Air Pollutions in Ports

Nussara Tieanklin
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, WA 98195
nussara@cs.washington.edu

Abstract

This project aims to use the open data to raise awareness of how tourism affects the air pollution. The City of Aarhus, the second largest city in Denmark, has one the largest industrial port in the northern Europe. The number of the visitors keeps growing, as the mass concentration of PM10 and PM2.5.This project aims to investigate if there is any correlation between the air pollution and the cruise ship emissions. With the current available measurements from sensors throughout the city, we do not have enough data to conclude whether there is any correlation between an increased in the mass concentration of the benzene ambient and cruise ship emissions.

1 Introduction

The *Air Pollution in Ports* project is one of many projects initiated by the European Union. They have been investigating how to use emerging technology like robots, data science, and machine learning/deep learning intervention to improve the quality of life for people in the EU. Since 2016, Open Data DK, one of many projects that Danish municipalities and regions have collaborated together aiming to support the smart city vision. Not only the real-time data helps increase the transparency, the open data project enhances the smart city services and gains insights for further improvements in healthcare, public safety, agricultural, economy, tourism, and etc. Being surrounded by seas, in 2019, the second largest city in Denmark, Aarhus, welcomed more than 70,000 visitors by cruises. And the number of visitors continues to increase, as does the air pollution. This project uses data measured by a total of 35 air quality sensors, devices that measure pollutants in the air such as the ambient concentration of benzene (e.g., PM2.5) throughout the city of Aarhus.

In this project, the data science practices were used to investigate whether there is any correlation between cruise ship emissions and air pollution in the city. Additionally, a machine learning model using multivariate linear regression was performed to forecast the mass concentration of the PM2.5 in the city. With limitations in getting weather data like wind speed and directions that could affect how much pollutants in air sensors were detected, current machine learning using multivariate linear regression was able to predict PM2.5 with large mean squared errors (MSE). Given the available data given the time of the project listed in section 2 Dataset , the evaluation shows that we do not currently have enough data to conclude whether there is any correlation between cruise emissions and pollutants in the air.

2 Dataset

2.1 Measurements from the sensors

There are a total of 49 CityProbe2 sensors measuring pollutants in the air. 35 of which were deployed in the City of Aarhus. The raw hourly measurement data were requested from the CityFLow API

provided by the Open Data Aarhus. There are 30 different measurements that the sensors collect such as the amount of rain, atmospheric pressure, temperatures, luminosity, the mass concentration of the PM2.5, and etc. Since ship schedule is only available from March 30 until May 24, 2022, the measurement data date is then restricted by the cruise ship schedule. Therefore, the hourly data for each sensors are being requested individually, containing the total of 17,823 elements from March 30 until May 24, 2022.

2.2 Ship Schedule

As mentioned prior, the ship schedules unfortunately do not have any achieve data earlier than the March 30, 2022. The data were being retrieved from visitaarhus.dk, the official tourism information website provided by the city of Aarhus. One thing to note is that not everyday the cruise ships come to the city. The maximum cruise ships per day was 2 ships. From March 30 until May 24, 2022, there are a total of 19 ships were being scheduled. The data contain such as date, the number of passengers and crews, nationality of the cruise, how many hours does the cruise park in the city and etc.

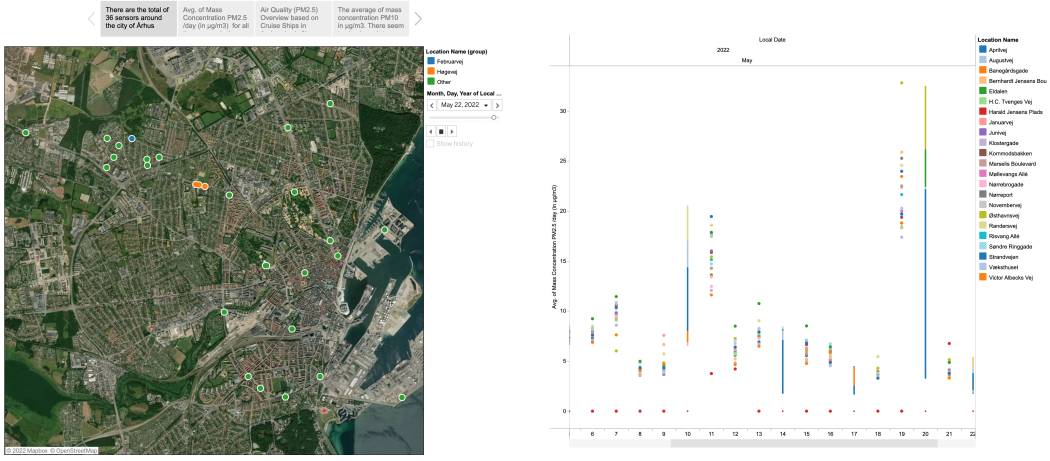
3 Methodology

The methodology that being used in this project follows the practices for data science and machine learning. The methods include find data quality issue (if any) as a pre-processing step, perform data exploration using data visualizations to gain more understanding about the data, and use a multivariate linear regression model to predict the mass concentration of PM2.5.

3.1 Pre Processing Data and Data Quality Issues

After performing pre-processing data, many data quality issues were discovered. Such as temperature of 0 Celsius, which become not realistic as Denmark was in Spring season even the past month. Although no null or N/A data was found in the data set, any outliers also found, the light luminosity was as high as 8,500 lux. The average rain include some negative numbers, which is impossible. I have reached out and notified the to the engineers of the CityProbe about the problems.

Figure 1: All of the 35 sensors, detecting pollutants in air, throughout city of Aarhus are being shown in the map on the left. The trend of the concentrations of PM2.5 over time is shown on the right.



3.2 Data Exploration and Visualizations

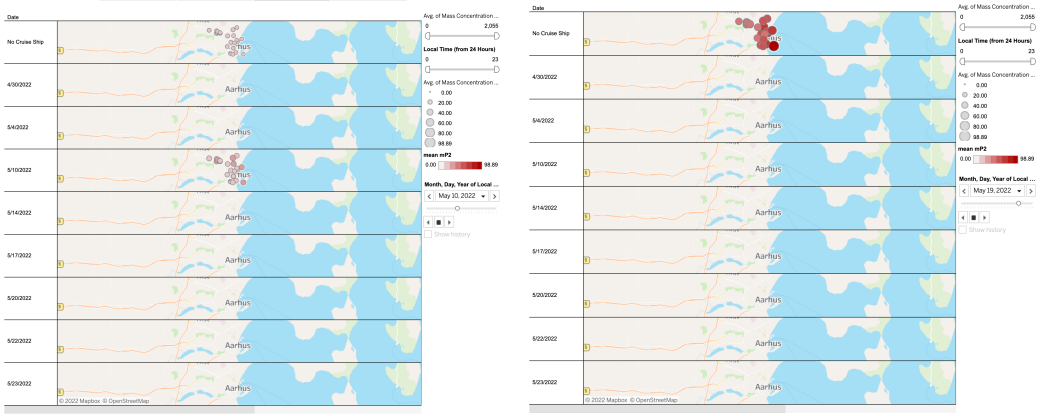
After the data is being processed, the measurements for every sensors for each hour everyday from March 30 until May 24, 2022 are being plotted. The interactive visualizations are available on Tableau Public

After performing preliminary analysis, the result shows that there were high surges occurred frequently in 2 sensor locations, *Høgevej* and *Februarvej*. The air pollutants were picked up on these two sensors 6-19 times higher than other sensors nearby. After the discussion with the CityProbe engineer team, they mentioned that the sensor in *Februarvej* location was placed too low on the station (rather than the normal height). Also the *Høgevej* location were surrounded by multiple construction sites, which may contribute to such high surges. With such circumstances and clarifications, these two sensor locations would not be able to measure air pollutants accurately. The two locations can be seen on the map encoded by blue and orange color in figure 11 on the left. Therefore, the two locations were excluded from the analysis in the following visualizations.

Figure 11 on the right indicates that there are, in fact, an increasing trends being observed on the day that cruise ships were in town and the increased in the mass concentration of the PM2.5. The days are indicated by straight lines being drawn. However, there were still some days that the mass concentrations of the PM2.5 increased, even though no cruise ships were in harbor.

We can see the information being shown clearer in the next following visualizations. The interactive visualizations in Tableau Public allow the audiences to explore further hour-by-hour how the mass concentration of PM2.5 change over time. The concentration of polluted air are being represented in circles. The larger and darker red the circles get, the higher PM2.5 is. The left plot of figure 2 2 shows the day that the cruise ships were in town. On the other hand, the right plot shows the one of those bad days that the city were highly polluted without any cruise ships nearby. Moreover, the similar trend is also shown in the mass concentration of PM10. More information please visit Tableau Public.

Figure 2: The left plot shows the day that the cruise ships were in town. The right plot shows the one of those bad days that the city were highly polluted without any cruise ships nearby.



3.3 Prediction

After understanding the data better from the data exploration, a multivariate linear regression model is used to predict the mass concentration of PM2.5. Model is based on variables listen on table 1 1, with the last variable (i.e. the mass concentrations of PM2.5) is the label and excluded before the training process. The test set is set aside the total of 5347 samples, and the model is being trained on 12,476 samples. Although the model was able to provide predictions for the mass concentrations of PM2.5 values, high mean squared errors (MSE) is also presented.

4 Evaluation

Given such limited dataset, even though we were able to predict the mass concentration of PM2.5, it results in a quite high value in return as well. The result indicates that we cannot conclude any correlation between the cruise ship emission and the increased in PM2.5 pollutant air with the available data at the moment.

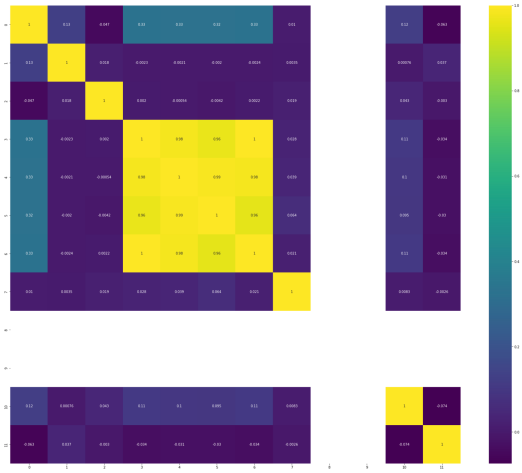
Table 1: The multivariate linear regression model is trained on the following variables. The last variable (i.e. the mass concentrations of PM2.5) is the label and excluded before the training process.

Index	Description	Unit
0	luminosity	lux
1	atmospheric Pressure	hPa
2	rain average	mV
3	number of cruise ships that day	ships
4	the total passengers of the cuise	persons
5	the total crew members of the cuise	persons
6	how many hours the cruis stays in town	hours
7	which day in the week does the cruise arrives	0 is Sunday
8	month number of when the cruise arrives	1 is January
9	date number of when the cruise arrives	N/A
10	the time of the day that the cruise arrives	24 hour time
11	the mass concentrations of PM2.5	μg/m3

To assure the assumption, the heat map of the correlation in Figure 3 3. This confirms that the available weather data (e.g., luminosity, average rain, and etc.) and cruise ships (e.g., size, date, and etc.) have very small, almost none correlation with how much the sensors were able to pick up the pollutants in the air. Although there has been attempts to find such data like wind speed and its direction, but no free achieve were found given the time constraint.

Therefore, any predictions in the future will required more extensive relevant achieve weather data.

Figure 3: The heat map of the correlation between all the features and the label (i.e. the last column/row).



4.1 Conclusions

Aiming to help improve the quality of lives of people in the city of Aarhus, Department of Culture and Citizens' Services attempts to use the available data from the Open Data Aarhus to Open Data Aarhus enhancing the smart city services and gaining insights for further improvements in various areas. The *Air Pollution in Ports* takes a part of the goal by using the available weather and air quality measurements to investigate whether there exist any correlation between the tourism by cruises and increased in pollutant air. However, with the current available measurements from the total of 35 sensors throughout the city, we do not have enough data to conclude whether there is any correlation between an increased in the mass concentration of the benzene ambient and cruise ship emissions. An extensive search of open weather data is needed and deep learning model to encode more data into neural networks can hopefully help improve the accuracy of the model.