|  |  |
| --- | --- |
| Image Source: ESRI  ACQUISITION AND EXPLORATION OF GEOSPATIAL DATA | MODULE ASSIGNMENT  Chaturvedi, H. (Harsh, Student M-GEO-WO)  S2330865  h.chaturvedi@student.utwente.nl |

***PART 1: DATA HANDLING***

1. **Cause-Effect Relationship, Ability to spread, Role and characteristics of Geodata for Air Quality Measurements.**

Air Pollution is caused by Aerosols which contain solid and liquid particles along with some gases suspended in the Air. Causes are Exhaust from automobiles(Chemical), Pollutants from factories by burning of fossil fuels such as coal or petroleum(Energy Use and Production), Dust, Pollen, Molds Spores(Biological), Hazards such as Wildfires or Ash from volcanic Eruptions, or even Chemical reactions in the Air of Sunlight and other chemical pollutants within Ground level ozone Layer.

Ability of pollutants to spread depends on size of compound and height at which it was emitted in air. Weather parameters affecting dispersions are Wind direction and speed, humidity, cloudiness and temperature. Cadmium(uncommon) spreads most easily in air (Margherita Ferrante et al.,2012).

The major effects of Air pollution are Smog (from Ground Level Ozone, affects visibility badly) and Soot(particulate matter in air, mostly health effects such as lung diseases(NO,SO,O3), Inflammation & Worsening of heart and lungs(Fine and ultra-fine particles) and preventing uptake of oxygen by the blood causing heart diseases(CO) ). Mercury, Lead, Dioxins and Benzene affect liver and harms immune nervous and endocrine systems. Molds and Allergens increase chances of Asthma attacks and allergies. GHG gases causes Sea Level rise, Heat related deaths and transmission of Lyme disease.

Geodata can be from Instruments on field(Ex: RIVM – high quality, with low cost sensors, more instruments needed) or Remote Sensing (Satellites Spectrometers: NASA’s Aqua, Terra and Suomi-NPP satellites, MAIA and ESA’s Copernicus Sentinel 5P(7km\*3.5km resolution)). Geodata from above sources can then be compared to health records or crowdsourcing data.

Note: The explanation in Part a) is in a holistic and general Sense. Further briefing and assumptions about assignment is inside Code file submitted as .txt format.

1. **Code + Summary – Provided in the Code file**

The details of all the data used throughout the assignment is provided in the figure below:-

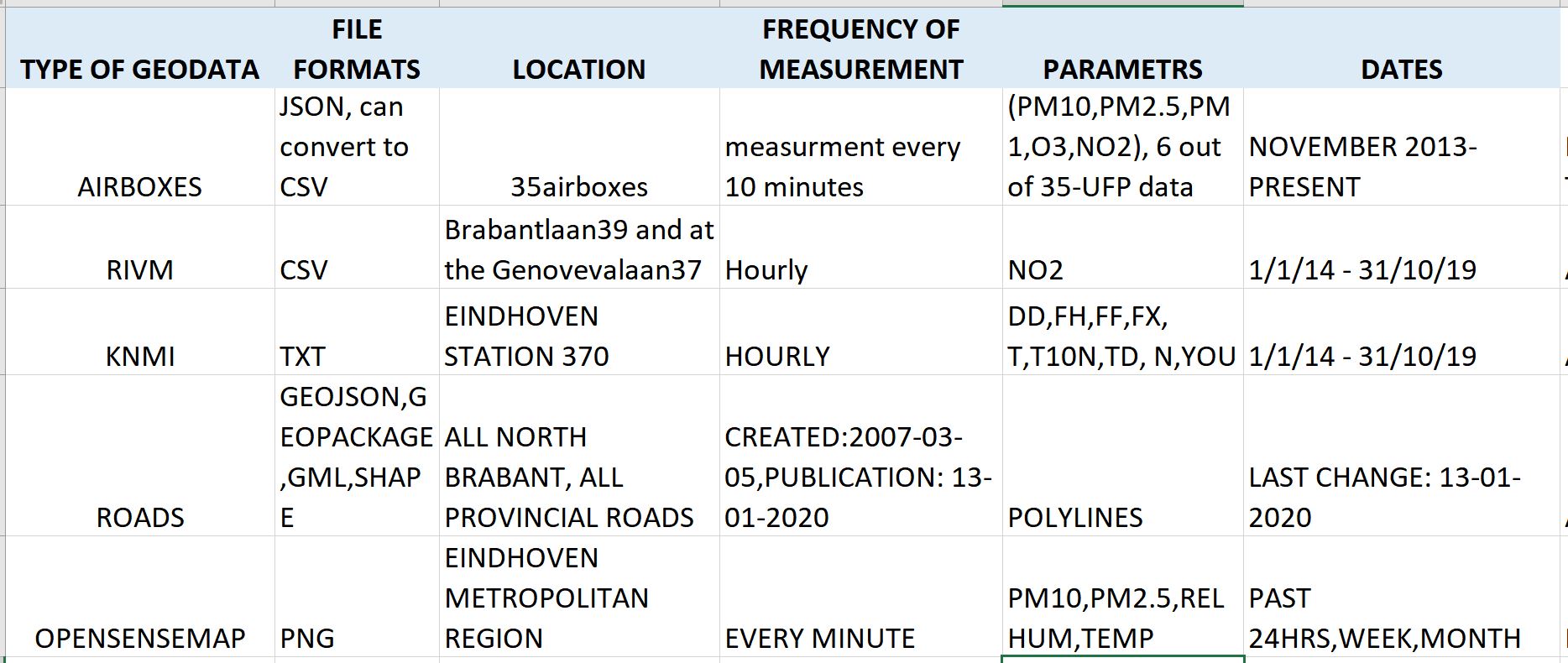


Figure 1: Types of Data used in Assignment

***PART 2: CROWDSOURCING***

1. **Other non-traditional sources of information:**

These other sources of information are obtained via web searches and does not include the data from which is already mentioned in form of a table. These are some volunteering initiatives that took place in Eindhoven along with crowdsourcing and information from NEWS Media and some research papers. The links are as follows:

<https://www.hackair.eu/hackair-a-review/> - Most important, Mobile Application

<https://strp.nl/program/pollution-explorers?thumb=aHR0cHM6Ly9zdHJwLm5sL3VwbG9hZHMvX3RodW1iL1BvbGx1dGlvbi1leHBsb3JlcnMucG5n> - volunteered

1. **Open Sense Map Data for Eindhoven**

**Eindhoven Metropolitan Region** have been considered for getting data from Open sense Map as there were no sensors situated connected to the service within city boundary.

**Utility: -** The parameters vary from basic sensors to advanced (2->4->5->7 respectively). In addition to that the quality and cost of the sensors that obtain this data is not as good as the AIREAS AIRBOXES and just meant for an idea of Ambient Air Quality outdoors. The sensors do not record precise measurements of other more important parameters such as NO2, O3, SO2, UFP etc. These sensors are capable of recording some weather data such as Temperature, Relative Humidity, Air Pressure, UV Intensity (UV Index) and Illuminance(lux). The data is downloadable in PNG file format which can be used for visual comparison with the results from the official datasets in Part 3. This will provide an indication of the validity of measurements from these sensors.

**Data Quality**: - The detail of data particularly from Open Sense Map can be described as very coarse and, in a way, seen as a point data. Also, this point data of Air quality parameters is mentioned to be using Inverse Distance Weighted IDW Interpolation Model on the Web Service, which can provide us with the approximate values near the point data and the farther the distance from point, the lesser the value. If the point data is combined with weather parameters (such as wind, temp, rel.hum. etc.) to generate complex models, then it can produce totally different results which questions the applicability of the correct model to get accurate results.

About the quality of the data, we can say that the measurements are regular and reliable if only the requirement is of within a short time frame (say, past month).

The details about them have been mentioned in a table below:

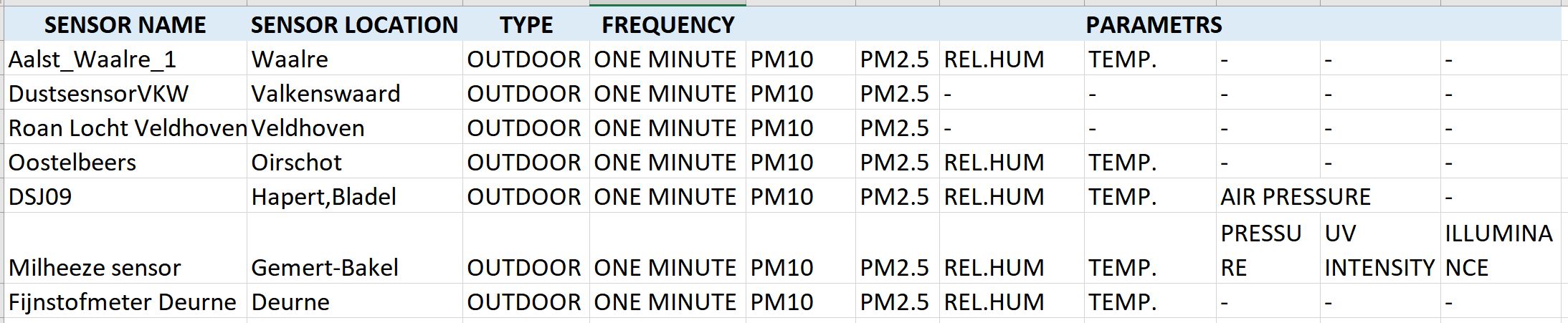


Figure 2: Details of OpenSenseMap Sensors

**Visual comparison with official dataset and results: -**

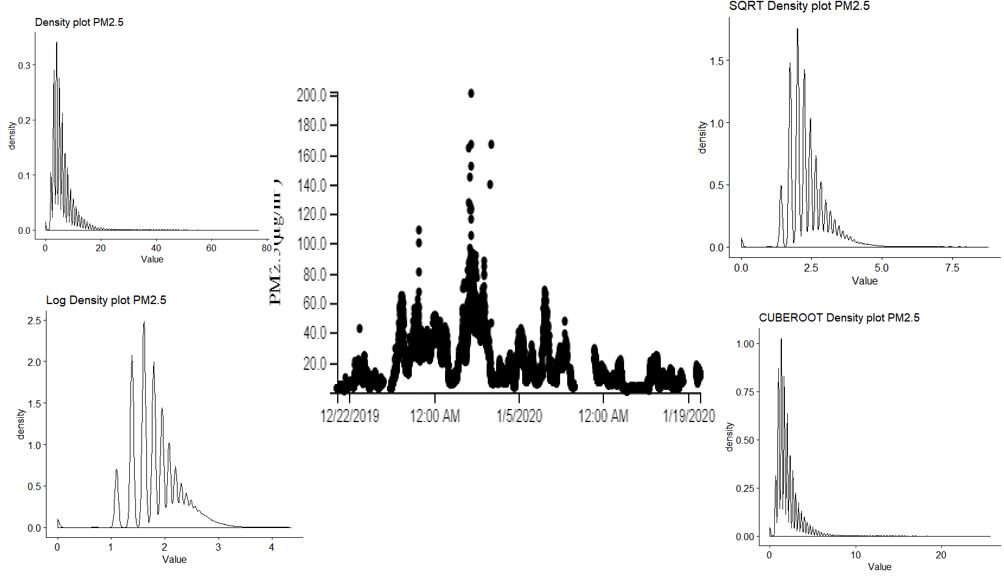


Figure 3: Visual Comparision: PM10

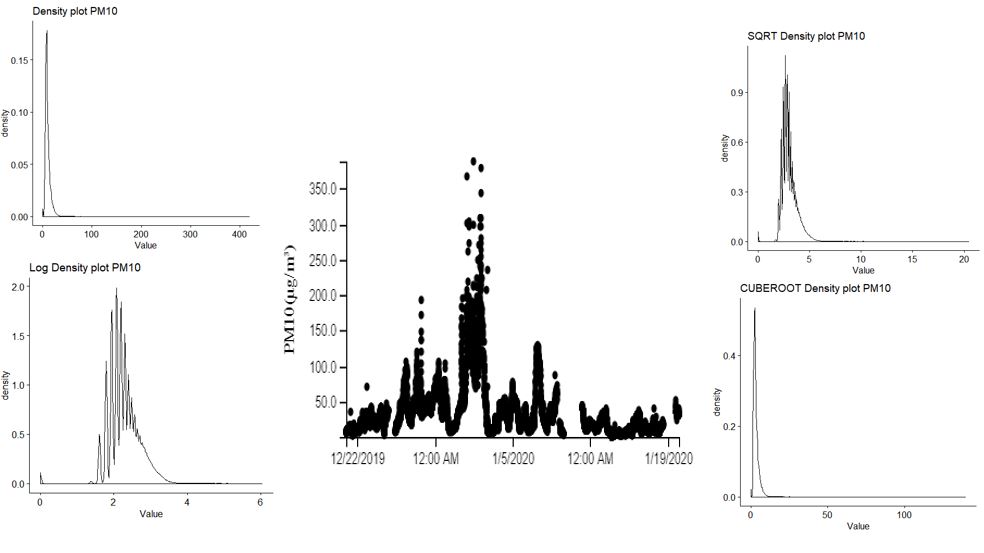


Figure 4: Visual Comparision - PM2.5

The Airbox chosen to compare the AiREAS data with the Open Sense Map data is located in Twickel 30 street. The reason for choosing the airbox was proximity as it was the closest AIRBOX to any Open Sense Map sensor. The timeline of the data was decided to be the past month from 21/12/2019 to 21/01/2020. The data for the Sensor located at Veldhoven had only two measurables PM 10 and PM2.5 which was available in PNG format. The distribution of data is clear in PNG format for the two measurables. PM10 and PM2.5 from Airbox (Twickel 30) were analysed which shows that after applying several transformations the distribution comes close to the Open Sense Map distribution. Square root transformation on Airbox data seems to be complimentary to the Open Sense Map data in visual terms.

1. **Existing and Own Citizen Science Projects Suggestion:**

**Existing:-**

All over the world for several Academic and Commercial Efforts have been made so far in this regard. Particularly for Eindhoven Projects Like hackAIR have worked out so far partnering with OpenAQ, OpenSenseMap, Lufdaten etc.

hackAIR is a collective effort of Six European Organisation for a EU funded project on ‘Collective Awareness Platforms on Sustainability and Social innovation. It was started in 2016, the prototypes and development was finished by 2017 and it was functional in 2018. Every detail can be found on their website from the source code to the toolkits and tutorials.

**Proposed:-**

The solutions on crowdsourcing projects in this situation can either be very innovative, uncommon approach(Huang et al., 2019) or a more practical and feasible approach to the masses (as suggested by Jonathan E. Thompson in his paper, Sensors attached to smart and portable devices or Initiative by AirVisual’s node device ). In both the cases there is a policy aspect with regulating bodies or interventions from Manufacturers of these systems. In my opinion, at present it is better to try and test the first approach as the same results can be reached by mounting fewer sensors on cars because they are dynamic in nature, also they can give real time updates on traffic and the quality of air at particular intersections on the signal waiting etc. Although less research has been made on this till now, but the results prove satisfactory and worth investing upon by policy makers or the manufacturers.

Requirements for Setting up a New Project: -

Target Group and How to involve them: - The target group can be people with automobiles especially Heavy Vehicles, it can also be mounted on motorbikes etc. Involving them would not require so much effort if they already come on mounted with the automobile. Only a small agreement can be made between user and producer for testing this technology. A total business ecosystem can be developed with the automobile manufacturer. Even heatmaps can be generated from vehicles movements with the corresponding measurement at that instant.

Sensors Used: - Usually the cost of sensors is around 200-250$, the sensors can be provided at a lower cost if sold in bulk from the manufacturer to the implementing agency.

Rewards for Making them stay for longer: - This approach also solves the need for the organisation needing this data to make the users to stay longer or do anything. Only while maintenance, the sensors can also be checked for re-calibration, defects etc. to maintain quality of data. Now for the part where the customer has to pay extra for the sensor can be compensated with free parking at few public places say for a few times per month which in the longer term would prove beneficial for the user to spend extra for sensor.

***PART 3: STATISTICS***

1. **Distributions and their Comparison of PM10 and O3 with Log Transformations**

**Assumption –** Air Pollution is maximum in winter months as proven by many studies, so January is taken as median winter month in the past year. A comprehensive analysis can be done for all months of a particular year or taking median month from each season and then time series analysis can be performed. Due to time constraints and data availability issues, I have chosen January 2019 as the time period for analysis which can give us a picture of the values on the higher side in a year.

For this part, all airboxes with available data were chosen for checking the distribution of data except Airboxes with ID 6,10,15,18,32,33(83% data was available,17% missing). The data consists of first 30 days(maximum allowable period to extract) of measurements from Airboxes instead of 31 days

The variable that were chosen to check the distribution were PM10 and O3 since these two along with NO2 come under the mandatory pollutants to be recorded for the Air Quality Index(EU standards).There are other categories of pollutants which includes Optional(PM2.5, CO and SO2) and Sub Optional(PM1, FP, UFP etc.) pollutants.

The number of observations are very high (1,51,200). Due to this large dataset, the graphs do not seem uniform and rather highly skewed. The reason to not remove the outliers was made so that the original recorded value can be shown. Some individual recorded values from Airboxes reached 45063, which depicts the fault in the instrument for that time period. The outliers can be removed, and the limits of x axis can also be restricted by xlim function. Below provided is a summary for the two variables, before and after applying log transformation:-

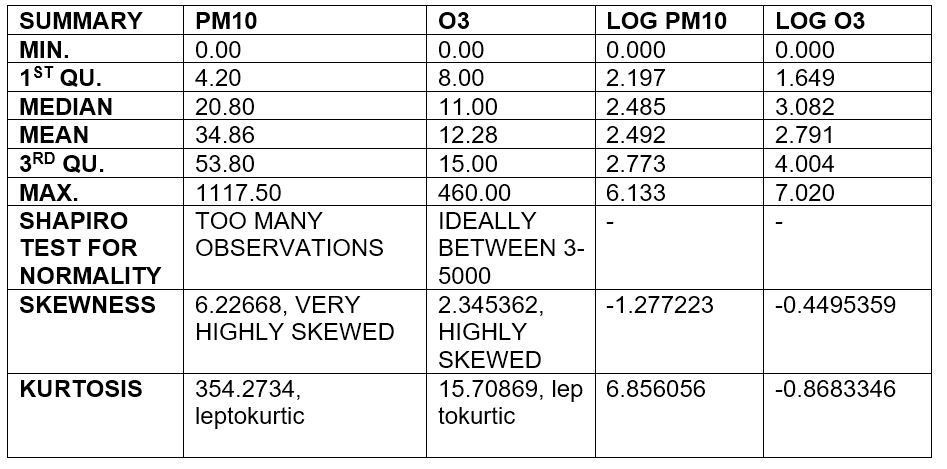


Figure 5: Descriptive Stats of PM10 and O3

The reasons for applying log transformations is that it makes highly skewed variable less skewed and tries to bring it closer to normal distribution. Since most of the assumptions of inferential statistics depend on Normal Distribution, it is recommended to transform the data to analyse it further. The plots coming further makes it easier for us to see the difference between original and transformed data. The variable before transformation as inferred from the table and charts are highly skewed in the positive direction (Right skewed - tail on right side. Also, the peak of the variable is very sharp before transformation. After applying the transformation, the skewness is drastically reduces and the spread of the data becomes more flat comparatively.

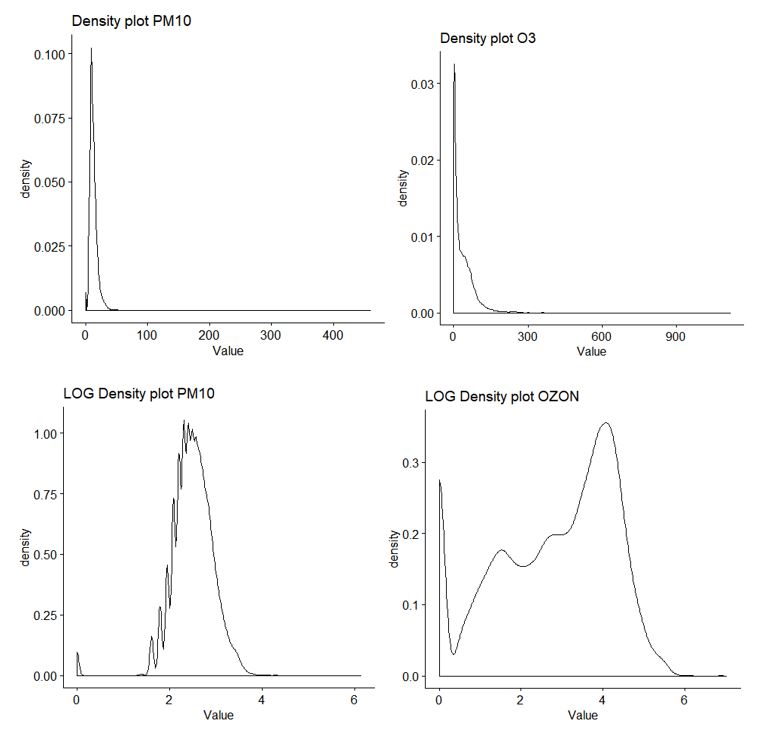


Figure 6: Distribution by Density Plots

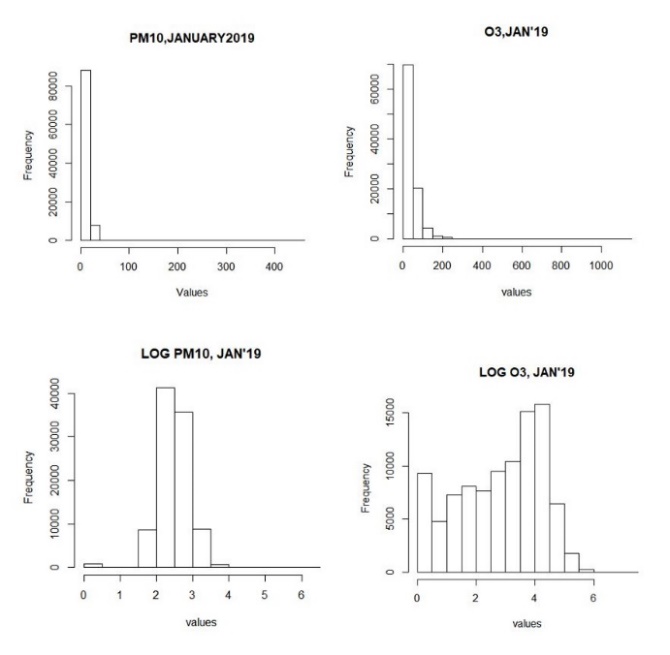


Figure 7: Distribution by Histograms

1. **Correlation and Regression (Simple Linear and Multiple Linear) between Airboxes and RIVM instruments and their comparision**

The same timeline of January 2019 have been followed for the analysis in this part as well. We already know which airboxes are missing data so we exclude them beforehand. This section focuses on modelling of relationship between the two RIVM airboxes and Aireas Airboxes. For that we need to choose two suitable Airboxes to model relationship with RIVM sensors. The criteria dor choosing those Airboxes were proximity, availability and correctness of data.

For the GenovevalaanRIVM instrument, the closest Airbox was on Maasteikstraat 7 (863 m) but all values corresponded null hence rejected, the next closest was Rijckwaartstraat 6 at 1200m with all data as null, further Airbox at GroteBeerlaan 15 was 1600 m far but most of the values were negative , finally at 2 km distance Airbox at Falstaff 8 was found to have correct data, thus chosen for modelling.

For the NoordBrabantlaanRIVM, the case was relatively very easier and the closest Airbox at Beukenlaan St. at 320 m held correct values and thus chosen.

In one Airbox, the ideal values for period of 30 days at 10 min intervals should be around 4320 but there was a level of incompleteness in the data. For Falstaff Airbox the recorded values for month was 4235(85 values not collected in json file), and for Beukenlaan 4230 values were recorded(90 values missing), even in the values recorded there were many values with NA, 0 and negative figures thus further complicating the analysis.

In the RIVM measurements the accuracy was relatively very high. Ideally it should collect 744 values per hour over spread of 30days, GenovevalaanRIVM had 728 values present and NoordBrabantlaanRIVM had 742 values present.Next task was to arrive at paired measurements to carry out regression and correlation analysis. There was a huge difference between the values of RIVM and Airboxes, so the values from Airboxes had to be averaged and aggregated on hourly basis to arrive at paired observations.

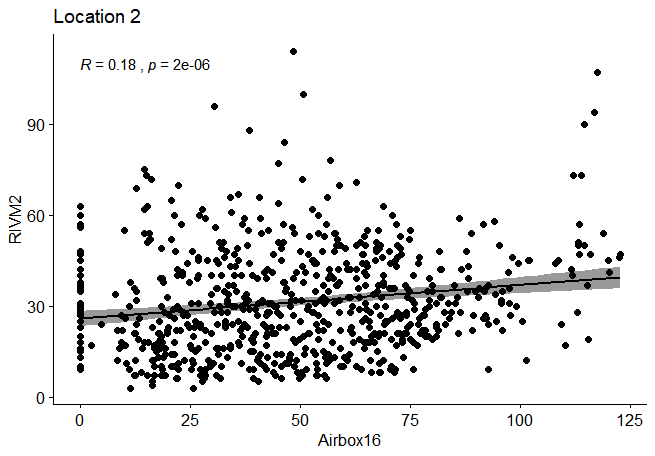
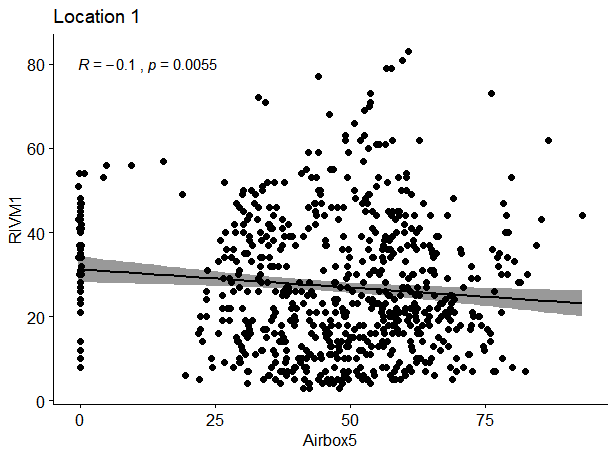
Since there were few values missing from the Airboxes the data for last 15 hours of the 30th day was not counted from both Airboxes and RIVM data (out of 31 days, the data for first 30 days chosen). Finally, we arrived at 705 paired observations. This step was the most crucial step for the proper analysis of data, in other case, if we had put all the unpaired values in the software, it will automatically pair the observations, but the integrity of corresponding hourly values will be disturbed, proving the results of the analysis to be heavily biased. With 705 paired observations, in further steps of the analysis NA values were omitted. On the basis of timestamps, this can be easily done in SQL by avg and group by substring functions. 

Figure 8: Correlation between Airbox and RIVM

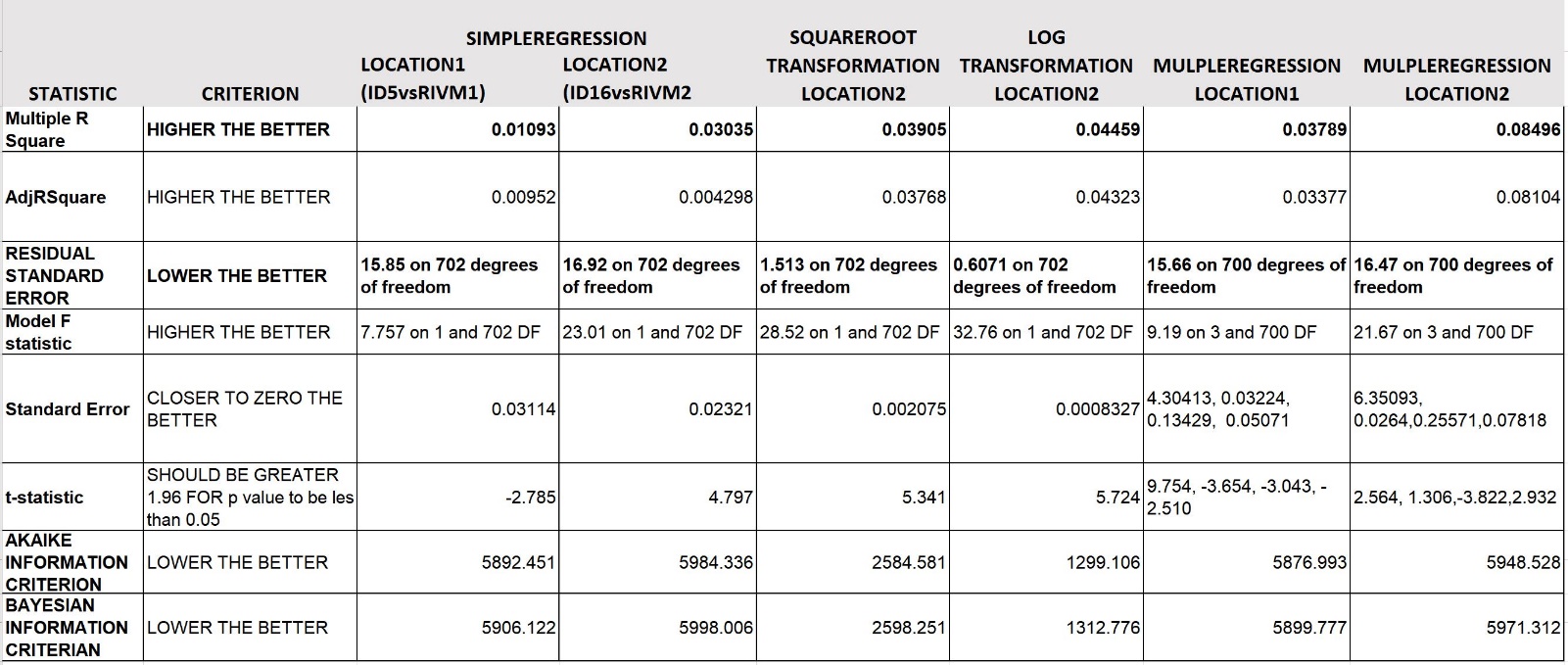


Figure 9: Criteria to compare Regression models

There are several criteria on which the regression model ‘s performance can be observed. Many of those criteria have been summarized in the table above . The most important criteria that we look at in the value of R squared and Adjusted R squared(**IN BOLD**) . Also for comparing models the values of Residual Standard error should be looked at, which tells us whether by applying transformations or changing univariate data to bivariate or multivariate, the error of the residuals have reduced or not. The table above from left to right is a sequential representation of the work in this section. First a simple regression for Location 1 and 2 was run and model 2 was found to show a positive relationship(also we can infer from the correlation graph previously), then square root and log transformations were applied on the model for Location 2 , which further increased the R squared value and decreased the residual standard error of the models. Finally, other variable such as temperature and Relative Humidity from Airboxes were includes for Multiple Regression, turns out that including those additional variable the model did not improve drastically but only to a small extent. The reason behind this can be the large number of values that were Negative, 0 or No data values. But still it showed satisfactory improvement in the model.

1. **Estimations and Predictions for NO2 data with their intervals**

Note: the used for previous analysis has been continues in this section as well to maintain the uniformity of Assignment. The csv file is attached with other requirements.

As it is mentioned in the exercise that the prediction and estimation are almost same, the only thing that differs is their intervals. Results are provided below: 1. Prediction intervals, 2. Confidence Intervals

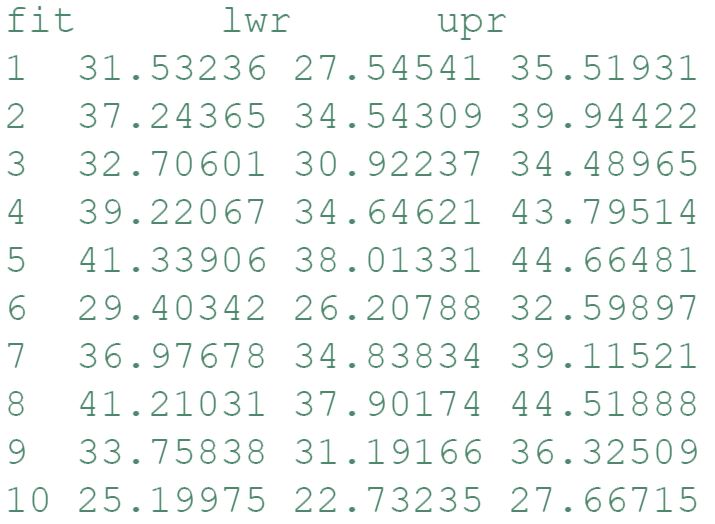
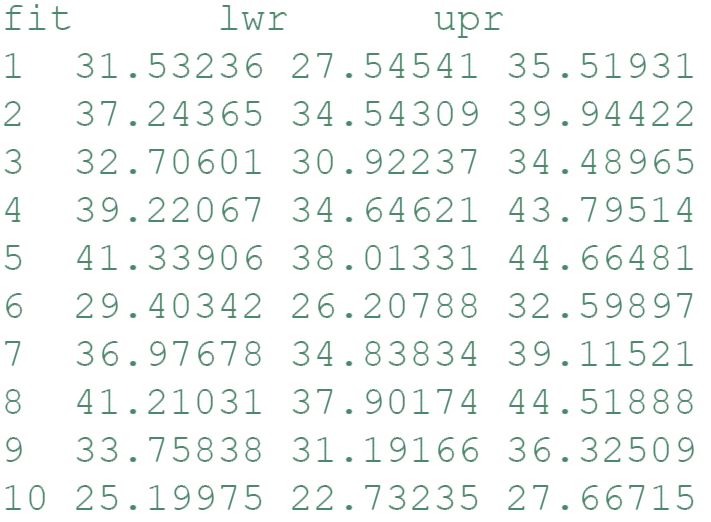


Figure 10: Prediction and Confidence Intervals

The confidence interval is giving a range for the estimated value of Y given X but the prediction interval is giving the range for the estimated value of Y itself. Estimating Y requires including the variance from the true error term also; hence the prediction intervals are wider than confidence intervals.

There are other measures also to assess the accuracy of predicted models such as MinMax Accuracy, MAPE (mean absolute percentage error), Mallows cp and Mean Square Error. Also the k-Fold Cross Validation method can be used to assess the robustness of the model under different circumstances.

1. **Changing the temporal resolution and Bias.**

The temporal resolution used for the analysis in second and third part is on Hourly basis. For making things easier and for analysis at different temporal scale, the daily average of the values from Airboxes as well as RIVM was calculated. The methods of measuring model’s accuracy have been mentioned above briefly. Since we already know that: -

Mean Square Error = Variance(a) + Bias²(a), a being the estimator

To get a holistic view of the change in performance of the model Mean Square Error (MSE) was calculated for simple linear models at different scales i.e. hourly and daily shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| LOCATION |  | HOURLY | DAILY |
| 1 |  | 250.5252 | 143.6546 |
| 2 |  | 285.4531 | 115.9172 |

The MSE value of the simple linear regression models at different temporal scale clearly shows that the error in the performance of the model at the scale of daily average is almost half as compared to Hourly average. That means the model performed almost twice as better at the coarser scale than the detailed scale.

***PART 4: CARTOGRAPHIC REPRESENTATION***

1. **Justification for choice of maps**

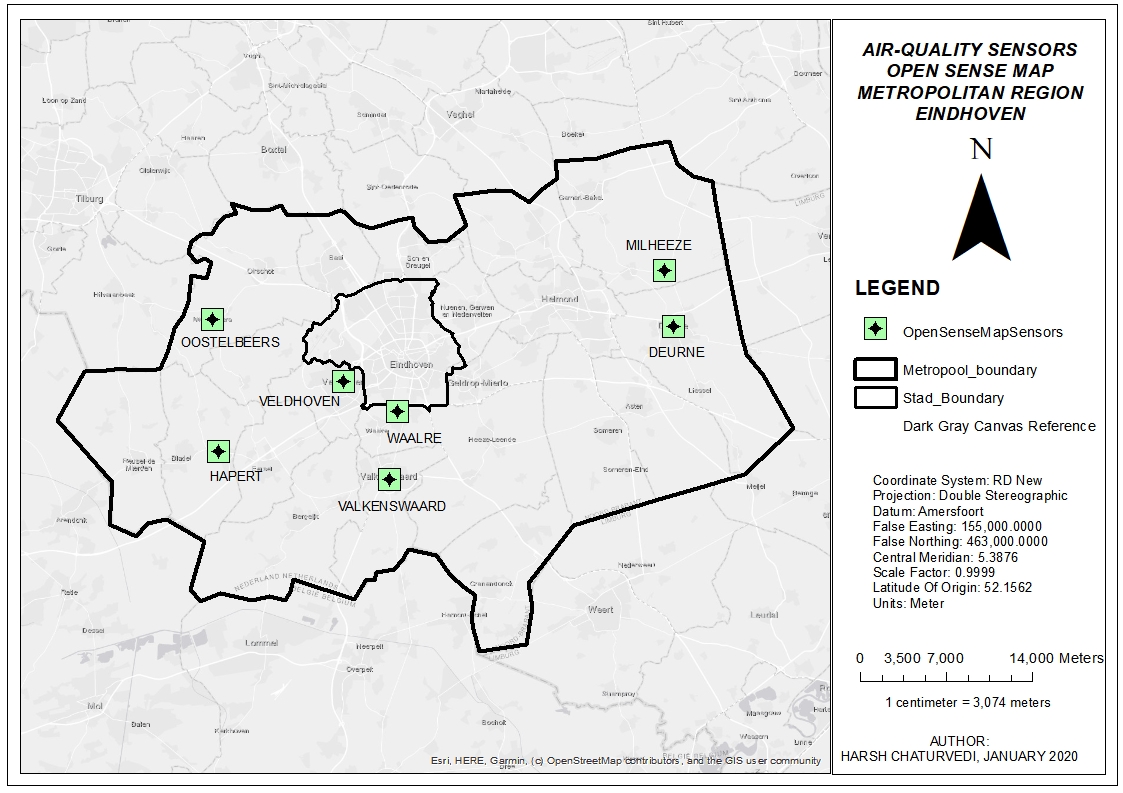
In the first map, data coming from Open Sense Map sensors have been depicted. Eindhoven Metropolitan region includes several municipalities within it and it is not just limited to Eindhoven city limits, which was the reason for showing a broader scale. Also due to no sensors present within the city limit, the decision to choose Metropolitan region was made for analysis. The green symbols in the map depict the location of sensors with respected municipalities. In Section2, the data from Veldhoven sensor (South-West to city boundary, also closest to any Airbox) was compared to Twickel 30 Airbox.

The second map depicts the location of all Airboxes within the Eindhoven Municipality boundary. These sensors include the location of RIVM high precision sensors (in purple pentagon) as well. The purpose of including this map was to observe the spread of the airboxes within city limits. Cluster analysis could have been implemented to check if the airboxes follow some patterns or not.

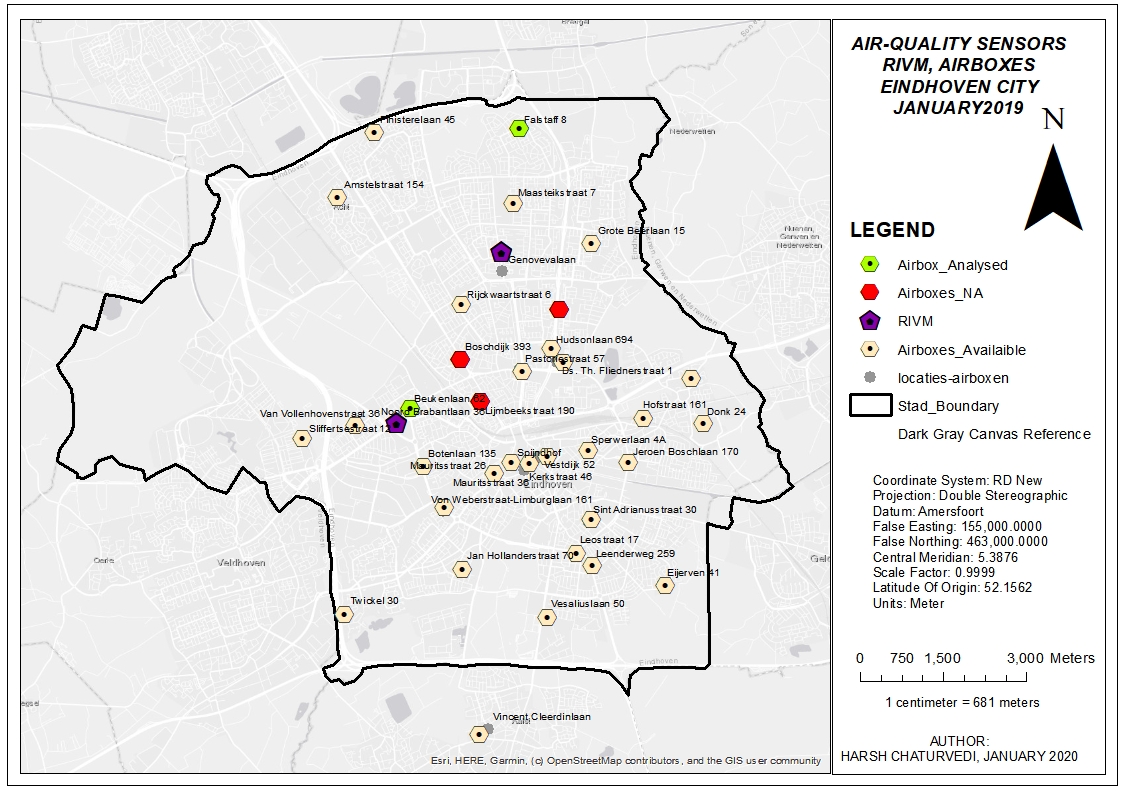
The fourth figure is a combination of Spatial and Temporal spread of the pollutants in the city of Eindhoven, during the month of January. The figure only depicts the average values from the analysis done between 1. Falstaff airbox vs GenovevalaanRIVM and 2. Beukenlaan Airbox vs Noord BrabantlaanRIVM. The figure gives us a general perception of the spread of pollutant in 2d space which is not the case in reality. The method used, decreases the intensity of pollutant as the distance from the Airbox increases. This model can be improved by including the average values of NO2 from all the sensors which provide the measure within that time frame. Open Sense Map also uses the same technique to show the spread of pollutants. We can notice the spread from hourly measurements is more than the aggregated daily values. This also takes us back to the 4th part of Statistics section where changing the temporal resolution of the data reduced the variation of model.

The final map depicts the green areas in the city which consists of parks, Enforcement areas and restricted areas which are essential to be depicted in any Air pollution study. These areas indicate the locations within a city that acts as Absorbing areas and affect the spread of pollutants (can be a thesis in itself, more relevant to Urban Planning). The enforcement areas consist of zones that are not allowed for heavier vehicles like Trucks, Trolleys to enter. These zones are implemented to maintain the air quality in the city centre. The enforcement areas consist of zones that are not allowed for heavier vehicles like Trucks, Trolleys to enter. These zones are implemented to maintain the air quality in the city centre. The Protected areas are the one that requires Special permits for any projects to start in vicinity.

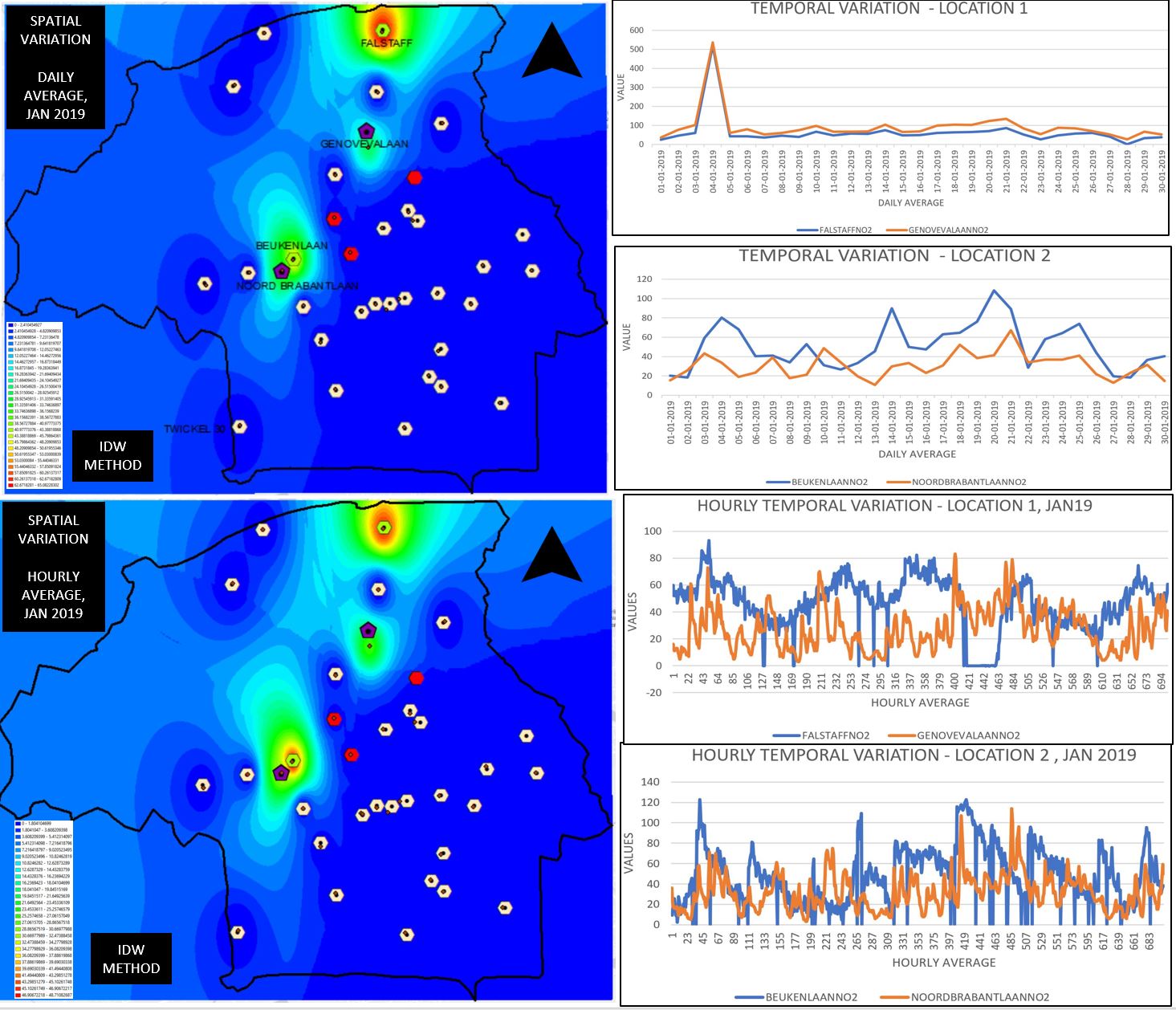
A complete webservice can be implemented which works on models including meteorological variables such as wind speed, wind direction, temperature etc. to visualize near real time concentration of pollutants by analysing measurements from the Airboxes and the RIVM instruments.



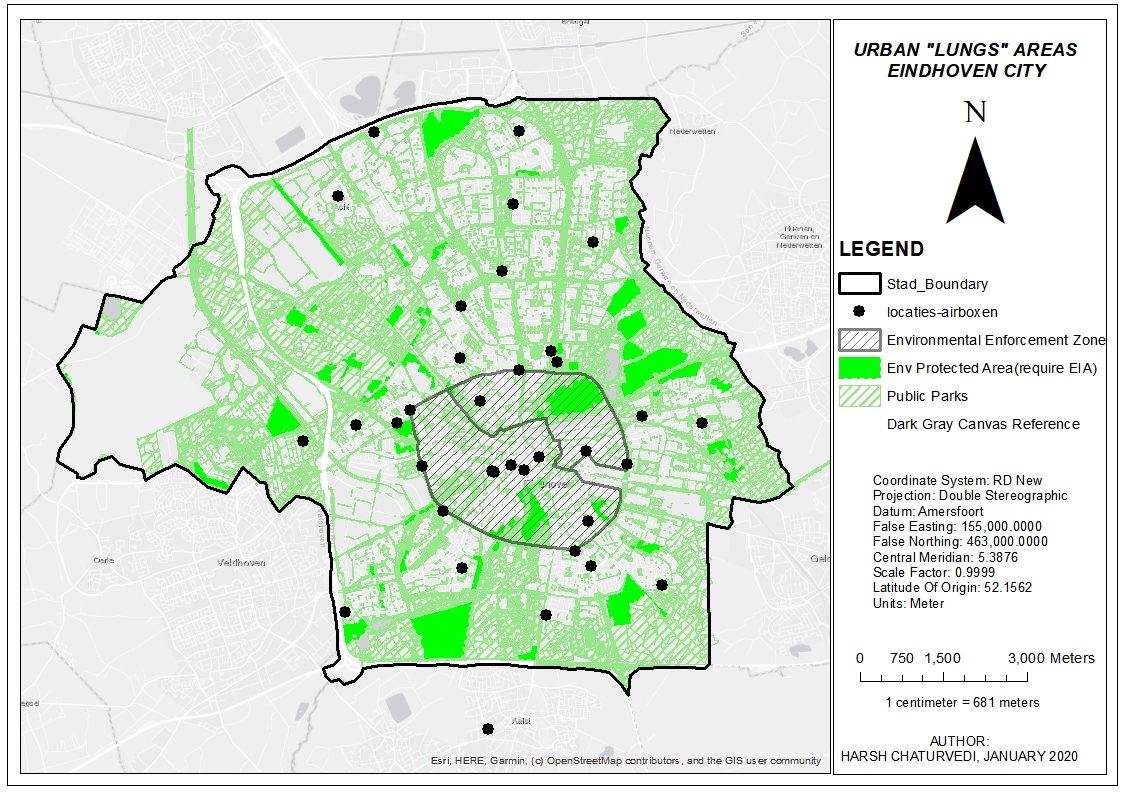
MAP 1: Sensors from open sense map



MAP 2: RIVM and Aireas Sensors



MAP 3: Spatial and Temporal Variation of NO2, January 2019



MAP 4: Absorbing "Lung “Areas

***REFERENCES***

Huang, J., Duan, N., Ji, P., Ma, C., Hu, F., Ding, Y., … Sun, W. (2019). A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments. *IEEE Internet of Things Journal*, *6*(2), 3240–3247. https://doi.org/10.1109/JIOT.2018.2881240

M.I.N.P. Munasinghe, G.I.U.S. Perera, J.K.W.D.B. Karunathilaka, B. C. S. C. and K. G. D. M., & A. (2017). Air Pollution Monitoring Through Crowdsourcing. *The Institution of Engineers, Sri Lanka*, (2017), 499–506. Retrieved from https://www.researchgate.net/publication/320608854\_Air\_Pollution\_Monitoring\_Through\_Crowdsourcing