Multiple regression and multi resolution analysis to identify factors affecting PM10 levels in Bangalore City (Sghalli Area)

Susmith Barigidad, Shashank Prasad, Vinayak Nayak and Sai Tarun Kumar N

https://doi.org/10.22271/maths.2018.v3.i5a.07

Abstract
Bangalore being one of the most growing cities in India has seen significant change after development. Air pollution has tremendously increased and could pose problems if not controlled. The current work focusses on understanding the factors which influence PM10 (Particulate Matter) Levels with focus on Sghalli locality. To understand the factors which affect, multiple regression analysis and multi resolution techniques were applied. Factors which were considered for the study include Sulphur oxides, Carbon monoxide, Nitrous oxides, Wind Speed, humidity, Wind Direction, Solar radiation, O3 and the response variable considered for the study was Logarithm of Pm10 levels.

Keywords: PM10, multiple linear regression, wavelets, backward regression, multi resolution analysis, haar wavelet, air pollution, NO2, NOx, SO2, NO, multivariate, principal component analysis, cooks distance

1. Introduction
The paper focusses on developing regression models for PM10 for Bangalore city. This work primarily focusses on identifying the factors which affect PM 10 for Sghalli location at Bangalore city. The focus is on developing a methodology to conduct regression analysis for air pollution data and accordingly obtaining the factors which affect PM 10. As stated in the [22] report on health effects of particulates PM10 include inhalable particles that are small enough to penetrate the thoracic region of the respiratory system. There are many factors which can affect PM10 which include the various air pollutants such as carbon monoxide (CO), Nitrogen dioxide (NO2), Nitrogen Oxide (NOx), Sulphur Dioxide (SO2). There are also other factors such as Meteorological Factors which include wind speed, wind direction, temperature and humidity. Thus, to develop a multiple linear regression model using these several factors would become complex, it is also important to note that not all factors would be affecting PM 10. To find out which factors play a crucial role backward regression analysis is needed.

Since the air pollution is a time series data and particulate matter may stay for some period, the data would be auto correlated. In such case applying regression models directly to time series data would not provide the correct results, since regression requires samples from a data set which has constant mean, minimal autocorrelation and constant variance. Therefore, it would be necessary to find data set which encompasses it. Several studies have been done on air pollution data and various models such as time series and regression analysis have been applied [18]. Developed an application to forecast the peak ozone levels with the aid of air quality and meteorological variables, in the Greater Athens Area and a linear model was used for regression and prediction of daily air pollution levels [3], focuses on the forecasting of daily Air Quality Index (AQI), based on previous day's AQI and meteorological variables using MLR and PCR model on the seasonal basis for a period of 5 years. The regression models for different seasons were developed using the MLR technique.
based on daily data. Results indicated that the MLR model is performing satisfactorily in all seasons and gives better results used multiple linear regression to investigate the complex relationships between meteorological and time parameters and forecast future PM10 concentrations. A study carried out by to investigate the performance of multiple linear regression method in predicting future PM10 concentration levels in Seberang Perai, Malaysia. This study also found that using meteorological parameters with gases as inputs worked better than meteorological parameters without gases. It also found that multiple linear regression method could be used for long term PM10 concentration predictions after assessing the model performance indicators (NAE, RMSE, PA, IA and R2). In Athens used MLR method to predict hourly PM10 concentrations for the next day and the result showed that multiple regression models can be used to predict PM10 24 hours in advance, conducted a study in Delhi, India and focused on using multiple linear regression models to understand the impact of local traffic flow on ambient concentrations of PM2.5, CO, NO and NO2 at a busy intersection.

2. Methodology

The methodology is shown in figure 1. In the current method Multiresolution Analysis (MRA) has been applied to the time series data and accordingly features have been extracted. The primary purpose of applying MRA is to obtain a data set which would exhibit constant variance and Constant $\bar{Y}$. The reason for choosing MRA is that regression models do not perform satisfactorily when there is change in variance, $\bar{V}$ and the calculation of $\beta_0$ and $R^2$ is based on constant $\bar{Y}$. Also, when the input variables are random variables then regression works on the assumption the response variables and predictor variables are jointly distributed hence the mean variance of Y and X should not change. Once the data set is obtained, the data is checked for autocorrelation, in case the data is not autocorrelated then the data can be concluded as stationary. Then use the sample stationary data, plot the correlation matrix, box plot for checking outliers and accordingly conduct backward regression by considering Variance inflation factor.

2.1 Data Collection

The data required for the study is extracted from the board will monitor the pollutant levels at five places across the Bengaluru but for the study SGhalli and BTM layout have been considered. The data collection is done at every half an hour basis and will be displayed in their website on real time basis. For this study, the daily average values are considered. For BTM layout PM 2.5 was monitored and for the remaining three places SGhalli, PM 10 was monitored.

3. Multiresolution, Statistical Analysis and Model adequacy check

For statistical analysis R and Minitab software were used.

2.1 Data Collection

The data required for the study is extracted from the . The board will monitor the pollutant levels at five places across the Bengaluru but for the study SGhalli and BTM layout have been considered. The data collection is done at every half an hour basis and will be displayed in their website on real time basis. For this study, the daily average values are considered. For BTM layout PM 2.5 was monitored and for the remaining three places SGhalli, PM 10 was monitored.

3. Multiresolution, Statistical Analysis and Model adequacy check

For statistical analysis R and Minitab software were used.

3.1 Multiresolution, Statistical Analysis and Model adequacy check for SGhalli

3.1.1 Stationary Series Identification using Multiresolution Analysis for SGhalli

Since the data shows there is variation in terms of parameters such as mean and variance as seen from the time series plot in Figure 2 and for modelling using regression it was necessary to identify the stationary series a multi resolution analysis (MRA) was done. The MRA was done using Haar Wavelet or DB2 and the scaling and detail coefficients are plotted. To conduct MRA using Haar wavelets, data was chosen based on dyadic length, and two data sets considered include the first 256 data and the next 128 data. Based on these two sets of data decisions were made, the MRA plots are shown in Figure...
Decomposition was done until scale 5, from the MRA plot of first 256 data, it can be seen that at scale 5 and scale 4 where most of the data at scaling level would be decorrelated, in other words a sound extraction of features of the stochastic process happens, it is seen that there is no mean changes from 193 to 256 data points and also looking at the detail coefficients from finer to coarser scale we see spikes at around data point 193 which indicate there would be have been change point, likewise the data from 193 onwards show minimal changes as seen in smaller values of detail coefficients from 193 onwards. Thus, based on this we can conclude the process from 193 exhibits stationary in mean and variance. However, autocorrelation needs to be checked. The autocorrelation plot show that there isn’t correlation, hence it can be assumed that the data is stationary within this region. To test for stationarity Dickey Fuller test and runs test were and is shown in table, it can be concluded that P value is low and hence reject null (Null implies data is Non-stationary) which indicates that data is stationary.

Likewise, it was found for the 128-data set which implies data set from 257 to 384, the data set from 66 to 97 in the graph or in other words from 323 to 354 does not have a change in mean. This was found by considering the original data set, scaling coefficients at scale 5 and based on the corresponding detail coefficients which does not show sudden changes at finer scale. The autocorrelation plot shows the autocorrelation value is less, however Dickey Fuller test was conducted to check for stationarity and the series was found to be stationary.

**3.1.2 Correlation plot for all variables**
The correlation plot in figure 3 for the stationary data set considered from 193 to 256 shows that variables which seem to have relation with Log (PM10) include RH with a negative correlation, NO2, NO, NOx, SR and to an extent Wind Speed. It is seen from the plot of Log (PM10) Vs. NO2, NOx there are two points at the extreme, however this cannot be eliminated as their other variables for which these large
values can be of use. Also, what is noticed after a certain value of NO2, NOx there is linear increase in value which would make more sense when these values are part of it. To check whether these points are influential observation or not, Cooks test can be conducted. For data set from 323 to 354 there is a negative correlation of Log (PM10) with RH, a positive correlation with NO2, NOx and CO and there is almost perfect correlation between NO2, NOx and 88% correlation of NO and NO2 and 93% between NO and NOx.

Fig 3: Correlation Plot for series 1 and series 2 Sghalli

3.1.3 Variance inflation factor (VIF)

It is seen from the correlation matrix NOx is strongly correlated with NO2, NO. Likewise there is strong correlation for NO2 and NO. By including NOx, NO2, and NO in the MLR model there would be issues of multicollinearity, hence including one of them would make more sense to the model, the choice of which depends upon looking at VIF and the dependency of the variables. Here both types of response variables (Log PM10, PM10) are considered and Tables below show that VIF Values and the R square values for both types of response variable [14, 19]. VIF values more than 10 indicates multicollinearity.

It can be seen from the data for both the response variables that the VIF values of NO2, NOx and NO are very high. In this NO and NO2 needs to be eliminated and NOx has been used as an independent variable or predictor variable since NOx represents various oxides of nitrogen. Also, Principal component analysis was conducted among these three variables and the new variable which encompassed the maximum amount of variance was considered as the independent variable. The PCA for these three variables were considered as shown in Table 1 and it is seen that first component encompasses 98.6%. Using this component and the relation between NOx, NO2 and NO, PCA 1 is calculated and used as predictor for Sghalli data. A comparison is done to show that by eliminating NO2 and NO and taking NOx as regressor provides the same result as having PCA1. Either of this sound enough as both fetch comparable results for model fitting. Due to these reasons NOx was used as a regressor variable.

For series 2 Log (PM10) was considered and in an analogous way analysis was conducted. The results are shown in Table 5. After considering VIF NO and NO2 were eliminated and NOx was used. It has been already seen for series 1 that using principle component analysis for these 3 variables and using that transformed variable as predictor is not making a difference in the model hence throughout this paper NOx has been used as the predictor. Table provides the initial results after eliminating of NO and NO2, VIF has been significantly reduced, as well as the R2 values. To obtain the appropriate model outliers, leverage and influential points needs to be looked.

Table 1: Regression Factors Details with all variables with Log PM10 as response variable Sghalli Series 1.
3.1.4 Elimination of outliers, leverage and influential points

Cooks distance and Standardized residuals are used together the remove outliers and leverage points. Any effortless way to eliminate is by obtaining the graphs of cook’s distance and standardized residual and identify points which are large in comparison to rest of points for cook’s distance and combine it with standardized residuals to see whether the residual values are large. Consider Sghalli series for the model which has been built using Log PM10 as response variable. After building the regression model, the cooks distance plot and residual plot is obtained and is shown in figure 4. It is seen from the cooks distance plot that data point 19 had high cooks distance value of 1.575, it was looked into what has caused such a high value and was found that the predictor CO has a very large value, data point at 18 of cooks distance which is have large value of around .193 and a large value of standardized residue value for data point 48 of value 2.4. The new model based on the data is shown in Table 2. It is seen that there is an increase in R square value.

3.1.5 Backward Regression for Sghalli series 1 and series 2

Backward regression for both the series have been conducted and the results have been tabulated in Table 3. To eliminate, large P values had been used and $R^2_{adj}$ was also taken into consideration when eliminating. Typically, $R^2_{adj}$ would increase when eliminating large P values, the iteration is stopped when $R^2_{adj}$ decreases. For Sghalli Series 1 initially Wind Speed is eliminated which has P value of .654 and next step wind direction was eliminated which has P value of .531 and the final model is shown in Table 3. SO2, CO, RH, SR and NOx are identified as significant factors. The regression model is used for identifying the factors and providing the relationship with log PM10. It can also be seen that $R^2_{pred}$ is around 57%, thus this model can be used for prediction.
Likewise, for series 2, backward regression is conducted and accordingly variables are eliminated, the steps of backward regression are shown in Table 4. Initially wind direction was eliminated due to its large p value and then Solar radiation was eliminated due to its large P value. In the next step even though wind direction has a p value of .138, it was not eliminated because when it is eliminated the $R^2_{adj}$ decrease to 63%. Hence the iteration was stopped at this level. The final model is shown in Table 4, factors which affect Log PM10 levels are SO2, CO, RH, WS and NOx. It is to be seen that for both the series the common factor’s which affect Log PM10 levels are SO2, CO, RH, NOX. In both the series CO seems to have a large coefficient value and may a play key role for further study.

### Table 3: Backward Elimination and Regression model using P value and $R^2_{adj}$ for Sghalli series 1

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Constant</th>
<th>SO2</th>
<th>CO</th>
<th>RH</th>
<th>SR</th>
<th>NOx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef</td>
<td>1.8655</td>
<td>-0.011202</td>
<td>0.15643</td>
<td>-0.007244</td>
<td>0.00018643</td>
<td>0.004914</td>
</tr>
<tr>
<td>SECoef</td>
<td>0.1161</td>
<td>0.004198</td>
<td>0.07735</td>
<td>0.001646</td>
<td>0.00009497</td>
<td>0.001774</td>
</tr>
<tr>
<td>T</td>
<td>16.07</td>
<td>-2.67</td>
<td>2.02</td>
<td>-4.4</td>
<td>1.97</td>
<td>2.77</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>0.01</td>
<td>0.048</td>
<td>0</td>
<td>0.054</td>
<td>0.008</td>
</tr>
<tr>
<td>VIF</td>
<td>1.1</td>
<td>2.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Model</td>
<td>Log PM10 = 1.87 - 0.0112 SO2 + 0.156 CO - 0.007244 RH + 0.00018643 SR + 0.004914 Nox</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Backward Elimination and Regression model using P value and $R^2_{adj}$ for Sghalli series 2

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Constant</th>
<th>SO2</th>
<th>CO</th>
<th>RH</th>
<th>WS</th>
<th>NOx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef</td>
<td>1.3292</td>
<td>-0.02406</td>
<td>0.18669</td>
<td>-0.004706</td>
<td>0.09764</td>
<td>0.007047</td>
</tr>
<tr>
<td>SECoef</td>
<td>0.1799</td>
<td>0.01149</td>
<td>0.07146</td>
<td>0.0009836</td>
<td>0.06335</td>
<td>0.002899</td>
</tr>
<tr>
<td>T</td>
<td>7.39</td>
<td>-2.09</td>
<td>2.61</td>
<td>-4.78</td>
<td>1.54</td>
<td>2.43</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>0.048</td>
<td>0.016</td>
<td>0</td>
<td>0.138</td>
<td>0.024</td>
</tr>
<tr>
<td>VIF</td>
<td>1.2</td>
<td>2</td>
<td>1.1</td>
<td>1.9</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>LogPM10 = 1.33 - 0.0241 SO2 + 0.187 CO - 0.00471 RH + 0.0976 WS + 0.00705 Nox</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.1.6 Model adequacy (Residual Analysis)

The check for normality assumption on the residues, independence and structure in residue was checked using Anderson darling test, also runs test above and below mean and plot of residue vs predicted values. For Sghalli series 1 and series 2 the test and the plots are shown below. It can be seen from the plots in Figure 5 for both series that residue follows normal distribution and based on runs test are random and based on plots there is no structure therefore constant variance.
4. Conclusion
From the study for Sghalli factors which affect Log PM10 levels are SO2, CO, RH, WS and NOx. It is seen that there are several factors which affect PM levels, the common factor which affect PM levels for Sghalli area in Bangalore are CO and WS whereas other factors which affect are not prominent. Therefore, to further understand about this study more areas across Bangalore city needs to be looked to get a thorough understanding of impact of these factors. This paper provides a methodology to obtain a sample of data and conduct regression analysis. Also understanding of CO level becomes crucial for the study and monitoring the levels and controlling it would reduce the level of PM values.

5. References
23. CPCB (www.cpcb.nic.in).