Repeated Measures Multivariate Generalized Linear Model Analysis of Vehicle Contribution on Particle Pollution for Mumbai Suburban Area and Particle Pollution Forecasting

By

Shramana Bhattacharya

Dissertation submitted for the partial fulfilment of the requirements for the

Award of Masters in Biostatistics and Demography

Academic year, 2018-20

Under the Supervision of

Dr. Suryakant Yadav

Department of Development Studies

INTERNATIONAL INSTITUTE FOR POPULATION SCIENCES

(Deemed University)

ACKNOWLEDGEMENT

In the present world of competition there is a race of existence in which those who are having will to come forward, succeed. Dissertation is like a bridge between theoretical and practical working. This dissertation is an individual work, yet, no hard work is successful without the help, guidance and support of a lot of people. I would like to express my sincere gratitude to these people without whom this dissertation could never be a success.

It is a genuine pleasure to express my deep sense of thanks and gratitude to my guide Dr. Suryakant Yadav (Assistant Professor, Department of Development Studies) for his constant guidance, help, support, monitoring and encouragement that provided inspiration and self-confidence to work and complete my dissertation. It has been a great experience to learn, explore and work under his supervision. The long journey could never have been easier without him. Thank you for allowing me to learn from my mistakes and being an astounding source of strength who never let me lose my confidence. Thank you Dr. K.S James, our director, for giving me the chance to express and apply my earned knowledge in form of the dissertation. I am deeply indebted to Dr. Murali Dhar and Dr. Preeti Dhillon (Masters in Biostatistics and Demography Course Coordinators), for lending me the opportunity to pursue my study. It is my pleasure to acknowledge the valuable suggestions of my evaluation committee members, Dr. Dhananjay W. Bansod , Dr. Laxmi Kant Dwivedi. I am thankful to all my respected faculties at IIPS, for their valuable suggestions and encouragement throughout my work.

A special thanks to my friend Mr. Sayak Banerjee for his assistance and cooperation which had a huge impact to uplift my spirits and carry on with the dissertation work. I would also like to thank my friends Miss Poulami Barman and Miss Nikita Patel for their support throughout. I am highly grateful to my parents, Mr. Somesh Bhattacharya and Mrs. Mitali Bhattacharya for their blessing, care and support.

Shramana Bhattacharya

CONTENTS

Acknowledgement *2*

List of Chapters *4*

List of Tables *5*

List of Figures *6*

Acronyms *7*

LIST OF CHAPTERS

CHAPTER 1: INTRODUCTION 9

CHAPTER 2: CONCEPTS 13

CHAPTER 3: LITERATURE REVIEW 20

CHAPTER 4: NEED FOR THE STUDY 23

CHAPTER 5: OBJECTIVES AND DATA DESCRIPTION 25

CHAPTER 6: METHODOLOGY AND FINDINGS 27

CHAPTER 7: CONCLUSION AND LIMITATIONS 49

REFERENCES 51

LIST OF TABLES AND RESULTS

***Table 1**: Correlation matrix between the variables PM10, PM2.5 and PM1, scooter_motorcycle_moped and car_jeep_van.*

***Table 2a**: Principal Component Analysis results displaying eigenvalues, difference, proportion and cumulative. Also showing the eigenvectors and unexplained proportion.*

***Table 2b**: Scoring coefficients of principal component.*

***Table 3**: Repeated measures multivariate GLM results.*

Results 4a, 4b and 4c are not in tabular form but they display the KPSS results obtained in R.

***Table 5**: Summary measures referring to minimum, first quartile, median, mean, 3rd quartile and maximum for PM10, PM2.5 and PM1.*

Results 6a, 6b and 6c depict the R results of the ARIMA model fitted to the train data of PM10, PM2.5 and PM1 respectively.

Results 7a, 7b and 7c depict the R results of the ARNN model fitted to the train data of PM10, PM2.5 and PM1 respectively.

Results 8a, 8b and 8c depict the R results of the hybrid ARIMA-ARNN model fitted to the train data of PM10, PM2.5 and PM1 respectively.

***Tables 9a, 9b, 9c:** These tables display the performance metrics. The rows represent the values corresponding to a model.*

LIST OF FIGURES

Figure 1: *Time series plot of PM10 for the year 2019 (hourly data).*

Figure 2: *Time series plot of PM2.5 for the year 2019 (hourly data).*

Figure 3: *Time series plot of PM1 for the year 2019 (hourly data).*

Figures i,ii,iii: *Normal Q-Q plots of the residuals for the train datasets of PM10, PM2.5 and PM1 respectively.*

Figures 4a, 4b, 4c*: Actual vs predicted forecasts for the test data sets of PM10, PM2.5 and PM1 respectively*

ACRONYMS

WHO	*World Health Organisation*
BMC	*Brihanmumbai Municipal Corporation*
SAFAR	*System of Air Quality Weather Forecasting and Research*
AQI	*Air Quality Index*
µm	*Micrometres*
PM	*Particulate Matter*
GAMs	*Generalized Additive Models*
PCA	*Principal Component Analysis*
ANN	*Artificial Neural Network*
SVM	*Support Vector Machine*
ARIMA	*Autoregressive Integrated Moving Average*
ARNN	*Autoregressive Neural Network*
GLM	*Generalized Linear Model*
ACF	*Autocorrelation Function*
PACF	*Partial Autocorrelation Function*
AIC	*Akaike Information Criterion*
BIC	*Bayesian Information Criterion*
MSE	*Mean Square Error*
ME	*Margin of Error*
MAE	*Mean Absolute Error*
RMSE	*Root Mean Square Error*
MAPE	*Mean Absolute Percentage Error*
MPE	*Mean Percentage Error*

CHAPTER 1

CHAPTER 1: INTRODUCTION

WHO's 2019 report states that seven million people are estimated to be killed as a result of air pollution worldwide every year. 9 out of 10 people inhale polluted air, according to the data published by WHO (**1**). Air pollution develops in two contexts: indoor and outdoor. Indoor air pollution affects the poor households across the world (**2**). Burning of solid fuel sources such as firewood, crop waste and dung is a major cause of indoor air pollution. It is responsible for about 1.6 million deaths each year, including premature deaths as per the study of Global Burden of Disease (**3**). Outdoor air pollution becomes worse for developing countries as they industrialize and transit from low to middle income. Long term exposure to air pollution can have serious impact on health and wellbeing of individuals such as cognitive function. Outdoor air pollution is a leading cause for stroke, heart disease, lung cancer and respiratory diseases such as asthma (**4**). India, being the world's second most populous country, attributes to high air pollution. The busy cities like Delhi, Mumbai, Bangalore, Chennai, Kolkata, etc. contribute largely to the pollution of the entire country affecting other areas lying close to them.

The city of Mumbai is known for its high population density. Presence of more number of, industries, companies, etc. attract people from across the globe to settle in Mumbai. Pollution in air is extremely high in Mumbai due to the presence of industries in the eastern suburbs and New Mumbai. Vehicles do not have sufficient control over emissions and there is occurrence of garbage burning by the BMC during the night hours at the 100 hectare Municipal Garbage Dump north of the Chembur-Vashi road. This affects Chembur, Ghatkopar, Mankhurd and New Mumbai. According to a recent study by the Environmental Pollution Research Centre (EPRC), 10% of Chembur population is suffering from bronchitis and respiratory distress caused due to pollution (**5**).

Solid or liquid matters suspended in the air refer to particulate pollution. Smoke, fumes, soot, and other combustion by-products constitute this type of air pollution. Primary particles come out of exhaust stacks and tailpipes; and secondary particles such as sulphates and nitrates are formed as a result of condensation of vaporized materials or from the by-products of the oxidation of gases in the atmosphere. They are a mixture of contaminants from a range of sources (**Douglas W. Dockery, 2012**)(**6**). Particles are characterized by their aerodynamic properties, measured as aerodynamic diameter measured in μm. The size of particles is directly linked to their potential for causing health problems. Small particles less than 10 micrometres in diameter pose the greatest problems, because they can get deep into your lungs, and some may even get

into your bloodstream. Exposure to such particles can affect both your lungs and your heart. Numerous scientific studies have linked particle pollution exposure to a variety of problems, including: premature death in people with heart or lung disease, nonfatal heart attacks, irregular heartbeat, aggravated asthma, decreased lung function, increased respiratory symptoms, such as irritation of the airways, coughing or difficulty breathing. People with heart or lung diseases, children, and older adults are the most likely to be affected by particle pollution exposure (7).

SAFAR, a project of the Indian Institute of Tropical Meteorology, Pune, determines in 2016, AQI in Mumbai at 10 different locations, identified sector-wise causes for air pollution in Mumbai. About 21.2% of pollution is due to the suspended dust that rises from unpaved roads. "As a vehicle speeds on this road, the dust rises. It might settle soon, but the road is never free of vehicles," Dr. Gufran Beig, the Project Director at SAFAR, says (8). Hypertension and associated risk of cardiovascular disorder has been found to be linked with exposure to polluted air (**Debasish Bandyopadhyay et al., 2014**) (9). Particles less than 2.5 micrometre in diameter are small enough to enter into respiratory system and cause fatal physiological consequences (10). Air pollution is indeed an alarming issue for the health of our future generations in India. We are aware of the ill effects of environmental pollutants and toxicants on health status of human as well as other living organisms and the environment (**Ghosh and Parida, 2015**) (11).

SAFAR provides location specific information on air quality in near real time and its forecast 1-3 days in advance for the first time in India. The stations of SAFAR are located at across Delhi, Pune and Mumbai. Air pollution data of Chembur station situated at International Institute for Population Sciences, Govandi Station Road, Deonar, Chembur, Mumbai-400088, is used here. SAFAR monitors pollutants: PM1, PM2.5, PM10, Ozone, CO, NOx (NO, NO2), SO2, BC, Methane (CH4), Non-methane hydrocarbons (NMHC), VOC's, Benzene, Mercury. Firstly, the effect of two-wheelers and three-wheelers (extrapolated data obtained from Censuses of 2001 and 2011) on particle air pollution (PM10, PM2.5 and PM1 data of SAFAR) is considered (12). A dataset is prepared based on SAFAR and Census data. PCA and then repeated measures multivariate GLM is applied on the data for analysis.

The current progress in the area of modern statistics and machine learning have equipped the forecasters with nonlinear forecasting tools such as ANN, deep learning, and SVM among many others. Combination of various forecasting models for time series prediction containing trend, seasonality, etc. has led to the development of hybrid models. Hybrid models are a combination of

classical models (namely time series regression and ARIMA model) and/or modern methods (Artificial Neural Networks). Real world data is not so simple. It comprises of patterns which cannot be extracted by a single linear or nonlinear time series model. Most of these time series data contain both linear and non-linear patterns that has to be taken care of. Studies therefore have constructed a hybrid forecasting model that combines ARIMA with ANN for short-term forecasting of time series. This proposed approach considers the linear and nonlinear patterns in the real data simultaneously so that it can capture more precise characteristics to describe the time series better. Another nonlinear forecasting model is ARNN which is sometimes stated as a better approach then ANN.

Analysing prediction models ARIMA, ARNN, hybrid ARIMA-ARNN and comparing the level of success of their application in prediction of data under consideration is the second goal of this research.

CHAPTER 2

CHAPTER 2: CONCEPTS

2.1. *PCA* is a statistical procedure that uses orthogonal transformation to convert correlated variables into a set of linearly uncorrelated variables called principal components. The transformation defines the component with the largest variance as the principal component which accounts for the majority of the data variability. Each component is orthogonal to the preceding component and captures the highest possible variance under the orthogonality constraint. The resulting vectors (each being a linear combination of the variables and containing *n* observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables leading to feature dimensionality reduction (**13**).

2.2. The *GLM Repeated Measures* procedure provides analysis of variance when the same measurement is made several times on each subject or case. This statistical technique considers a dependent correlated or non-independent variable. Commonly used when measuring the effect of a treatment at different time points. The independent variables may be categorical or continuous. The GLM Repeated Measures procedure provides both univariate and multivariate analyses for the repeated measures data. Both balanced and unbalanced models can be tested. A design is balanced if each cell in the model contains the same number of cases. In a multivariate model, the sums of squares due to the effects in the model and error sums of squares are in matrix form rather than the scalar form found in univariate analysis. These matrices are called SSCP (sums-of-squares and cross-products) matrices. In addition to testing hypotheses, GLM Repeated Measures produces estimates of parameters (**14**).

2.3. *ARIMA* is a popular linear time series model, used for tracking linear tendencies in stationary time series data. ARIMA model is denoted by ARIMA(p, d, q). The parameters p and q are the order of the AR model and the MA model respectively, and d is the level of differencing (to be used for stationarity). ARIMA model can be mathematically expressed as follows:

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q},$$

where yt denotes the actual value of the variable under consideration at time t, εt is the random error at time t, φi and θj are the coefficients of the model. The basic assumption made by ARIMA model is that εt−l (εt−l = yt−l − yˆt−l) follows zero mean with constant variance, and satisfies the i.i.d condition. Building an ARIMA model for any time series data set requires to follow the basic three iterative steps. The steps are as follows: model identification (achieving stationarity), parameter estimation (the ACF and the PACF plots are used to select the AR and MA model parameters, respectively), and model diagnostics checking (finding the 'best' fitted forecasting model using AIC and/or the BIC) (**15**).

2.4. *ANN* are computing systems which can be related to the biological neural networks. Without being programmed with task-specific rules such systems are trained or they learn to perform certain tasks. ANN are multi-layer fully-connected neural networks or in other words are Feedforward Neural Networks. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. The figure on next page illustrates an ANN model.

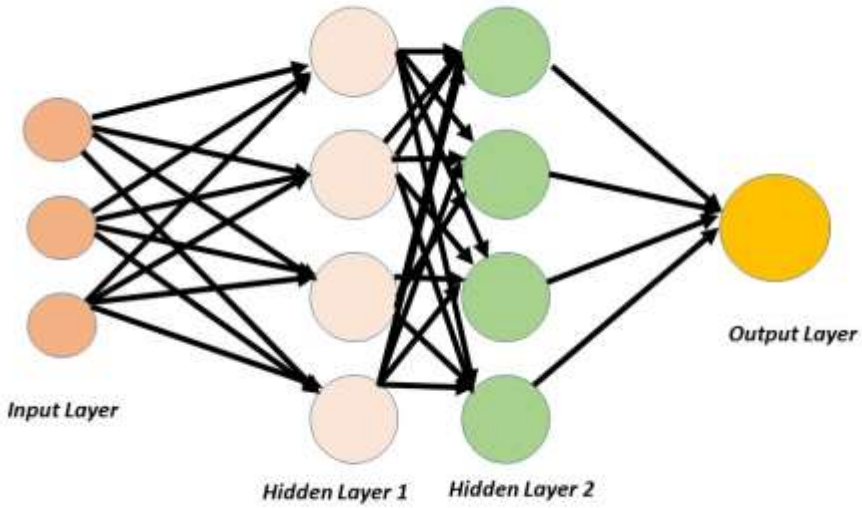
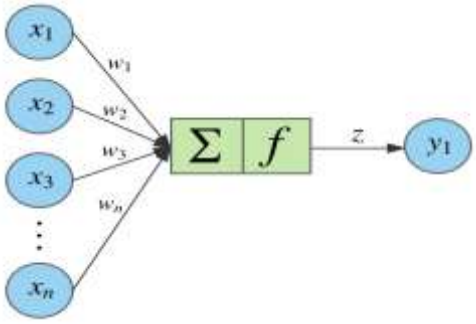


Fig: Artificial Neural Network

One of the hidden nodes can be picturised as shown:



A given node takes the weighted sum of its inputs, and passes it through a non-linear activation function. This is the output of the node, which then becomes the input of another node in the next layer. The signal flows from left to right, and the final output is calculated by performing this procedure for all the nodes. Training this deep neural network means learning the weights associated with all the edges (**16**).

2.5. *ARNN* model is less complex and easy to implement as compared to ANN model. We take a simple network with no hidden layer. Linear regression with four predictors can be versioned as a neural network and is shown in the figure on next page. The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a "learning algorithm" that minimises a "cost function" such as the MSE. Of course, in this simple example, we can use linear regression which is a much more efficient method of training the model.

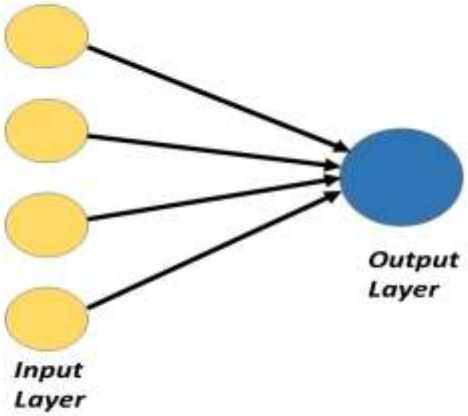
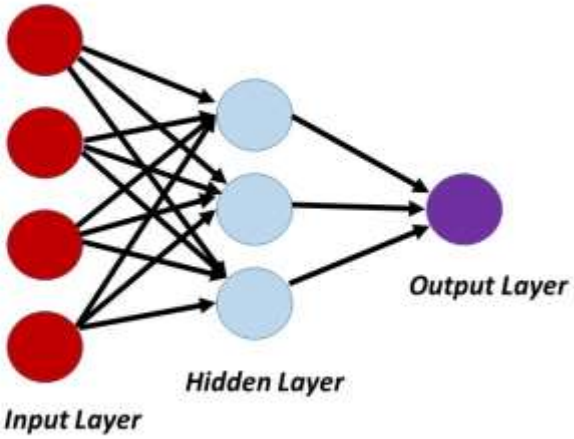


Fig: Neural network version of linear regression

Once we add an intermediate layer with hidden neurons, the neural network becomes non-linear.

A simple example is shown in Figure next page.



This is known as a multilayer feed-forward network, where each layer of nodes receives inputs from the previous layers. The outputs of the nodes in one layer are inputs to the next layer. The inputs to each node are combined using a weighted linear combination. The result is then modified by a nonlinear function before being output (**17**). This model takes in p lagged values of the time series as inputs to the model and has only one hidden layer with k nodes. The model is depicted as ARNN or NNAR(p,k) where p is the number of lags for AR(p) model and k=((p+1)/2).

2.6. Margin of error (ME) tells us how many percentage points results will differ from the real population value (**18**).

M.E = $\sqrt{\frac{p(1-p)}{n}}$ p : sample proportion, n: sample size

2.7. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are (**19**).

RMSE = $\sqrt{\frac{\sum(e_i-\bar{e})^2}{n}}$, e_i : residuals , n: sample size

2.8. Absolute Error is the amount of error in your measurements. It is the difference between the measured value and "true" value. The Mean Absolute Error (MAE) is the average of all absolute errors. The formula is:

MAE = $\frac{\sum |x_i - x|^2}{n}$, x_i : experimental value, x: true value, n: sample size (**20**)

2.9. The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values (**21**).

MAPE = $\frac{1}{n} \sum \left| \frac{A_t - F_t}{A_t} \right|$, A_t : actual value and F_t : Forecast value

2.10. Mean percentage error (MPE) is the computed average of percentage errors by which forecasts of a model differ from actual values of the quantity being forecast.

The formula for the mean percentage error is:

MPE = $\frac{100\ \%}{n} \sum \frac{A_t - F_t}{A_t}$, A_t : actual value and F_t : Forecast value (**22**)

CHAPTER 3

CHAPTER 3: LITERATURE REVIEW

There have been many researches on air pollution and its effect on health. Researchers have used techniques such as multivariate analysis employing correlation analysis, principal component analysis (PCA), and cluster analysis (CA) resulted in establishing a correlation between different pollutants (**David et. al., 2019**) (**23**).

Statistical and substantive contributions to time-series analyses of air pollution and health outcomes produced improved semiparametric regression models for time series analyses of air pollution and health; Bayesian hierarchical models for producing national, regional, and city estimates of the relative risk of mortality associated with concentrations of particulate matter 10 µm or smaller in aerodynamic diameter (PM10) and other pollutants for the 88 largest urban centers in the United States (**Dominici et al. 2000a, 2002a, 2003a,b; Samet et al 2000a,b,c**) (**24**).

In another study, the changes of the concentration of air pollutants were examined, based on air pollution, meteorological and climatologic data gathered over an interval of two years (2012-2013) by the regional measuring station located in the basin, as well as the sources of the air pollutants were studied with the help of factor analysis and the correlations between pollutants (**Reka Keresztes et. al.,2017**) (**25**).

The problem of Air pollution is very severe in a majority of Indian cities. In the study, the level of Air pollution with respect to urban and vehicular population is studied. The tools used for the study are Karl Pearson's Coefficient of Correlation and Principal Component Analysis. The result of the study reveals that urban population and vehicular population strongly positively related. Hourly PM2.5 concentrations at 35 air quality monitoring (AQM) stations in Beijing between 2013 and 2014, and daily meteorological data and geographic information during the same time period have been used to develop a two-stage method comprising a dispersion model and a generalized additive mixed model (GAMM) to estimate the traffic and non-traffic contributions to daily PM2.5 concentrations separately (**Xin Fang,2018**) (**26**).

Descriptive analysis and predictive analysis have been used to study the trends of various air pollutants in Delhi like sulphur dioxide (SO2), nitrogen dioxide (NO2), suspended particulate matter (PM), ozone (O3) carbon monoxide (CO), benzene, and forecast the future trend. Predictive analytics is the use for statistics and machine learning techniques to predict about the future (unknown data). The goal is to predict the probable future through past experience. In the analysis, predictive analysis has been done by implementing time series regression forecasting.

Another research contains an analysis of the performance and effectiveness of the Long Short-Term Memory (LSTM) neural networks, as a recurrent neural network type suitable for solving the type of data obtained from city of Skopje, Macedonia; prediction problems. Parametric methods widely used for times series analysis and forecasting of air pollution data are the autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) models (**Gocheva-Ilieva et. al.,2013**) (**27**). Recently, combined models with both linear and nonlinear models have greater attention. In some research, ARIMA, linear ANN, multilayer perceptron (MLP), and radial basis function network (RBFN) models are considered along with various combinations of these models and are compared to see the best among them (**Atilla et. al., 2007**) (**28**).

CHAPTER 4

CHAPTER 4: NEED FOR THE STUDY

Air pollution is an alarming cause of various health problems which needs to be taken care of with immediate effect. Mumbai, with its overpopulation is a highly affected area. There are about 10.2 lakh private cars on Mumbai's roads, comprising about 28 percent of the city's total number of vehicles. Mumbai's western suburbs have about five lakh registered cars, while eastern suburbs have 1.7 lakh private cars. The number of registered vehicles across the financial capital of India was almost three million in 2016. Mumbai was the most car-congested city in the country in 2019. In recent years, the density of privately-owned vehicles increased by 18 percent in the city (**29**).

It is important to understand the relationship between the vehicles and the rising particulate pollution in the city suburbs which will help curb the pollution by taking proper actions such as commuting in public transport or shares. One type of particulate matter is the soot seen in vehicle exhaust. Fine particles pose a serious threat to human health, as they can penetrate deep into the lungs. PM can be a primary pollutant or a secondary pollutant from hydrocarbons, nitrogen oxides, and sulphur dioxides. Diesel exhaust is a major contributor to PM pollution.

The objective of a predictive model is to estimate the value of an unknown variable. A time series has time (T) as an independent variable and a target dependent variable. The output of the model is the predicted value for Y at time T. Whether we wish to predict the trend in financial markets or electricity consumption, time is an important factor that must be considered. It would be interesting to forecast the hourly air pollution for the next 12 hours, for example. Being able to get an idea about the future trend of data, steps can be taken today to prevent it.

CHAPTER 5

CHAPTER 5: OBJECTIVES & DATA SOURCE AND DESCRIPTION

5.1. OBJECTIVES

1. To understand the vehicular emission effect on particulate matter pollution in Mumbai Suburban area during the years 2015-2018.
2. To establish a forecasting model on hourly particle pollution data of SAFAR, Chembur 2019.

5.2. DATA SOURCE AND DESCRIPTION

1. Hourly data for PM10, PM2.5, PM1 ($\mu g/m^3$) is obtained from SAFAR, Chembur 2015-18. Census'01 and Census'11 assets and amenities file has been used to get the extrapolated data on scooter_motorcycle_moped and car_jeep_van for Mumbai Suburban for the years 2015-18. The assembly constituency list of Mumbai Suburban is used to get the areas falling under Mumbai Suburban. Chembur is one such assembly constituency. The SAFAR station at Chembur is assumed to capture the air pollution data of the Mumbai Suburban area (**30**).

   Annual average value of particle pollution and extrapolated four year values for vehicles are considered. These values are repeated over the assembly constituencies of the Mumbai Suburban area. Z-scores are calculated and used for analysis. A Z-score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean. Z-scores are a way to compare results to a "normal" population. Results from tests or surveys have thousands of possible results and units; those results can often seem meaningless. A z-score is therefore used.

2. Hourly data of particle pollution from SAFAR, Chembur 2019 is available from 1st January to 30th April, 2019. This data is used for forecasting.

CHAPTER 6

CHAPTER 6: METHODOLOGY & FINDINGS

6.1. METHODOLOGY

Firstly, our dataset 1 has three dependent variables namely PM10, PM2.5 and PM1 and two independent variables- scooter_motorcycle_moped and car_jeep_van. We will obtain a correlation matrix to find if there exists any linear association between the variables. Table 1 shows the correlation matrix. Stata is used.

Since the four-year data is repeated over the assembly constituencies across Mumbai Suburban, we cannot use simple linear regression.

Principal Component Analysis (PCA) is done on the independent variables. This is done due to presence of multicollinearity. Table 2a displays the results of PCA. Table 2b refers to the scores obtained after predicting the principal component/s. Results are obtained using Stata.

Next, repeated measures multivariate Gauss Markov Linear Model (GLM) is performed on the available three dependent variables and the principal component/s. Table 3 refers to the estimates obtained after doing repeated measures multivariate GLM in SPSS.

Secondly, in dataset 2, we consider 2742 cases by hours (year 2019) without considering any missing value. Figures 1, 2 and 3 represent the time series plots for PM10, PM2.5 and PM1 respectively obtained in Excel. We will check whether the data is stationary or non-stationary by using KPSS test in R. In econometrics, Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests are used for testing a null hypothesis that an observable time series is stationary around a deterministic trend (i.e. trend-stationary) against the alternative of a unit root. Results 4a, 4b, 4c show the findings of KPSS test for PM10, PM2.5 and PM1 respectively. A summary() is used to get the descriptive statistics. Table 5 displays the summary.

There are three basic forecasting models available:- Holt Winter's Model, ETS(Error Trend Seasonality) Model and ARIMA Model.
The first two give one step ahead forecasts only. In this case, we want 12 step ahead forecasts, that is, forecast for the next 12 hours.

So, we divide our dataset into train and test.

We will discuss and use ARIMA Model as our first forecasting model. Results 6a,6b,6c depict the ARIMA results for PM10, PM2.5 and PM1 respectively obtained from R.

ANN (Artificial Neural Network) is a popular machine learning model, highly useful for sophisticated nonlinear time series forecasting. ARNN (Autoregressive Neural Networks) overcomes the problems of fitting ANN for time series data sets like the choice on the number of hidden neurons, and its black box nature. ARNN model uses lagged values of the time series as inputs to the model. ARNN(p,k) is a nonlinear feed-forward neural network model with one hidden layer (having p lagged inputs) and k hidden units in the hidden layer.
We have used ARNN model. Results 7a, 7b, 7c depict the ARNN results for PM10, PM2.5 and PM1 respectively obtained from R.

This dissertation considers a hybrid approach that studies the relationship between linear and nonlinear components of the time series. The proposed hybrid methodology assumes an additive relationship between linear and nonlinear models with the assumption that different models can separately model the linear and nonlinear patterns of a time series, and then the forecasts can be combined. In the first phase of the proposed model, an ARIMA model is applied to catch the linear patterns of the data set. Residual error values of the ARIMA model are calculated and restored for further modelling. In the next stage, a nonlinear ARNN model is applied to capture the nonlinear trends in the data set using the residual values obtained from ARIMA. We call this two-step approach as 'hybrid ARIMA-ARNN' model. Results 8a,8b,8c depict the hybrid ARIMA-ARNN results for PM10, PM2.5 and PM1 respectively obtained from R. Note that ARIMA results are same as obtained before.

Performance metrics refer to the values ME, RMSE, MAE, MPE, MAPE. Figures 4a, 4b, 4c represent the actual vs predicted forecasts for the test data sets of PM10, PM2.5 and PM1 respectively, also, tables 9a, 9b, 9c represent the performance metrics for the same.

6.2. FINDINGS AND INTERPRETAIONS

Dataset 1:

Table 1 (Stata results):

	z_pm10	z_pm25	z_pm1	z_scooter_motorcycle_moped	z_car_jeep_van
z_pm10	1.000				
z_pm25	0.996	1.000			
z_pm1	0.735	0.746	1.000		
z_scooter_motorcycle_moped	0.757	0.739	0.955	1.000	
z_car_jeep_van	0.783	0.770	0.967	0.998	1.000

Interpretation: The dependent variables have high correlation among themselves and also with the independent variables. Hence repeated measures linear regression is possible.

But, the independent variables have a very high correlation as well. Therefore, problem of multicollinearity arises. Hence, PCA needs to be performed.

Note: Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. This violates our assumption of uncorrelated independent variables in the classical linear regression model. Therefore, multicollinearity has to be taken care of. Variables can be dropped to reduce the effect of multicollinearity. But, here, we are dealing with only two independent variables obtained from Census data extrapolation and hence dropping of which variable will be logical is unknown. Therefore, PCA can be a better solution to the problem of multicollinearity here.

Table 2a:

Principal Component Analysis (Stata results):

Rotation: (unrotated = principal)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.99815	1.9963	0.9991	0.9991
Comp2	.00185044	.	0.0009	1.000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
z_scooter_motorcycle_moped	0.7071	0.7071	0
z_car_jeep_van	0.7071	-0.7071	0

Interpretation: Here, we have 2 components since we performed PCA on our two independent variables.

The eigenvalues and eigenvectors can be considered to be the PCA core. Eigenvectors or the principal components determine the direction of the new feature space and eigenvalues are nothing but their corresponding magnitudes.

Here, the eigenvalue for component 1 is 1.99815 and for component 2 is 0.00185. The difference between the eigenvalues is 1.9963. Component 1 has captured 99.91% of the total variation of the data. Thus, component 2 has almost zero significance here. Cumulative column just displays the cumulative proportions.

Another goal of PCA is to reduce dimensionality of the original feature space by projecting it onto a smaller subspace, where these eigenvectors form the axes. In our case, dimensionality reduction is not very important because our main focus is to remove multicollinearity.

The first eigenvector is the first principal component which captures the maximum overall variance. The second eigenvector is the second principal component which has the second maximal variance, uncorrelated to the first principal component. The eigenvectors with the lowest eigenvalues are dropped as they contain least information.

From this table, we can easily understand that Component 1 (Let it be pc1) contains the maximum information. So, we will consider only pc1 for further analysis.

Table 2b:

Stata results for predicting the principal component:

Scoring coefficients

sum of squares (column-loading) = 1

Variable	Comp1	Comp2
z_scooter_motorcycle_moped	0.7071	
z_car_jeep_van	0.7071	-0.7071

Interpretation: Scores are nothing but the loadings obtained before.

The corresponding predicted pc1 is obtained in the data editor window of Stata. This pc1 is used in repeated measures GLM analysis as an independent variable.

The concept of PCA states that the obtained principal components are uncorrelated with one another. Hence, doing PCA automatically removes multicollinearity. Also, only pc1 is taken. This pc1 is the first principal component that has the maximum data information. The data represents independent variables (scooter_motorcycle_moped and car_jeep_van). Therefore, pc1 is such a component that can represent both the independent variables containing their maximum information. Using this component will help us get an idea of whether the representative component of two-wheelers and three-wheelers affect particle pollution. In other words, we can find the effect of principal component of vehicles on particle pollution by using pc1 instead of the two collinear independent variables.

Table 3:

Repeated measures multivariate GLM (SPSS results):

Parameter estimates

Dependent Variable	*Parameter*	*B*	*Std. Error*	*t*	*Significance*
z_pm10	Intercept	0.518	0.039	13.273	0.000
	pc1	0.325	0.028	11.712	0.000
z_pm2.5	Intercept	0.861	0.034	25.566	0.000
	pc1	0.267	0.024	11.161	0.000
z_pm1	Intercept	0.798	0.017	47.427	0.000
	pc1	0.406	0.012	33.891	0.000

Interpretation: All the beta coefficients and intercepts are statistically significant (less than 0.05 for 5% level of significance).

The repeated measures multivariate GLM has taken three dependent variables and have regressed the pc1 on these variables to obtain the regression coefficients (B), the standard error, t statistic and significance.

For one-unit increase in pc1, z_pm10 will increase by 0.325, z_pm2.5 will increase by 0.267 and z_pm1 will increase by 0.406.

Since pc1 has been used instead of the independent variables, we cannot directly say that increase in number of scooter_motorcycle_moped and car_jeep_van will lead to the increase in particle pollution. It is difficult to comment in this case. Our pc1 carries 99.91% of the total data variation. We can thus say that a huge portion of the information contained in the original data is present in pc1. It is clearly seen that pc1 positively affects particle pollution. Therefore, we might; intuitively; conclude that the increase in number of vehicles (expecting pc1 to carry maximum information of number of vehicles i.e. two and three wheelers) will increase particle pollution.

Dataset 2:

Figure 1:

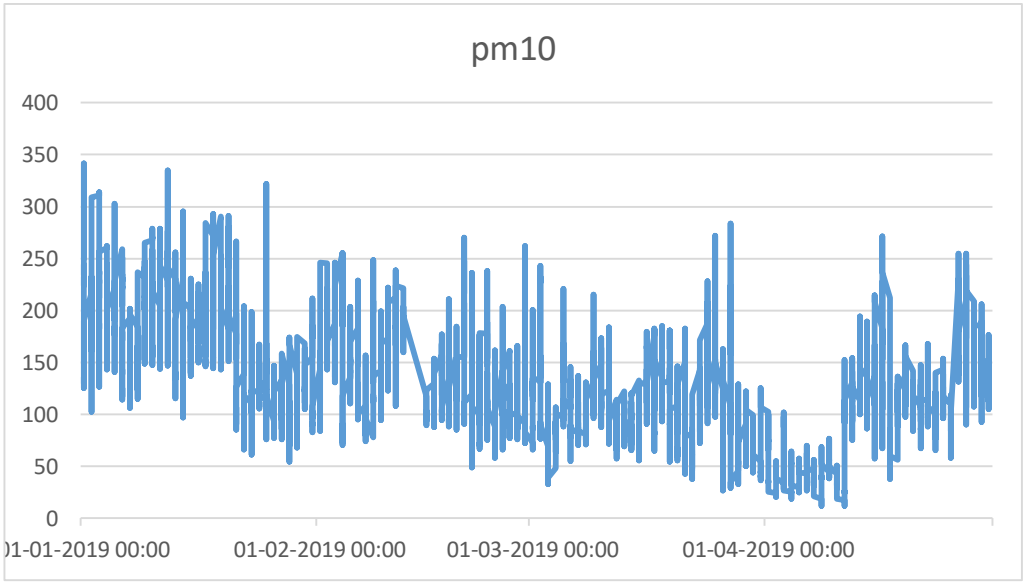


Figure 2:

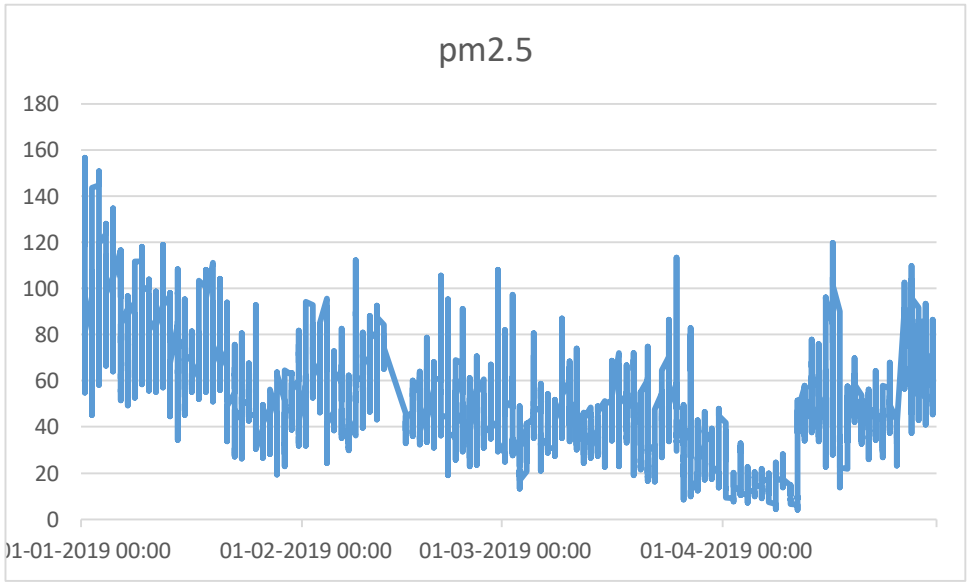
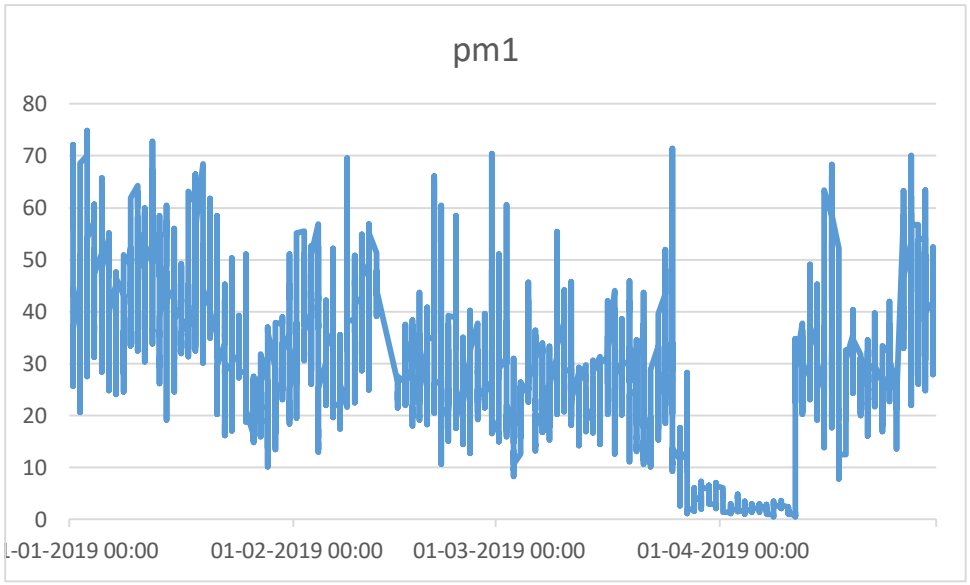


Figure 3:



Interpretation: All these plots show non stationary time series. There is a drop during the period 27th March to 11th April,2019 for the three particles. We cannot comment on trend or seasonality from these graphs.

Result 4a:

KPSS Test for Level Stationarity

H0: The data is stationary vs. H1: Data is non-stationary

KPSS Level = 11.255, Truncation lag parameter = 9, p-value = 0.01

P value is less than 0.05. Hence, the data is non stationary.

Result 4b:

KPSS Test for Level Stationarity

H0: The data is stationary vs. H1: Data is non-stationary

KPSS Level = 10.467, Truncation lag parameter = 9, p-value = 0.01

P value is less than 0.05. Hence, the data is non stationary.

Result 4c:

KPSS Test for Level Stationarity

H0: The data is stationary vs. H1: Data is non-stationary

KPSS Level = 8.852, Truncation lag parameter = 9, p-value = 0.01

P value is less than 0.05. Hence, the data is non stationary.

Note: decompose function of R reports that time series has no or less than 2 periods. This implies absence of seasonality in data. Hence, decompose function could not work.

Interpretation: The KPSS test has already been explained in the methodology section. It is performed to check for non-stationarity in the data which is visible from the time series plots displayed above. The above KPSS test results obtained in R confirm that the data for PM10, PM2.5 and PM1 are non-stationary. Non-stationarity is important to carry on with the time series predictive analysis.

Table 5 (R results):

Summary

	Pm10	**Pm2.5**	**Pm1**
Minimum	11.49	3.93	0.55
1st Quartile	89.78	34.48	19.83
Median	121.43	47.20	27.82
Mean	128.45	50.60	28.20
3rd Quartile	165.07	64.97	37.27
Maximum	341.73	156.61	74.90

Interpretation: Summary measures of any data helps to get an overall picture of the entire data. Mean value is maximum for PM10 and minimum for PM1. Throughout PM10 has the highest value and PM1 the lowest for all the separate measures.

We can understand, the most released particle pollutant is PM10 and the least released is PM1 in the year 2019.

Result 6a:

ARIMA(2,1,1)

Coefficients:

AR1		**AR2**		**MA1**	
1.0338		-0.2309		-0.9670	
S.E	0.0196	S.E	0.0192	S.E	0.0073

Result 6b:

ARIMA(2,1,1)

Coefficients:

AR1		**AR2**		**MA1**	
1.0625		-0.2437		-0.9708	
S.E	0.0193	S.E	0.0190	S.E	0.0065

Result 6c:

ARIMA(2,1,1)

Coefficients:

AR1		AR2		MA1	
1.0390		-0.2266		-0.9625	
S.E	0.0197	S.E	0.0191	S.E	0.0077

Interpretation: In the above results, we have used a non-seasonal ARIMA model since seasonality is not visible in the plots. The ARIMA(p,d,q) model obtained for each of the three variables is same. p= 2, d=1, q=1; i.e. order of Autoregressive Model is 2, order of differencing is 1 and order of Moving Average Model is 1. Differencing is done to make the data stationary. Autoregressive models are same as regression models except that the regressors are the lagged values of the dependent variable. Moving average gets the average of the points in a series for a specific lag.

In all the three cases we can see that the standard errors for AR2 is less than for AR1. The model has considered p=2. Standard error of MA1 is also given.

ARIMA model forecasting equation is basically used for linear stationary time series data. Now, the data considered here may not be completely linear. Non-linear time series can be modelled with the help of nonlinear models like ARNN (concept discussed in Chapter 2). Next, we will thus fit an ARNN model to the data.

Result 7a:

Series : trainpm10

Model: NNAR(27,14)

Average of 20 networks, each of which is a 27-14-1 network with 407 weights.

Result 7b:

Series: trainpm2.5

Model: NNAR(27,14)

Average of 20 networks, each of which is a 27-14-1 network with 407 weights.

Result 7c:

Series: trainpm1

Model: NNAR(27,14)

Average of 20 networks, each of which is a 27-14-1 network with 407 weights.

Interpretation: The concept of ARNN or NNAR model has been discussed in Chapter 2. It is a popular machine learning model which is used for non-linear time series forecasting. Here NNAR(p,k) model is thus used. We can see for all the three datasets, p=27 and k=14 have been considered. Each dataset displays similar result. An average of 20 different neural networks have been used, where each is 27-14-1 network i.e. 27 lagged inputs, one hidden layer with 14 nodes and 1 output and 407 weights.

In methodology, a hybrid ARIMA-ARNN has been explained. This model is used to model the linear and nonlinear components of the time series datasets in the next step.

Results 8a:

This adjacent normal Q-Q Plot of residuals confirm that non-linearity is present in the residuals which can be modelled using ARNN.

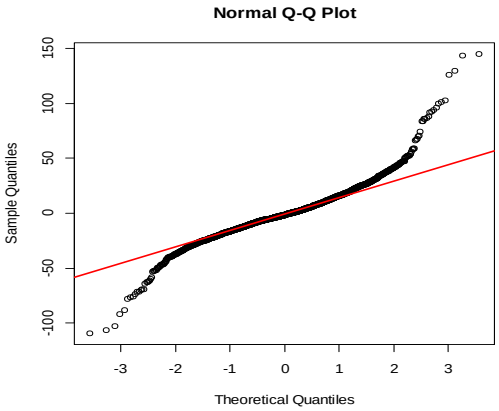


Figure i

ARNN Model

Series: trainpm10

Model: NNAR(34,18)

Average of 20 networks, each of which is a 34-18-1 network with 649 weights.

Result 8b:

This adjacent normal Q-Q Plot of residuals confirm that non-linearity is present in the residuals which can be modelled using ARNN.

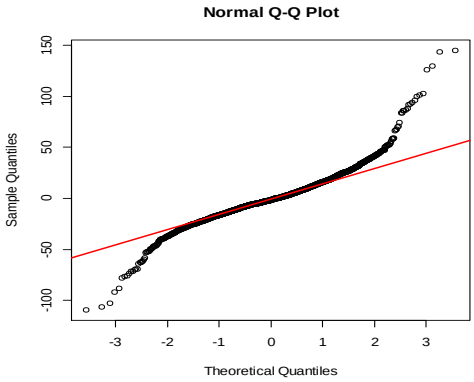


Figure ii

ARNN Model

Series: trainpm2.5

Model: NNAR(34,18)

Average of 20 networks, each of which is a 34-18-1 network with 649 weights.

Result 8c:

This adjacent normal Q-Q Plot of residuals confirm that non-linearity is present in the residuals which can be modelled using ARNN.

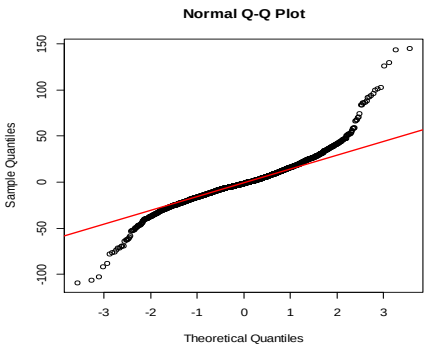


Figure iii

ARNN Model

Series: trainpm1

Model: NNAR(34,18)

Average of 20 networks, each of which is a 34-18-1 network with 649 weights.

Interpretation: An ARIMA model has already been fitted to the datasets and we know the results- ARIMA(2,1,1) model has been used. Now, the residuals of this ARIMA model have been obtained in R. These residual's normal Q-Q plots have been obtained as shown. The residuals show non-normal behaviour confirming that nonlinear pattern is present in the residuals. If the residuals were completely linear, it would have been white noise i.e. should have followed standard normal distribution.

As residuals are nonlinear, we fit the residuals to the nonlinear ARNN model as discussed before. We obtain NNAR (34,18) using the average of 20 networks, each of which is a 34-18-1 network and has 649 weights.

This is the hybrid ARIMA-NNAR model.

Next, the performance metrics of each of the above models have been obtained using R, whose results are shown. We take the model as our best performing model that shows minimum values.

Table 9a:

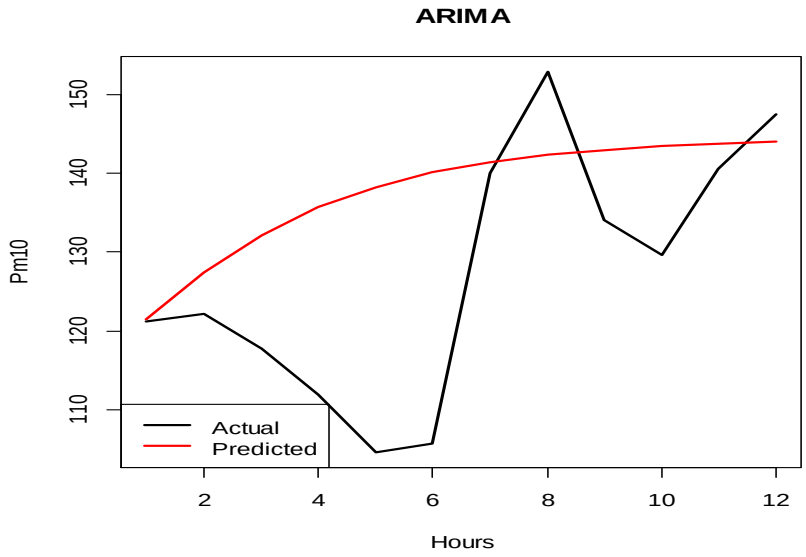
PM10

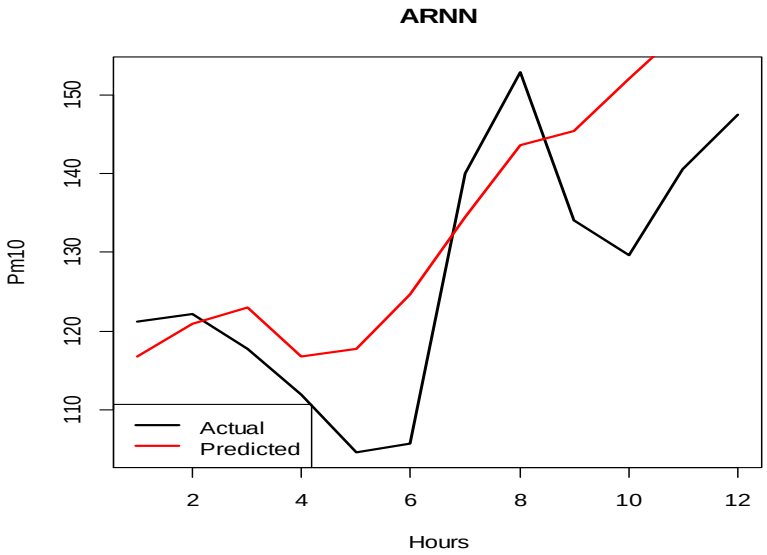
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
ARIMA	10.42316	17.13712	12.75278	7.576519	9.208704
ARNN	7.457451	12.66089	10.84276	5.261487	7.804284
Hybrid ARIMA-ARNN	9.889459	17.41201	13.50649	6.993316	9.745678

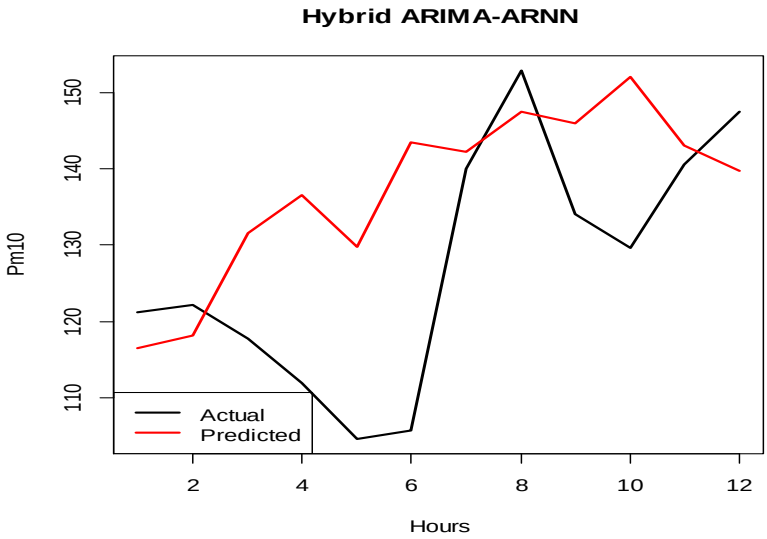
Interpretation: It is clearly visible, that for each of the performance metric discussed in Chapter 2; model ARNN performs best (minimum value). Hence, for PM10 dataset, we consider this model to forecast.

The plots represent the actual vs predicted forecasts for the test data of PM10 with respect to each model.

Figure 4a







Interpretation: The ARNN plot illustrates that the forecasted values (predicted values) fits best to the actual values or test values depicting that ARNN model performs best among the three for PM10.

Table 9b:

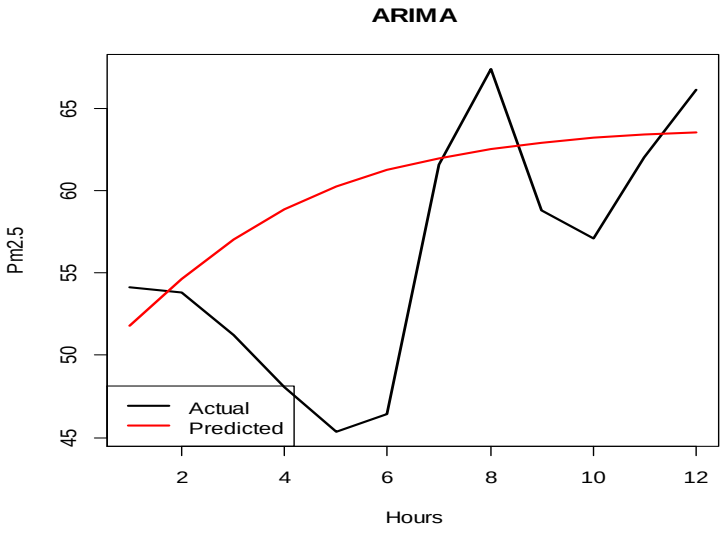
PM2.5

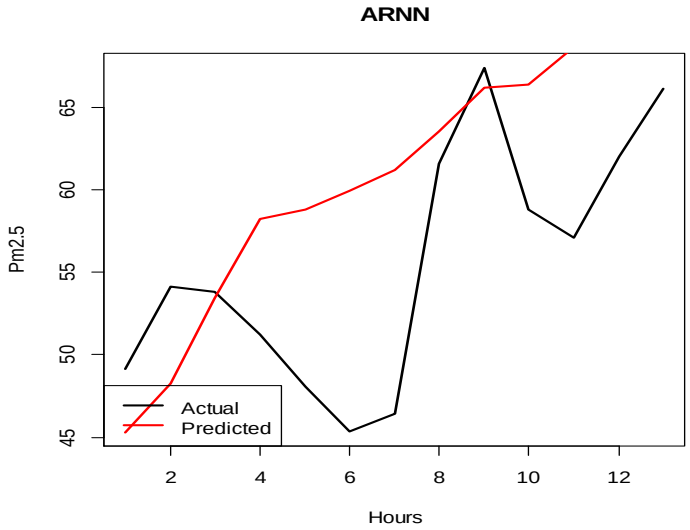
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
ARIMA	4.109319	7.536724	5.733877	6.797226	9.511875
ARNN	7.912073	9.772696	8.088844	12.18858	12.5218
Hybrid ARIMA-ARNN	3.195631	6.156759	5.002156	5.918386	9.100771

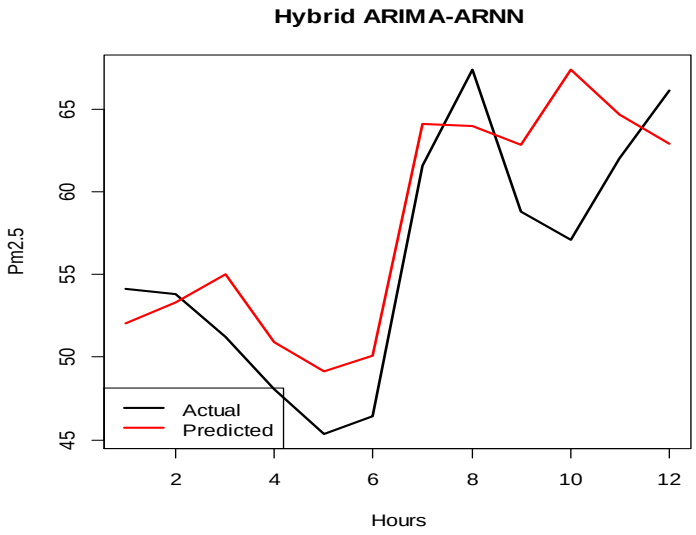
Interpretation: It is clearly visible, that for each of the performance metric discussed in Chapter 2; model Hybrid ARIMA-ARNN performs best (minimum value). Hence, for PM2.5 dataset, we consider this model to forecast.

The plots represent the actual vs predicted forecasts for the test data of PM10 with respect to each model.

Figure 4b







Interpretation: The Hybrid ARIMA-ARNN plot illustrates that the forecasted values (predicted values) fits best to the actual values or test values depicting that it performs best among the three for PM2.5.

Table 9c:

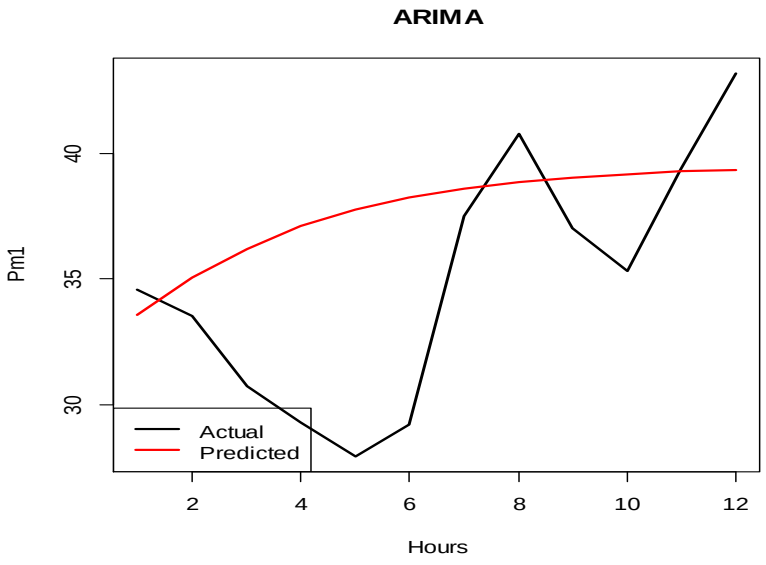
PM1

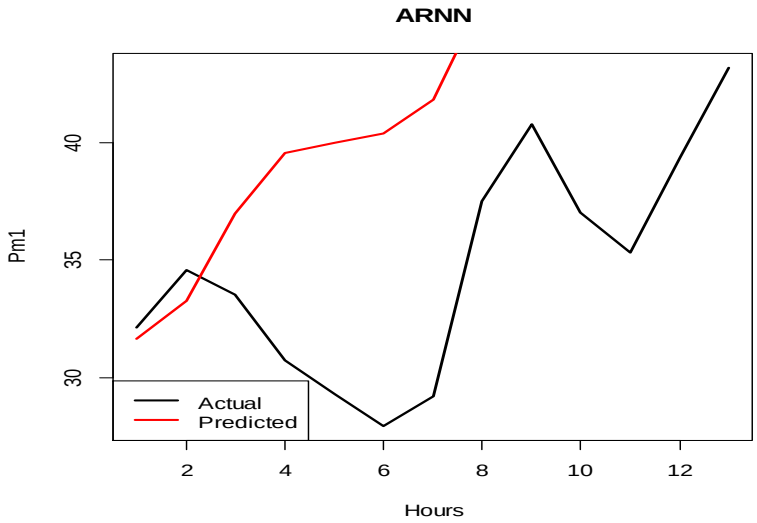
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
ARIMA	2.822029	5.098318	3.965638	7.525233	10.51318
ARNN	9.103894	10.133	9.103894	20.22193	20.22193
Hybrid ARIMA-ARNN	1.039206	4.330194	2.241506	6.917591	8.97523

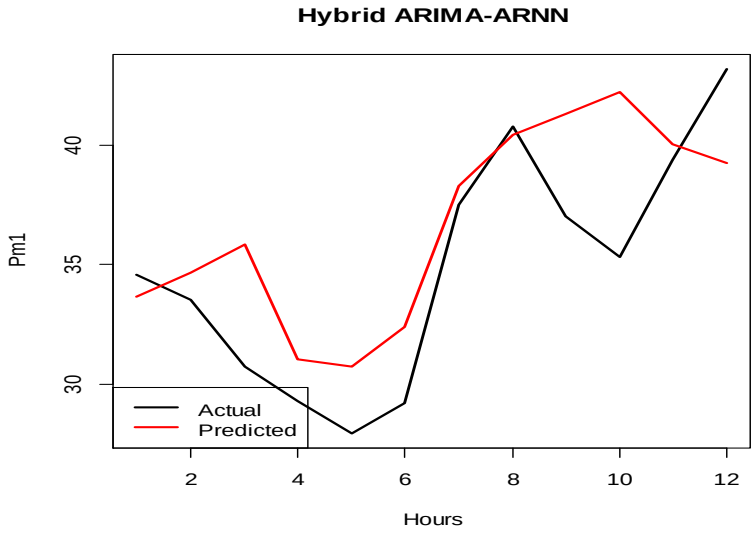
Interpretation: It is clearly visible, that for each of the performance metric discussed in Chapter 2; model Hybrid ARIMA-ARNN performs best (minimum value). Hence, for PM1 dataset, we consider this model to forecast.

The plots represent the actual vs predicted forecasts for the test data of PM1 with respect to each model.

Figure 4c







Interpretation: The Hybrid ARIMA-ARNN plot illustrates that the forecasted values (predicted values) fits best to the actual values or test values depicting that it performs best among the three for PM1.

CHAPTER 7

CHAPTER 7: CONCLUSION AND LIMITATIONS

7.1. CONCLUSION AND DISCUSSION

The findings and interpretations establish a linear relationship existing between the dependent and independent variables. Repeated measures multivariate GLM confirm that the independent variables together (pc1) have a significant effect on the dependent variables. This means that particle pollution depends largely on the number of vehicles (two wheelers and three wheelers) possessed by the people residing in a particular area (in this case, Mumbai Suburban). Vehicular emission is responsible for releasing particulate matter into the air which causes air pollution and in turn is responsible for various airborne diseases confronted by us.

Decreasing the number of vehicles on road is the need of the hour. We can resort to travelling by public transport or use car pools. Delhi has tried to decrease their vehicle count on road each day by introducing the Odd-Even rule. This rule was made to reduce the pollution in the capital city after it saw a major bump in PM10 and PM2.5 levels post Diwali. The Odd-Even rule is a space rationing scheme that determines which vehicles will ply on the roads on specific days. According to the scheme, odd-numbered and even-numbered vehicles will ply on the roads on alternate days. Vehicles with registration numbers ending in odd numbers will be allowed on the roads on odd days and even-numbered vehicles will be allowed on the roads on even days. Such initiatives can be a very good option. This data, though, focuses on the total number of vehicles possessed by people and not on the number of vehicles each day on road. Still, it is quite clear that if the number of vehicles belonging to the population increases, this will definitely mean more vehicles on road and more pollution. If number of vehicles on road can be restricted, this will also bring down the number of cars owned.

A proper forecasting model is needed to get an estimate of the next few hour particle pollution. A proper time series modelling is important because it is used to study the past behaviour of the phenomena under consideration. It is used to compare the current trends with that in the past or the expected trends. Thus it gives a clear picture of growth or downfall. After comparing three types of models, we can conclude that ARNN performs best for PM10 while hybrid ARIMA-ARNN performs best for PM2.5 and PM1. The applications of these models can be used for future prediction of the time series data for the year 2019.

7.2. LIMITATIONS OF THE STUDY

The data for repeated measures multivariate GLM has been obtained only for four years for the area of Mumbai Suburban due to lack of data availability. This cannot give us a broader picture of the current scenario. We have used extrapolated Census data which might not match with the true values. Only association between vehicles and particle pollution is considered, though there are various other sources that emit particulate matter like industries.

Moreover, the independent variables show multicollinearity, therefore we used PCA and tried find the effect of pc1 on particle pollution. From this, we cannot separately obtain how scooter_motorcycle_moped and car_jeep_van individually is responsible for increasing particle pollution.

There are other time series forecasting models available which could have given better results than the three considered here. Since, we wanted to propose a hybrid model, so we have taken those discussed models for analysis

REFERENCES

(1) https://www.who.int/health-topics/air-pollution#tab=tab_1
(2) https://ourworldindata.org/air-pollution
(3) https://ourworldindata.org/indoor-air-pollution
(4) https://ourworldindata.org/outdoor-air-pollution
(5) http://theory.tifr.res.in/~sgupta/bombay/amenities/sanitation/air-pollu.html
(6) Dockery et. al.; Health Effects of Fine Particulate Air Pollution: Lines that Connect (2005)
(7) https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm 3
(8) https://citizenmatters.in/mumbai-delhi-air-pollution-particulate-matter-aqi-environment-12940
(9) http://www.medicinenet.com/script/main/art.asp? articlekey=105529 5
(10) http://www.ibtimes.co.uk/world-environment-day-10-most-polluted-cities-world-1504260 6 3
(11) Ghosh and Parida; Air Pollution and India: Current Scenario (2015)
(12) http://safar.tropmet.res.in/
(13) https://towardsdatascience. com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
(14) https://www.theanalysisfactor.com/repeated-measures-approaches/
(15) Tanujit Chakraborty et. al.; A Hybrid Model for European Unemployment Rate Forecasting and Its Asymptotic Behavior (2019)
(16) https://otexts.com/fpp2/
(17) https://otexts.com/fpp2/nnetar.html
(18) https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/margin-of-error/
(19) https://www.statisticshowto.datasciencecentral.com/rmse/
(20) https://www.statisticshowto.datasciencecentral.com/absolute-error/
(21) https://www.statisticshowto.datasciencecentral.com/mean-absolute-percentage-error-mape/
(22) https://www.totalassignmenthelp.com/blog/percentage-error-formula/

(23) David et. al.; Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region (2019)
(24) Dominci et. al.; Time-Series Analysis of Air Pollution and Mortality: A Statistical Review (2004a)
(25) Reka et. al.; Statistical analysis of air pollution with specific regard to factor analysis in the Ciuc basin, Romania (2017)
(26) https://openarchive.ki.se/xmlui/bitstream/handle/10616/46343/Thesis_Xin_Fang.pdf?sequence=3&isAllowed=y
(27) Gocheva-Ilieva et. al.;Time series analysis and forecasting for air pollution in small urban area: An SARIMA and factor analysis approach (2013)
(28) Atilla et. al; Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting (2007)
(29) https://www.moneycontrol.com/news/india/mumbai-is-indias-most-car-congested-city-has-more-vehicles-than-delhi-and-pune-3694881.html
(30) https://ceo.maharashtra.gov.in/Downloads/Extent_of_ACs_in_Mumbai_City_and_Suburban_corrected_on_3_8.pdf