How to Cite:
Sudha, G., & Suguna, S. (2022). Health Hazard: PM2.5 forecast - a visual analytic framework using ARIMA. *International Journal of Health Sciences*, *6*(S2), 630–642.
https://doi.org/10.53730/ijhs.v6nS2.5066

Health Hazard: PM2.5 Forecast - a Visual Analytic Framework Using ARIMA

G. Sudha
Part Time Research Scholar, Department of Computer Science, Madurai Kamaraj University, Madurai, Tamilnadu, India

S. Suguna
Assistant Professor, PG and Research Department of Computer Science, Sri Meenakshi Government Arts College for Women (A), Madurai, Tamilnadu, India

***Abstract*---**Health of every living being is dependent on level of pollutant in the environment, weather condition, etc, which are interrelated and influencing each other. Presence of pollution particles in high percentage such as particulate matter of size 10 and 2.5 (PM10, PM25) microns in creates higher risk in lung functionality, respiratory disease and reduces cardio functionality; also affects the quality of water sources. Rainfall, technically known as precipitation, occur at non season also create health hazards, seasonal diseases. In our previous work, the relationship between precipitation and pollution were analysed; it is observed that pollution attribute particulate matter, PM2.5, influences precipitation. Pollution data as well as numerical weather data are time series data in nature. This research work analyses the time series data (day wise observations) of pollution and weather, recorded at Chennai city at the year 2017. ARIMA model is constructed after applying required transformations on underlying data. Trend and seasonality found in underlying time series data set are unique since time series components are data-dependent and autocorrelated with previous day observations. Improved forecasting of PM2.5 and precipitation are obtained through ARIMA modelling. This ARIMA model is used as modelling component in Visual Analytic approach and the model performance, prediction reliability is assessed through visualizations as per visual analytic approach. This work has shown improved results, predictions over the existing work.

***Keywords*---**Health Hazards, pollution, time series analysis, trend, seasonality, PM25, precipitation, visual analytics, Autocorrelation (ACF), Partial Autocorrelation (PACF), transformation, ARIMA, prediction.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Corresponding author: *Sudha, G.; Email: g.sudha79@rediff.com*
Manuscript submitted: 27 Nov 2021, Manuscript revised: 09 Feb 2022, Accepted for publication: 18 March 2022

Introduction

Time series data is series of observations recorded at every specific time interval under the application domain. The interval may be every hour, every six hours, daily, weekly, monthly, quarterly, half yearly and annual. Weather recordings and air pollution observations fall under the category of time series. Such continuous recoded observation highlights the following most important insights and it forms the major component of time series data.

- Trend
- Seasonality

Trend can be identified by observing the upward or downward scenario of recordings found at specific period. Seasonality refers the periodic fluctuations or instabilities identified in the particular short period. For a sample, Online sales has remarkable increase trend during COVID period. In our case, PM2.5 has increased during festival since the excessive firing of crackers. Time series is said to be stationary if the mean and variance are merely constant across the observations. Time series analysis and modelling is applicable if the underlying time series data is stationary. However, the real time series recordings do not possess stationarity property. There are several techniques recommended by researchers to transform the time series data to be stationary.

(a) Time Series Analysis
Every real time recordings tends to be a time series data since the observations are recorded in regular interval period. Time series data hence hold trend, seasonality along with few noises or random or error. Analysing time series data is quite difficult since it is not potentially stationary. Several transformation techniques are available to make the data stationary.
Autocorrelation is the similarity between observations as a function of the time lag between them and this constitutes Auto Regression (AR) component in time series data. The average of recordings tends to vary according to trend and seasonality; this forms Moving Average (MA) component in time series data. Time series can be modeled using moving average model or exponential smoothing or ARIMA.

(b) Visual Analytics
The visual analytic technique is the blend of data modelling and visualizations which enable researchers to identify unidentified insights about underlying application domain. First step in visual analytic approach is to visualize the dataset to get basic statistical insights about attributes. Then in the next phase, the data model is constructed in par with visualizations.
The output given by model is highly evidenced by the visualizations. The data model entertains user feedback or any parameter adjustments; this is once again fed into modelling construct, once again the data model and its visualizations reflect the changes.
This will be repeated until required output is achieved. The visual analytic approach benefits us in fine tuning the output based on visualizations generated in par with model. This research work implements ARIMA model as modelling component and visualizations are generated in parallel.

(c) ARIMA Model
The AR model and MA model are integrated to get Auto Regressive Integrated Moving Average (ARIMA) model. It is well established, result-proven model to handle time series data. Since the time series data at any point depends on previous observations, auto correlation and partial auto correlation is also computed to decide the dependency on the past-observations; this dependency period is technically called as lag, playing a vital role in ARIMA model.
ARIMA model is constructed based on three parameters (P, D, Q); where P specifies the lag order, D specifies the degree of differencing, Q specifies the size of the moving average window and future prediction is represented by equation (1)

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q} \qquad (1)$$

Here, $\hat{y}_t$ is the prediction or estimate on required attribute at time 't', which is predicted using values recorded beyond 'p' time slots; those values are represented as y_{t-p} . ϕ represents the Auto Regressive(AR) coefficients and θ represents Moving Average(MA) coefficients. The term 'μ' represents the constant; term 'e_{t-1}' denotes the noise concern at time (t-1).

Background study

Ha-young kwak et al [7] have analyzed the way in which weather conditions such as precipitation, pollutants have affected the road transport; this work is carried out with IoT sensor data. This work has assessed the direct and indirect effect on transport by investigating correlations among various weather and pollution attributes and summarized their insights. Andreea Badicu et al [8] have presented a research work to forecast the PMx concentration using ARIMA and achieved 89% of accuracy using correction algorithm based on correlation, to adjust the attribute value with pollution monitoring instrument readings. Utpal Barman et al [9] have constructed Auto Regressive (AR), Moving Average (MA) and Auto Regressive Moving Average (ARMA) models on the precipitation data recorded from 1901 to 2017 at Assam and Megalaya state of India. They constructed ACF, PACF and concluded that ARMA model with auto regressive with order eight and moving average of one performs well for underlying dataset. Paul Dankwa et al [10] have forecasted precipitation for 19 years using seasonal exponential smoothing method and compared results with ARIMA model. N. B. M. Khairudin et al [11] have developed Five machine learning models and presented the conclusion that LSTM performs well. They have evaluated their models based on Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Lanyi Zhang et al [12] have constructed ARIMA (6,1,1) model for forecasting PM2.5 of Fozhu, China; They also performed correlation analysis of pollution dataset and presented the insights over weather parameters. They have concluded that the marginal variation in forecast is due to seasonal variation. Afan Galih Salman and Bayu Kanigoro [13] have suggested as a part of weather forecasting, ARIMA gives the best results to forecast the visibility attribute (at airport), with lowest MSE value below 0.003 and coefficient of variation value less than 0.004. S. Mahajan [14] have constructed and discussed about Holt-Winters method, ARIMA model and Neural Network Auto Regressive Model for forecasting PM2.5. Mossamet Kamrun Nesa et al [15] have developed ARIMA model to forecast the number of

COVID victims in categories such as infected, deaths, and recoveries in Bangladesh during October 2021. Sudha G et al [16-19] have analyzed weather and pollution attributes and pointed out the way in which pollution related with weather attributes and insights, results have been presented using visualizations. The survey work [20] has presented the summary of application of visual analytics in health care domain.

Dataset and preprocessing

This research work has been used real time dataset from various real time observation recordings available at portals. The numerical weather observations [5,6] are derived from NOAA and IMD portals observations which are primarily recorded by ground stations. The pollution dataset [1-4] is utilised from portals of Central pollution control board, Department of Environment (ENVIS) – Government of Tamilnadu and Tamilnadu Pollution Control Board and Central Control Room for Air Quality Management.

This work has used the weather and pollution data observed at Chennai city of Tamilnadu, India, recorded in the year 2017. This particular year observation is chosen, since this year predominantly passthrough regular activities, events and the weather, pollution parameters are realistic. Later year suffer from COVID pandemic restrictions and parameters have drastic changes in its observations. The next year, 2018 observations are used for testing the predictions delivered by the proposed model and thus serve the performance analysis. Data derived from multiple sources enables prediction of missing recordings. Statistical methods are used to integrate the recoding and single data set is obtained by combining weather and pollution observations in daily basis, fused on date & time. The correctness of estimate on missing observations is also statistically tested [16].

Proposed work

Once the dataset is constructed the stationary property of forecasting attribute of underlying dataset is tested. The data transformation is also done to make it stationary. After deriving stationary data, the multiple ARIMA model is constructed to predict the precipitation and major pollution role playing attribute PM25.

This work has used augmented dickey fuller test to check the stationarity since it handles larger, more complex models than dickey fuller test. If the p-value of this test is less than 0.05, infers that the null hypothesis (H0) can be rejected and shows the stationarity since data does not have a unit root. The entire research work has been implemented using statistical package R.

Algorithm: Stationarity_Check_VAmodel is developed to test the underlying data is stationary. This algorithm uses visual components as evidence of underlying activities. This algorithm recommends various basic statistical data transformation methodologies such as inverse, inclusion additive components, multiplicative components, logarithmic transformation, differential transformation and combination of transformation methodologies. The inverse functions are applied to get the real value of prediction.

```
Algorithm: Stationarity_Check_VAmodel(D(A))
                    # A is forecast Attribute of Dataset
i. P <- time_series_plot(D)
ii. Check for the time series components
          a. c(trend) <- trend(P)
          b. c(seasonality) <- seasonality(P)
          c. c(observed) <- observed(P)
          d. c(random) <- error(P)
iii. while ( D(A) is stationary ) loop
       a. p_val <- ADF_test (D(A))
           # Augmented Dickey Fuller Test
         b. examine alternate hypothesis
         c. if alternate hypothesis is achieved and
            p_val <= 0.05 then
                   return (D(A))
                    break
            else
                    D(A) <- transform(D(A))
                    Change the lag order
           Repeat loop
```

Algorithm Construct_ARIMA_VA uses visual analytic approach in which recommended statistical test are being visualized through suitable plots. This algorithm uses ACF, PACF (Auto Correlation and Partial Auto Correlation) to show the regression component present at attribute, over the underlying time series, considered in chosen lag. The change of lags and data transformation is repeatedly applied to enhance the better ARIMA model construction. Moreover, the three parameters of the model is also varied to give better prediction. The goodness of models is tested with Ljung box method, to ensure the errors are not correlated or presence of white noise. This also is evidenced by constructing histogram visualization on residual of underlying model.

```
Algorithm: Construct_ARIMA_VA(D(A))
    i. A <- forcasting_attribute (D)
    ii. D1<- Call Stationarity_Check_VAmodel(D(A))
    iii. acf_plot <- (ACF_model <- ACF(D1))
                      # Autocorrelation
    iv. pacf_plot <- (PACF_model <- PACF(D1))
                     # Partial auto correlation
    v. p_val<- ADF_test(M1)
    vi. if  acf_plot and pacf_plot is well fitted
           and p_val <0.01
                    # lags with in reference lines
         Repeat until better model is constructed
              M2 <- ARIMA(P, D, Q)
              Form train_set and test_set from D1
                   # 80% data for training data
              P1_val<-Ljung_Box_test(residual(M2))
              Hist_plot <- Histogram_plot(residual(M2))
              If P1_val < target and Hist_plot is bell shaped
```

```
        Then
          print(predict(M2, test_set))
          print(accuracy(M2(test_data)))
                print(residual(M2(test_data)))
        Else
          Change the parameters (P,D,Q)
Repeat loop
```

Results and Discussion

Various well proven statistical transformations are applied on the input time series data as mentioned in table I and table II; of course the inverse transformations are made during interpreting the future forecast on respective attributes. This research work has constructed several ARIMA models with various transformations on PM2.5 and precipitation time series data; best three among them are presented in table I and table II respectively to forecast PM2.5 and precipitation. The results of Augmented Dickey Fuller test value are considerably very less, lag ranges from 7 to 10 and P-value is less than 0.01 which has proved that the null hypothesis is rejected and underlying data is stationary. The performance of these models is measured using Mean error (ME), RMSE, MAE, Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE) and p_value of Ljung box test as listed in Table I and table II. This work is compared with the research work [8] and the results have been shown in figure 8. The proposed work has used different data pre-processing methodology and data transformation techniques and differs from the existing work.

(a) ARIMA model to predict PM2.5

The figure 1 shows the box plot illustrating the relationship between precipitation and PM2.5; this visualization shows that higher rainfall is evidenced during lower PM2.5 level. Figure 2 shows the time series components of PM2.5 observations recorded throughout the year and it has evidence of both trend and seasonality; a gradual increasing trend is observed from June to October and sudden increase is identified during October-November; sudden decrease is seen at November – December (due to winter season). Performance of model 3 with the transformation (log(log(time series data))) performs well and it is evidenced by figure 3 illustrating the prediction. And residual component of underlying model has produced good bell-shaped histogram visualization as shown in figure 4.

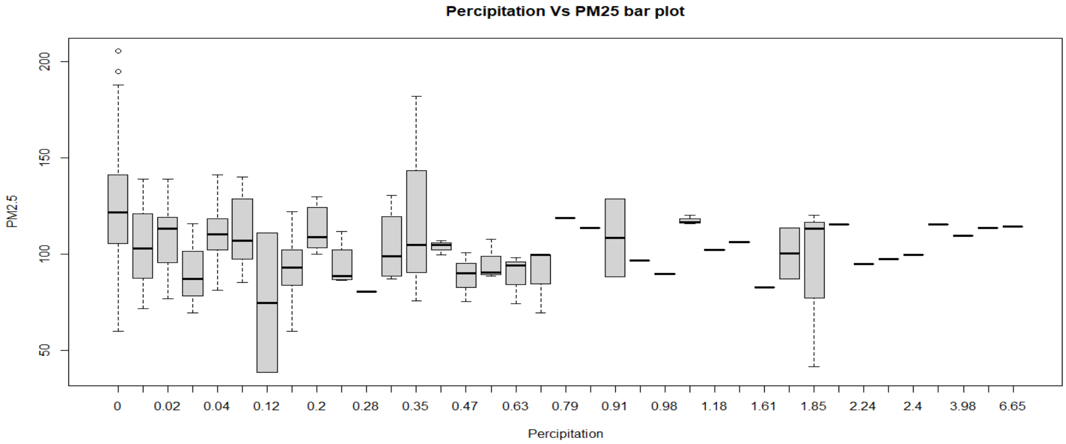

Figure 1. PM2.5 Vs precipitation of observations recoded in the year 2017, Chennai city, Tamilnadu

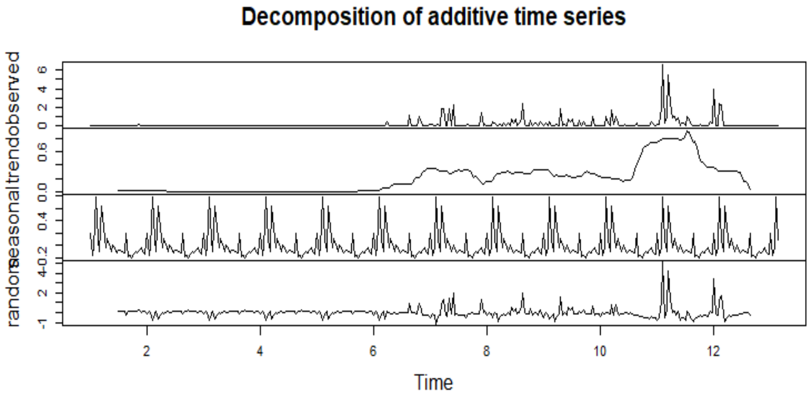

Figure 2. Time series components of PM2.5

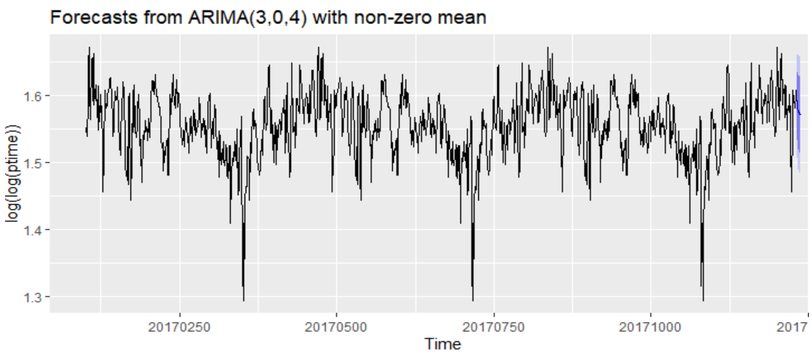

Figure 3. Prediction of PM2.5: ARIMA (3,0,4) model

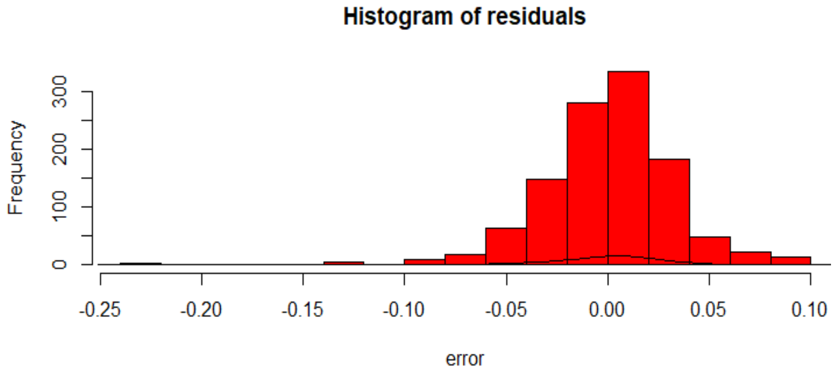

Figure 4. Histogram of residual of PM2.5: ARIMA (3,0,4) model

Table 1
ARIMA model on Pm2.5

Model Transformation	Performance parameters						
	ARIMA (P,D,Q)	*ME*	*RMSE*	*MAE*	*MPE*	*MAPE*	*Ljung box test P value*
Diff(f(x))	(3,1,1)	0.0001	0.15	0.11	-0.1	2.39	0.01
Diff(diff(f(log(x)))	(2,0,3)	-0.0003	0.15	0.11	-0.11	2.37	9.7e-08
f(log(log(x)))	(3,0,4)	-7.9e-05	0.03	0.02	-0.05	1.52	1.5e-07

From the summary results shown by table I, the better one is model with log (log(time_series_data)) transformation, which has lesser Ljung box test value. Similarly, the other performance evaluating parameters such as ME, RMSE, MAPE supports the wellness of the model.

(b) ARIMA model to predict Percipitation

To forecast the precipitation several ARIMA models have been constructed in this research work with various suitable data transformations. Figure 5 shows the observed, trend, seasonality and random components of precipitation time series data. Among the several models, better three ARIMA models along with underlying data transformation, parameter values and its performance are summarized in Table II. From table II, the third model ARIMA (2,1,2) performs well since its prediction rate is 85%, MAPE – Mean Absolute Percentage Error value is 15 (which is less than 20% as recommended by most research works) suggested as better forecasting model. Model-1, ARIMA (5,0,0) with diff(diff(1/1+x))) is also equally good model providing 83% of accuracy and more reliable.

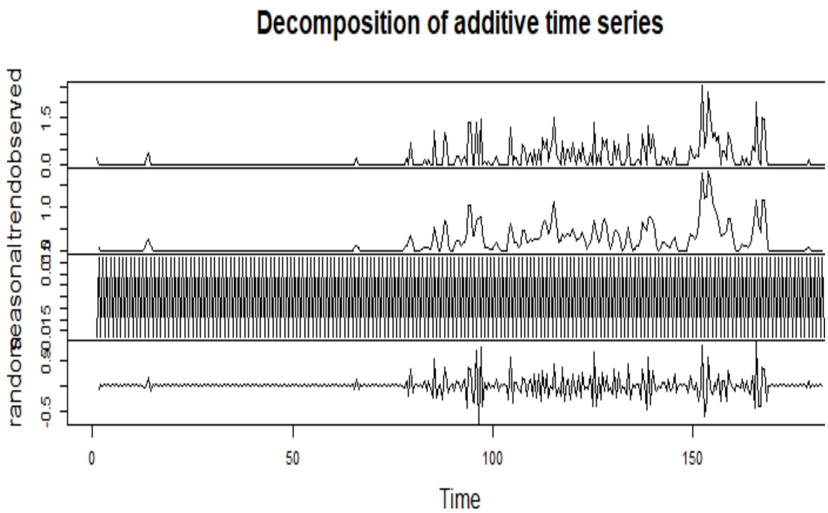

Figure 5. Time series components of precipitation

Figure 5(a) shows the auto regressive and moving average roots are residing within unit circle; this visualization is the evidence of good fitting of ARIMA model. Figure 6(a) shows the autocorrelation lags are falling with in the reference level shown by blue dotted lines, which also evidences the good model fit. Figure 6(b) shows the forecasting of precipitation of next month. Figure 7(a) shows Q-Q plot showing the line merging with the majority of estimated points and denotes the better model for prediction. Figure 7(b) shows the residuals form a bell-shaped histogram around the mean (zero); this visualization also suggests that the underlying ARIMA model predicts well and residuals are highly near to zero.

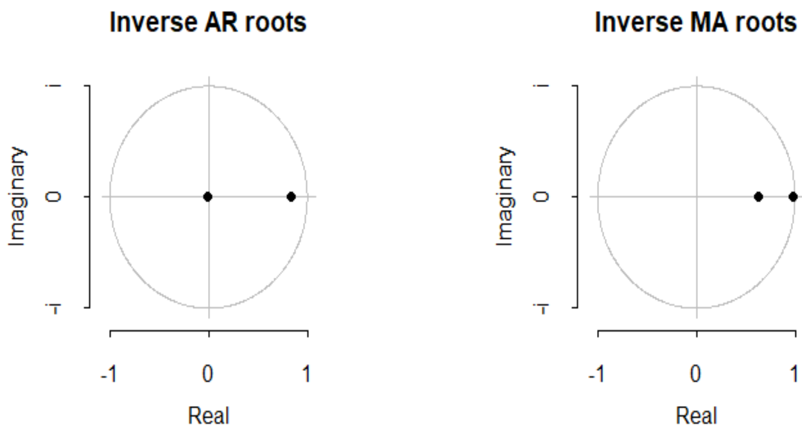

Figure 5a. AR and MA Roots fall inside unit circle: ARIMA (2,1,2)

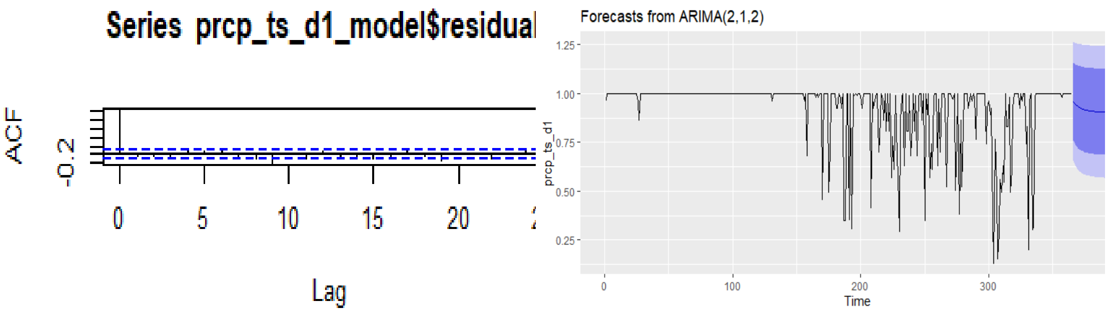

Figure 6. ARIMA(2,1,2): (a) ACF of model residual shows the lags are inside the reference line (b) forcasting of precipitation

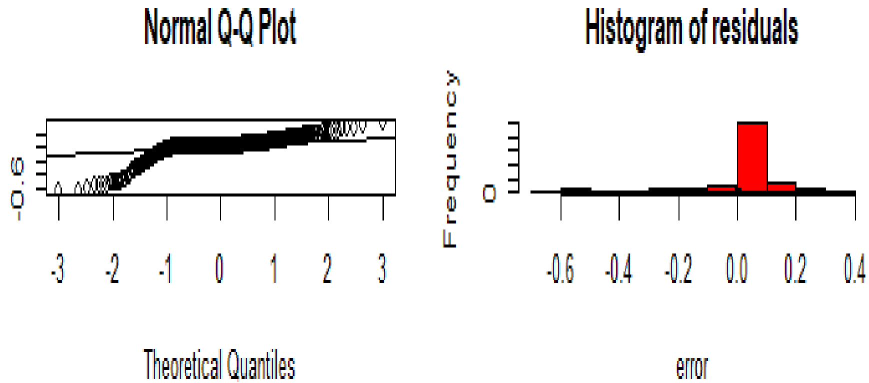

Figure 7. (a) Normal Q-Q plot. (b) Bell shaped histogram of residual of percipitation ARIMA (2,1,2) model

Table 2
ARIMA model on Precipitation

Model Transformation	Performance parameters					
	ARIMA (P,D,Q)	*ME*	*RMSE*	*MAE*	*MASE*	*Ljung box test P value of residual*
Diff(diff(f(1/1+x)))	(5,0,0)	-0.0002	0.17	0.09	0.31	2.3e-08
Diff(log(1/1+x))	(0,0,2)	9.3e-05	0.26	0.12	0.5	7.7e-08
f(1/(1+x))	(2,1,2)	-0.005	0.15	0.08	0.12	2.1e-08

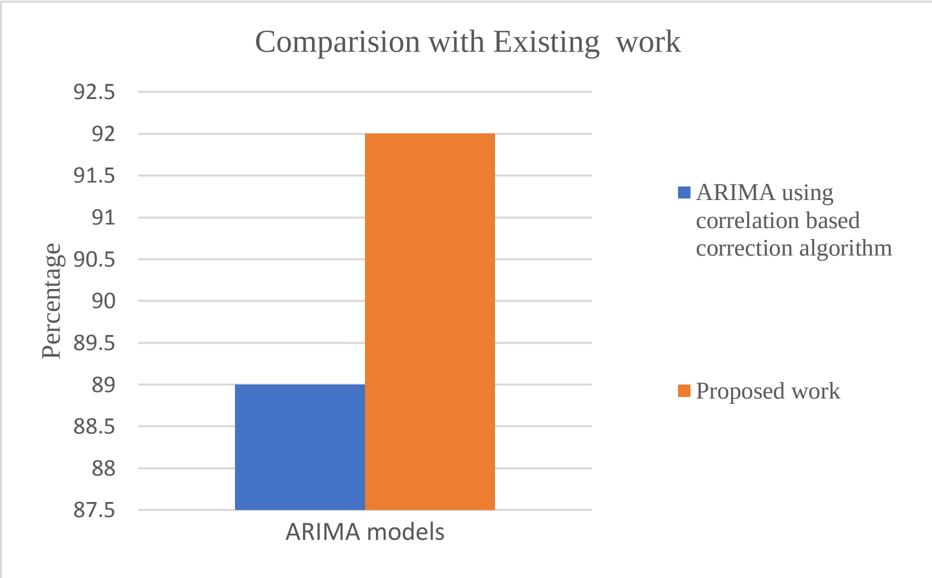

Figure 8. Comparison with existing work

Conclusion and Future Work

This research work has constructed ARIMA model to predict PM2.5 and precipitation. The stationarity of the underlying real time data is evaluated with Dickey fuller test and the models are evaluated through Ljung Box Test for its correctness and reliability. Moreover the output parameters are also validated through most promising error assessment parameters. This research work has adopted several well proven statistical transformation techniques on underlying time series data and brings out better models for the prediction with side-by-side visualizations generated for better understanding. The visual analytic approach enhanced the reliability of the models at each step of entire process by incorporating suitable visualizations. Both numerical output and visualizations collaboratively contributes more concrete insights towards good forecasting.
Even though ARIMA model forecasts well, it lags at assigning weightage to near past observations; if so, the predictions will be more accurate. The future work can be implemented using deep learning techniques to make accurate predictions by assigning weightage on recent observations.

References

[1] http://www.tnpcb.gov.in/air-quality.php
[2] https://smartcities.data.gov.in/
[3] http://tnenvis.nic.in/
[4] https://app.cpcbccr.com/
[5] http://imdchennai.gov.in
[6] https://www.tn.gov.in/
[7] Ha-young kwak, Joonho Ko, Seungho Lee and Chang-Hyeon Joh,"Identifying the correlation between rainfall, traffic flow performance and air pollution concentration in Seoul using path analysis", World

conference on transport research Procedia 25, Elsevier B.V., 2017, 3552-3563.
[8] A. Badicu et al., "PMs concentration forecasting using ARIMA algorithm," 2020 IEEE 91st Vehicular Technology Conference, 2020, pp. 1-5, DOI: 10.1109/VTC2020-Spring48590.2020.9129390.
[9] U. Barman, R. D. Choudhury, A. Ekbal Hussain, M. J. Dahal, P. Barman and M. Hazarika, "Comparative Assessment of AR, MA and ARMA for the Time Series Forecasting of Assam and Meghalaya Rainfall Division," 2020 International Conference on Computational Performance Evaluation, 2020, pp. 507-511, DOI: 10.1109/ComPE49325.2020.9200014.
[10] Paul Dankwa, Eugene Cudjoe, Ebenezer Ebo Yahans Amuah, Raymond Webrah Kazapoe, Edna Pambour Agyemang, "Analyzing and forecasting rainfall patterns in the Manga-Bawku area, northeastern Ghana: Possible implication of climate change" Environmental Challenges, Elsevier, Volume 5, 2021, 100354, https://doi.org/10.1016/j.envc.2021.100354.
[11] N. B. M. Khairudin, N. B. Mustapha, T. N. B. M. Aris and M. B. Zolkepli, "Comparison of Machine Learning Models For Rainfall Forecasting," International Conference on Computer Science and Its Application in Agriculture, IEEE, 2020, pp. 1-5, DOI: 10.1109/ICOSICA49951.2020.9243275.
[12] Lanyi Zhang, Jane Lin, Rongzu Qiu, Xisheng Hu, Huihui Zhang, Qingyao Chen, Huamei Tan, Danting Lin, Jiankai Wang, "Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model", Ecological Indicators, Elsevier, Volume 95, Part 1, 2018, Pages 702-710, ISSN 1470-160X, https://doi.org/10.1016/j.ecolind.2018.08.032.
[13] Afan Galih Salman, Bayu Kanigoro, "Visibility Forecasting Using Autoregressive Integrated Moving Average (ARIMA) Models", Procedia Computer Science, Elsevier, Volume 179, 2021, Pages 252-259, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.01.004.
[14] S. Mahajan, L. -J. Chen and T. -C. Tsai, "An empirical study of PM2.5 forecasting using neural network," IEEE (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1-7, DOI: 10.1109/UIC-ATC.2017.8397443.
[15] Mossamet KamrunNesa, Md. RashedBabu, Mohammad TareqMamun Khan, "Forecasting COVID-19 situation in Bangladesh", Volume 4, Issue 1, pg 6-10, Feb 2022, Elsevier BV, https://doi.org/10.1016/j.bsheal.2021.12.003.
[16] G.Sudha, Suguna Sangaiah, "Analysing the impact of air pollutants on precipitation: a Visual Analytic approach", Journal of the Maharaja Sayajirao University of Baroda (UGC Approved Journal), ISSN: 0025-0422, Volume-55 No.3 2021, pg: 83-88.
[17] G.Sudha G.Arumugam, Suguna Sangaiah, "Weather Analysis and Prediction: A Survey with Visual Analytic Perspective", International Journal of Research and Analytical Reviews, 2018/7, 2348 – 1269, Pp: 66 – 73, Atman Publishing Academy, ISSN 2349 – 5138, UGC recognised journal.
[18] Sudha G., Thangaraj M., Sangaiah S. (2020) "Numerical Weather Analysis Using Statistical Modelling as Visual Analytics Technique", Advances in Intelligent Systems and Computing, vol 1054. Springer, Singapore. https://doi.org/10.1007/978-981-15-0135-7_9
[19] G.Sudha, Suguna Sangaiah, "Insights through Visualizations of Air Quality at Chennai City" , Proceedings of International conference on data science

and information ecosystem'21 (ICDIE'21), ISBN: 978-93-91373-04-7, Pg.135-138.

[20] Kumar, S. (2022). Strategic management of carbon footprint using carbon collectible non-fungible tokens (NFTS) on blockchain. Academy of Strategic Management Journal, 21(S3), 1-10

[21] Kumar, S. (2021). Review of geothermal energy as an alternate energy source for Bitcoin mining. Journal of Economics and Economic Education Research, 23(1), 1-12

[22] Ritika Malik, Aarushi Kataria and Naveen Nandal, Analysis of Digital Wallets for Sustainability: A Comparative Analysis between Retailers and Customers, International Journal of Management, 11(7), 2020, pp. 358-370.

[23] Aarushi, Naveen Nandal, Parul Agrawal. AN EXPLORATORY RESEARCH IN PRODUCT INNOVATION IN AUTOMOBILE SECTOR. JCR. 2020; 7(2): 522-529. doi:10.31838/jcr.07.02.98

[24] Sudha, G. and Suguna, Sangaiah, "A Survey on Contribution of Visual Analytics in Health Care Domain", 2018 IADS International Conference on Computing, Communications & Data Engineering (CCODE), http://dx.doi.org/10.2139/ssrn.3165309