Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 373–381

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Time Series Data Prediction using IoT and Machine Learning Technique

Raghavendra Kumar[a*] Pardeep Kumar[b] Yugal Kumar[b]

[a,b]*Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, 173234, India*

[a]*Department of Information Technology, KIET Group of Institutions, Ghaziabad,201206, India*

Abstract

Time series analysis and prediction have been widely accepted in various domains from last two decades. Business analytics, Medical drugs & pharmaceutical, Dynamic Marketing, Weather forecasting, Pollution measures, financial portfolio analysis and Stock market prediction are the favorite domains among research communities under time series analysis. Since air quality is one of the paramount factors which make life possible on earth and monitoring air quality data as time series analysis is a one of prime area. The most affected air quality parameters on health are carbon monoxide (CO),carbon dioxide (CO2), Ammonia(NH3) and Acetone ((CH3)2CO). In this paper we have taken the sensor's data of three specific locations of Delhi and National Capital Region (NCR) and predict air quality of next day using linear regression as machine learning algorithm. Model is evaluated through four performance measures Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The study further assesses with benchmark model and obtains significant results.

Keywords: Time series; Regression Model; ARIMA, Machine Learning;

10.1016/j.procs.2020.03.240

1. Introduction

Time series data is a sequential data at regular time interval in a given period. G.E.P. Box et al. (2008), emphasis on time series prediction for decision making using historical data pattern [23]. A prediction model usually justifies the mean of regression model as time series trends for future values.

Nomenclature

IoT	Internet of Thing
ARIMA	Auto Regressive Integrated Moving Average
CO	Carbon Monoxide
CO2	Carbon Dioxide
NH3	Ammonia
(CH3)2CO	Acetone

Availability of substance in the air which generates higher risk to living being is called Air pollution. Automobile, real estate and in-organic agriculture are the major contributors are in air pollution [1]. However, it cannot be denied that weather and climate change are also play significant role in Air pollution [2]. Proposed system uses gas sensors (MQ135) and microcontroller (Arduino-Uno) to fetch the data. Imputation techniques are used to clean data by removing non-finite values in the dataset as they can interfere in the analysis process. Level of air quality for proximity is calculated by comparing data with Break Points for Comparing AQI in India (Table-1) [3] and ATSDR-Toxicological Profile [4]. Linear regression predictive analytics and machine learning are used to predict air quality of coming day using the data of present day. The results of analysis are displayed on a mobile application using Firebase for data storage. Mobile application also suggests ways to limit the harmful effects of air pollution in that area. Wong Tze Wai (2009) created a system for reporting air pollution index. The Air Pollution Index (API) Reporting System can prove to be a significant tool for communication of probable risks [5].

Nashwa El-Bendary et al. (2013) developed a Smart Environmental Monitoring system. In the system sensor nodes are used to sense the data and then transmit this data to the microcontroller. This chapter provides an understanding about the techniques and features which researchers and system designers should keep in mind for successful deployments and operation of real smart environmental monitoring systems [6]. Anjaiah Guthi(2016) implemented a Noise and Air Pollution Monitoring System using Internet of Things (IoT). This adaptable and distributive system monitors the fluctuation of noise and air pollution level and best suits for infrastructural environment [7]. In 2000 World Health Organization (WHO) published a report on the importance of the information hidden on air pollutants. The principles stated in this report are projected to encourage progressive modification of the networks that keep an eye on air quality, and enhance their utility for assessment of health impact [8]. Riteeka Nayak et al. (2017) implemented a system where a LCD displays the PPM data of harmful gases which air quality goes down from a particular level [9]. Fuzzy inference systems have been introduced by José Juan Carbajal Hernández et al. (2012) to assess air quality status in urban areas. They used fuzzy inference system to monitor air pollution and found efficient and more accurate results. [18].Wang, D. et al. (2016) proposed an ELM and decomposition based hybrid model to predict non-stationary, irregular and random data series. They used to forecast stock price as well as PM2.5 winds speed [17]. Liu B-C et al. (2017) conducted AQI experiments on three cities (Beijing, Tianjin and Shijiazhuang) and reached on conclusion that support vector regression model is reliable as well as strong enough for AQI prediction [16]. Bagirov et al. (2017) developed the cluster-wise linear regression technique to predict monthly rainfall in Victoria city, Australia. Authors had the fusion of clustering and regression techniques and found that more climate parameters help to improve the performance [19]. Lee M. et al. (2017) selected high high-density and high-rise city for spatial variability in air pollution concentrations and reached to long term exposure for public health threat due to population. They used land used regression modeling (LRU) [20]. Suling Zhu et al. (2017) proposed two hybrid model (EMD-SVR and EMD-IMF) for regional air quality indexes (AQI) and found forecasting accuracy is better than ARIMA, SVR and other GRNN models like EMD-GRNN and Wavelet-GRNN for air pollution time series data [21]. Ji Jia et al. (2018) emphasis on missing data imputation of air pollutants in time series. They used LSTM imputation method to improve PM2.5 concentration prediction accuracy [22].

Table-1. Break Points for Comparing AQI, India [1]

AQI Category	PM10 (24 Hr)	PM2.5 (24 Hr)	NO2 (24 Hr)	O3 (8 Hr)	CO (8 Hr)	SO2 (24 Hr)	NH3 (24 Hr)	Pb (24 Hr)
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
moderate polluted (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10-17.0	381-800	801-1200	2.1-3.0
Very Poor (301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1200-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

As per Jiaming Zhu et al. (2018), There are various traditional statistical models, (i.e. linear regression (LR), multiple linear regression (MLR), principal component regression (PCR) technique and the non-parametric regression (NR), auto-regressive integrated moving average (ARIMA)), AI techniques (i.e. support vector regression (SVR), artificial neural networks (ANN), Recurrent Neural Network(RNN)) and hybrid models designed and implemented for accurate air quality index(AQI) prediction [21][24]. Kar Yong Ng et al. (2018) proposed model for prediction of PM_{10} concentration using multiple linear regression (MLR) model. The MLR includes MLR1, MLR2 and regression with time series error (RTSE) [27].In another study Jaehyun Ahn et al. (2017) designed a sensors based chip and designed a model on deep learning to analyze the indoor air quality [28]. Remainder of the paper organizes the content as section 2 includes the dataset collection. It describes the IoT arrangements and data collection & visualization. Section 3 depicts the methodology contains model description and role of performance measures for air quality index. Section 4 concludes the result and contains comparison with benchmark model with analysis of air quality index.

2. Dataset Collection

In this research paper, air quality data is measured from temperature and gas sensors [8][9]. Data is collected in the month of July-September 2018 at national capital region New Delhi. This real time dataset which is comprised of values AQI component and ppm concentration of different polluting gases, i.e., carbon monoxide, carbon dioxide, ammonia and acetone. The IoT system helps to collect data with the help of sensors without human intervention. Dataset is collect by an IoT arrangement that comprises of an MQ-135 gas sensor, DHT-22 humidity and temperature sensor and Arduino-Uno microcontroller. Missing data is an important issue in environment data that has to be addressed while taking data from sensors [25]. It is the cause of having insufficient sampling or measurement errors to implement time series prediction. To ensure the correct form of data we adopted to go for regression based imputation using EM algorithm that works between missing data and available data [26].

2.1. Components of IoT Arrangement

MQ-135 Module sensor (Fig. 1) is a low cast sensors and used for smoke and harmful gas (i.e. Ammonia, Carbon Dioxide, Acetone, Carbon monoxide) to monitor different applications. The DHT22 known to measure humidity and low cost digital temperature. The humidity sensors measure the surrounding air and records humidity and temperature for further analytics. Arduino is a combination of software and hardware. It reads input data like sensors light, text data or signals and then producing output using LED.

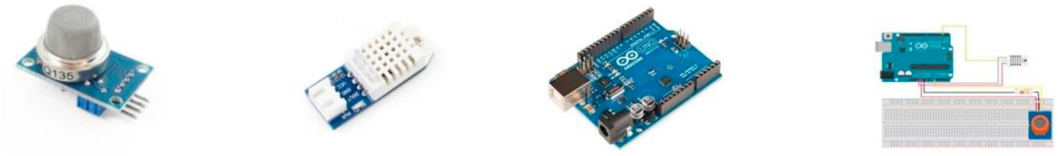

Fig.1. MQ-135 Gas Sensor [8] DHT-22 Humidity-Temperature Sensor [9] Arduino-Uno Microcontroller [10] Circuit Diagram

2.2. Data Extraction

MQ-135 sensor gives an analog value which is required to be corrected by incorporating humidity and temperature values which are collected by DHT-22 sensor. This value is further required to be processed in order to get carbon dioxide, carbon monoxide, ammonia and acetone gas concentrations using ppm vs. Rs/Ro graph for MQ-135 [12]. Following notations have been taken for computing slopes for different gases in the above graph:

Rs = Resistance of Sensor
Ro = Resistance of Sensor in Environment of 100ppm Ammonia at 20°C/65%RH
M = Slope (log-log graph)
X_1 = Initial ppm Value
Y_1 = Initial Rs/Ro Value
X_2 = Final ppm Value
Y_2 = Final Rs/Ro Value

$$M = log\ (Y_2 / Y_1)/ log\ (X_2 / X_1) \tag{1}$$

From Eq. (1) slope of the line of a gas is calculated. It can be now used to calculate the concentration of the gas using the given formula.

$$F_1(X) = (F_0(X) / X_0^M)\ X^M \tag{2}$$

Since *Y* is a function of *X,* therefore we can deduce the following equations:

$$F_0(X) = Y_0 \tag{3}$$
$$F_1(X) = Y_1 \tag{4}$$

Put Eq. (3) and (4) in Eq. (2) and then take inverse of the function to get *X* as the function of *Y*. This gives ppm concentration for a gas.

$$X^M = Y_1 / (Y_0 / X_0^M) \tag{5}$$
$$X = (Y_1 / (Y_0 / X_0^M))^{1/M} \tag{6}$$

X in the Eq. 6 gives the ppm concentrations for above stated gases. Calculation of M and other constants for all the gases lead to the following equations for calculating their ppm concentration.

$$ppm_{co2} = 113.9691 * (Rs/Ro)^{-2.9145}$$
$$ppm_{co} = 660.7628 * (Rs/Ro)^{-4.0043}$$
$$ppm_{ammonia} = 101.7891 * (Rs/Ro)^{-2.4683}$$
$$ppm_{acetone} = 33.7822 * (Rs/Ro)^{3.3785}$$

Weekday is also concatenated to the data to keep record of day on which the data was recorded.

3. Methodology

Data collected by the sensor is cleaned by removing non-numeric values. Box plots of AQI component, carbon dioxide, carbon monoxide, ammonia and acetone are observed and it can be concluded that all five parameters under consideration do not show much variance. So, the mean of these parameters can be used for further analysis. Means of upper stated five parameters are compared with Break Point for Computing AQI in India [1] and ATSDR-Toxicological Profile [2] to determine level of air pollutants for readings of a day. Further, the AQI break points and means are plotted to identify level of pollutants. It is observed that Mean of AQI data shows that it falls in moderately polluted level. Mean of CO2 data shows that it falls in moderately polluted level. Mean of CO data shows that it falls in moderately polluted level. Mean of Ammonia data shows that it falls in moderately polluted level. And mean of Acetone data shows that it falls in good level.

3.1. Data Factored by Days of Week

In order to observe trend in AQI data on the basis of days in a week, a simple points graphs is plotted having reading of each day in which days are on x-axis and corresponding AQI value on y-axis. It can be inferred from the change in slope of blue line which is connecting the AQI values of ten weeks that there is a pattern in changing AQI values over the course of a week (Fig. 2). A conspicuous up and down takes place in AQI values during a week for different weekdays. For any week, Sunday shows the least value of AQI. On the other hand, Monday gives the highest value of AQI. Tuesday and Wednesday show a drop in AQI values corresponding to their previous days. Especially, there is a significant fall in AQI value of Wednesday corresponding to Tuesday. Thursday, Friday and Saturday show a considerable increase in AQI value from their respective previous days. These observations are significant for predictive analysis.

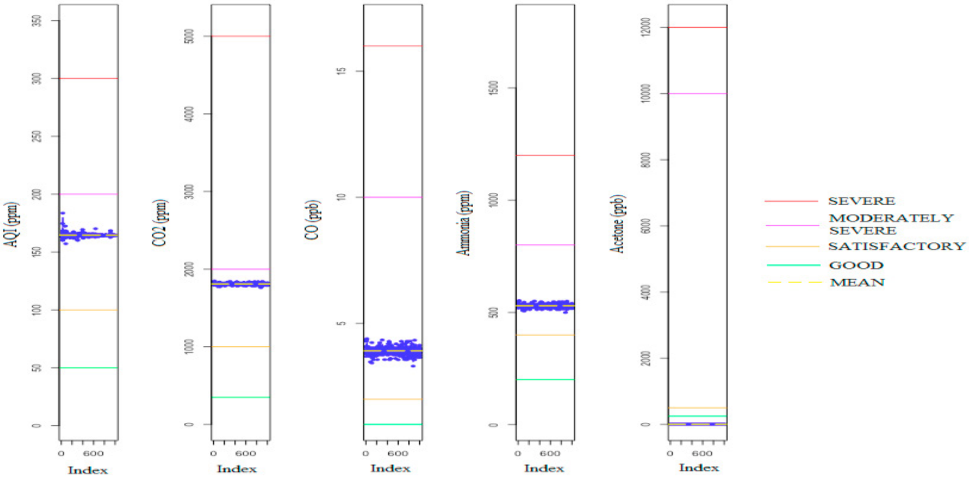

Fig. 2. Level of Pollutants for a Day

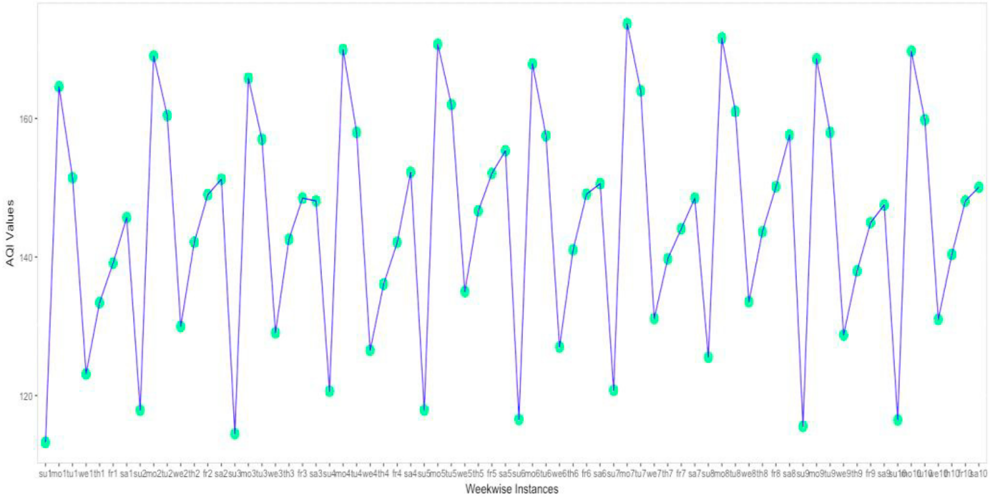

Fig. 3. Trend of AQI values during a week

By observing the AQI trend graph (Fig. 3) can be seen that there is a pattern in changing AQI values for consecutive days over the course of a week. It might be possible that AQI value of a day affects the AQI value of the next day. In order to investigate this fact, the linear regression graphs for each set of consecutive days are plotted. In all graphs (Fig. 4), it can be observed that there is a conspicuous linear relationship between consecutive days. This information can be used to fit the linear regression model for prediction.

3.2. Prediction Model

Linear model is fit for each pair of consecutive days (Fig. 4) to predict the PPM value of a day using the AQI value of the earlier day. Both Training Data and Testing Data are kept completely separate from each other during the entire course of analysis. Here, we use data of Sunday to predict AQI reading for Monday.

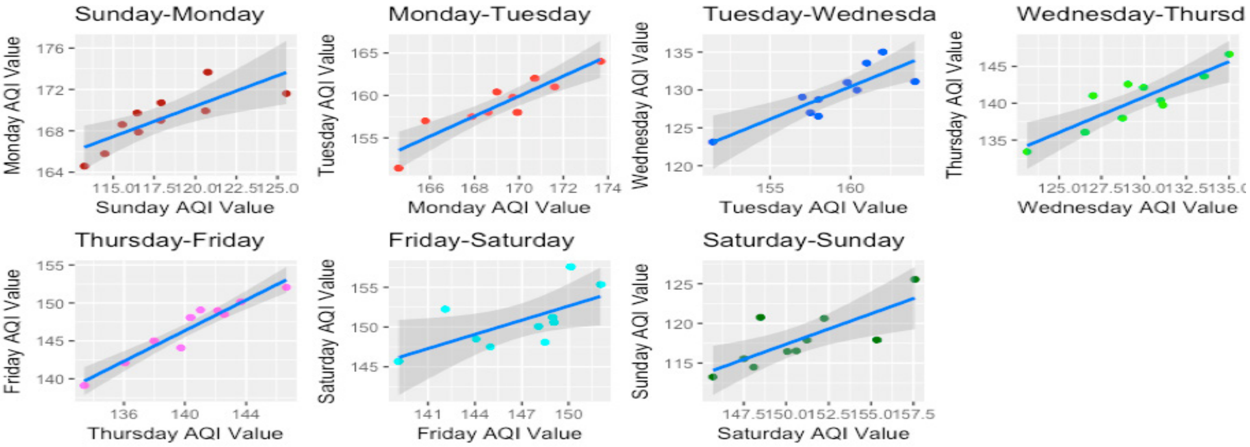

Fig.4. Linear Regression Graphs for Two Consecutive Days

A linear regression line is fitted for the Training Dataset of Sunday and Monday (Fig. 5(a)). This line shows a positive slope which tells that as AQI value increases on x-axis, i.e., independent variable, the value for dependent variable also increases. Difference between the original value *y* and predicted value *yhat* is known as Residual. This set of residuals is observed with respect to training data to look for any anomaly. In case of anomaly, prediction model is cast off. In the given case there is no anomaly in the Residual Observation on Training Data (Fig. 5(b)). Residuals are also examined alongside independent variable to find any pattern. A residual is the difference between the observed value (from scatter plot) and the predicted value (from regression equation line).It shows vertical distance between actual plotted point and the point on the regression line.

$$\text{Residual } (e_i) = X_i - X_i` \tag{7}$$

3.3. Performance Measures

We selected MAE, MSE, RMSE and MAPE as performance evaluator to calculate forecasting error and judge prediction model. As we know MAE, MSE and RMSE are scale dependent measures, based on absolute and squared value while MAPE is based on percentage error and known as scale independent. MAE is known as Mean Absolute Error it gives less weight to outliers.

$$MAE = 1/n \sum_{i=1}^{n} |Xi - X'i| \tag{8}$$

While MSE measurement works on bias and variance for forecasting and referred as Mean Squared Error. RMSE brings the unit back to original unit by taking Root of MSE. RMSE is used to measure the error gap analysis between the actual and estimated values. The Root Mean Square Error (RMSE) for Training Dataset is 1.29 ppm which is small in comparison to dependent variable and therefore, it is acceptable.

$$RMSE = \sqrt{\sum_{i=1}^{n} (Xi - X'i)^2/n} \tag{9}$$

MAPE known as Mean Absolute Percentage Error, which is normalized by true observation and have significant role to bring accuracy in time series data like air quality index.

$$MAPE = 1/n \sum_{i=1}^{n} |Xi - X'i/Xi| * 100 \tag{10}$$

4. Result Representation

It is established that Linear Regression Model can be used for predictive analysis. Linear regression line of training data can also be plotted for testing data to have a pictorial understanding of working of prediction model. The value of Mean Absolute Percentage Error (MAPE) must be measured in range of .05~.09 [18] while observed values of MAPE at different locations (Table-2) are satisfactory improved as compare to its benchmark model proposed by Bing-Chun Liu et al. (2017).

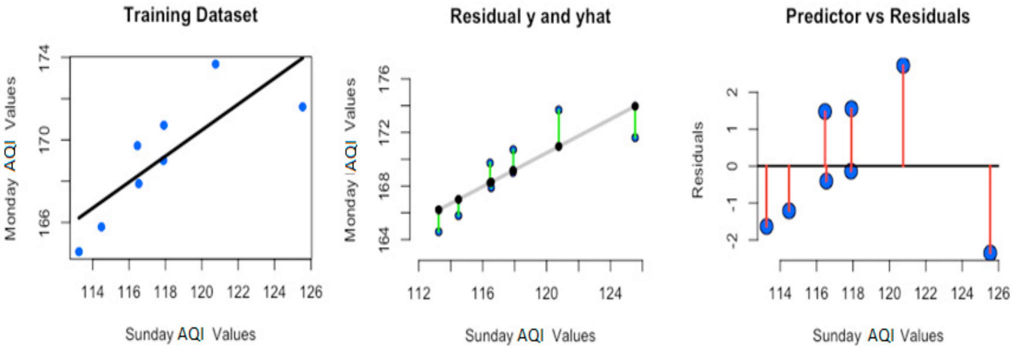

Fig. 5. (a) Fitted Linear Regression Line (b) Residual Observation on Training Dataset (c) Residuals vs. Independent Variable Graph

If a pattern exists then the model is not a good fit. Here, no observable pattern exists in Residual vs. Independent Variable Graph (Fig. 5(c) by fitted linear regression line on training dataset (fig. 5(a), residual observation on training dataset (fig. 5(b) and residual vs. independent variable graph (fig. 5(c)).

Table-2. Results Comparison of evaluation parameters at different locations of Delhi & NCR

Location	MAE	MSE	RMSE	MAPE
Location-1	7.55	115.38	10.48	0.0813
Location-2	7.72	106.25	10.23	0.0765
Location-3	7.44	102.23	10.18	0.0729

Bing-Chun Liu et al. (2017) proposed a model for Urban Air Quality Index (AQI) prediction in China based on Support Vector Regression (SVR). This study is done in multiple cities of china (Beijing, Tianjin and Shijiazhuang) with multi-dimensional regression [18]. The model is considered as benchmark and compared in Table-3 for the all obtained performance measures. Proposed model obtains better result in MSE, RMSE and MAPE.

Table-3. Results Comparison with Benchmark Model

Models	MAE	MSE	RMSE	MAPE
Benchmark Model	7.2675	108.5975	10.355	0.0881
Proposed Model	7.57	107.9533	10.29667	0.0769

5. Conclusion and Future Work

Time series prediction helps in decision making with historical data pattern using air quality index (AQI). Efficient identification of pollution level in living vicinity can help to take appropriate steps for minimizing ppm concentration of gases in the associated area and hence decrease the damage. Pattern of change in AQI level in a week time can be used to predict the AQI level of very next day using AQI level of present day. This paper concludes the arrangement of IOT generated sensor's data of various locations of National Capital Region (NCR).Sensor's data predict air quality of next day using linear regression model. Proposed model obtains forecasting accuracy of different locations as mean absolute error (7.57), root mean square error (10.29) and mean absolute percentage error (0.07). This paper considers the prediction of AQI based on generated data through IOT arrangements. However, there can be other factors like traffic density, dramatic changes in weather, area specific constraints etc which may affect the accuracy of model can be considered to produce more accurate result.

References

[1] Bartra, J., Mullol, J., Del Cuvillo, A., Dávila, I., Ferrer, M., Jáuregui, I.,Valero, A. (2007). Air pollution and allergens. *J Investig Allergol Clin Immunol*, *17*(Suppl 2), 3-8.

[2] Jacob, D. J., & Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric environment*, *43*(1), 51-63.

[3] Beig, G., Ghude, S. D., & Deshpande, A. (2010). *Scientific evaluation of air quality standards and defining air quality index for India*. Indian Institute of Tropical Meteorology.

[4] ATSDR-Toxicological Profile https://www.atsdr.cdc.gov/ToxProfiles/tp.asp? id=5&tid=1 accessed on 20/9/2017.

[5] Wai, W. T., San, W. T. W., Shun, M. A. W. H., Hon, A. L. K., Ng, M. S. K., Yeung, M. D., & Ming, W. C. (2012). A study of the air pollution index reporting system. *Statistical Modeling*, *13*, 15.

[6] El-Bendary, N., Fouad, M. M. M., Ramadan, R. A., Banerjee, S., & Hassanien, A. E. (2013). Smart environmental monitoring using wireless sensor networks. *Wireless Sensor Networks: From Theory to Applications; El Emary, IMM, Ramakrishnan, S., Eds*, 731-755.

[7] Guthi, A. (2016). Implementation of an efficient noise and air pollution monitoring system using Internet of Things (IoT). *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(7), 237-242.

[8] Breuer, D., & Bower, J. (Eds.). (1999). *Monitoring ambient air quality for health impact assessment* (Vol. 85). WHO Regional Office Europe.

[9] Xiaojun, C., Xianpeng, L., & Peng, X. (2015, January). IOT-based air pollution monitoring and forecasting system. In *2015 International Conference on Computer and Computational Sciences (ICCCS)* (pp. 257-260). IEEE.

[10] MQ-135 gas sensor image https://potentiallabs.com/cart/air-quality-control-gas-sensor-mq135 last accessed on 20/9/2017.

[11] DHT-22 Humidity-Temperature Sensor image https://www.aliexpress.com/ item/DHT22Temperature-Humidity-Sensor.html accessed on 12/8/2017

[12] Simić, M., Stojanović, G. M., Manjakkal, L., & Zaraska, K. (2016, November). Multi-sensor system for remote environmental (air and water) quality monitoring. In *2016 24th telecommunications forum (TELFOR)* (pp. 1-4). IEEE.

[13] Yaswanth Sai P., (2017) An IoT Based Automated Noise and Air Pollution Monitoring System. International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Issue 3, March.

[14] Schwartz, J. (1994). Air pollution and daily mortality: a review and meta analysis. *Environmental research*, *64*(1), 36-52.

[15] Briggs, D. J., de Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., & Smallbone, K. (2000). A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of the Total Environment*, *253*(1-3), 151-167.

[16] Juan Carbajal J. ,Hernández, Luis P. Sánchez-Fernández , Jesús A. Carrasco-Ochoa, José Fco. Martínez-Trinidad (2012) "Assessment and prediction of air quality using fuzzy logic and autoregressive models", Atmospheric Environment 60 37e50.

[17] Wang, D., et al. (2016) "A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine", Science of Total Environment, 12,018.

[18] Liu B-C, Binaykia A, Chang P-C, Tiwari MK, Tsao 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.

[19] Bagirov, A. M., Mahmood, A., & Barton, A. (2017) "Prediction of monthly rainfall in Victoria, Australia: Cluster wise linear regression approach". Atmospheric Research, 188, 20–29.

[20] Lee, M., Brauer, M., Wong, P., Tang, R., Tsui, T. H., Choi, C. et al. Barratt, B. (2017) "Land use regression modeling of air pollution in high density high rise cities: A case study in Hong Kong". Science of the Total Environment, 592, 306–315.

[21] Suling Zhu, Xiuyuan Lian, Haixia Liu, Jianming Hu, YuanyuanWang , Jinxing C.(2017) "Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution" 231, 1232e1244.

[22] Jia,Yiwen Zhang J. ,Yuan,Guoming Xu H., Yao Z. (2018) "Imputation of Missing Data in Time Series for Air Pollutants Using Long Short-Term Memory Recurrent Neural Networks", Singapore, Singapore © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-5966-5/18/10.

[23] Box G.E.P., Jenkins G.M., Reinsel G.C. (2008) "Time Series Analysis, Forecasting and Control", 4th ed., Wiley Series in Probability and Statistics.

[24] Zhu, J., Wu, P., Chen, H., Zhou, L., & Tao, Z. (2018). A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. *International journal of environmental research and public health*, *15*(9), 1941.

[25] Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, *14*(5), 853-871.

[26] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895-2907.

[27] Ng, K. Y., & Awang, N. (2018). Multiple linear regression and regression with time series error models in forecasting PM 10 concentrations in Peninsular Malaysia. *Environmental monitoring and assessment*, *190*(2), 63.

[28] Ahn, J., Shin, D., Kim, K., & Yang, J. (2017). Indoor air quality analysis using deep learning with sensor data. *Sensors*, *17*(11), 2476.