Final Report - Team 15: Compiling…

Mansi Gupta, Ria Verma, Raghav Raj Mittal, Arnav Jeurkar

I. Introduction - Motivation

Data from WHO shows that 9 out 10 people breathe air high in pollutant levels. According to data compiled in IQAir AirVisual's 2019 World Air Quality Report, India has 21 of the world's 30 cities with the worst air pollution [20]. The pollution levels in these cities, including the capital New Delhi is expected to worsen given increasing economic activity and heavy industrialization throughout the country. In such a scenario, we were motivated to build an informative tool that could be used to understand the pollution levels of India for past years, as well as make predictions for the future.

II. Problem Definition

Air pollution is a growing concern, especially in growing countries like India. In order to take decisive action and funnel resources, the Indian government first needs to know regions that are most at-risk in order to acquire funding. Furthermore, pollution trends are of great interest to environmentalists and residents of India. Rather than creating a tool that shows India's air pollution data consolidated over different regions over different years, our tool aims to allow the user to see air pollution data in different states spread out over different years. In addition to visualizing the trends of air pollution data in India, we also predict future pollution levels at both city and state level granularity. We hope to make a useful tool that can be used by the Ministry of Environment and Forests and Central Pollution Control Board of India while allocating funds and research resources towards tackling the problem of increasing air pollution in India.

III. Survey

Air pollution kills an estimated seven million people worldwide every year [19], and according to the World Health Organization (WHO), 13 of the world's 20 cities with the highest levels of particulate matter < 2.5 μm in diameter (PM2.5) are in India [1]. According to Balakrishna et al. [8] , nearly 76% of rural households in India are dependent on solid biomass as cooking fuels, which increases household air pollution (HOAP) levels past those recommended by WHO.

In 1987, the US Environmental Protection Agency defined the National Ambient Air Quality Standard (NAAQS) for particulate matter less than 10 μm in diameter (PM10) [9]. This cutoff was based on the similar reasoning that particles of such a size would be deposited in and affect the lower airways and the gas-exchanging regions of the lung [11].

The Air Quality Index (AQI) reports how clean the air in a region is. The AQI is calculated for four major air pollutants as regulated by the Clean Air Act: ground-level ozone, particle pollution (given by a specified PM value), carbon monoxide, and sulfur dioxide. The AQI can take a value of 0 to 500, where a higher AQI value represents a higher degree of air pollution. According to the NAAQS, AQI values of 100 or greater are considered to be unhealthy for sensitive groups of people, and even higher AQI values for everyone [12]. Air Quality Research Subcommittee of the Committee on Environment and Natural Resources CENR reviewed the need for Air Quality Forecasts, existing techniques, understanding different elements involved, ways to improve prediction, and federal programs associated with research needs for air pollution forecasting methods [6]. It summarized current methods for prediction like Classification and Regression Tree (CART), Regression analysis [17], Artificial Neural Networks, Emissions models [18], Meteorological models, Chemical models etc. It familiarizes you with a number of factors important for building predictive models for air pollution forecasts. This report can be used as a guide when deciding on how to proceed with our model building process.

Machine learning approaches to predict air pollution levels by multiple regression and cumulative modelling method have been implemented and described for affordable data collection [5]. Rybarczyk et al. [5] talks about these models that make use of real-time traffic monitoring data to correlate traffic density with air pollution. The model described in this book takes into account chemical parameters along with meteorological factors like RH, SR, Xwind, and T. Both real-time traffic monitoring data and meteorological factors lack in our dataset thus, the method suggested by this book cannot be implemented in our project. The take home from this read was understanding the importance of PM 2.5 concentrations in air pollution prediction. Addressing issues like difficulty in attaining accurate air pollution forecasts and chaotic and non-stationary behavior of data, Zhu et al. [7] talks about the shortcomings of existing model methods like multiple linear models, ARIMA [16] and SVR. It proposes two novel, effective methods to forecast air pollution indexes - EMD-SVR-Hybrid and EMD-IMFs-Hybrid [10] on the basis of AQI data. There is a comparative study highlighting the efficiency and performance of these models over the classical regression techniques. Considering Bai et al. [2] review work on classical and hybrid methods of air pollution forecasting, the article states that traditional AI methods perform much better than statistical methods, but worse than the hybrid models. The article compares the three categories of air pollution forecasting methods: potential forecast model, three dimensional forecast methods [3], and hybrid system in depth, and concludes with advantages and disadvantages of each method.

In conclusion, forecasting air pollution levels in different areas containing different pollutants made the process of building models and the prediction methods much more complex [2]. Therefore, there is a need to choose different forecasting methods based on area and the contributing polluting factors. There is no one best manner to select and predict the air pollution levels.

IV. Proposed method

1. Intuition - why should it be better than the state of the art?

The state-of-the-art visualizations on air pollution range from static visualizations to plots that show air pollution at a country level granularity. Our goal with this project is to create an interactive visualization that empowers the user to learn about different air pollution trends - both past and predicted trends for the future. We are especially excited to create a tool that can be used by environmentalists and government officials because they are a group of changemakers that can especially make a major impact on lowering air pollution. To engage them and create a visualization that is better than the state of the art we have focused on allowing the user to have easy and intuitive access to explore air pollution at various levels of city and state levels of granularity.

While looking through research papers and existing tools on visualizing air pollution trends, we did not come across other tools that visualize at such a fine-level granularity. Our intuition indicates that in order to take meaningful steps towards reducing the ever rising air pollution trends, policymakers need access to air pollution at a local level. For example, after using our tool, a policymaker will be able to pinpoint which village, town or city is causing a lot of air pollution. Furthermore, with our time range selector tool, which enables the user to see the air pollution trends over a time period of several years, the policymaker will be able to proactively see which towns have increased air pollution trends. Also, the prediction model makes predictions about the AQI levels for each city in the years to come, thereby empowering policy makers to identify regions where action needs to be taken immediately and enables them to channel their resources effectively to create the most impact.

2. Description of your approaches: algorithms, user interfaces, etc.

For the visualization part, we plan to create an interactive choropleth map of India air pollution levels of

all the states at a glance. It would show all states of India with their capitals pinned. The states will be shaded based on an AQI legend with darker colors indicating high pollution level and lighter colors indicating lower pollution level, for the given default range of years i.e. 2015 to 2030. The users would be given an option to change the range of years using a slider which ranges from the beginning of available data to 2030. This slider would allow the user to select an average of two years at minimum and select as big a range as desired. When zoomed on a state, cities other than the capital city will be shown. Cities would be clickable which would give a graph which shows a trend of air pollution levels for the selected range of years.

The ARIMA (AutoRegressive Integrated Moving Average) model which we are using comes under the class of statistical models which mainly cater to analyzing and forecasting time series data. We vary the 3 parameters of the model - **p**, the lag order, **d**, the degree of differencing, and **q**, the order of moving average. We will also experiment with a linear regression model to see if it gives more accurate predictions for the pollutant levels in future years. We try to keep the model extremely lightweight so that it can be easily integrated with the UI to obtain quick results, as and when it is requested by the user.

A type of regression analysis in which there is a nth degree polynomial relation between the independent input variable x and the dependent variable y broadly comes under the polynomial regression technique. Depending upon the degree of the polynomial that is fit over our data, the regression will fit a nonlinear relationship between the input and given/predicted output. It is important to note that even if the polynomial regression technique fits a nonlinear model to the data, the regression function is of a linear nature for the unknown variables which are estimated values or predictions. Hence polynomial regression is considered to be a special case of linear regression executing multiple iterations. For this project we have designed a pseudo ensemble technique for predicting the future values of the AQI/pollution indexes. We are using the polynomial regression calculated over multiple degrees and then taking the average of it. This ensures that the predictions have some component of the regular and outlier data points. We have plotted the prediction graph at state level granularity which will be displayed once the user hovers/clicks on the state.

Our main platform will also include a pie chart, showing the different pollutant levels and how they contribute towards the final Air Quality Index (AQI). The aim is to help the users understand the contribution of different pollutants to overall pollution level. We also plan to integrate population levels for each state. When the user hovers over a state, it will show a pop-up containing the average AQI and the population for that state for the selected range of years. The population and air pollution in a region directly shows how many people are being impacted and to what degree. In order to make an interactive map, we used MapBox API which gave us access to the geographical data for the city and state borders of India. We also used the MapBox api to create the map legend and enable different states to be different shades of color - darker color represents more pollution. We used the D3 JavaScript library to load the air pollution data, as well as implement the interactive components of the tool - mouseover functionality, pop-up functionality etc. We got the map data from the Kaggle dataset “India Air Quality Data”. We used CSS to make the web page visually appealing.

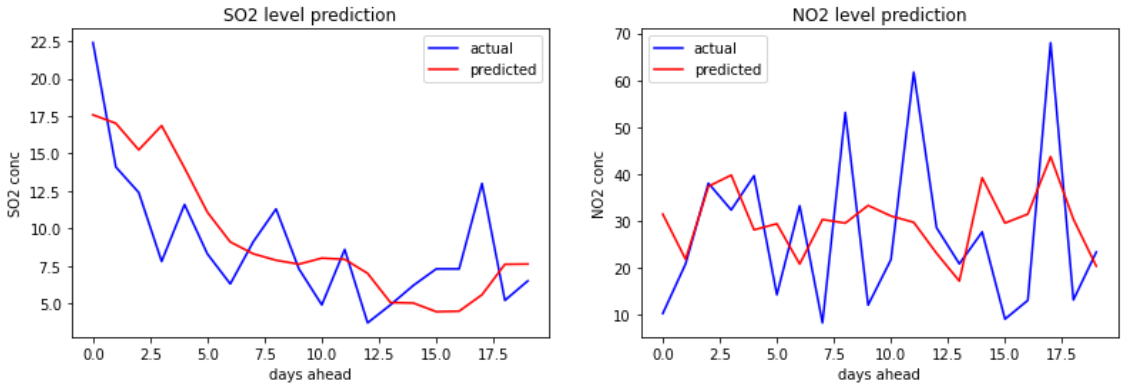
3. Innovations in our Product

1. Since we are including more information i.e. air trends at different levels of granularity, our intuition indicates that we should leverage interactive aspects for the user to better navigate this data in an intuitive manner. Thus, we’ve incorporated capabilities that enable zooming in on cities/states of interest to show pollution stats in depth.

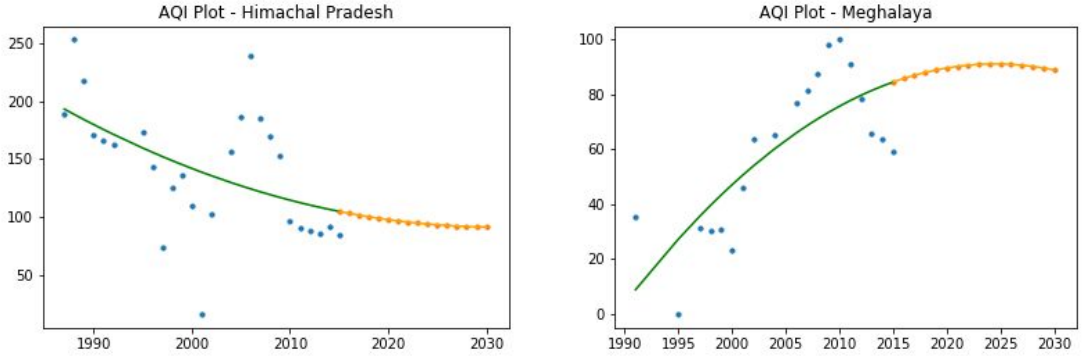
2. We also process numerous data entries to calculate and show stats on average air pollution components - sulfur dioxide levels, carbon dioxide levels, fine particle radius etc. The user has the ability to change the year range, with the default being set from 2010-2015.

3. In order to make our product more educational, we show a pie chart that shows how much each pollutant level contributes to the total AQI. Users can also learn about the causes and sources of that air pollutant by clicking on a particular section of the pie chart, while looking at the impacted regions in real-time.

V. Experiments/ Evaluation



The following two graphs plot the actual vs predicted values from the baseline ARIMA model for two of the pollutant levels - SO2 and NO2. The model is trained using measures from 80 consecutive data points for a specific city (in this case, Hyderabad), and then tested on the next 20 data points. The RMSE for SO2 model is 12.862, a low value given that it is a baseline model. The RMSE for the NO2 model is much higher at 273.386, which could mean that the ARIMA model is incapable of representing the underlying trend in the NO2 levels. We also experiment with linear regression models to see if a better prediction model can be built.



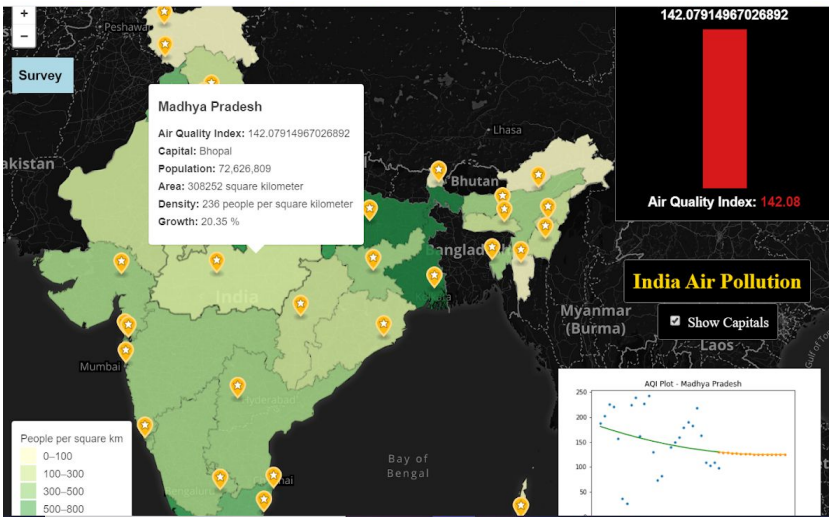
The following two graphs plot the actual vs predicted AQI values from the ensemble of the two polynomial regression models. The model is trained using averaged measures of AQI values for each given year, and for a specific state (in this case, Himachal Pradesh and Meghalaya), and then extrapolated further to get the predictions for 2015 to 2030. tested on the next 20 data points. As we can see, by averaging two models (one linear and one quadratic), we still obtain a quadratic prediction line. While the

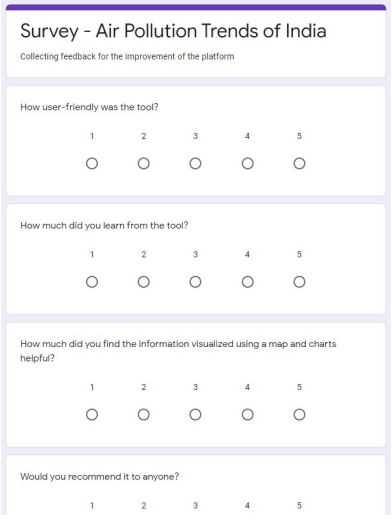
predictions do seem reasonable for future years, it is imperative to understand that these prediction models give us the trend line, and might not accurately capture year-to-year variations. For example, given the current lockdown in India where people are social distancing as a precaution for COVID-19, the AQI level in India has almost halved to a value of 82, as compared to the levels before this year [21].

We have a feedback section that aims to gain information on how helpful our platform was for the target audience. The feedback would be in the form of a small survey with users giving rating as answers to the following questions:

1. How user-friendly was the tool?
2. Did they learn anything new?
3. Did they find the information visualized using a map and charts helpful?
4. Would they recommend it to anyone?
5. Did they find it helpful for all ages of the audience?
6. Did their awareness regarding the air pollution level in India increase?
7. Did it compel them in any manner to change their actions and attitude towards the environment and nature?
8. A free-response comment section, for any additional comments or suggestions

We asked our friends to test this tool and give us feedback on the tool, if they found the tool easy to explore and if the information shown was easy to understand. This was just to get preliminary feedback on our tool so that we could improve it further. We were informed that the tool was very user-friendly, they found some of the features like the bar graph to be easily perceived. This positive feedback helped us move forward with our tool and we added the final touches to it.





VI. Conclusion

Currently, our map is an interactive platform which gives useful information about India's air pollution and population to the user. This integration of these two datasets will give users an intuitive relation between population and pollution of each state. The final tool shows all the states and their respective capital cities along with a pop-up informative tooltip for each state. For each state, this tooltip includes information like AQI, Capital, Population, Area, Density, Growth. When hovered over a state, a bar graph shows the level of the average AQI for that state and a trendline plot shows the AQI levels and how the levels have changed over the past 40 years. The AQI plot also contains the predicted model data till 2030,

which will help the user get a good idea of how pollution levels will increase over time.

The legend of the map shows the number of people per square kilometer. So, a darker green color attributes to a state having a higher population density. We chose to show this metric to help viewers of our visualization tool who are unfamiliar with the different states of India. By understanding how the population densities are spread throughout India, the users get a better understanding of how many people are impacted by the different air pollution trends and statistics displayed. The tool also includes a feedback form under the "Survey" section. This involves questions given in the evaluation section to assess how user-friendly and helpful our tool was for the target audience.

VII. Distribution of Team Effort

All team members have contributed a similar amount of effort.

VII. Citations:

[1] Gordon, T., Balakrishnan, K., Dey, S., Rajagopalan, S., Thornburg, J., Thurston, G., et. al. (2018, October). Air pollution health research priorities for India: Perspectives of the Indo-U.S. Communities of Researchers.

[2] Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. International journal of environmental research and public health, 15(4), 780.

[3] Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of PM2. 5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, *107*, 118-128.

[4] Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE 13(3): e0194889. https://doi.org/10.1371/journal.pone.0194889

[5] Rybarczyk, Y., & Zalakeviciute, R. (2018). Regression models to predict air pollution from affordable data collections. Machine Learning—Advanced Techniques and Emerging Applications.

[6] Ammonia, A. Sources and Fate A Review of Ongoing Federal Research and Future Needs Air Quality Research Subcommittee Meeting Report (Notes from the October 1999 meeting of the CENR Air Quality Research Subcommittee) Prepared by Committee On The Environment And Natural Resources Air Quality Research Subcommittee, June 2000.

[7] Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., & Che, J. (2017). Daily air quality index forecasting with hybrid models: A case in China. Environmental pollution, 231, 1232-1244.

[8] Balakrishnan K, Sambandam S, Ramaswamy P, Ghish S, Venkatesan V, Thangavel G, Mukhopadhyay K, Johnson P, Paul S, Puttaswamy N, Dhaliwal RS, Shukla DK, SRU-CAR Team, 2015. Establishing integrated rural-urban cohorts to assess air pollution-related health effects in pregnant women, children and adults in Southern India: an overview of objectives, design and methods in the Tamil Nadu Air Pollution and Health Effects (TAPHE) study. BMJ Open 5 (6).

[9] Environ. Prot. Agency. 1987. 40 CFR Part 50. Revisions to the national ambient air quality standards for particulate matter: Final rules. Fed. Regist. 52 (126):24634-69

[10] Inman, R. H., Pedro, H. T., & Coimbra, C. F. (2013). Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, *39*(6), 535-576.

[11] Richard B. Schlesinger. 1999. Toxicology of Sulfur Oxides. Air Pollution and Health, pages 585-602.

[12] Wong TW, Tam WWS, Lau AKH, Ng SKW, Yu ITS, Wong AHS, Yeung D.: A Study of the Air Pollution Index Reporting System, Final Report, The Chinese University of Hong Kong, 2012.

[13] Qin, S., Liu, F., Wang, J., & Sun, B. (2014). Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. *Atmospheric Environment*, *98*,

665-675.
[14] Zhou, Q., Jiang, H., Wang, J., & Zhou, J. (2014). A hybrid model for PM2. 5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Science of the Total Environment*, *496*, 264-274.
[15] Grivas, G., & Chaloulakou, A. (2006). Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric environment*, *40*(7), 1216-1229.
[16] Hong, W. C., Dong, Y., Zheng, F., & Lai, C. Y. (2011). Forecasting urban traffic flow by SVR with continuous ACO. *Applied Mathematical Modelling*, *35*(3), 1282-1291.
[17] Sykes, A. O. (1993). An introduction to regression analysis.
[18] Nie, B. (2008). Introduction of Domestic and Foreign Common Air Quality Model. *HAIYANG JISHU*, *27*(1), 132.
[19] https://www.who.int/health-topics/air-pollution#tab=tab_1
[20]https://www.cnn.com/2020/02/25/health/most-polluted-cities-india-pakistan-intl-hnk/index.html
[21]https://economictimes.indiatimes.com/news/politics-and-nation/covid-19-impact-delhi-breathes-third-week-of-clean-air/articleshow/75001969.cms?from=mdr

VIII. Appendix

Final plan of activities:

<table>
<tr><th></th><th>Ria Verma</th><th>Mansi Gupta</th><th>Raghav Mittal</th><th>Arnav Jeurkar</th></tr>
<tr><td>Week 1</td><td colspan="2" rowspan="3">Created basic map;
Added slider for selecting range of years;
Added legend for shaded states;
Added capitals and other cities;
Added the zoom in and out feature;
Created the pie chart for pollutant levels;</td><td colspan="2" rowspan="3">Pre-process the dataset by filling in null values, extracting useful columns, and filtering it based on a given state and city;
Build a baseline ARIMA prediction model;
Calculate AQI value from the available data;</td></tr>
<tr><td>Week 2</td></tr>
<tr><td>Week 3</td></tr>
<tr><td>Week 4</td><td colspan="4">Progress Report</td></tr>
<tr><td>Week 5</td><td colspan="2" rowspan="3">Add the hover over feature;
Add clickable option for each city;
Add bar graph/line chart for average AQI each city;
Integrate the results from prediction model</td><td colspan="2" rowspan="3">Experiment with linear regression models
Obtain state and national AQI values by combining city-level data;
Integrate the results with the front-end;</td></tr>
<tr><td>Week 6</td></tr>
<tr><td>Week 7</td></tr>
<tr><td>Week 8</td><td colspan="4">Final Report and Poster</td></tr>
</table>

Initial timeline (before mid-term progress report):

<table>
<tr><th></th><th>Ria Verma</th><th>Mansi Gupta</th><th>Raghav Mittal</th><th>Arnav Jeurkar</th></tr>
<tr><td>Week 1</td><td colspan="2" rowspan="2">Create basic map and bar graph visualization for historic data</td><td colspan="2" rowspan="2">Pre-process the data and work on a baseline model</td></tr>
<tr><td>Week 2</td></tr>
<tr><td>Week 3</td><td colspan="4">Spring Break!</td></tr>
<tr><td>Week 4</td><td colspan="4">Progress Report</td></tr>
<tr><td>Week 5</td><td colspan="2" rowspan="3">Develop the interactive component and integrate with the results from the prediction model</td><td colspan="2" rowspan="3">Train prediction models and validate results</td></tr>
<tr><td>Week 6</td></tr>
<tr><td>Week 7</td></tr>
<tr><td>Week 8</td><td colspan="4">Final Report and Poster</td></tr>
</table>