**Time Series Forecasting of Air Pollution Data using ARIMA model**

CHEPPALA SATWIK  
(117EI0429)

ELECTRONICS AND INSTRUMENTATION

NIT ROURKELA

**ABSTRACT**: The Air pollution is a serious concern in All the countries. The problem should be tackled in an efficient manner as All the governments and citizens of the countries are showing high interest for the same. Air Quality index (AQI) is a measure of pollution in air. Due to industrialization and increase in fossil fuels there is a tremendous increase in Air pollution in past few decades. The social and economic activities include transportation, construction...etc. The pollutants such as particulate matter, Nitrogen dioxide, sulphur dioxide is being released in vast amounts into atmosphere. The urban cities are more effected by Air pollution [1]. Use of computational intelligence like machine learning is important for dealing with huge amounts of data. Machine learning approaches are considered as an efficient and cost-effective method to monitor air quality index and are widely used. The general regression model techniques are found to be inadequate and insufficient when dealing with forecasting regular time series models. The present study is mainly focusing on using Box-Jenkins ARIMA (Auto Regressive Integrated Moving Average model) for predicting a pollutant. stochastic ARIMA model has a strong potential for short term prediction. In this study we would be applying time series analysis on data from UCI website. This paper presented a general model of Time series with observed periodicity and seasonality for the data taken from Italy vehicular pollution. The order of best ARIMA model has been found out by carrying out different combinations of Akaike Information’s criterion, Bayesian Information criterion and prediction error along with auto correlation function and partial auto cross correlation function. RIMA model assumes that Time series is Linear and the residual terms should follow a specific distribution known as NORMAL distribution. With help of ARIMA model the behavioural dynamics can be adjusted into single equation.

Keywords: ARIMA, CSS-RME, FORECASTING, Time Series, Box-Jenkins Methodology.

**1.INTRODUCTION:**

All the countries (developing and developed) are facing a serious concern of Air pollution due to rapid technological and industrial development. The health of citizens is at serious risk due to the harmful pollutants being released into atmosphere.

There are many pollutants that causes diseases in humans. Among them, Particulate Matter (PM), particles of variable have very small diameter, penetrate the respiratory system via inhalation, causing respiratory and cardiovascular diseases, central nervous system dysfunctions, and cancer. Ozone(O3) even though it protects from ultra violet radiation, has an enormous negative effect when present in atmosphere[2].It is observed that air pollution displays more intensity in urban areas.[3]The road traffic and construction has shown clear impact on people living in towns and cities rather than villages. The health of existing heart and lung conditions have been detreated more severely[4].It is estimated that 98% of cities in low and middle income countries with more than 100,000 inhabitants do not meet the World Health Organization (WHO) air quality guidelines [5].Increased humidity has shown to make particles heavier causing more density in air and clogged pollutants in atmosphere[6].Many a time it happens that the sensors located at different places are remote locations and data cannot be accessed certain times, so there is a data sparsity and it can be solved by using computational mechanisms.so The air pollution can be predicted by chemical transport methods(numerical),but it is difficult to update source list depending on reliability.[7] Unlike statistical Models. The regression approach methods are also used (most generally Fourier transform) but this technique requires predicted variable.[8] An observed time series has generally two parts: series generated by real process and noise; the elimination of noise is main aim of time series analysis. The selection of model depends on several factors like availability of historic data. The lack of data available may be a main contribution for reduced accuracy in time series analysis techniques. The studies have shown that Box -Jenkins(1994) is the method provided best forecasts for majority of tests[9].the cost is also a leading factor for selection of any model and it is found out that ARIMA is best one[10].In a study after analysing the prediction performances between regression and univariate ARIMA its found that standard error deviation measure is best for ARIMA model[11].

integrated moving average (ARIMA) methodology has been successful in analysing and forecasting the air pollution prediction in the past. It is also used for various other predictions like stock predictions [12], water quality time series predication.[13] For the prediction purpose one or both of two types of models, usually known as structural regression models and time series models are often used in practice. For structural regression methods often, dependent variables are required for estimation. On the other hand, time series

analysis, especially Box-Jenkin type ARIMA models, let the data handle on itself. The future values of time series are predicted using past values and present values. (Box and Jenkins, 1976). Among the stochastic time series models ARIMA types are very powerful and popular as they can successfully describe the observed data in a more user understandable way and can make forecast with minimum forecast error[14].In this model the variable term t is a function of past time lags like t-1,t-2…so the role of predictor variable can be used implicitly using ARIMA stochastic modelling

IN this study the pollutant is predicted using Univariate Arima stochastic modelling method.

**2. Description of Product Model design:**

The System consists of four parts 1) Application layer: This layer is used as customer end data presentation from the data collected from lower layer.

2)Middleware layer (cloud server and Database)-The Cloud data base used in the project is MongoDB which is a NoSQL database supported by PARSE server.

3)Communication Layer (Http)-This is used as a connection between Application layer and sensing layer through middleware layer.

4)sensing layer (rpi + ADC + sensor): The sensing element contains various types of electro chemical sensors which gives analog output.

The processing unit is Raspberry pi3 which processes data with help of an ADC(ARPI-600)

The communication method is NB-IOT and WIFI /Ethernet-IOT module provides facilities like LTE/GSM/NB-IOT.

Power supply unit consists of a step-down circuit, a battery, a charge controller, and power

input to the controller

The proposed software model is combination of cloud data base unit and user interface unit.

1)The sensing layer detect various amount of pollutant values and sends it to cloud where the data would be stored. The techniques mentioned in this study are used for analysing the data and predicting the future values of the same. We can predict the data at various intervals of time like hours, days, months, years.’

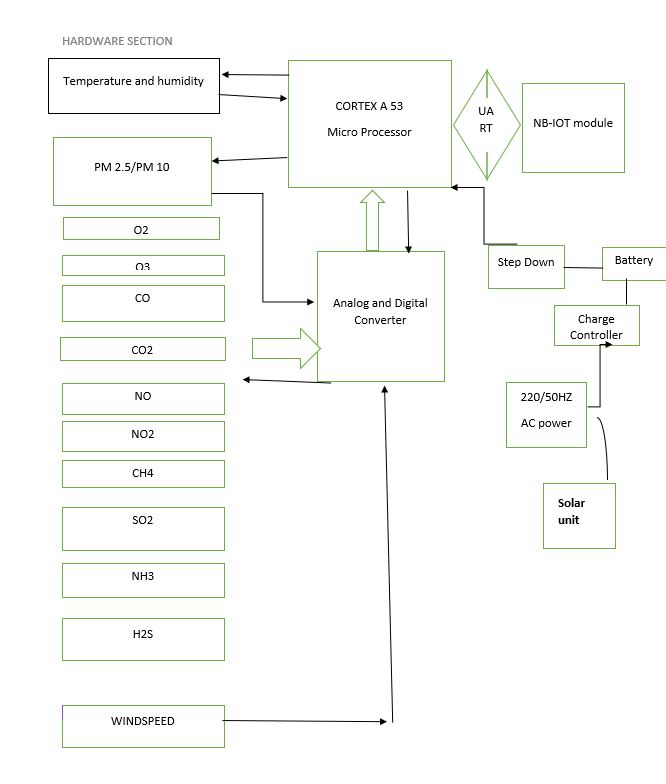


Figure Hardware Diagram.

Application Layer

(Android/WebAPP)

Middleware Layer

(Cloud server and DB)

Communication Layer

(HTTP)

Sensing Layer

(RPI+ADC+Sensor)

Figure 2 Block diagram

**3. FLOWCHART AND ALGORITHM**

ARIMA- these models are normally used to forecast a timeseries (union of set of observations on values that the random variable takes at different time periods) which can be turned into stationary series. A stationary random variable is one whose mean, variance and other statistical properties are constant over time. The correlation with its past values remains constant over time. The ARIMA model is an equation containing lags of variable itself or lags of forecast errors.

Predicted value = weighted sum of past values and/or weighted sum of errors.

The ARIMA model can be classified as Autoregressive model and Moving Average model

In Auto regressive Model The future random variable only depends on its past values

The moving average process only follows random error terms which follows white noise (zero mean, constant variance, uncorrelated random variables) process

Here +

Here the

In general, these are denoted by p, q respectively and d is difference is number of seasonal auto regressive terms and q is number of seasonal moving average terms.

Hence the equation as a whole for ARIMA model is given by

-( +)

1)stationarity: This feature is the initial testing if we have to proceed to predict a time series using ARIMA model. If the series is not stationary that means that statistical properties are changing and there is no particular pattern can be visible and many statistical tests rely on it.

Usually time series, showing trend or

seasonal patterns are non-stationary in nature. In such cases, differencing and power

transformations are often used to remove the trend and to make the series stationary. In this study the test is performed using augmented dickey fuller test (unit root test) and the necessary conclusion can be drawn from p-value obtained. For a p value less than 0.5 is considered to be significantly stationary. In this study we obtained p value as 0.00 from which we can assume series to be stationary. Initially we consider a null hypothesis that series is not stationary then after checking the p value we reject the null hypothesis**. NO**

2)difference transformation:

Stationary?

In case if the series is not stationary, we would be going for differencing with respect to lags of random variable (1,2, 3…so on) so to make series stationary the d parameter in ARIMA (p, d, q) represents these lags.

There are 3 implementations that need to be performed while stating ARIMA mode

1. Model identification: using plots of data of autocorrelation graphs, partial auto correlation graphs and other info, a set of parameter values are initialized for p, q .in our study we obtained p, q values as 5,2 respectively
2. Model estimation: In our study we have used CSS-MLE (conditional sum of square-maximum likelihood estimator) for estimating parameters checking values of AIC (AKAIKE INFORMATION CRITERION) AND BIC (BAYESIAN INFORMATION CRITERION) respectively, for an optimized model, the values of AIC and BIC are as low as possible.
3. Diagnostic checking: The plots of ACF and PACF are observed and we can ensure p, q values as 7,3 respectively.

The block diagram shows the process of fixing ARIMA parameters

Time series

Stationary?

Power transformation differencing

ACF & PACF

**P, d, q**

**YES**

ARIMA coefficient Estimation

Diagnostic Check?

**YES**

ARIMA model

For diagnostic checking following table is considered with respect to identifying whether the model is pure AR or pure MA or ARMA (or ARIMA) model

Table Identification of Model

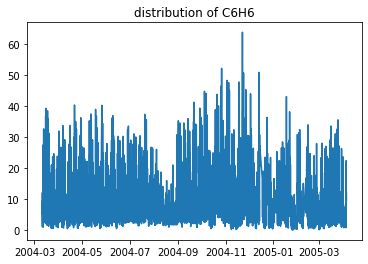
|  |  |  |
| --- | --- | --- |
| MODEL | Auto Correlation Function | Partial Auto correlation Function |
| Auto regression (AR) | Spike decays towards zero | Spike cut off to zero |
| Moving Average (MA) | Spike Cut-off to zero | Spike decay towards zero |
| ARIMA | Spike decay towards zero | Spikes decay towards zero |

**4.A COMPARISION WITH EXISTING MODEL:**

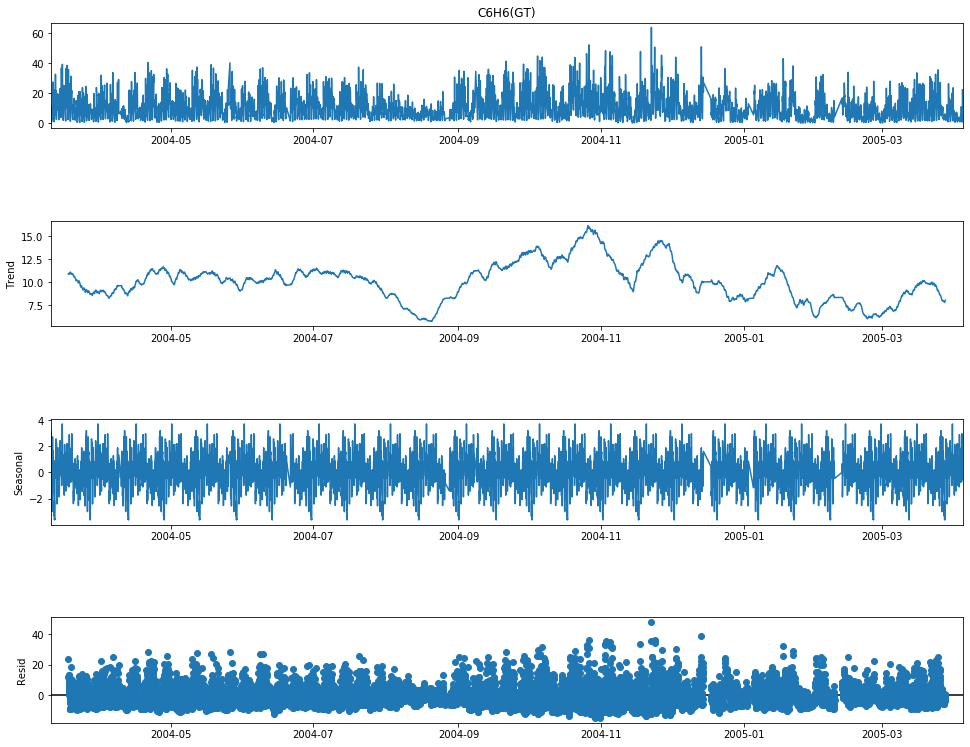
Artificial neural networks are data driven and self-adaptive methods with fewer assumptions. They are good at learning from previously obtained results, so they are not so reliable when there is no specific information on previous standard results. Many a time we don’t have access to accurate results and series is so continuous that only predicting from the series itself turns to be a better choice. Hence ARIMA modelling outdraws the fact that there is no need of a dependent variable and the random variable can safely depend on its past values. Artificial neural networks are mainly relied on cross sectional data, in ANN the modification can be seen in features.as ARIMA is univariate it will not exploit explanatory variables.

**4.B IMPLEMENTATION OF ARIMA**

The vehicular pollution data set is taken from UCI website and it belongs to Italy. The data set is properly cleaned and null values are substituted with a certain lag values for specific variable. Care is taken such that the seasonality is not lost. From the dataset benzene(C6H6) is taken for an example (although any of 13 pollutants can be taken) the data is plotted initially to see the pattern and also with help of seasonal decompose from stats model library.



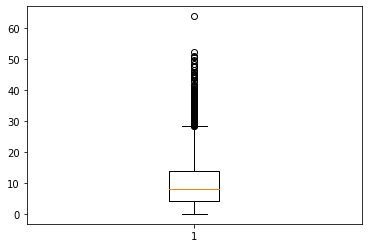
The following shows seasonal decomposition of the c6H6 dataset



The mean, variance, standard deviation is calculated for making sure it isn’t white noise and we can proceed further. As the values are not zero.

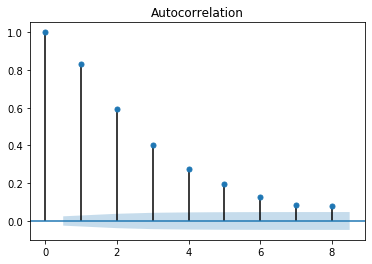
Mean = 10.08310532754978

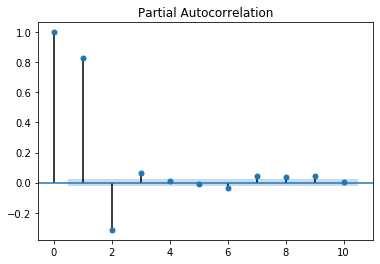
Variance = 7.449405393614615

The boxplot from sea born library is taken to check the distribution of values.

The training and test sets are distribution of 70:30 ratio using data split in python.

The auto correlation and partial auto correlation plots are plotted for finding the values of p, q precisely





Augmented dickey fuller test for stationarity is carried out by using adfuller function from stats model.

As we know more negative is statistic value more likely that we will be rejecting the Null hypothesis (the series is stationary). also, p value less than 0.05 states that the series is stationary and also as statistic value (-8.717916

) is less than 1% value (-3.4313950452870032) so we can confirm it.

The following table shows the values of Akaike Information Criterion(AIC) and Bayesian information criterion(BIC),Root Mean square Error obtained using Conditional Sum of squares-Maximum Likelihood Estimation for ARIMA.For different p,q,d values The values of the AIC,BAE,Rmse.hence the p value is 0.00 there is no need for differencing.also clearly this is not only AR or only MA.so It must be ARIMA

Table ARIMA PARAMATER SELECTION TABLE

|  |  |  |  |
| --- | --- | --- | --- |
| (p, d, q)  Model | AIC | BIC | RMSE |
| (1,0,1) | 35514 | 35541 | 9.05 |
| (1,0,1) | 35487 | 35521 | 9.07 |
| (1,0,2) | 35485 | 35519 | 9.07 |
| (4,0,0) | 35445 | 35492 | 9.13 |
| (5,0,2) | 35431 | 35491 | 8.97 |
| (6,0,2) | 35483 | 35551 | 9.07 |
| (7,0,3) | 35461 | 35542 | 9.09 |
| (2,0,1) | 35487 | 35521 | 9.07 |
| (4,0,1) | 35445 | *35492* | 9.13 |
| (5,0,1) | 35489 | 35543 | 9.07 |
| (5,0,3) | 35478 | 35546 | 9.06 |

From the table we can consider that (p, q, d) of (5,0,2) is most appropriate and it aligns with the plots of auto correlation and partial auto correlation functions as there are sudden fall in values of p, q in

auto correlation and partial auto correlation plots respectively.

**5. RESULTS AND CONCLUSION**:  
For a the ARIMA model designed the with (p, d, q) parameters of (5,0,2). The values are trained using model\_fit from ARIMA module of statsmodels and the predicted values are compared with test values giving a root mean square value of 8.94.

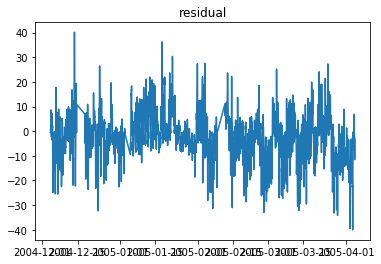
The residual plot (plot for difference of predicted values and actual values) for a proper model should be a normal distribution (random noise) and our resultant output aligns with this statement

Mean of residuals = -2(which is almost close to 0)

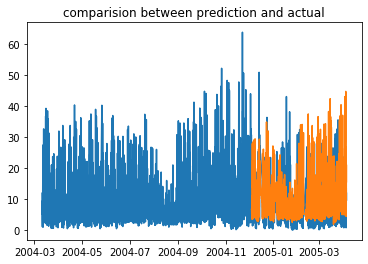
Variance = 75(which is comparatively very much close to 1)

So, it goes with normal distribution of mean =0 and variance =1.

The following diagram shows the residual plot:



The following shows the comparison of predicted and actual values:



**CONCLUSION:** Time series model here is an important tool for observing and control pollutant values hence controlling the air quality condition. It is very much useful to take immediate actions by watching out the data trend beforehand by taking precautionary measures and the expected values can always be handy. The ARIMA model is very much elastic in nature and it can be extended to Seasonal Auto Regressive Integrated Moving Average (SARIMA).And Vector Regressive Integrated Moving Average (VARMA). Hence the Appropriate forecasting model is a prerequisite for developing an appropriate pollution control.

REFERENCES AND CITATIONS:

[1] **J.** Y. Zhu, C. Sun and V. O. K. Li, "An extended spatio-temporal Granger causality model for air quality estimation with heterogeneous urban big data", IEEE Trans. Big Data, vol. 3, pp. 307-319, Sep. 2017.

[2] AUTHOR=Manisalidis Ioannis, Stavropoulou Elisavet, Stavropoulos Agathangelos, Bezirtzoglou Eugenia TITLE=Environmental and Health Impacts of Air Pollution: A Review

[2]missing data: **11.** Y. Zhou, S. De, W. Wang, R. Wang and K. Moessner, "Missing data estimation in mobile sensing environments", *IEEE Access*, vol. 6, no. 1, pp. 69869-69882, Dec. 2018.

[3] **2.** J. Y. Zhu, C. Sun and V. O. K. Li, "An extended spatio-temporal Granger causality model for air quality estimation with heterogeneous urban big data", *IEEE Trans. Big Data*, vol. 3, pp. 307-319, Sep. 2017.

[4] **3.** *Pollution Hotspots Revealed: Check Your Area*, 2018, [online] Available: https://www.bbc.co.uk/news/science-environment-42566393.

[5] World Health Organization, Media Centre (2016). Air pollution levels rising in many of the world’s poorest cities. <http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/>.

[6] Y. Li, Q. Chen, H. Zhao, L. Wang, and R. Tao, “Variations in pm10, pm2.5 and pm1.0 in an urban area of the sichuan basin and their relation to meteorological factors,” *Atmosphere*, vol. 6, no. 1, pp. 150–163, 2015. View at: [Publisher Site](https://doi.org/10.3390/atmos6010150) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Variations%20in%20pm10,%20pm2.5%20and%20pm1.0%20in%20an%20urban%20area%20of%20the%20sichuan%20basin%20and%20their%20relation%20to%20meteorological%20factors&author=Y.%20Li&author=Q.%20Chen&author=H.%20Zhao&author=L.%20Wang&author=&author=R.%20Tao&publication_year=2015)

[7] X. Xi, Z. Wei, R. Xiaoguang et al., “A comprehensive evaluation of air pollution prediction improvement by a machine learning method,” in *Proceedings of the 10th IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2015 - In conjunction with ICT4ALL '15*, pp. 176–181, Hammamet, Tunisia, November 2015. View at: [Publisher Site](https://doi.org/10.1109/SOLI.2015.7367615) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20comprehensive%20evaluation%20of%20air%20pollution%20prediction%20improvement%20by%20a%20machine%20learning%20method&author=X.%20Xi&author=Z.%20Wei&author=R.%20Xiaoguang%20et%20al.)

[8] Badra, Niveen & Sabry, M & Abdel-Latif, Hatem. (2007). Comparison Between Regression and Arima Models in Forecasting Traffic Volume. 1. 126-136.

[9] Layton, Defris and Zehnwirth (1986)

[10] Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis

and Empirical Results Billy M. Williams, M.ASCE1 and Lester A. Hoel, F. ASCE2

[11] Comparison Between Regression and Arima Models in Forecasting Traffic Volume Niveen Badra, Hatem Abdel-Latif

[12] A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, 2014, pp. 106-112, doi: 10.1109/UKSim.2014.67.

[13]A hybrid neural network and ARIMA model for water quality time series prediction-general uses of arima Author links open overlay panel[DurduÖmer Faruk](https://www.sciencedirect.com/science/article/pii/S0952197609001390?casa_token=ucP4-oZAlMgAAAAA:Lfk4Fp8ER13osQHKeWKB5RV6T2f3QUJ_jiFBBXsjk3WBdIgawZ67FRuucVyG1oFAF9irSxuwhYE#!)

[14] Modelling of Rice Production in Punjab using ARIMA Model N. N. JambhulkarCentral Rice Research Institute, Cuttack – 753006

[15] **Seasonal ARIMA for Forecasting Air Pollution Index: A Case Study** 1Muhammad Hisyam Lee, 1Nur Haizum Abd. Rahman,2Suhartono, 3Mohd Talib Latif, 1Maria Elena Nor and 1Nur Arina Bazilah Kamisan

1Department of Mathematics, Faculty of Science, Universiti Technology Malaysia, 81310, Skudai, Malaysia

[16] K. Krishna Rani Samal, Korra Sathya Babu, Santosh Kumar Das, and Abhirup Acharaya. 2019. Time Series based Air Pollution Forecasting using SARIMA and Prophet Model. In *Proceedings of the 2019 International Conference on Information Technology and Computer Communications* (*ITCC 2019*).

[17] Data set: <https://archive.ics.uci.edu/ml/datasets.php>

[18] White Noise Test: detecting autocorrelation and non-stationarities in long time series after ARIMA modelling Margaret Y Mahan\_, Chelley R Chorn‡, Apostolos P Georgopoulos