A Data-Driven Approach for Region-wise Environmental Health and COVID-19 Risk Assessment Scores

Sanjana Pai Nagarmat, Saiyed Kashif Shaukat
Research & Development Centre
Hitachi India Pvt. Ltd.
Bangalore, India
email: {sanjana,saiyed.shaukat} @hitachi.co.in

***Abstract*—High population density in India has led to rapid influx of citizens into urban cities. With urbanization, cities are facing massive issues in terms of improving citizen health and living conditions, medical infrastructure facilities and overall environmental conditions. The recent COVID-19 pandemic further threw light on the need to solve these pre-existing challenges. Academic scholars, researchers and industries across the country have developed several platforms and smart city applications to help city planners. However, the data sources largely remain segregated. Due to this, applications have restricted ability to provide interpretable results and quantified scores at a granular level. Our paper introduces a novel data-driven approach that helps city planners easily comprehend the current situation and further prioritize action plans to solve urban challenges of environmental health and medical infrastructure facilities in city wards. Data from several sources are combined to provide scores at a granular ward level that is indicative of ward environmental health and ward level risk in situations like the pandemic. Health score provides recommendations to improve the environmental health while the risk score helps in identifying critical zones and predicting the number of active patients in the region. Several Machine Learning based scoring models to predict risk scores for wards are considered. Risk scores are analyzed on a spatio-temporal basis and results are validated with the actual data from Pune smart city. The model introduced in this paper based on Gradient Tree Boosting Algorithm successfully predicts the ward risk scores with 94.55% accuracy.**

***Keywords*—COVID-19; air quality index; tree cover; health score; risk score.**

I. INTRODUCTION

The urban population in India is one of the fastest growing in the world. With more than half of world's population living in cities, there is tremendous attraction and focus towards planning and development of cities. There is increased connectivity as people and cities are getting smarter. However, this rapid urbanization is causing several problems including environmental degradation, excessive air pollution, deforestation, insufficient water availability, waste-disposal problems, poor healthcare facilities and so on. These issues are further strengthened by the significantly increasing population density, unprecedented events and requirements of the urban cities. There is an ardent need for solutions that quantify and provide insights about these issues at a granular level. These solutions can further aid city planners prioritize and resolve the issues faster.

With increased population in India, Health care systems in particular faced tremendous challenges during the *COVID-19* pandemic [1]. Providing medical care to the citizens became challenging due to grappling shortage of medical facilities and infrastructure [2]. The need to primarily prioritize health care facilities in cities became crucial. However, no method provided granular information in the cities to help anticipate the required medical facilities. Also, there were no solutions or applications which helped city planners predict the risk associated with anticipated number of patients in the cities.

The Government of India has already started several initiatives to promote urbanization and growth in a sustainable manner. Smart city project is one such initiative that is focussing on the development of 100 smart cities across the country. Pune, a sprawling city in the western Indian state of Maharashtra has received funding from the Government of India to develop under this initiative. To tackle the problems of urbanization via adoption of next generation technologies and solutions, the *Pune Smart City Development Corporation Limited (PSCDCL)* was formed on 23rd March 2016 in Pune. Together with the *India Urban Data Exchange (IUDX)* [3], an open source data exchange platform aimed at enabling cities to harness full potential of the enormous data being generated in the smart cities, Pune launched the *Pune Urban Data Exchange (PUDX)* [4] pilot and exposed about 850 databases via the IUDX APIs. Apart from PUDX, the *Pune Municipal Corporation (PMC)* [5] has also created the Pune datastore with around 450+ datasets related to varied sources including but not limited to information and technology, environment, transport, healthcare facilities etc. The intention behind this is to mainly promote problem-solving, co-creation among stakeholders to create sustainable outcomes in Pune.

In-spite of several data platforms and data sources being available, the data is segregated and siloed. Data aggregation becomes necessary to gain deeper and meaningful insights from it.

 ISBN: 978-1-61208-863-1

This paper presents a novel data-driven approach that combines several relevant data sources, creates data rich models to improve the environmental conditions of the city, its medical infrastructure and to provide better quality of life to its citizens. The work is based on the data from PMC region in Pune. PMC is the civic body that governs the inner limits of Pune spread over an area of 331.26 sq.km. Based on the 2011 census, the data from 144 wards of the PMC region is considered in this study [6]. However, the same approach can be applied to other smart cities as well. Contributions of this work are as follows:

- **Combining and utilizing rich feature set to provide ward-wise environmental health scores and recommendations**: Rapid urbanization has led to drastic deterioration of Pune's natural heritage. Low-density suburbs rich in greenery have morphed into neighbourhoods congested with high-rises [7]. This has negatively affected environment, climate, overall health and biodiversity in Pune. To help authorities get a wholistic view of the environmental health at a granular ward level in PMC-Pune [8], we provide a data-driven scoring and recommendation model. Environmental attributes like air quality, tree cover and population of wards in the PMC region are used to formulate a score that indicates ward level environmental health. Relevant data features from Pune sources are selected and sub scores are calculated to compute the final health score. Further, based on the health score a recommendation is provided to increase the green tree cover of the region. We also developed an interactive user dashboard to visualize the results.
- **Investigating the environmental attributes, their trends and variations during the COVID-19 pandemic**: Impact of the COVID-19 pandemic over the various environmental attributes including Air Quality Index (AQI), sound, ozone, $N0_2$ and particulate matters (PM_{10}, $PM_{2.5}$) are studied. Variations of these attributes are quantified over time and relevant screens are implemented to help user easily understand and interpret the results.
- **Aggregated data rich models that provides dynamic ward level risk scores**: PMC faced grappling shortage of staff and medical facilities during the COVID-19 pandemic due to lack of planning [2]. A novel data-driven scoring mechanism to indicate ward level risk is implemented. Rich data sources that provides ward level information related to air quality levels, tree cover, population density, hospital and medical infrastructure facilities and pandemic related information are combined, relevant features are selected and latest data available is utilized to formulate the score dynamically. The numerical score provided by the system represents the medical health status of the ward which is indicative of the number of predicted active COVID-19 cases in the ward. The risk scores not only indicate the ward status but can also be used to compare wards and gain deeper insights.

Overall the work in this paper provides granular ward level details and localized information dynamically considering the change in the attributes. The scores proposed in this work can help provide adequate information to concerned authorities to plan and implement control measures. It will also guide them to prioritize and identify potential high-risk wards. Further, map-based representation of wards to effectively provide spatio-temporal information about ward health, environmental attribute variations and risk levels is provided to relay information in an easily understandable manner to the users.

The remaining part of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes the technical information pertaining to the data sources, score computations, recommendations, detailed analysis with suitable graphs and plots. Section 4 presents the results and discussions while Section 5 describes conclusion of our approach.

II. Related Work

Related works that analyse the environmental attributes like the air quality index and the tree cover independently in Pune exist. These works separately analyse the attributes and present approaches to optimize the environmental conditions. Kane et al. [9] studied the scope and opportunity in Pune to understand the green cover for maintaining a balanced floral diversity. While National Research Development Corporation [10] studied the air quality in Pune to develop an air information response plan. These previous approaches study the environmental attributes independently and do not provide any kind of scoring mechanisms to quantify the findings. In our work, we provide a wholistic view of the environmental health at a *granular ward level* in Pune. We formulate a relative health score for the ward by combining several attributes – *air quality index, tree* and *human population*. Morani et al. [11] provide a planting index for a location by analysing the associated tree cover. It makes use of Geographic Information System (GIS) based visual approximation techniques to estimate the green tree cover. Our work utilizes the air quality, population and the surveyed tree data from PMC-Pune to provide an index that indicates the ward health. The publicly available geo-sensor based surveyed tree information is used in the calculation of green tree cover rather than the approximation techniques used in [11]. We then aggregate the tree information at a ward level, combine several tree attributes to score the trees and utilize this sub score in computing the environmental health score. Based on the ward health, our method goes one step ahead by providing recommendations on how the score can be used to further improve the green tree cover in the area.

Sharma et al. [12] studied the effect of restricted emissions during COVID-19 on air quality in India. In our work, we investigate the trends of air pollutant concentrations at a granular ward level in PMC-Pune during the COVID-19 pandemic. Such granular details at ward level will help ward authorities take specific actions on areas under their jurisdiction.

Yaro et al. [13] studied and evaluated the impact of some selected demographic and environmental variables to identify potential risk areas and hotspots for COVID-19 transmission. The primary focus of this work was to identify statistically

 ISBN: 978-1-61208-863-1

significant hotspots for further transmission of COVID-19 specifically in Nigeria. In our work, we make use of the ward level medical infrastructure details, ward demographics, environmental variables like air quality, tree data and the existing hotspots information to collectively provide a single numerical indicator of risk at a ward level. Our work is based on data from PMC-Pune region.

III. Methodology

This section provides details on the data sources, formulation of scores along with dashboard visualizations.

A. Environmental Health Score Formulation

1) Data Sources: To provide a relative measure of understanding the environmental ward health, health scores are computed by combining air quality, tree details and population data.

- *Air Quality data* [14]: The PUDX platform provides data related to the air quality sensors installed in several parts of Pune city. 50+ sensors installed across the city actively measure the concentrations of pollutants. The sensors provide AQI readings every 15 minutes. A pipeline is created and a job to collect this data on a regular basis is scheduled. The collected sensor data is then mapped to the appropriate wards.
- *Pune Tree Census data* [15]: PMC has mapped over 40 lakh geo enabled trees using GIS technologies and made data available to the public. It includes data of tree attributes like height, girth, width, botanical names, ownership details, location information and so on.
- *PMC Population breakup of administrative wards* [16] : Census is conducted once in every ten years in India. It provides information at the root level. The Pune city census office is responsible to carry out the census activities effectively and efficiently in the PMC area. The administrative ward level detailed census data in Pune is made available to the public by the PMC.

2) Score Formulation: Health score is formulated using air quality, population and tree data. The trees in the PMC-Pune region are studied. Based on their importance and contributions towards the environment, individual tree scores are calculated. From the literature survey, we found out that evergreen trees planted in rows can capture up to 85% of the particulate air pollution blowing through their branches. In addition, a single tree produces nearly three-quarters of the oxygen required for a person and a canopy of trees in an urban environment can slash smog levels up to 6% [17]. We also studied the local tree species, their ideal height and girth requirements to come up with weights that will help in scoring the individual trees. Based on our investigation, four important tree attributes i.e. height of the tree (in meters), its physical condition (poor, average, healthy), phenology (seasonal, evergreen) and its canopy diameter (in meters) are selected and used to score each tree (Tree_Value).

$$Tree\_Value = w1(Height) + w2(Canopy) + w3(Phenology) + w4(Condition) \quad (1)$$

Here weights w1, w2, w3 and w4 are calculated with respect to Table I. We have assumed that taller and healthier evergreen trees with wider canopies provide more contribution towards a green buffer zone. Also, evergreen trees are more economical requiring less maintenance when compared to seasonal trees.

The trees are mapped to appropriate wards based on geo coordinates. According to an Indian Institute of Science (IISc) report [18] from 2014, the ideal tree-human ratio should be seven trees for every person. With this reference, if a ward has **n** trees we calculate the Tree_Score as a ratio of sum of all the individual tree scores in the ward to the ideal tree score (Ideal_Tree_Score).

$$Tree\_Score = \frac{\sum_{i=1}^{n} Tree\_Value(i)}{Ideal\_Tree\_Score} \quad (2)$$

$$Ideal\_Tree\_Score = Max\_Tree\_Score * Ward\_Population * 7 \quad (3)$$

Max_Tree_Score is the maximum score that can be assigned to a tree based on Table I. The upper bound for the Tree_Score of a ward is set to 1. In several smart cities, health check for the trees are done by authorities on a regular basis. They periodically make a note of tree health and changes in their conditions if any. Such periodically updated information can be used in calculating more appropriate tree scores.

The AQI score is calculated by utilizing the dynamic AQI values provided by PUDX every 15 minutes. For wards with sensors installed, we take an exponential moving average [19] (decay constant is calculated based on the number of AQI observations in a span of one month) of AQI values for individual sensors, calculate the mean of this value for all the sensors available in that ward and then normalize the mean value with respect to the maximum AQI value (considered as 500).

$$AQI\_Score = \frac{(Max\_AQI\_Value - AQI\_Value)}{Max\_AQI\_Value} \quad (4)$$

For wards where air sensors are not available, we have used approximations to get the AQI value at a ward level. We identify five nearest sensors (based on PMC sensor locations, five sensors are available on an average in the vicinity of a ward) to these wards, take an average of their AQI value and assign it as the AQI value of the ward. Geo-coordinates and spherical boundary calculations are used here to identify the nearest sensors [20].

To account for the increase in population over the years, in our work we have studied the average population growth in Pune and made interpolations to the actual population census data collected in 2011. For calculating the population score, we first calculate the population density of the wards. Then, the value is normalized with respect to the ward that has maximum population density in the PMC region (Max_Density).

 ISBN: 978-1-61208-863-1

TABLE I
WEIGHTS FOR TREE ATTRIBUTES

w1		w2		w3		w4	
Height	***Score***	***Canopy***	***Score***	***Condition***	***Score***	***Phenology***	***Score***
0-3m	0.2	0-1m	0.2	Poor	0.25	Seasonal	0.3
3-5m	0.4	1-3m	0.4	Average	0.5	Evergreen	0.7
5-8m	0.6	3-5m	0.6	Healthy	0.75		
8m+	0.8	5m+	0.8				

$$Population\_Score = \frac{(Max\_Density - Ward\_Density)}{Max\_Density} \quad (5)$$

Once the scores (Tree_Score, AQI_Score and Population_Score) are calculated for all the wards, they are combined with a weight factor [11] and normalized on a scale of 0 to 100 as shown in (6).

$$Health\_Score = (Tree\_Score * 30) + (AQI\_Score * 40) + (Population\_Score * 30) \quad (6)$$

3) Plots and Visualization: Figure 1 shows the health scores for wards in PMC region. Health scores are numeric values in the scale of 0-100. Wards are categorized into zones (Extremely poor (0-20), Poor (20-40), Moderate (40-60), Satisfactory (60-80) and Good (80 and above)) based on the values of the health score.

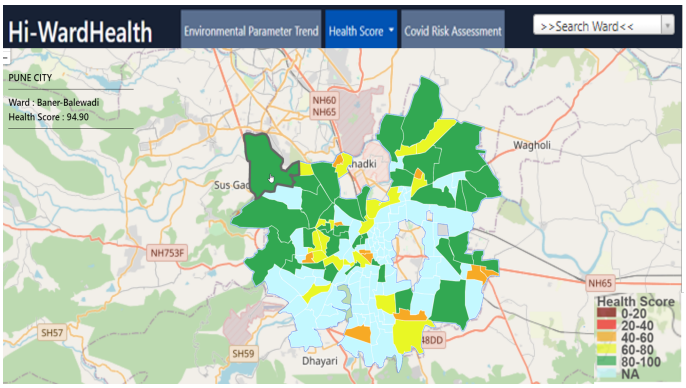

Figure 1. PMC-Pune Ward Health Scores.

4) Green Buffer Zone recommendation and example: For any selected location within the PMC region, recommendations to further improve the green cover of that place is provided. For a selected location and a boundary area around it, the attributes (air quality, tree count, population density) are analysed to compute an environmental health score based on (6). Our technique also goes one step beyond this by providing recommendations for the construction of green buffer zones. The green buffer zone recommender system at the backend compares the health score of a place with a pre-defined threshold value (80 by default). If the health score is less than the threshold and Tree_Score is less than 1, recommendation to further improve the green tree cover of the place is provided to the user. To improve the health score of a place to at least the predefined threshold, the recommender system provides insights on the number of trees that must be planted (until Tree_Score upper bound of 1 is not reached) as shown in Figure 2.

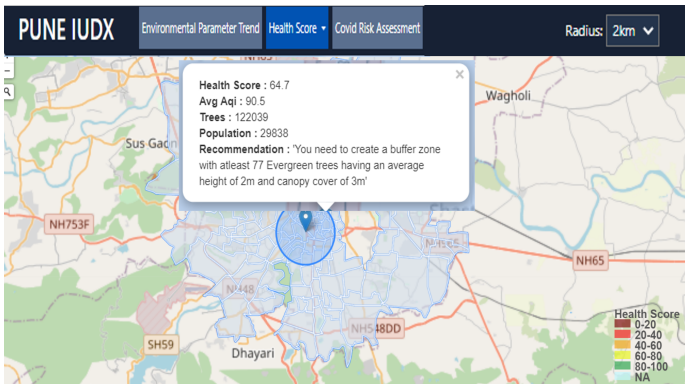

Figure 2. Green Buffer Zone Recommendation.

5) Health Score usability: We have calculated the health scores for PMC-Pune wards where data is available. Higher value of health score indicates better ward health. Table II shows health scores, tree count, population and AQI levels for some wards.

TABLE II
HEALTH SCORES OF PMC-PUNE WARDS

Ward Name	Health Score	AQI	#Trees	Population
Baner-Balewadi	94.8	54.1	207671	150190
Ved Bhavan	92.2	80.4	115324	171906
Lohagaon Vimantal	91.9	111.2	293974	163064
Mundhvagaon	91.0	94.4	160422	90815
V Mahavidhyalaya	90.5	98.0	81776	68110
S Mahavidhyalaya	90.0	106.3	19766	73811
Fergussion College	89.0	114.9	50105	76911
Magarpatta Hadapsar	89.0	83.8	25203	193003
Sadhana Vidhyalaya	88.5	93.5	22666	135103
Kothrud Gaon	86.8	103.2	72851	90671
RajBhavan	85.6	123.6	5084	88484
Shanivarwada	82.3	133.8	54499	61042
PhuleNagar Yerwada	81.7	160.2	125254	78268
Koregaon Park	81.1	111.7	5458	71299
AundhGaon	68.2	93.2	7663	83859
D.P Dattawadi	65.8	83.5	42050	83026
S.G Rugnalya	63.7	103.3	3071	86535
Tingre Station	62.2	83.2	4336	111015

 ISBN: 978-1-61208-863-1

Of the 63 wards, we noticed that 12 wards had moderate health scores, 15 had satisfactory scores and 36 wards had good scores. With the health score, it becomes easy to identify wards with poor environmental health. Health score can help:

- Real estate owners decide selling and buying price of properties across wards depending on its environmental conditions.
- Government to impose certain restrictions on vehicular movements in highly polluted wards.
- Authorities to grant or deny permissions for further construction activities in wards based on their health conditions.

B. Study of variations of environmental attributes during COVID-19

With the onset of COVID-19 cases in India, the Government of India imposed a complete nationwide lockdown [21] starting from March 25th, 2020 which restricted people from stepping out of their homes. All transport services - road, air and rail were suspended during the lockdown phase. We studied the various environmental parameters like AQI, Sound (in db), Ozone, Particulate Matters ($PM_{2.5}$, PM_{10}) and NO_2 across thirty-one PMC-Pune wards over a period of four months (January 2020 to April 2020) and quantified their variations. Average parameter variations observed across PMC-Pune wards include:

- 41.0% decrease in AQI level
- 3.0% decrease in Sound level
- 70.3% decrease in PM_{10} level
- 83.7% decrease in $PM_{2.5}$ level
- 82.5% decrease in NO_2 level
- 77.5% increase in Ozone level

This significant decrease in various attributes can be contributed to suspension of activities and restrictions imposed on citizens during the lockdown. Residential areas showed significant reduction in attributes like AQI, sound and pollutant concentrations with the imposition of lockdown. However, since few essential services continued their operations during this phase, we could see variations in reduction of attributes across wards. Based on the data avaiable from PUDX, sample AQI variations of one of the residential wards: Phulenagar both day-wise and month-wise is shown in Figure 3 and Figure 4 respectively.

This study helps concerned authorities plan and prioritize control measures to improve the quality of air in PMC-Pune. Figure 5 shows a dashboard that helps users visualize the variations and trends of these attributes. They can compare wards and visualize the trends of the environmental attributes over time using this dashboard. Figure 6 shows a dashboard that helps users visualize the ward-level attribute variations and trends in further detail. They can also analyse the ward level variations on a daily, weekly or monthly basis.

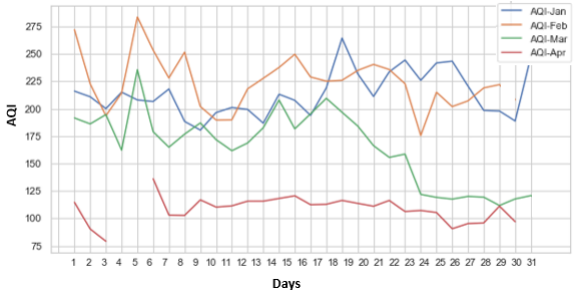

Figure 3. Day-wise AQI variation in Phulenagar.

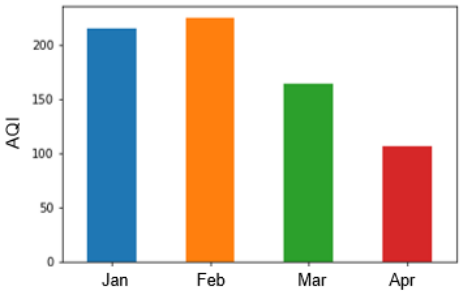

Figure 4. Month-wise AQI variation in Phulenagar.

C. Covid Risk Assessment Score

During situations like the pandemic, providing primary facilities like healthcare to all the citizens becomes the need of the hour. With the steady increase in the COVID-19 cases, PMC faced grappling shortage of healthcare staff and hospitalization facilities over time [2]. With the influx of patients from rural areas into the cities, the hospital management faced tremendous pressure in providing medical help. Thus, we propose a solution that will help concerned authorities get ward level insights for planning further development of medical facilities in wards and to be better prepared in situations like this. Our solution assesses the overall health and medical infrastructure facilities available at a ward, formulates a risk score and provides ward specific insights. A prediction model learns from rich sources of data and predicts a risk score that is correlated to the anticipated number of active patients in the ward. Finally, based on the risk scores the wards are categorized into risk zones. This categorization helps concerned authorities take prioritized decisions with respect to ward administration.

1) Data Sources:

- *COVID hospitals and bed details* [22]: A detailed list of hospitals and the number of beds allocated to treat COVID-19 patients that is updated on a day-to-day basis is collected using web scrapping techniques.
- *Additional hospitals* [23]: Provides a list of additional healthcare facilities and the number of beds available in PMC-Pune region.
- *Demographics* [24]: Publicly available information like - number of literates, number of children below age 6,

 ISBN: 978-1-61208-863-1

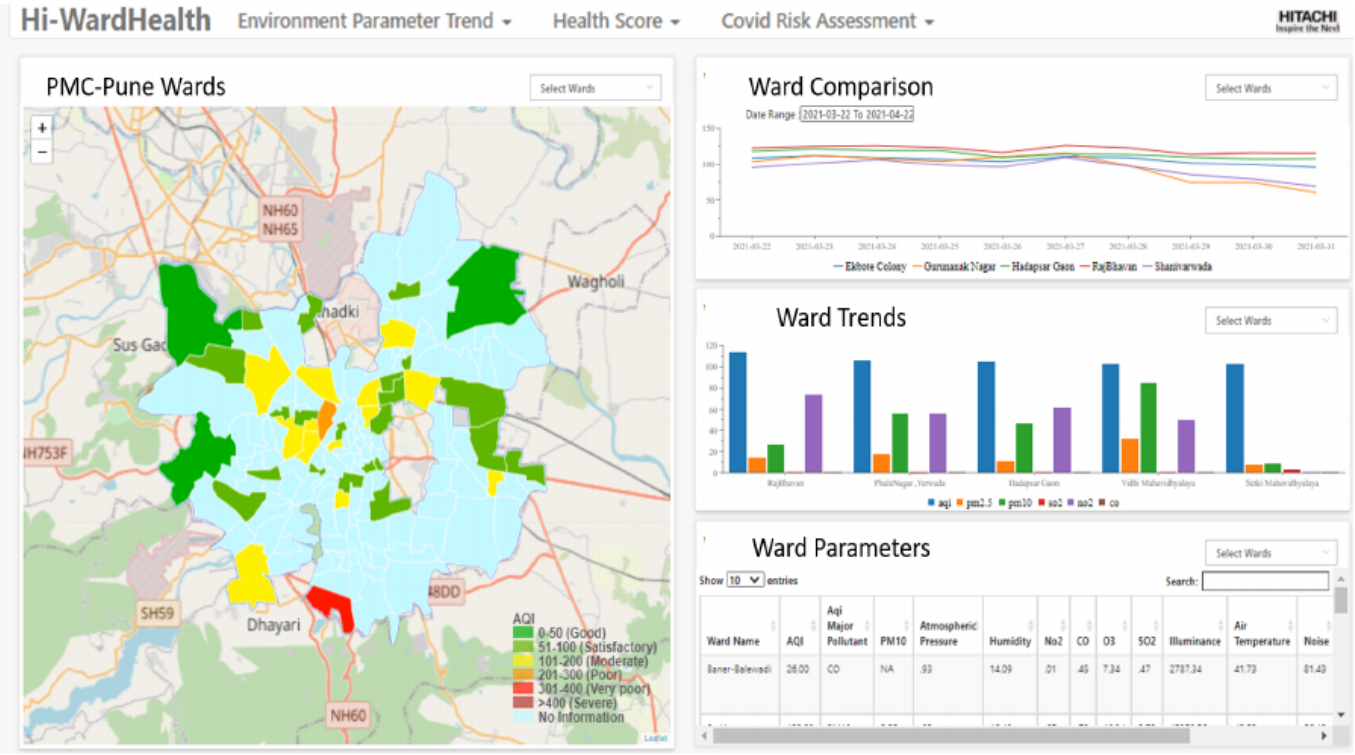

Figure 5. PMC-Pune city Environmental Attributes trend and variations.

working population and average family size in a ward is collected from the data source.

- *Hotspots* [25]: Information about the critical zones in the PMC-Pune region with COVID-19 cases is provided. Data is extracted using relevant APIs.
- *Environmental health scores*: The ward wise health scores formulated based on AQI, tree and population count mentioned in Section 3.1.3 is used.

2) Modelling System and Score Formulation: To predict the ward risk score, firstly, a pipeline is created, and a job is scheduled for periodic data collection. This is followed by data pre-processing, data aggregation and model building. In the first step, data is collected from all the above-mentioned data sources on a day-to-day basis. Next, in the pre-processing step, the geo-coordinates and addresses available in the collected data is mapped to appropriate ward IDs using geocoding and mapping techniques. Further, in the data aggregation step, based on ward IDs the data is aggregated at a ward level.

We studied an extensive set of environmental and demographic factors that are important attributes in contributing to ward risk at times like the pandemic. Based on the literature survey, thirteen relevant features as shown in Table III are selected to train our model. However, it must be noted that no single attribute or feature can individually explain the measure of ward risk. Grietens et al. [26] in their work show how factors such as high human mobility rate, age structure, poverty level, high illiteracy and population density as well as dependency ratio have been influencing the transmission dynamics of diseases. As per a study conducted in New York, USA, using Kendall and Spearman rank correlation test, it was found that temperature and air quality had a significant association with the COVID-19 pandemic (Bashir et al. [27]). Air quality, population density along with the green tree cover is formulated as a health score and is used as an important factor in assessing the ward risk (Feature 1) in Table III. Dalziel et al. [28] and Casanova et al. [29] showed that virus transmission can be influenced by several geographical factors such as climatic conditions (temperature and humidity) and population density. Also, as we are aware, the primary means of transmission of the virus is through physical contact as shown in work by Pung et al. [30]. The spread of infection is usually accelerated in crowded conditions.

To account these notable factors and observations, this paper considers demographic factors at ward level that include the number of houses (Feature 2), literate population (Feature 3), population below the age group of 6 (Feature 4), average family size (Feature 5) and working population (Feature 6). Apart from this, the preparedness of any ward in situations like the pandemic, depends upon its ability to provide the required medical facilities. To account this, the information of hospitals (Feature 7) and the bed facilities: beds with oxygen (Feature 8), beds without oxygen (Feature 9), ICU beds with ventilator (Feature 10), ICU beds without ventilator (Feature 11), additional non COVID-19 beds available in medical facilities across PMC-Pune (Feature 12) to treat patients is

 ISBN: 978-1-61208-863-1

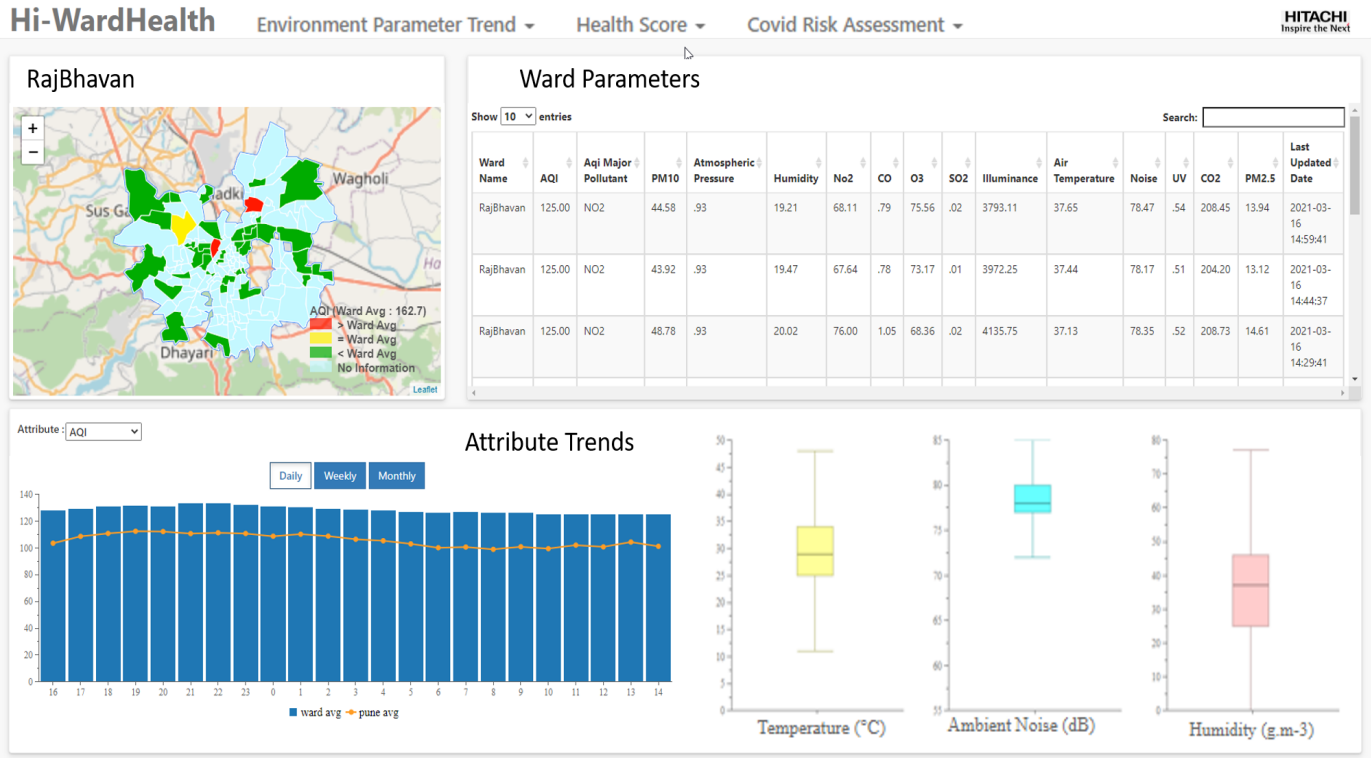

Figure 6. Ward level Environmental Attributes trend and variations.

used. All the features are normalized with respect to the ward population.

TABLE III
FEATURES USED TO CALCULATE THE COVID-19 RISK SCORE

Feature#	Feature Name
1	Health Score
2	Houses
3	Literate Population
4	Population under 6
5	Family Size
6	Working Population
7	Hospitals
8	Oxygen beds
9	Beds without oxygen
10	ICU ventilator beds
11	ICU beds without ventilator
12	Additional beds
13	Hotspots

The data collected for these features over a period of time (November 2020 - December 2020) is used to train the model and the data from January 2021 is used for testing. Models are built to predict ward-wise Risk Assessment Score (RAS) or risk score which is an indicator of anticipated number of active patients for the next day in the ward. Several Machine Learning based prediction algorithms like Linear Regressor, Random Forest, K-Nearest Neighbor and Gradient Tree Boosting are used to predict the risk scores based on the thirteen input features selected. In any ward, more the number of active COVID-19 cases, more is the risk. With this analogy, we calculated the number of active patients per day in a ward, formulated the risk score as a function of category of active patients. Weights are assigned to each of the category of active patients as shown in (7). This is then normalized with the ward population (8) and used as the Initial Risk Assessment Score (Initial_RAS) while training the model.

$$\begin{aligned} Risk\_Sum = 0.1 \ * \ (Patients\ without\ oxygen) \\ + 0.2 \ * \ (Patients\ with\ oxygen) \\ + 0.3 \ * \ (Patients\ without\ ventilator) \\ + 0.4 \ * \ (Patients\ with\ ventilator) \end{aligned} \tag{7}$$

Here, more importance is given to patients currently under ventilator support in the ICUs (weight factor: 0.4) followed by those in ICUs without ventilator (weight factor: 0.3), followed by patients in regular wards with oxygen supply (weight factor: 0.2) and finally the regular patients (weight factor: 0.1).

$$Initial\_RAS = \frac{Risk\_Sum}{Ward\_Population} \tag{8}$$

Based on the thirteen input features (normalized with ward population) and Initial_RAS as the output, the model is trained to predict the Risk Assessment Score (Predicted_RAS) to anticipate the number of active patients on the next day. For evaluation, based on thirteen feature values on current date, the model predicts Predicted_RAS score (which is an indication of active patients anticipated for the next day). This is then compared with Initial_RAS score (8) which is computed based

 ISBN: 978-1-61208-863-1

on the actual patient numbers on the next day. The Root Mean Squared Error of the prediction (Predicted_RAS) with respect to the Initial_RAS score is then compared to select the best performing model. The predicted risk scores are normalized on a range of 0-100, where 0 indicates minimum and 100 indicates maximum relative risk score. Table IV shows Gradient Tree Boosting based predicted risk scores for the wards in Pune. A sample of the risk scores for the wards on January 31st, 2021 is shown. Wards are categorized into five zones based on the value of risk scores - Severe (100-80), High (80-60), Moderate (60-40), Low (40-20), Very Low (20-0). Higher the risk value, greater is the risk.

TABLE IV
WARD LEVEL COVID-19 RISK ASSESSMENT SCORES

Ward Name	Risk Score	Risk Zone
Dhanori	1.9	Very Low
Bopodi	3.3	Very Low
Yerwada Prison Press	6.7	Very Low
Bibvewadi	14.5	Very Low
Renuka Swarup Prashala	16.1	Very Low
Kharadigaon	19.2	Very Low
Yashwantrao Chavan Natyagraha	23.2	Low
Bharti Vidhyapeeth	23.5	Low
Dinanath Mangeshkar Rugnalaya	42.7	Moderate
Baner-Balewadi	81.9	Severe

Based on the calculated risk scores as of January 31st, 2021 1 ward showed severe risk, 1 showed moderate, 2 showed low while 18 wards showed very low risk. These relative ward level risk scores can help the authorities take prioritized actions by focussing on wards that require immediate attention. The risk scores of wards also vary over days. Some wards show significant variations while some show minimal variations in the risk levels.

3) Plots and Visualizations: The dashboard shown in Figure 7 help users visualize the ward risk scores and categorize them into zones. Feature values across wards in PMC-Pune along with the hotspot case progression is shown.

The user can also analyse details at a ward level and gain insights on the various features, their values and compare it with the overall PMC-Pune average as shown in Figure 8. With such comparisons, it becomes easier to understand the relative preparedness of the wards.

4) Evaluation: The performance of several Machine Learning models was considered to predict the risk scores. Table V shows prediction errors in risk scores with various Machine Learning algorithms.

The predicted risk score values and their variations for wards across days is studied in detail. The average prediction error rate of models using various machine learning algorithms is compared. Table V shows a sample reference of various models and their performance for 15 days. Gradient Tree Boosting based Regressor shows 5.45% error in prediction when compared to Random Forest Regressor (5.92%), K-Nearest Neighbors (6.01%) and Linear Regressor (22.40%).

TABLE V
PREDICTION ERRORS WITH MACHINE LEARNING MODELS: LINEAR REGRESSION (LR), K-NEAREST NEIGHBORS (KNN), RANDOM FOREST (RF), GRADIENT TREE BOOSTING (GTB)

Dates	LR	KNN	RF	GTB
12-01-2021	23.46	5.49	6.23	5.55
13-01-2021	23.69	3.68	4.38	3.56
14-01-2021	26.11	5.98	6.57	5.87
15-01-2021	23.31	4.72	5.23	4.39
16-01-2021	22.38	4.96	5.53	4.66
17-01-2021	23.70	5.44	6.06	5.35
27-01-2021	32.51	12.75	11.77	10.93
28-01-2021	31.55	11.38	10.34	9.57
29-01-2021	19.67	6.48	4.45	4.18
30-01-2021	19.75	6.79	5.3	4.78
31-01-2021	20.14	7.60	6.11	5.6
01-02-2021	19.52	6.22	4.73	4.19
02-02-2021	16.21	2.82	3.97	4.1
03-02-2021	16.92	3.43	4.57	5.02
04-02-2021	17.11	2.49	3.63	4.07
Average error	**22.40**	**6.01**	**5.92**	**5.45**

Based on the results, Gradient Tree Boosting algorithm is selected to predict the ward level risk scores. The results are further validated with the initial hypothesis which indicated that risk score of the ward is correlated to the number of active cases in the ward. A sample result is shown in Table VI.

TABLE VI
CORRELATION BETWEEN RISK SCORE AND ACTIVE PATIENTS

Ward Name	Predicted_RAS	Active Patients*
Dhanori	1.9	7
Bopodi	3.3	23
Bibvewadi	14.5	43
Yashwantrao Chavan Natyagraha	23.2	76
Dinanath Mangeshkar Rugnalaya	42.7	145
Baner-Balewadi	81.9	599

*per 10000 population.

Ward-wise risk scores can help authorities take planned actions and be better prepared to provide required medical facilities to its citizens in times of need.

IV. RESULTS AND DISCUSSIONS

In our work two scores to indicate the environmental health as well as the risk score for wards in PMC region of Pune are formulated. Out of 144 wards in PMC as per 2011 census, based on data availability, data for 63 wards was collected and scores were derived for them. In case of health scores, AQI information was collected from 33 sensors, tree data from 73 wards and population census data from all the 144 wards. AQI data was interpolated to calculate the health scores of the wards. Out of the 63 wards, 12 showed moderate scores, 15 satisfactory and 36 showed good health scores. Additionally, a few locations in PMC-Pune were also selected to check the environmental health status. Some regions already showed good health scores while some showed satisfactory and poor conditions. For those locations appropriate recommendations

 ISBN: 978-1-61208-863-1

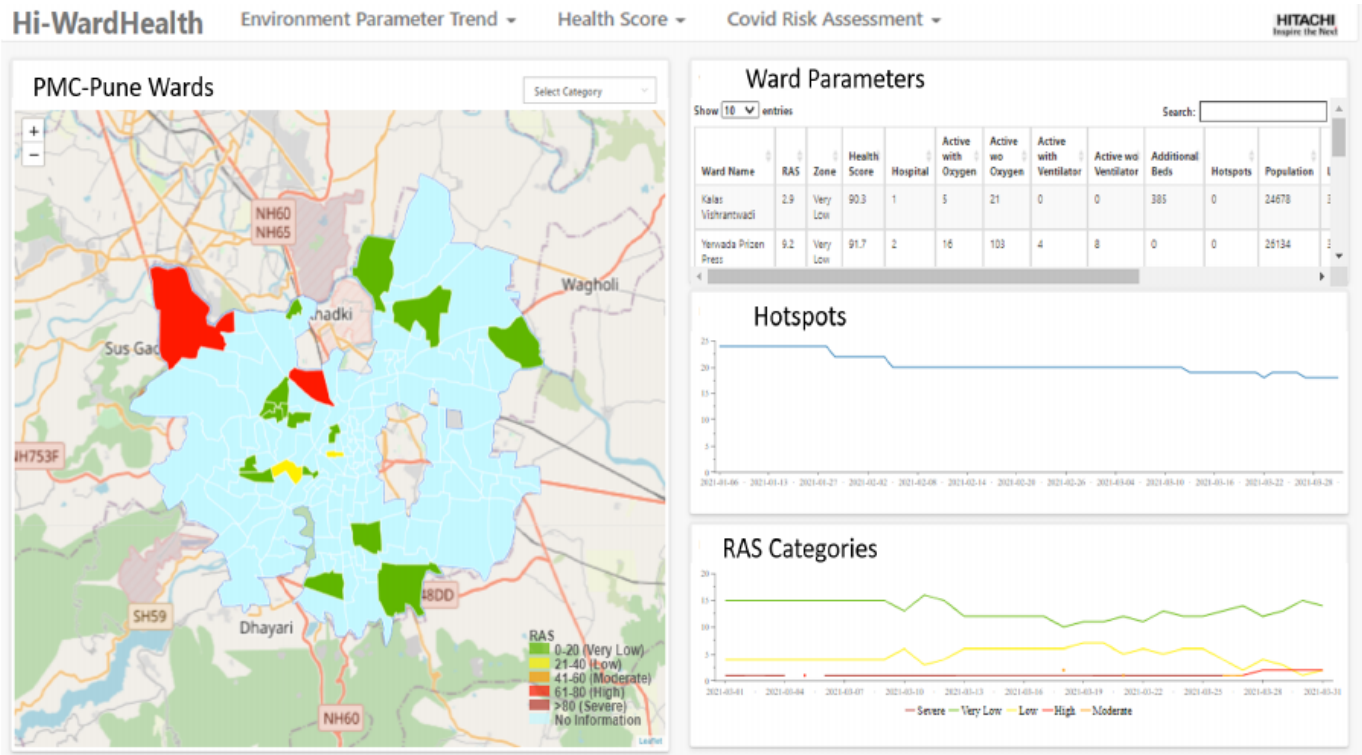

Figure 7. PMC-Pune COVID-19 Risk Assessment Scores.

on the number of trees further needed to improve the health score was provided.

Based on our detailed analysis of environmental parameters during lockdown, a significant decrease in level of pollutant concentrations in Pune was observed. AQI, $PM_{2.5}$, PM_{10}, NO_2 and sound levels showed decrease of 41.0%, 83.7%, 70.3%, 82.5% and 3.0% respectively while Ozone showed an increase of 77.52%. According to a study [31] a high NO_x level reacts with Ozone and mops it up. The Ozone that escapes to cleaner areas has no NO_x to further cannibalize it and as a result, Ozone concentration builds up in these areas. This explains the increase in concentrations of Ozone in the atmosphere.

In case of risk scores, we combined several data sources and collected data for PMC-Pune wards. The combined data available for 22 wards was used to calculate the risk scores. Majority of wards showed low risk scores while a few showed medium risk scores. These relative scores help compare wards and take further decisions to improve the ward level infrastructure and medical facilities.

V. Conclusion and Future Work

This paper provides a data-driven method of modelling environmental health and ward level risk of a city considering the available infrastructure, demographics and the environmental health conditions. Our method utilizes the most recently available data to provide dynamic scores to wards. These numerical scores can be used as an indicator in prioritizing the wards that need more focus than the others. It also gives an understanding of wards at a granular level. A dynamically changing health score index makes an attempt to quickly assess the ward environmental health. It serves as an important indicator that helps city planners improve the environmental conditions of the ward. Our paper also studied the variations of environmental parameters in Pune over a stipulated period of time. Further, a dynamic risk score prediction model is introduced in this paper that combines several relevant sources of data to predict the risk levels of a ward. When the overall condition of a ward improves, its ability to handle situations like the pandemic increases. With better facilities, the living conditions in the ward can improve and the ward can gradually attain self-sufficiency.

The model can be further enhanced by adding valuable features like average hospitalization duration of patients, number of direct contacts and other COVID-19 transmission related information based on data availability. The visualizations presented in this work can be further enhanced by making use of Human Computer Interaction techniques. We also plan to implement a simulator that would help authorities foresee the risk scores in wards with gradual changes in the ward facilities over a period of time. Overall, the scores proposed in this work and its evaluation was done based on the data available in PMC-Pune. However, the same can be extended to other smart cities where similar data is available.

References

[1] COVID-19. https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome_coronavirus 2021.05.24.

 ISBN: 978-1-61208-863-1

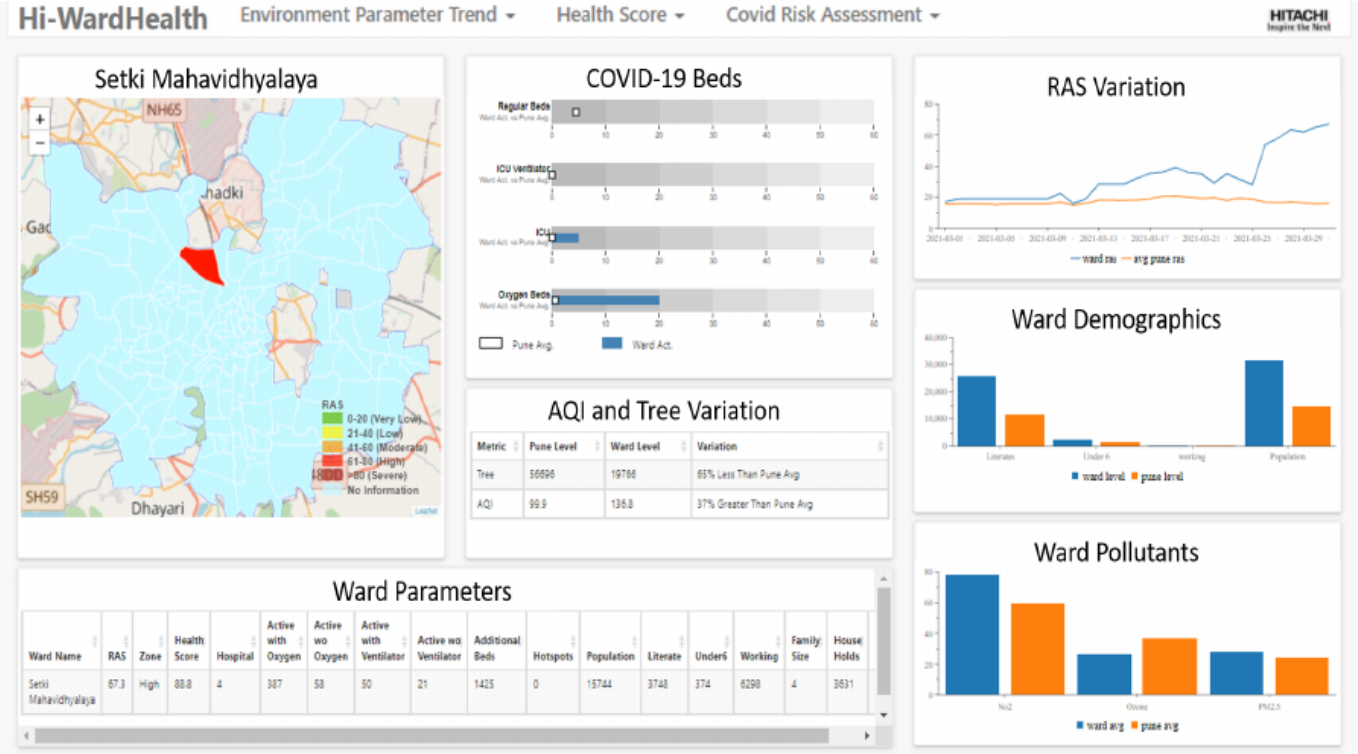

Figure 8. Ward level COVID-19 Risk Assessment Scores.

[2] PMC Healthcare. https://indianexpress.com/article/cities/pune/amid-healthcare-staff-shortage-and-rising-fatigue-overburdened-pvt-hospitals-in-pune-at-breaking-point-6591101/ 2021.05.24.

[3] IUDX. https://iudx.org.in/ 2021.05.24.

[4] Pune IUDX Catalog. https://pune.catalogue.iudx.org.in/ 2021.05.24.

[5] PMC. https://pmc.gov.in/mr 2021.05.24.

[6] Pune wards. https://ourpuneourbudget.in/know-your-city/pre-2012-144-wards/ 2021.05.24.

[7] PMC-Pune ward details. https://www.pmc.gov.in/en/all-ward-office 2021.05.24.

[8] What is Pune losing. https://www.civilsocietyonline.com/cover-story/what-is-pune-losing-and-how-fast/2021.05.24.

[9] A. Kane, V. K. Ganesan, M. Sardesai, and M. Shindikar, "Green Cover Analysis using Tree Census Data to Optimize the Bio-diversity in Pune Municipal Development Area", UDS 2020, vol. pp. 45–56, 2020. doi: http://ceur-ws.org/Vol-2557/.

[10] NRDC. https://www.nrdc.org/sites/default/files/media-uploads/pune_air pollution_ib.pdf 2021.05.24.

[11] A. Morani, D. J. Nowak, S. Hirabayashi, and C. Calfapietra, "How to select the best tree planting locations to enhance air pollution removal in the MillionTreesNYC initiative". Environ Pollut. 2011;159(5): pp. 1040-1047. doi:10.1016/j.envpol.2010.11.022.

[12] S. Sharma et al., "Effect of restricted emissions during COVID-19 on air quality in India. Sci Total Environ. 2020;728:138878. doi:10.1016/j.scitotenv.2020.138878.

[13] C. A. Yaro, P. S. U. Eneche, and D. A. Anyebe, "Risk analysis and hot spots detection of SARS-CoV-2 in Nigeria using demographic and environmental variables: an early assessment of transmission dynamics", International Journal of Environmental Health Research. doi:10.1080/09603123.2020.1834080.

[14] IUDX dataset. https://pune.catalogue.iudx.org.in/search/dataset/items 2021.05.24.

[15] PMC Open Data Store. https://opendata.punecorporation.org/Citizen/CitizenDatasets/Index?categoryId=24&dsId=483&search=tree%20census 2021.05.24.

[16] City census department. https://www.pmc.gov.in/en/census 2021.05.24.

[17] Evergreen tree benefits. https://canopy.org/tree-info/benefits-of-trees/ 2021.05.24.

[18] IISC survey. https://www.asianage.com/metros/mumbai/211217/one-tree-for-every-four-persons-bmc-census.html2021.05.24.

[19] Exponential moving average. https://www.investopedia.com/ask/answers/122314/what-exponential-moving-average-ema-formula-and-how-ema-calculated.asp 2021.05.24.

[20] Finding points within a distance of a latitude/Longitude using bounding coordinates. (n.d.). J. P. Matuschekpage. https://JanMatuschek.de/LatitudeLongitudeBoundingCoordinates 2021.05.24.

[21] Lockdown in India. https://en.wikipedia.org/wiki/COVID-19_lockdown_in_India 2021.05.24.

[22] COVID hospital data. https://www.divcommpunecovid.com/ccsbeddashboard/hsr 2021.05.24.

[23] Additional hospitals. http://opendata.punecorporation.org/Citizen/CitizenDatasets/Index?categoryId=35&dsId=446&search=hospital 2021.05.24.

[24] Demographic data. https://indikosh.com/city/588016/pune 2021.05.24.

[25] Hotspots in India. https://www.covidhotspots.in/ 2021.05.24.

[26] K. P. Grietens et al., "Characterizing types of human mobility to inform differential and targeted malaria elimination strategies in Northeast Cambodia". Sci Rep. 5(1):16837. doi:10.1038/srep16837.

[27] M. F. Bashir et al., "Correlation between climate indicators and COVID-19 pandemic in New York, USA". Sci Total Environ 728:138835. doi: 10.1016/j.scitotenv.2020.138835.

[28] B. D. Dalziel et al., "Urbanization and humidity shape the intensity of influenza epidemics in U.S cities". Science 362: pp. 75–79, 2018. doi:10.1126/science.aat6030.

[29] L. M. Casanova, S. Jeon, W. A. Rutala, D. J. Weber, and M. D. Sobsey, "Effects of air temperature and relative humidity on coronavirus survival on surfaces". Appl Environ Microbiol. 76(9): pp. 2712–2717, 2010. doi: 10.1128/AEM.02291-09.

[30] R. Pung et al., "Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures". Lancet. 395(10229): pp. 1039–1046, 2020. doi:10.1016/S0140-6736(20)30528-6.

[31] Air Pollutants COVID-19. https://indianexpress.com/article/explained/covid-19-lockdown-air-pollution-ozone-6479987/ 2021.05.24.

 ISBN: 978-1-61208-863-1