM 10 is (4,1,4) and Box-Cox test provided the lambda value for PM2.5 pollutant is 0.2043267 and for PM10 pollutant is 0.168829 which have been passed through auto.arima(). Lambda value from Box-Cox test represents the values of transformation required to stabilize the variance. In our case value is close to cube root transformation. Then we had forecasted the PM10 and PM2.5 air pollutants value for next 15 days, and forecasted value is extracted in csv format. Then we had checked that GARCH model is applicable in our case or not. For that, we had checked the residual plot using QQ plot of volatility, further, we observe the square residual

plot, as ARCH/GARCH model should be used to model the volatility, that should reflect more recent changes in the series. Finally, we had plotted ACF and PACF plot of squared residuals that helps to confirm that residuals are not independent and can be predicted. (Pham; 2018). Implementation for GARCH on top of ARIMA will be the part of future work.

We had also implemented TBATS model using tbats() function. tbats() function is an inbuilt function in R. It takes less time to build, but it takes a long time to train, TBATS is a combination of many other models and it must check several values likes arma error, lambda values, Fourier tera, and damping parameter. For forecasting the PM2.5 and PM10 air pollutant, data is passed to the tbats() function and results were extracted in csv format. Then we had implemented Feedforward Neural Network with single layer. To apply this model, we had used caret library in which nnetar() function is available in R to perform Feedforward Neural Networ with a single layer. Then we had produced the next 15 days forecast of PM10 and PM2.5 pollutant and with the help of autoplot we had visualized the forecast values and using summary() function we received the RMSE and MAE value for each model. To calculate the execution time of each model we had used proc.time().

To check the performance of models we had used RMSE, MAE and execution time of the models. For the validation of time series models 80-20 split was used and suitable plots and statistics had been considered where required. Which are discussed in evaluation section.

5 Evaluation and Result

We had evaluated the applied models using RMSE, MAE and the execution time of the model. In each model, light blue line is representing the forecasted value based on 80 percent confidence and dark blue representing the forecasted values based on 95 percent confidence interval.

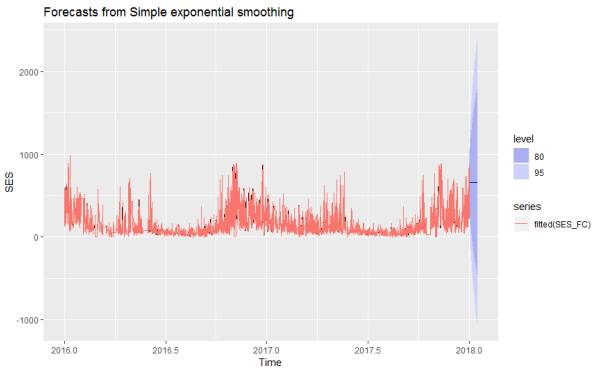

Figure 8: SES Forecasting Plot of PM2.5

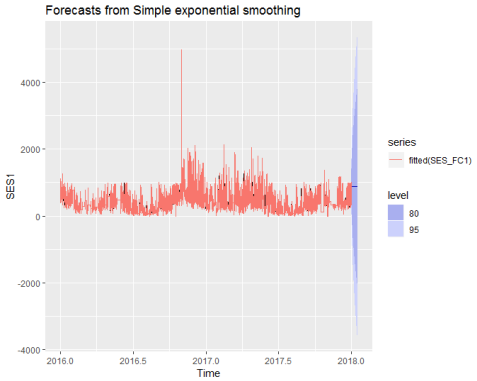

Figure 9: SES Forecasting Plot of PM10

The forecasted result from SES model shown in fig:8 and fig:9. As shown in figures it is projected straight line i.e. mean of all the predicted values. Therefore, this model is not being an effective one to forecast the PM10 and PM2.5 air pollutants in Delhi. fig:10 and fig:11 shows the forecasting values of holt model. As, holt model is double exponential smoothing that works good with trend. Since our data doesn't show much

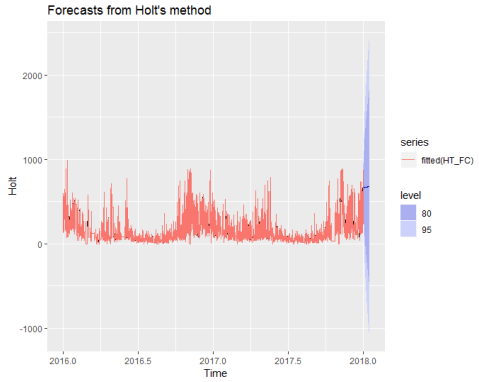

Figure 10: HOLT Forecasting Plot of PM2.5

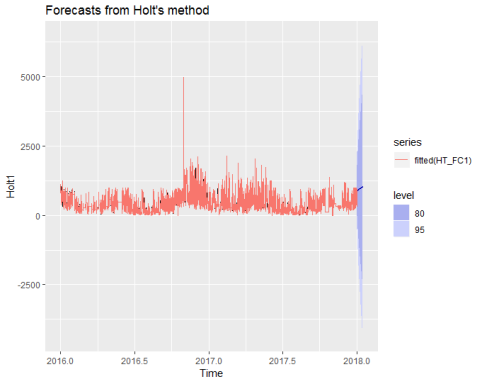

Figure 11: HOLT Forecasting Plot of PM10

trend variation hence result of forecasted values are straight. Therefore, Holt model is also not being suitable for this research.

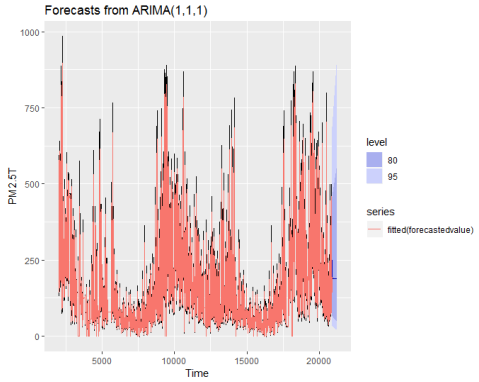

Figure 12: ARIMA Forecasting Plot of PM2.5

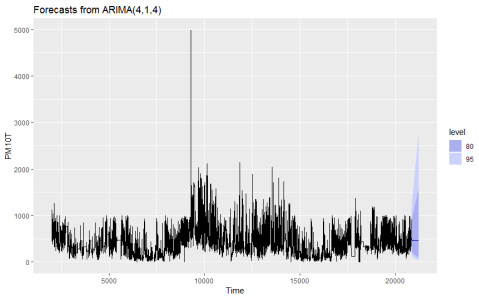

Figure 13: ARIMA Forecasting Plot of PM10

ARIMA plot in fig:12 and fig:13 shows the closest pattern of forecasted values with parameter (1,1,1) for PM2.5 and (4,1,4)for PM10. First variable indicates the last observation is used as predictors in regression equation, second variable(1) indicates that differencing has been done t make data stationary, and third variable indicates the lagged error that have been used in regression equation.

Below fig:14 and fig:15 showing the result of TBATS model. It is considered many combinations of models. For PM2.5 TBATS (1,2,5–,8766,1) is showing the best model where first variable 1 representing the box cox transformation value, then next two variables representing the p,q value of ARMA, here 2 represents the last observation is used as predictor is regression equation and last 5 past lagged error used in regression equation. - represents the damping parameter, and next value 8766,1 represent the seasonality at 8766 values that handled by one Fourier form. Whereas for PM10 best fit model is (1,4,4,–8766,1).

The output of Feedforared Neural Network shown below in fig:16 and fig:17. In NNAR(p,k) represented as P=lagged inputs, K= nodes in the single hidden layer.

In this value of p, k are not specified, they are automatically selected by function NNAR(). For PM2.5 (p,k) value are 42 and 22, whereas for PM10 (p,k) value are 41 and 21.

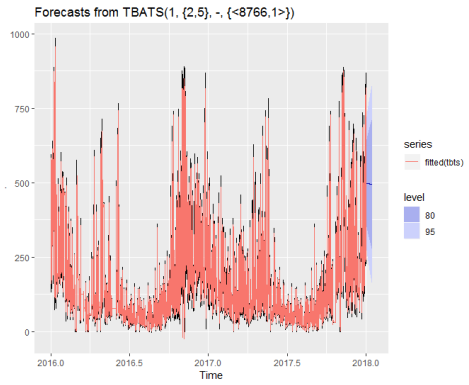

Figure 14: TBATS Forecasting Plot of PM2.5

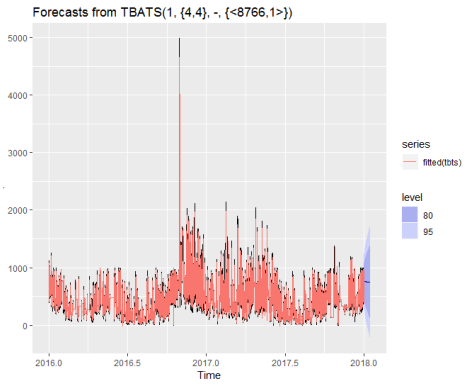

Figure 15: TBATS Forecasting Plot of PM10

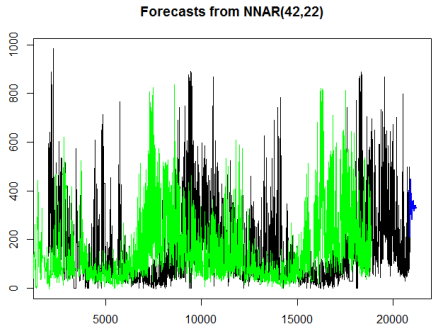

Figure 16: Neural Network Forecasting Plot of PM2.5

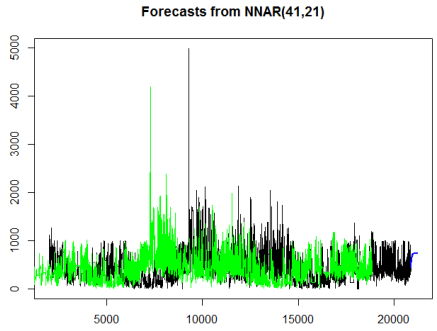

Figure 17: Neural Network Forecasting Plot of PM10

	PM2.5			PM10		
MODEL	**RMSE**	**MAE**	**Execution Time(sec)**	**RMSE**	**MAE**	**Execution Time(sec)**
SES	47.06	23.92	13.16	120.933	61.30	8.99
HOLT	47.021	23.94	14.81	121.069	61.88	13.36
ARIMA	46.67	25.34	7.98	116.52	64.22	7.87
TBATS	45.557	24.34	34.30	116.37	63.79	36.68
NN	39.34	22.43	501.15	103.107	61.06	478.27

Table 1: Performance Comparison

In the above table, we had compared the performance of applied time series models. So, from comparison table, it is clear that Feedforward Neural Network performed better that any other applied model, but its execution time is very high. TBATS and ARIMA performed better that SES and HOLT model. Based on RMSE, and MAE we can say that Neural network is the best model to predict the Delhi air pollution, but for long term data, it is not a suitable model, as its execution time is very large. For long-term data, ARIMA performed better than any other applied model as its RMSE, MAE and execution time is better than other models. We can also use TBATS as well because its RMSE and MAE are lower compared to ARIMA but execution time is large compared to ARIMA. So, its depends on a business requirement to choose the appropriate method.

6 Conclusion and future work

Overall, the main objective of this research is to implement advanced time series model to predict the PM10 and PM2.5 air pollutants and compared the performance of applied models. Based on the research we can say that ARIMA model is performed better for long term data whereas feed forward neural network is performed good for short term data, as its execution time for training is high. This is the research where TBATS, and Feed forward neural network have been used to predict the Delhi air pollution and these two-models performed well. Simple exponential smoothing and Holt method are not a good option to predict the Delhi air pollution.

There is lot of scope in this study to carry out further, first, in future, weather and traffic data also can be included, due to limitation of time these parameters are not considered. These parameters can help to improve the forecasting. Also, GARCH model can be implemented due to non- linearity and volatility nature of data. That helps to improve the performance of ARIMA. However, this research is only restricted to the dataset of two year that can be extended up to 10 years that helps to analyze data more accurately. Also, methods like Deep learning especially stacked auto encoder, LSTM and ensemble models can be used that would elongate to improve the models predictive accuracy.

Acknowledgment

I would like to express my sincere and faithful gratitude to my Supervision Mr. Vikas Tomer, who has been constantly supporting me and providing his valuable feedback in each stage. He also motivated me during every phase of the Research. His guidance helped me a lot to learn immensely during very a short time. Also, I would like to thank all the Data Analytics faculty, especially Dr. Anu Sahni.

Finally, I would like to thank my parents and friends for always encouraging and supporting me to achieve a goal.

References

Bhardwaj, R. and Pruthi, D. (2016). Time series and predictability analysis of air pollutants in delhi, *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*, IEEE, pp. 553–560.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*, springer.

Buzzi-Ferraris, G. and Manenti, F. (2011). Outlier detection in large data sets, *Computers & chemical engineering* **35**(2): 388–390.

Castro, J., Castillo, O., Melin, P. and Rodriguez-Diaz, A. (2008). A hybrid learning algorithm for interval type-2 fuzzy neural networks in time series prediction for the case of air pollution, *Fuzzy Information Processing Society, 2008. NAFIPS 2008. Annual Meeting of the North American*, IEEE, pp. 1–6.

Chan, K. Y. and Jian, L. (2013). Identification of significant factors for air pollution levels using a neural network based knowledge discovery system, *Neurocomputing* **99**: 564–569.

Chandra, B., Sinha, V., Hakkim, H., Kumar, A., Pawar, H., Mishra, A., Sharma, G., Garg, S., Ghude, S. D., Chate, D. et al. (2018). Odd–even traffic rule implementation during winter 2016 in delhi did not reduce traffic emissions of vocs, carbon dioxide, methane and carbon monoxide., *Current Science (00113891)* **114**(6).

De Livera, A. M., Hyndman, R. J. and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing, *Journal of the American Statistical Association* **106**(496): 1513–1527.

de Mattos Neto, P. S., Cavalcanti, G. D. and Madeiro, F. (2017). Nonlinear combination method of forecasters applied to pm time series, *Pattern Recognition Letters* **95**: 65–72.

Fotopoulou, E., Zafeiropoulos, A., Papaspyros, D., Hasapis, P., Tsiolis, G., Bouras, T., Mouzakitis, S. and Zanetti, N. (2016). Linked data analytics in interdisciplinary studies: The health impact of air pollution in urban areas, *IEEE Access* **4**: 149–164.

Gorai, A. and Mitra, G. (2017). A comparative study of the feed forward back propagation (ffbp) and layer recurrent (lr) neural network model for forecasting ground level ozone concentration, *Air Quality, Atmosphere & Health* **10**(2): 213–223.

Gore, R. W. and Deshpande, D. S. (2017). An approach for classification of health risks based on air quality levels, *Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on*, IEEE, pp. 58–61.

Hasham, F. A., Kindzierski, W. B. and Stanley, S. J. (2004). Modeling of hourly no x concentrations using artificial neural networks, *Journal of environmental engineering and science* **3**(S1): S111–S119.

Hassani, H., Silva, E. S., Gupta, R. and Segnon, M. K. (2015). Forecasting the price of gold, *Applied Economics* **47**(39): 4141–4152.

Jiang, P., Dong, Q. and Li, P. (2017). A novel hybrid strategy for pm2. 5 concentration analysis and prediction, *Journal of environmental management* **196**: 443–457.

Jiang, Z., Meng, X., Yang, C. and Li, G. (2008). A bp neural network prediction model of the urban air quality based on rough set, *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, Vol. 1, IEEE, pp. 362–370.

Kumar, A. and Goyal, P. (2011). Forecasting of air quality in delhi using principal component regression technique, *Atmospheric Pollution Research* **2**(4): 436–444.

Kurt, A. and Oktay, A. B. (2010). Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks, *Expert Systems with Applications* **37**(12): 7986–7992.

Li, X., Peng, L., Hu, Y., Shao, J. and Chi, T. (2016). Deep learning architecture for air quality predictions, *Environmental Science and Pollution Research* **23**(22): 22408–22417.

Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K. and Tsao, C.-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang, *PloS one* **12**(7): e0179763.

Nagpure, A. S., Gurjar, B. R. and Martel, J. (2014). Human health risks in national capital territory of delhi due to air pollution, *Atmospheric Pollution Research* **5**(3): 371–380.

otexts (2018).
URL: *https://www.otexts.org/*

Pham, L. (2018). Arch/garch model in r.
URL: *https://talksonmarkets.files.wordpress.com/2012/09/time-series-analysis-with-arima-e28093-arch013.pdf*

Pollutedcities (2018).
URL: *https://www.weforum.org/agenda/2018/05/these-are-the-worlds-most-polluted-cities*

Salcedo, R., Ferraz, M. A., Alves, C. and Martins, F. (1999). Time-series analysis of air pollution data, *Atmospheric Environment* **33**(15): 2361–2372.

Sindhwani, R. and Goyal, P. (2014). Assessment of traffic-generated gaseous and particulate matter emissions and trends over delhi (2000–2010), *Atmospheric Pollution Research* **5**(3): 438–446.

Stationarity and differencing (2018).
URL: *https://people.duke.edu/ rnau/411diff.htm*

Taneja, S., Sharma, N., Oberoi, K. and Navoria, Y. (2016). Predicting trends in air pollution in delhi using data mining, *Information Processing (IICIP), 2016 1st India International Conference on*, IEEE, pp. 1–6.

Tzima, F. A., Mitkas, P. A., Voukantsis, D. and Karatzas, K. (2011). Sparse episode identification in environmental datasets: the case of air quality assessment, *Expert Systems with Applications* **38**(5): 5019–5027.

Wang, H. and Zhao, L. (2018). A joint prevention and control mechanism for air pollution in the beijing-tianjin-hebei region in china based on long-term and massive data mining of pollutant concentration, *Atmospheric Environment* **174**: 25–42.

Wang, W. and Niu, Z. (2010). Data analysis in los angeles long beach with seasonal time series model, *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, IEEE, pp. 113–120.

Wang, Z., Gong, Z., Zhu, W. and Zhao, W. (2009). A rough set based pso-bpnn model for air pollution forecasting, *Natural Computation, 2009. ICNC'09. Fifth International Conference on*, Vol. 3, IEEE, pp. 357–361.

Wei, W. W. (2006). Time series analysis, *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Wendel, J. (2014). Air pollution ranks as largest health risk, *Eos, Transactions American Geophysical Union* **95**(14): 120–120.

wikipedia (2017).
URL: *https://en.wikipedia.org/wiki/Feedforward*$_n$*eural*$_n$*etwork*

Yang, G., Huang, J. and Li, X. (2018). Mining sequential patterns of pm2. 5 pollution in three zones in china, *Journal of Cleaner Production* **170**: 388–398.

Yoon, H.-J., Xu, S. and Tourassi, G. (2016). Predicting lung cancer incidence from air pollution exposures using shapelet-based time series analysis, *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*, IEEE, pp. 565–568.