Air Quality Improvement via Urban Green Belts

Komal[1] Saloni Gupta[2] Phalguni Goel[17] Ritika Gautam[22] Manpreet Kaur[29]

Indira Gandhi Delhi Technical University for Women, Delhi, India
{phalguni7d,abc11guptasaloni,ritika100898,komalb1409,manpreet0519}@gmail.com

Abstract. With the ever growing increase of pollutants level in air, one of the feasible solutions to combat it is to plant trees.The aim is to find the trees as per the local availability and APTI (Air Pollution Tolerance Index) that could be beneficial in purifying maximum air. For this we have selected some of the states whose cities are major sources of pollution and analysed their air quality levels since 2000 to 2020. Also total tree cover is analysed for the same to interpret an estimate of total requirements in terms of how many and what kind.

Keywords: AQI (Air Quality index) · APTI (Air Pollution Tolerance Index) · Tolerant Species · Green Cover

1 Introduction

Climate change repercussions are faced by everyone in today's world. There is no need to tell that humanity can soon face extinction due to this. Trees which are the soul of our earth are either burned or cut down at rapid pace by us. Though as per researchers trees are no longer enough to revive the poor air quality ,since there is not enough space to plant. But the good news is that young plants are very effective in absorbing pollutants through gas exchange processes. And suggesting the right trees could definitely help in greatest way possible.

Problem Statement. *Studying air pollution levels of certain regions and finding trees that could be planted there to minimise pollution. The goal is to find the common plants available in a region that could provide maximum absorption of pollutants as per worst to best air qualities indications.*

2 Research Questions

1. Was the tree cover as per the AQI of the city is enough?

2. How much tree cover is actually required to get the desired AQI?

3. Is the required tree cover feasible to be planted?

4. What kind of trees could be planted in a region to control the pollution?

3 Dataset

The project is data-driven, the dataset is explained as :

3.1 Dataset Description

We have two categories of datasets:

1) Air Quality Index Dataset
It consists of air pollutant levels for the year 2000 to 2015 state wise [1] and for the year 2016 to 2020, also state wise [2].

2) Trees related Dataset
It consists of Tree Cover state wise [3], Forest Cover state wise [4], Plants available in a state [5] and APTI values of different plants in a region [6].

Datasets are collected from Government sites mainly, as stated, except the APTI values which are referred from a research paper. For the record we have selected 7 states (Delhi, Gujarat, Kerala, Maharashtra, Tamil Nadu, Haryana, West Bengal).
Dataset's data attributes are as follows:

Table 1. Details of Data Attributes of Air Quality Index Dataset.

Data Attributes	**Brief Explanation**
Year	Year whose AQI is mentioned
State	Selected states of India
$SO_2 AnnualAvg$	Average of $SO_2 producedAnnually$
$NO_2 AnnualAvg$	Average of $NO_2 producedAnnually$
PM2.5 Annual Avg	Average of PM2.5 produced Annually
Overall AQI	Overall Air Quality Index

AQI is calculated using Equation 1

$$(\frac{SO_2}{SO_{2am}} + \frac{NO_2}{NO_{2am}} + \frac{PM2.5}{PM2.5_{am}}) * 100 = AQI \quad (1)$$

,where am refers to Ambient air quality.

[1] https://www.kaggle.com/india-air-quality-data
[2] https://www.aqicn.info/historical
[3] http://www.frienvis.nic.in
[4] https://data.gov.in/catalog/district-wise-forest-cover
[5] https://www.biodiversityofindia.org/
[6] http://www.cwejournal.org/vol13no1/a-review-on-air-pollution-tolerance-index–apti–and-anticipated-performance-index–api-/

Table 2. Details of Data Attributes of Tree Cover Dataset.

Data Attributes	**Brief Explanation**
State	Selected states of India
Year	Total tree cover for that year

Table 3. Details of Data Attributes of Forest Cover Dataset.

Data Attributes	**Brief Explanation**
State	Selected states of India
Geographical Area	Total area of the region in km^2
Year and Type	Total Forest cover of that type in the given year

Table 4. Details of Data Attributes of Species Available Dataset.

Data Attributes	**Brief Explanation**
Species	Name of the Plant Species
Common Name	Local name of the Species
States	States in which these plants are found

Table 5. Details of Data Attributes of APTI values Dataset.

Data Attributes	**Brief Explanation**
Year	In which APTI is calculated
Region	Area in which plants are locally available
Species	Name of the plant species
APTI Values	Air Pollution Tolerance Index

APTI values are based upon the biochemical parameters such as ascorbic acid content (AA), leaf Relative Water Content (RWC), leaf extract pH and total leaf chlorophyll (TCh). Its composition is as per the Equation 2

$$\frac{[AA(TCh + pH)] + RWC}{10} = APTI \tag{2}$$

More the APTI value, tolerant the species is to air pollutants making it ideal for green belt and less the APTI value, sensitive it is to pollutants and can only act as bioindicators.

3.2 Data Pre-processing

1) Air Quality dataset is obtained directly in the form of CSV files from the sites mentioned above for the year 2000 to 2020 for respective states.

2) Tree Cover and Forest cover dataset is scraped out from the site using *BeautifulSoup* library. Since the data is available for every alternate year, we average out the next and previous year data to get the missing year values.

3) Using the same library commonly found plant species are retrieved.

4) In the case of APTI values, data is present in the form of a jpeg image. To retrieve it, an api called *OCR(optical Character Recognition) and cv2 library* is used. Using this api we can easily read the text from the image and retrieve it in a string format.

Null attributes are either replaced by average values or their entire rows are deleted. For Outliers we used the Inter-quartile range to detect them and checked for any data value outside the range to filter them out. Also while converting to CSV files there is a problem of variable array length which causes list values of a column to be truncated. To avoid this *zip_longest function of itertools library* is used. It fills the blanks with na.

4 Data Visualisations

4.1 Relationship between APTI and corresponding AQIs

In the APTI values dataset we have gathered APTI values cooresponding to 8 locations which are Andhra Pradesh, Orissa, Uttar Pradesh, Rajasthan, Tamil Nadu, Maharashtra, West Bengal and Delhi, over the span of corresponding different years. The average AQIs of these locations as per the years along with the average APTI values of the plant species available there were calculated to derive the relation among them.

Pearson Correlation Coefficient is calculated among the two which returns a

value between -1 to 1 ,that gives us a interpretable correlation as shown in 3.

$$\frac{covariance(X,Y)}{std(X) * std(Y)} = [-1, 1] \tag{3}$$

, where

$$Covariance(X,Y) = \frac{\sum_{i=1}^{n}(x - mean(X)) * (y - mean(Y))}{(n-1)} \tag{4}$$

The APTI average was calculated using two ways:

1) Normal Average :

$$\frac{\sum_{i=1}^{n} APTIs}{n} = AvgAPTI \tag{5}$$

, where n is the total number of APTIs available in the region.

The Correlation relation value of this comes out to be **0.7901765228495925**

2) Assuming that at an average there are 50,000 trees in one sq km area, we multiplied it with total green cover (GC) to get an estimate of number of trees and the average APTI and averaged it out with the regions geographical area (GA).

The Correlation relation value of this comes out to be **0.7514628614677439**. Both these values indicates a positive correlation. The scatterplot for the same is in fig 1.

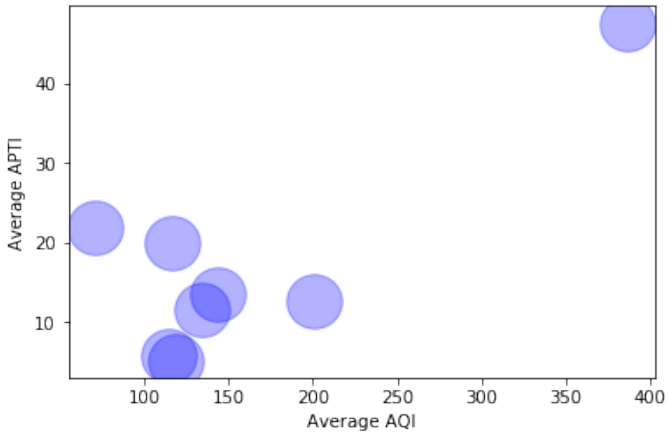

Fig. 1. APTI vs AQI

4.2 Relationship between Green Cover and AQIs

To calculate this we picked the green cover (GC) values of the region along with their geographical area (GA), to average them out and get a percentage as shown in 6

$$\frac{GC * 100}{GA} = GreenCoverPercent \tag{6}$$

Also the average calculated AQIs were taken for the respective regions and their region and the correlation between the two is plotted using eq as shown in 3

The correlation value between the two is **-0.20229019817427843**. This indicates that both these values are negatively correlated. The scatterplot for the same is shown in fig 2

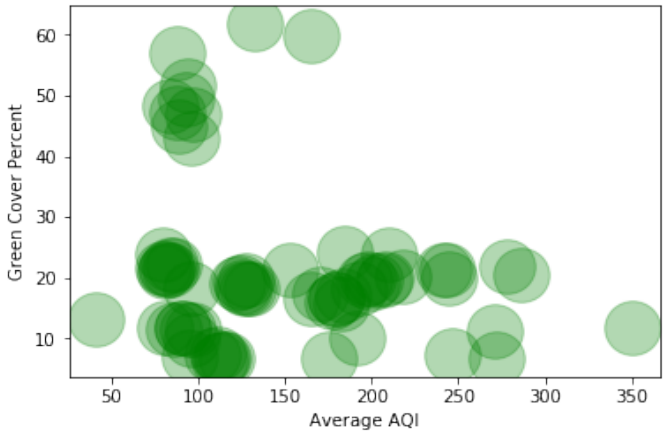

Fig. 2. Green Cover vs AQI

5 Proposed Approach

5.1 K-means Clustering

1) This unsupervised algorithm is used to classify the APTI values of different plant species in separate groups. The elbow point for the K-means is found to be 3.

Thus three clusters were formed whose centroid values are as follows:

c1 = 47.59, c2 = 9.76 and c3 = 21.13
Plants having high APTI values are more tolerant towards air pollution and thus can be used to control the pollution. In comparison, plants having low tolerance index are ineffective towards controlling pollution and can only be used as indicators to the rate of pollution.

Thus on the basis of these centroids we can labels our plant species as **Tolerant, Moderately Tolerant and Sensitive**. The data points around c1 are marked as the Tolerant species, around c2 are marked as Intermediate species and around c3 are marked as Sensitive species which is in accordance with the fact that APTI range of 1-16, 17-29 and 30-100 is classified as intermediate, sensitive and tolerant respectively.

This classification thus further helps us to select Tolerant plant species region wise for controlling air pollution.

2) Next we have done an analysis of aqi and green cover as data points for the year 2001 and 2015 and observed the shift in the cluster to mark how the tree plantation over the year has increased or decreased. The elbow point for the same is shown below in the figure 3 and corresponding k means clusters in the figure 4 and figure 5.

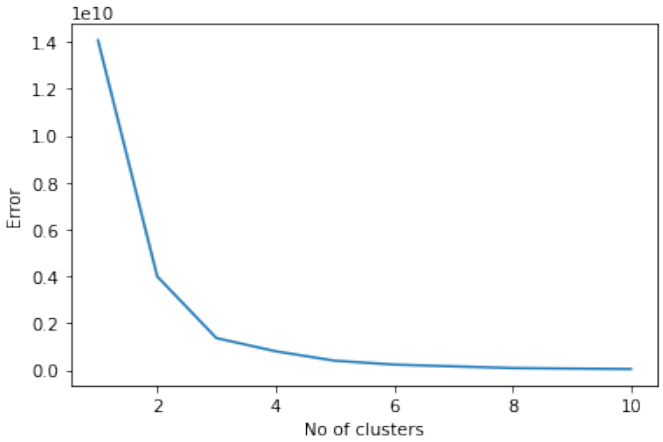

Fig. 3. Elbow Method

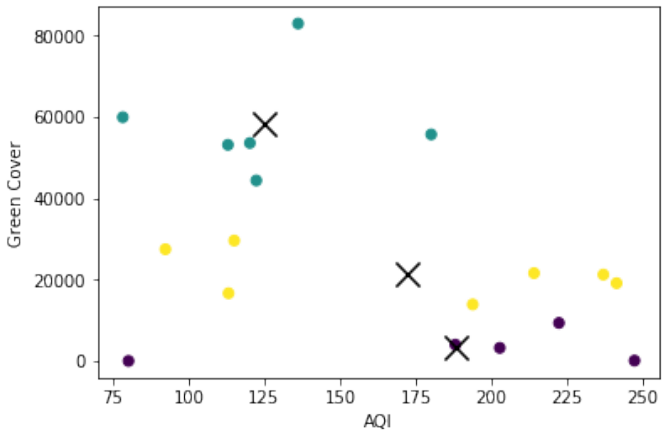

Fig. 4. K-means cluster for year 2001

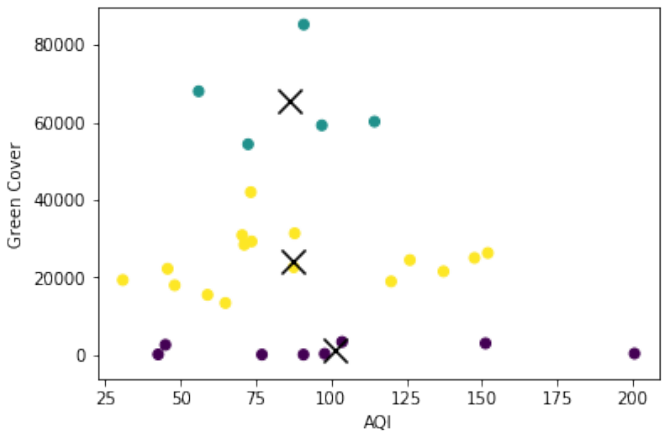

Fig. 5. K-means cluster for year 2015

5.2 Linear Regression

It is a supervised learning algorithm that helps us to predict values y for some variable x given some previous values of x and y. Our goal using the same algorithm is to predict the required green cover given the values of current green cover and their AQIs.

1) We have taken AQI as the x-axis and the Green Cover as the y-axis.

2) To discover the best polynomial fit we ran the values in a 80:20 split and found the corresponding R^2 for each degree. The corresponding figure for the same is shown below in fig 6

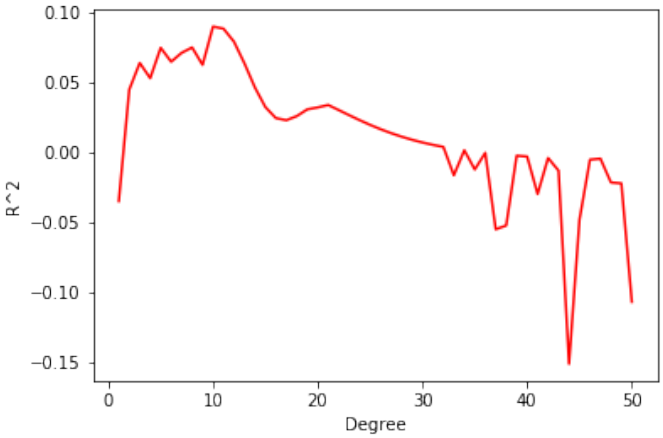

Fig. 6. Best Polynomial fit

The best R-Square value comes out to be **0.08810811402176788** for degree 9, which is calculated using eq 7.

$$\frac{Varianceexplainedbymodel}{TotalVariance} = R^2 \tag{7}$$

3) For various configuration versus accuracy rate, following is the figure we got in fig 7. The accuracy is same as R^2.

4) The Intercept and the coefficients of the model are :

$\beta_0 = 11437.823473508895$

$$\begin{aligned}\beta_i's = [\ & 1.78420966e-08, && 5.09557625e-04,\\ & 8.30627795e-05, && 2.58137237e-03,\\ & -5.32190975e-05, && 4.44086894e-07,\end{aligned}$$

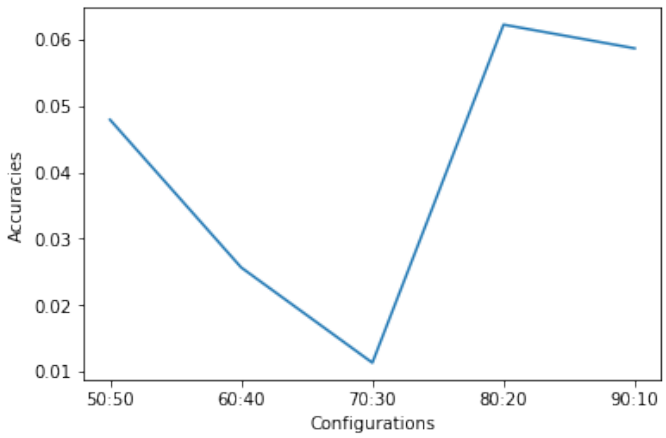

Fig. 7. Configurations vs Accuracy

$-1.86503053e-09$, $3.92290545e-12$,
$-3.29272200e-15$].

$$y = \beta_0 + \sum_{i=1}^{9} \beta_i * X \tag{8}$$

When Intercept and Coefficient is put together in eq 8 we can now predict the value of y based upon x. x here, a single valued attribute is transformed and made multi dimensional as per the requirement of polynomial fit into X.

5) The y predicted as compared to y tested came out is shown in table 6 for few values and corresponding bar graph below in fig 8.

Table 6. Actual Values vs Predicted Values

Actual Values	**Predicted Values**
40718.0	31658.523782
25837.0	9744.792788
3425.0	24455.297877
41765.0	29254.092655
83519.0	28336.148145
84269.0	29818.289454
151.0	9872.319454
28665.0	31792.374326

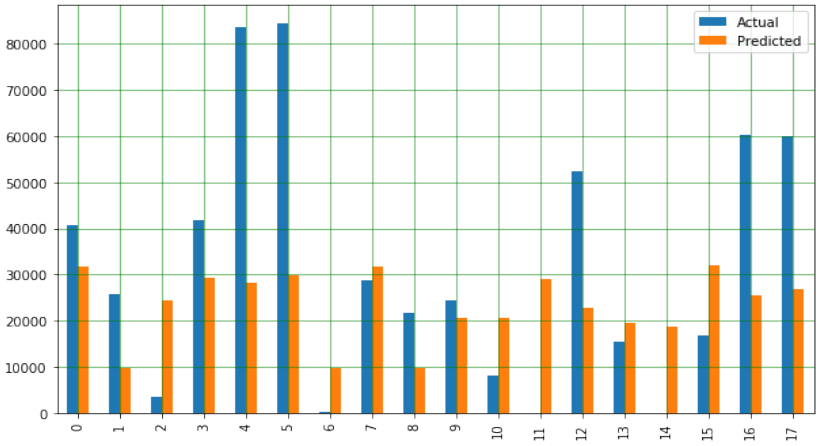

Fig. 8. Actual vs Predicted

6) Regression line for the same is in fig 9

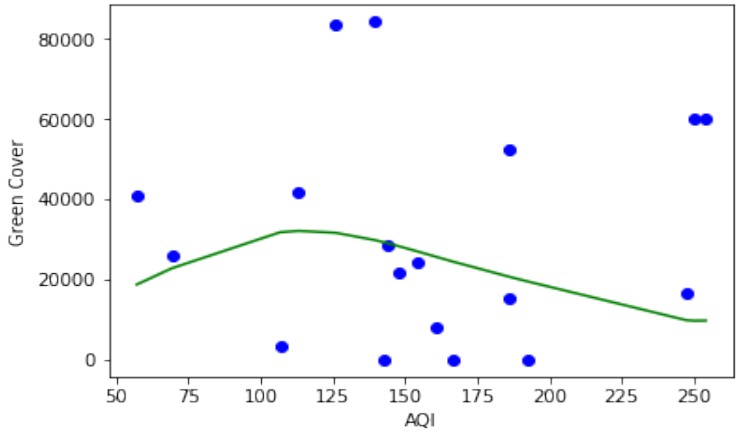

Fig. 9. Linear Regression Curve

6 Results

The Visualisations in the section 4 shows us that more the AQI, more APTI index plant species are required. As higher APTI index species are not only more tolerant towards high levels of AQI but also they are compatible towards

controlling that.
Second Visualisation in the same section indicates a negative correlation, stating that as the green cover of an area increases the AQI level of that area decreases. Thus providing good air quality. Also the data is not highly negatively correlated which further states that to actually moderate out the current AQI a much larger portion of geographical area need to be taken under green cover.

Thus the currently available tree cover is not enough.

In the K-means algorithm we can see the cluster movement of the trees from high to moderate and low regions since 2001 to 2015.
In the linear regression curve, the graph between actual vs predicted shows us the area that have viable green cover as per the AQI and those also that don't have such as compared to AQIs. Thus using these ratios we can actually tell what average amount of green cover would be required as per the AQI.

7 Future Work and Conclusion

1)Enough APTI values and their comparative AQI are not available, because of which proper correlation between the two values can't be devised. With proper set of values this correlation can further assist to suggest effective species.

2)Also the APTI's need to be calculated on the basis of certain biochemical parameters available in the plants, since that information was not available we have taken into account only limited plant species whose APTI was precomputed.

Availability of those parameters can help to calculate APTI of any species and we can tell on the basis of individual plant itself, if it's suitable for controlling that area's pollution or not.

References

1. Begum, A., Harikrishna, S.: Evaluation of some tree species to absorb air pollutants in three industrial locations of south bengaluru, india. Journal of Chemistry **7** (12 2010). https://doi.org/10.1155/2010/398382
2. Panda, L.L., Aggarwal, R., Bhardwaj, D.: A review on air pollution tolerance index (apti) and anticipated performance index (api). Current World Environment **13**(1), 55 (2018)
3. Tripathi, A., Gautam, M.: Biochemical parameters of plants as indicators of air pollution. Journal of Environmental Biology **28**(1), 127 (2007)