**ASSIGNMENT 13**

**Final Manuscript**

**DS-670**

Neha Kumari

**Abstract**

Smart city is all about Internet of things (IoT), as it produces large volumes of data. This data if harnessed properly can lead to remarkable insights, which is of high business value. Air pollution is a major problem in smart city, which has harmful effects on human health and climate.

Air pollution can damage the environment, causes global warming, acid rain, haze and smog. Air pollution measure is a complex task as it depends on local weather conditions, emissions both by traffic and industries and other local emissions.

This means, the pollution level changes many times a day in a region and to predict these levels with accuracy is challenge. Besides, we rely on sensor data, which may not always provide accurate pollution measures. According to WHO reports, 92 percent of the world population is exposed to air pollution above threshold levels and it has caused 7 million premature deaths globally. Air pollutants are the result of local emissions as well as pollution transported by the winds. Air pollution particles have different physical and chemical properties. Some particles are a result of direct emission whereas some are the result of chemical reactions with other particles or in a suitable weather conditions. Many pollutants remain in the lower levels of atmosphere for long; it then gets carried away by the winds/storms and thus travels near by places. Sensor technologies allow collecting the real time data and sharing this data across wide area networks.

Smart cities are well-connected cities with sensors and they deal with the urban problems smartly. It aims to build real time applications that can solve the smart city problems like pollution, traffic congestion, and waiting times, smart parking. These applications can help people in planning and decision making like what route to take, expected exposure to pollution on a given day, expected waiting time at some place. Not just the people but also even the local government can use these applications to take future preventive measures to check pollution in the cities.

In this paper we used K-means clustering to identify the healthiest and unhealthiest area in the city as well as time series forecasting using ARIMA model to forecast the pollution levels for the next day in the city of Aarhus.

**Work by competitors**

There are many research articles on air pollution analysis, which focuses on some highly polluted cities across the world mainly in Asia and Middle East. Beijing, Allahabad, Delhi, Riyadh, Al Jubail, Raipur, Patna, Gwalior are some of the most populated cities in the world according to WHO. In Beijing where the levels had reached to dangerous levels, the government of China has initiated many research and measures to control it. Artificial smog chambers are one example where the motive is to reduce smog and haze is studied.

In many cities, pollution levels have crossed the threshold values and reached to alarming levels. Tableau is a visualizing tool popular for exploratory analysis.

Both supervised and unsupervised learning algorithms have been used to analyze the pollution data.

* Artificial Neural network has learning capabilities. It works like human brain and learns from training data. ANNs have also been used to study the pollution sources. When a hidden layer in the neural network is introduced, it produces high accuracy rates.
* Classification algorithms like support vector machines, K-means clustering, Naive Bayes, and logistic regression models are used to classify the pollution into levels.
* Fuzzy logic (FL) calculates approximate results rather than accurate results. It is computationally expensive algorithms and therefore not very popular. In fuzzy model top-down tree is used to generate rules.
* Hidden Markov model (HMM) is also used for time series forecasting, based on Markov process. HMM are also used in speech and handwriting recognition, DNA sequencing.
* A project called Green horizon, in Beijing by IBM is to study the chemical reactions of air mixtures under various conditions. The model is then used to forecast the pollutants for several days. IBMs cognitive computing approach developed for Watson supercomputer using natural language processing and statistical techniques. This will be very useful in building a decision system. It will help in suggesting possible ways to take preventive steps in controlling the pollution levels, if the model forecasted high pollution in an area. Measures like building low emission zones, restricts the entry of heavy vehicles.
* The University of California has been researching the various aspects of air pollution – one is low visibility due to haze, examining the structure of these particles and effective measures to reduce the emissions that decrease the visibility. Second the agricultural emissions, the waste from livestock, farming and cleaning. Thirdly studying the atmospheric constituents that may lead to global climate change and also ocean and atmospheric systems.
* Data mining is a fully knowledge discovery. The main aim is to know the sources or pattern in the pollution. Patterns can be seasonal or specific to locations, weather conditions and traffic.
* PFCM (probabilistic fuzzy c means), it calculates the mean pollutant level in a location. This helps in setting a threshold contingency level of the area.
* If the data has timestamp or date then we can employ forecasting which can predict the levels for the next couple of hours or the next day. Auto regressive models rely on the past values for predicting future values.
* The topic of Air pollution is a flag-ship in Aarhus University, meaning it has high priority and intensive research is being conducted to understand the effects of particles under different conditions.

**Contribution**

The first type of analysis is the exploratory and descriptive data analysis. Here the pollution levels of various pollutants are studied, summary, skewedness, mean, median etc. Histograms, boxplots and scatter plot gives insights of the data set structure. Bar plot suggested that the range of the particles is nearly same on AQI. Also it indicates of any transformations that might be needed to work with the data. Our data did not need any transformations.

Next, we used K-means clustering to identify the healthy area in the city. The data set had five air pollutants- ozone, carbon monoxide, sulfur dioxide, nitrogen dioxide and carbon monoxide but we used only ozone level concentrations. It takes two parameters the data and the k, which is the number of clusters. The k-means works on similarity basis such that the similarity within the cluster is high while between the clusters is low.

First it randomly choses K objects, each of which initially defined as cluster mean or cluster centroid. For the remaining of the objects each object is assigned to these K objects to which it is most close (The closeness is measured in terms of Euclidean distance). It then re-calculates the cluster mean for each cluster also called the centroid. The process is repeated until there is no major change in the mean value of cluster. This phenomenon is called convergence.

Using ARIMA model needs some basic steps: first is to plot our data and look for patterns like seasonality, cycle or some linear trend. Next create a time series object of the data. We also checked if the series is stationary or not using Auto co-relation and partial auto co-relation diagnostic plots. Choose the right order of ARIMA model, create diagnostic plots for the residuals of the model and finally use it for forecasting values.

We also used time series forecasting using ARIMA model to predict the levels of pollution for the next couple of hours.

So, the study will aim to serve two purpose- first it will identify most polluted and cleanest area and second to forecast the pollution levels for the next day.

**Data**

The data is from the Smart city project, city pulse (<http://iot.ee.surrey.ac.uk:8080/datasets.html>). It is from the city of Aarhus in Denmark. Many sensors have been deployed to capture the pollutant levels. The data is generated from the sensors at every 5 minutes from August 1st 2014 to October 1st 2014. Each sensor assigns a value between 25 and 100, and after 5 minutes the value will be updated. If the value was less than 20 earlier then, then a new values between 1 and 10 plus the last value. If the value was high then last value minus random number 1 and 10. Air pollutant measurement needs very sophisticated instrumentation to meet accuracy requirements. Air pollution sensors are still in their early stage development. The monitoring site, that is the place where the sensors are placed is of great importance to capture the right pollutant levels

The data has total 449 files; each file is the air pollutant level at a particular location at 5 minutes gap. So we have the data for about 449 locations in Aarhus. In total there are nearly 1750000 records.

The data set has – Ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide, longitude, latitude and timestamp.

* Ozone: There are two types of ozone – healthy and unhealthy. Ozone occurring naturally in the atmosphere is called the good ozone, whereas ozone at the ground level (at troposphere) is considered as bad ozone. Ozone is created due to chemical reactions of the gases emitted by vehicles. Also under sunshine and without sunshine the reactions differ. It causes health issues among adults and children like respiratory diseases, congestion, and chest pain throat irritation.
* Particulate matter: Particulate pollution can cause heart diseases, ling cancer and asthma attacks. PM usually comes in two size ranges – PM10 and PM 2.5. Diesel is the main source of PM in vehicle emissions; wildfires smoke is another cause of increased PM levels in the environment.
* Carbon monoxide: The burning of vehicular fuels produces it. Breathing carbon monoxide can cause headaches, nausea, dizziness, vomiting. It is said that high CO levels causes slow poisoning in the humans when exposed for a long time because it reduces the amount of oxygen reaches to human tissues.
* Sulfur Dioxide: The main source of Sulfur dioxide in the atmosphere is fossil fuel burning and industrial pollution. It can combine with other particles in the environment and can cause haze. Haze reduces the visibility. It also is harmful for old monuments and stones, as it stains and damages them. On human health it can cause wheeze, cough, chest-tightness. It can aggravate the respiratory problems.
* Nitrogen dioxide: It is also responsible for lung diseases in humans particularly children and older adults. It is highly reactive gas. When combined with water, oxygen and other chemicals it causes acid rain.

**Method**

We used Zeppelin notebook for this implementation. Main languages we worked on were R and Pyspark.

Data Introspection/ preparation:

* As a first step data was loaded using spark data frame. It contained all the 449 files in a single data frame.
* The data was clean with no missing values. This made the data suitable for k-means and forecasting.
* The initial study showed that data was nearly normally distributed. This was evident from the histograms of all the five air pollutants
* The boxplots revealed, that data was not skewed. The mean and median were within the same range. Also the outliers were very few.
* There was slight co-relation between nitrogen dioxide and sulfur dioxide. This was evident from the co-relation plot. The co-relation was between 0.4-0.6. There was no co-relation between any other pollutants.
* The scatter plot also suggested that co-relation between other pollutants were minimum.
* The mean values of all the pollutants were in the range of 110-113 and the median were between 110-113. The measurements were made in AQI (Air quality Index)

Min. : 15.0 Min. : 15.0 Min. : 15.0 Min. : 15.0<br />

1st Qu.: 66.0 1st Qu.: 64.0 1st Qu.: 65.0 1st Qu.: 65.0<br />

Median :112.0 Median :110.0 Median :111.0 Median :110.0<br />

Mean :112.8 Mean :112.1 Mean :112.4 Mean :111.9<br />

3rd Qu.:159.0 3rd Qu.:159.0 3rd Qu.:159.0 3rd Qu.:159.0<br />

Max. :215.0 Max. :215.0 Max. :215.0 Max. :215.0<br />

nitrogen\_dioxide longitude latitude<br />

Min. : 15.0 Min. :10.06 Min. :55.99<br />

1st Qu.: 64.0 1st Qu.:10.15 1st Qu.:56.14<br />

Median :110.0 Median :10.18 Median :56.16<br />

Mean :111.7 Mean :10.18 Mean :56.16<br />

3rd Qu.:159.0 3rd Qu.:10.20 3rd Qu.:56.18<br />

Max. :215.0

* To make the data suitable for k-means we aggregated the data and applied group by operations. The data was grouped by the longitude and latitude. In this way we obtained a mean value for the pollutants at every location. So now we have 449 values, each value represented the mean pollutant level of a location. This was done to make data suitable for k-means.

X = ∑ xi / n

Where,

X = arithmetic mean of an air pollutant in certain location

n = number records in each file for that location

*x1 = sqldf("select longitude, latitude, avg(ozone) as avg\_ozone, avg(particullate\_matter) as particulate from tbl group by longitude,latitude")*

K-means clustering:

* K-means is an unsupervised learning algorithm for classification. Here we wanted to classify the locations based on air pollution levels.
* We used ozone levels as the deciding factor in classifying the areas.
* We used the algorithm with a starting value of k=4 and continued till k=10, until convergence was achieved to get the best results.
* At first we took the data with all five pollutants and applied k-means, where k changed from 4 through 10.
* At k=10 we got best results where we found the minimum levels of ozone at some latitude and longitude
* The other approach we used was using only ozone in the data set
* We applied k-means again from k=4 through 10, to get the results.
* This time our results improved as we got lower levels of ozone in our cluster, which we classified as the healthiest area in the city.

Forecasting using ARIMA time series model:

* Next after identifying the healthiest and unhealthiest areas in the city we wanted to forecast the pollution levels for the next day.
* For this we used the data corresponding to the healthiest area in the city and used it for forecasting.
* Our data was from 1st August 2014 to 1st October 2014 at an interval of 5 minutes.
* As a first task we separated the timestamp field in the data set into two different columns namely – date and time.
* Now we plotted the data to look for any patterns. Seasonal component refers to any fluctuation in the data according to calendar cycles. Usually seasonality is fixed and it’s either quarterly or annually. Next we look for trend of the series, to check weather there is an overall increase or decrease in the series. The other component is cycle, which shows increase or decrease in the series but not seasonally. The part of the series, which is not explained by any of these components, is considered residual or error. There was no trend, seasonality or cycles detected in the data.
* From the initial plot it looked that the series was non-stationary.
* ARIMA models needs the series to be stationary. A series is called stationary if it’s mean; variance and auto covariance is not dependent on time.
* The ADF test (augmented Dickey-Fuller) is a statistical technique to test if the series is stationary or not. The null hypothesis is that the series is not stationary while the alternate hypothesis is that the series is stationary. If the change in dependent variable is explained by lagged value is non-significant we reject the null hypothesis.
* If the series is non-stationary like in our case we can use a simple transformation like differencing. It helps in removing any trends and cycles. The main idea is if the original series change over time then the difference in their values might be constant. The difference is calculated by subtracting the current value with the previous value.
* The auto correlation graphs and partial auto correlation graphs are used to visually determine if the series is stationary or not. Apart from this, these graphs also helps in choosing the order parameters in ARIMA model.
* Interpreting ACF is very important, this graph display the correlation between the time series and lags. Also it helps in choosing the differencing order, which is usually 1 or 2. PACF displays the co-relation between time series and lags that are not explained by previous lags.
* There are many criteria to determine if the model is a good fir or not. The Akaike information criteria (AIC) and Baysian information criteria (BIC). The lower these values are the better the odel is. These values indicates how much information is lost, if this model is chosen.
* The data was converted to time series object using ts method
* Next we use the auto co-relation (ACF) and partial auto co-relation function (PACF).
* The ACF graph further confirmed that the series was non-stationary. The ACF graph was decaying very slowly and was above the significance range. There was high co-relation between the lags.
* So this indicated that we must include differentiating factor and we decided to use ARIMA(p,d,q) model.
* Next choosing the value of p (AR order), d (differencing factor) and q (moving average) was crucial.
* The auto regressive (p) component means using the past values in the regression. If we use only one value it means the AR order is 1,AR (1) or ARIMA (1,0,0) if two pasts values are used then it is an AR(2) model or equivalently ARIMA(2,0,0). A moving average MA(q), combines the past errors, q determines the number of error terms of the past models to include in the current model. The d value represents the degree of differencing (I (d)). It simply means the number of subtracting the past values from current. Usually the value of d doesn’t go beyond 2.
* We used auto.arima() function to find the best ARIMA model suitable for our data. This gave us the best ARIMA model as ARIMA (1,1,0).
* After using this chosen model, we need to understand if this model will produce good forecasts. So we plotted the ACF and PACF plotting of the residuals. If the model is good we would expect no significant autocorrelations between the lags.
* We compared the AIC and BIC values with other ARIMA models with different combinations of parameters p, d and q. We found the minimum AIC and BIC values at order (1,1,0).
* We plotted the ACF of the residuals and it did converge to 0. The levels were within the significance levels.
* We also plotted the residuals and it did look like white noise.
* Next we used this model for forecasting for the next 500 points.

Notable packages used:

* **ggmap, maps ,ggplot2**– These packages were used for visualizing the data on maps. The get\_map function was used to plot the clusters on the map using longitude and latitude.
* **Sqldf** : This package offers sql like queries which are helpful in aggregating and merging the data
* **Readr, dplyr**: hese packages were needed to combine the data in 449 files into one single data frame
* **Forecast**: This package was used for time series modeling. Arima() and auto.arima() functions were used to fit the model.

**Results**

K-means clustering:

Took 10 clusters, grouped data on latitude and longitude

K = 4

longitude latitude avg\_ozone particulate

1 10.17102 56.16822 94.32454 117.74445

2 10.18229 56.16382 113.46153 117.07341

3 10.17705 56.17164 130.43995 111.18956

4 10.17615 56.16417 109.35957 99.12508

K = 5

longitude latitude avg\_ozone particulate

1 10.17102 56.16822 94.32454 117.74445

2 10.18229 56.16382 113.46153 117.07341

3 10.17705 56.17164 130.43995 111.18956

4 10.17615 56.16417 109.35957 99.12508

K =6

longitude latitude avg\_ozone particulate

1 10.18324 56.17770 132.73874 113.9244

2 10.17545 56.16624 89.56244 123.9848

3 10.17683 56.16007 106.34493 114.8368

4 10.18525 56.16414 116.16317 121.0616

5 10.17839 56.16568 116.44108 101.5405

6 10.15835 56.16847 96.50405 100.9700

K = 7

longitude latitude avg\_ozone particulate

1 10.17647 56.16176 105.60458 115.36039

2 10.18355 56.15940 116.92120 96.96951

3 10.18582 56.16181 115.38675 124.21001

4 10.17874 56.17044 117.82135 110.40380

5 10.17218 56.16558 88.92402 124.20092

6 10.15835 56.16847 96.50405 100.97003

7 10.18290 56.17964 135.92837 115.61803

K = 8

longitude latitude avg\_ozone particulate

1 10.18023 56.17058 118.01701 112.23172

2 10.15081 56.16349 95.25544 99.63985

3 10.17543 56.16871 87.21657 123.12371

4 10.17732 56.16306 105.34921 113.45741

5 10.18570 56.16142 114.37760 98.11670

6 10.17707 56.17570 132.45073 124.08558

7 10.18096 56.15995 110.75201 125.16890

8 10.18214 56.17365 132.71768 105.81515

K =9

longitude latitude avg\_ozone particulate

1 10.18442 56.16306 116.07195 123.69064

2 10.16102 56.17356 136.82826 125.90213

3 10.18583 56.15928 115.45212 96.45877

4 10.18759 56.17984 133.05192 107.95164

5 10.15406 56.16399 95.34281 100.19310

6 10.17816 56.17024 117.89435 110.68795

7 10.17013 56.17043 101.31702 123.55242

8 10.18032 56.17061 81.92253 121.50843

9 10.18094 56.15773 106.83063 111.85106

K=10

longitude latitude avg\_ozone particulate

1 10.17052 56.17031 100.98485 123.39569

2 10.18366 56.16245 115.32841 124.68005

3 10.17410 56.15099 114.33345 89.32272

4 10.18570 56.16588 111.94914 102.53156

5 10.18032 56.17061 81.92253 121.50843

6 10.18209 56.17875 137.23991 119.00751

7 10.15168 56.16368 94.99104 101.03512

8 10.17870 56.16995 127.04528 103.73866

9 10.18328 56.17163 118.22879 112.93678

10 10.17883 56.15726 106.61785 112.47661

With K = 10, for cluster 5, the ozone concentration is minimum at longitude 10.18032 and latitude 56.17061.This approach gave us minimum ozone levels as 81.92253 and maximum ozone levels: as 136.89854

Time series forecasting:

We used three ARIMA models

ARIMA (1,1,1)

|  |  |
| --- | --- |
| Coefficients | -0.0085 |
| Standard Error | 0.4434 |
| AIC | 90856.23 |
| ME | -0.00231288 |
| RMSE | 3.211871 |
| MAE | 2.769834 |
| MPE | -0.1918734 |
| MAPE | 4.219638 |
| MASE | 1.00131 |
| Log likelihood | -45425.11 |

ARIMA(1,1,2)

|  |  |
| --- | --- |
| Coefficients | -0.0088 |
| Standard Error | NA |
| AIC | 90857.19 |
| ME | -0.002332561 |
| RMSE | 3.211777 |
| MAE | 2.769982 |
| MPE | -0.1930519 |
| MAPE | 4.219568 |
| MASE | 1.001364 |
| Log likelihood | -45424.6 |

We used finally, ARIMA (1,1,0) model for forecasting ozone level for next 500 points. This model gave us the minimum AIC value. The results are below:

|  |  |
| --- | --- |
| Coefficients | -0.0176 |
| Standard Error | 0.0075 |
| AIC | 90854.27 |
| ME | -0.00231218 |
| RMSE | 3.211875 |
| MAE | 2.769806 |
| MPE | -0.1918289 |
| MAPE | 4.21961 |
| MASE | 1.0013 |
| Log likelihood | -45425.13 |

Note we did not use moving average, MA(0) . This might indicate that our model did not need smoothing. Auto.arima function is used to get best ARIMA model confirmed by ACF, PACF plots.

**Discussion**

This study helped us understand the air pollution conditions in the Aarhus city, Denmark. There is an on going research in the Aarhus University on air pollution and it’s effect on health mainly in Denmark, Europe and Arctic environment. Denmark has been very active in reducing the pollution levels since early 90s. The government has initiated programs like taking bicycle to work and even created separate zones for cyclists. This has created lot of public awareness and helped in controlling the traffic pollution.

Air pollution levels in Denmark have been decreasing since the year 2000 as a result of these measures and in Europe because of reduction in emissions and government initiated programs. It has a solid funding of nearly 600 billion euros per year in Europe, out of which 4 billion euros alone in Denmark. It is also a part of EU policy.

The ultimate goal would be to use this study into a decision support systems that aims to improve the quality of life in an urban city.

The smart city project can be integrated further to understand the impact of road traffic and on future urban planning and development.

It should fulfill the following needs:

* Monitor real time air pollution.
* Pollution exposure for residents, pedestrians and drivers on an average day.
* Generate reports to take decisions on urban planning and support.

The benefits we can expect are:

* Optimize road traffic in areas with high pollution and improve quality of life.
* Take informed decisions on urban planning.
* Increase public awareness and pollution reduction initiatives.

According to an analysis by WHO, almost 90% of urban population today breathes in polluted air such that the pollutant level is much higher than the recommended thresholds. It poses great health issues among children and adults. Also according to UN by 2030 nearly 60% of world population will live in urban areas. The reports also suggests that cities only occupy 2% of the earth’s surface but are responsible for nearly 70% of the carbon dioxide emissions in the global environment.

The other area of work in this field can be accurate forecast of air pollution levels. The sensors collect real time data and can be used for real time series forecasting. Deep Learning frameworks like Tensor Flow, azure, theano, Torch, Keras, Caffe, with computer vision are powerful tools used for real time analysis. Deep Neural Networks, artificial intelligence can precisely predict the pollution levels in several day advances. The neural network having more layers, and each layer are linked which helps in accurate predictions. ELM is a two-layer neural network that works well for predictions. First layer is random and second layer is trained. It is used for classification, regression, feature selection etc. Bu there are certain disadvantages also of using ANN, poor generalization of training data or over-fitted values.

The training process involves hundreds of iterations wherein the algorithm learns and reduces the error, meaning the difference between the forecasted values and actual values. The weather, road traffic along with air pollution data can be needed for accurate forecasts as it also affects the pollution levels.

Further, there are many research going on, one of them is to collect highly localized data, meaning instead of collecting data from sensors, sensors being attached to the wrists of people who travel across the common routes like subway, trains, cars, pedestrian walking. This study can be immensely helpful to understand the exposure to air pollution of common people at different time of the day. The other is the government has approved building Smog Chamber in Beijing, China. The chamber would emit different pollutant mixtures and create conditions similar to smog, to study them. Like the chemical reactions that take place under sunshine and without sunshine to control heavy smog.

In our project findings, we predict the cleanest area for the day as per the data available from the sensors. In this study ozone was the deciding factor, other pollutants can also be used to see and investigate their effect. The data set was pretty clean and structured, well suited for K-means and time series analysis. The results may vary for each run, as the initialization of K-means is different for every run. But as per our observation in most runs the best results were obtained with K=9 or K=10.Also time series forecasting was used for the location having minimum ozone levels. We got our basic time series running. The next we can discuss about is to ways how we can improve it. ACF and PACF are some diagnostic plots for studying the series and choosing the right ARIMA model that can produce some meaningful predictions.

Other forecasting techniques like exponential smoothing can make the model more accurate where seasonality, trend and historical values are used as weighted combinations. More complex models like ARMAX and dynamic regression, which allow inclusion of other predictor variables.

Now the big question is how can we use these results in improving quality of life in a Smart City. We need to build the responses, preventive measures and decision systems as per the predictions of these models. For example if our model predicts high pollution level in certain region, the place could raise the congestion charge, block heavy vehicles like trucks or encourage people to use public transports, electric cars, car sharing, creating green areas initiatives, mobile apps that can direct to the greenest route for that day just like Google maps for fastest route. On a larger level where the air pollution has already reached to alarming levels like in Beijing some preventive measures are taken like the regulation to control air pollution is done. Every city will have a yearly quota to limit the pollutant emissions; the purpose is to gradually control the pollutant levels and in particular smog in these areas. Other measures like Building low emissions zones (LEZ) is also becoming popular in many parts across the world. Smart parking apps can serve multiple purpose of reducing traffic congestion, fuel consumption and subsequently emissions. It is a step to control pollution in urban cities and usually high pollutant emission vehicles are not allowed in these areas.

The effect of industrial gases in industrial areas can be dealt separately.

In fact many experiments and research are going on in collaboration with companies like Intel, Siemens and government organizations with the universities on this subject. There are real time Air Quality Index apps using deep learning frameworks like Tensor Flow, theano, being developed locally. In many cities like Chicago, Pittsburg, Oregon, cities in Germany the air monitoring systems are working that provide the air quality in real time for different locations in the city. Cameras along with sensors let people know the source of pollution like traffic, industrial sources, Carbon dioxide emissions etc. Further the effect of green house gases on the pollution level can be studied. It’s been on research level and the next step would be to build these data driven systems that can provide some useful insights on which decision systems can be made. The Smart City is a connected city so these findings can be useful for integrating some other smart city projects working on different or similar data sets like weather or road traffic.

**Conclusion**

* K-means gave us the most environmentally clean are at longitude 10.18032 and latitude 56.17061 where ozone level was minimum, 81.92253 on Air Quality Index.
* The unhealthiest are was found to be at longitude 10.18209 and latitude 56.17875 with ozone levels 137.23991
* Time series forecasting gave us results with 95 percent confidence levels for the next two days (to be precise next 41 hours).
* According to the table below that provides the AQI levels and it’s meaning, we conclude that the unhealthiest are we found had the levels 136.89854, which may affect people in the sensitive group and not the general public. The overall air quality index of the Aarhus city is good and healthy for the general public living.

|  |  |  |
| --- | --- | --- |
| AQI index level for health | Value | Meaning |
| Good | 0-50 | Air Quality is satisfactory with no or little risk |
| Moderate | 51-100 | Air Quality is acceptable. Moderate health concerns for people sensitive to air borne allergens |
| Unhealthy for sensitive groups | 101-150 | Sensitive people are affected but not harmful for general public |
| Unhealthy | 151-200 | The general public may experience little health issues |
| Very unhealthy | 201-300 | Health warnings where the entire population is at risk |
| Hazardous | 301-500 | Health alert where entire population is at risk of serious health problems |