Dyson School of Design Engineering | MEng Design Engineering

Module Exam

Module code and Name	DE4-SIOT Sensing & IoT
Student CID	01065771
Assessment date	**10th Jan 2019**

Presentation URL (publicly accessible link):
http://youtu.be/ZUo2nYQ2qIo?hd=1

Code & Data (publicly accessible link):
Google Spreadsheet for Data Collection
https://docs.google.com/spreadsheets/d/1Q7huLmlmSzNR7hQMpo2Y38iL7hkVvBlJPPjcK9VUXV4/edit?usp=sharing

Github repo for data collection, processing, and webpage code
Note: API Credential files were excluded due to security.
https://github.com/gracechin/siot_coursework

Website Platform
https://gracechin.github.io/siot_page/siot_page

Coursework 1: Sensing

Introduction and objectives

According to an article (Ogden, 2016), "In London alone, bad air quality is thought to kill nearly 10,000 people a year." Currently, it is estimated that air pollution causes 15% of Chronic Obstructive Pulmonary Disorder which is predicted to the third leading cause of premature death worldwide by the World Health Organisation. A King's College London study estimates that London's two main pollutants (nitrogen dioxide, NO, and fine particulates) are responsible for 5, 900 premature deaths a year (Excell, 2015).

In this part of the project, the relationship between the how much people care about air pollution and the air quality throughout time is explored. The number of tweets about air pollution over a period is used to represent the awareness and care of the people. A periodic measure of the different pollutants in the air is used to represent the changes in air quality throughout time.

Data Collection

The data sources used are: (i) OpenAQ API, which provides open air quality data (OpenAQ, 2018). This is used to collect real-time pollutants concentrations of different locations. (ii) Twitter Developer's Standard Search API, which searches through tweets in Twitter (2018), a popular social media platform. This is used to collect real-time twitter posts with the key phrase "air pollution". Data collected from both sources can be viewed via this link in separate sheets.

Timestamp	Location	pm25	pm10	no2	o3	co
2019-01-04T14:00:00.000Z	Causeway Bay	46.7	64.3	165.8	2.7	852.8
2019-01-04T14:00:00.000Z	New Territories	53.7	95.2	102.5	2	877.6

Table 1: *Sample of the data collected from OpenAQ API*

Air quality being location dependent, different locations were selected as focus for measuring the air quality. The locations were selected based on the places that have higher tweeting frequencies (more tweets over a period) so that the changes and trends in tweeting frequency can be more obvious, and places that may yield more interesting data about pollution. Data collection occurred throughout ten days (day and night) due to the time dependency of tweet rates, since there are various time differences in the selected locations.

Timestamp	Text	User	Tag	Location	Specific Location	General Location
Fri Jan 04 14:45:13	RT @BkPhilanthropy: Mercury pollution in the air we breathe m is down 81% because of regulations. https://t.co/uhXmCAW5dm	KCSunshineMom		Geeks Resist HQ	other	-
Fri Jan 04 14:45:02	Air is OK near Croydon - Park Lane (Pollution Low : 1)	breathinglondon		London	London	UK

Table 2: *Sample of the data collected from Twitter API*

Ideally, the location where each tweet is posted is recorded as the associated location. However, the Twitter API can only provide that information for geo-tagged posts, which is the minority of posts. Instead, the location that is set on the tweeter's profile, which is provided for each tweet, is used. Further programming was used to categorise the tweets to the chosen different general locations (UK -United Kingdom, HK- Hong Kong, US – United States, IN-India), which is used to relate with the locations used for air quality data collection.

To prevent aliasing and to fulfil Nyquist's Criteria, the quantities of different air pollutants data are sampled at the maximum sampling frequency rate, which is a sampling period of one hour. OpenAQ API only provides new measured values every hour. The tracking of tweets about air pollution is recorded in real-time with a time stamp. This means that the fixed period chosen to define the tweet rate (number of tweets across a fixed period) can be decided after the data collection.

The data were stored using the Google Sheets API to automatically store the (near-) real-time streaming information into a Google Spreadsheet, which is stored in the cloud.

Time-series Data Analysis

For initial data processing, the different pollutants data for the individual locations were grouped to give average data for their respective countries/large city. For further simplicity, the measured pollutant concentrations were converted into an Air Quality Index (AQI) for each four general locations. Inspired by how the U.S EPA AQI was calculated: the largest IAQI (Individual AQI), which is the index for each pollutant, was chosen on an hourly basis as the current AQI (EPA Victoria, 2015). IAQI was calculated as such $IAQI = \frac{pollutant\ concentration}{mean(pollutant\ concentration)} x100$. Missing data were filled with the most recent previous data. For large data gaps in the data collection due to travelling and lack of Internet connection, as suggested by Pandas (n.d.) as the most common ways of handling these cases, the data gaps were removed, and the remaining data sets were re-indexed.

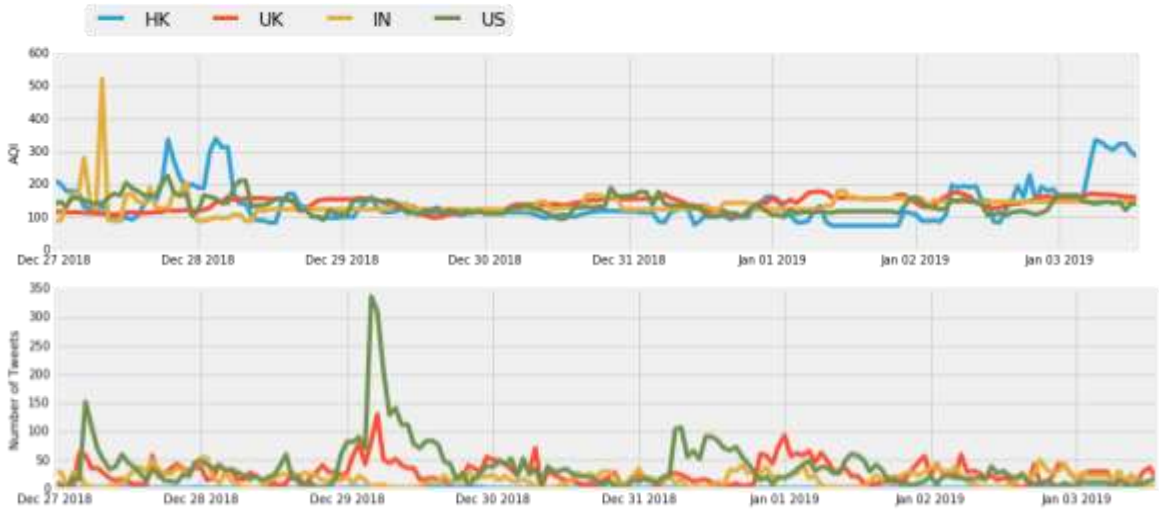


Fig.1: *Graphs showing data collected for AQI and Number of Tweets per hour in different locations*

There is no clear periodicity across the seven days. However, there are some fluctuations suggesting that more periodicity may be shown if more data were to be collected across a year monthly. Due to limited data, a conclusion as to whether there is a periodic pattern is unknown. There appears to be two relatively high peaks for both graphs suggesting anomalies. From looking at the data collected, the spike for the number of US tweets was found to be caused by a viral political tweet message, "The Trump Administration is attempting to dismantle the valuable work the EPA has done to protect the air we breathe. You can submit a public comment against this proposal by emailing a-and-r-docket@epa.gov. Include Docket ID No. EPA-HQ-OAR-2018-0794 in the subject line." The other spike for India's AQI was caused by a high concentration of PM10 pollutants in Delhi. Although the spike is indicated to be on the 27th December on the graph, due to time difference, the high AQI occurred on the 26th December India time. That is the day at which the FIATA World Congress event was being hosted in Delhi (Simhan, R., 2018).

Using Python, the different time-series datasets can be decomposed into trend, seasonality, and residual. From the graphs, it can be suggested that there is no clear trend, and a lot of residual (random variations) for both datasets, showing stochastic processes. Although there is no monotonic trend, the fluctuations show that the statistical properties (e.g. mean, variance, autocorrelation) change through time and is a non-stationary series. The lack of a clear trend can also be due to the lack of data. HK, which is relatively small compared to the other compared places, is excluded from further analysis due to the lack of tweet data.

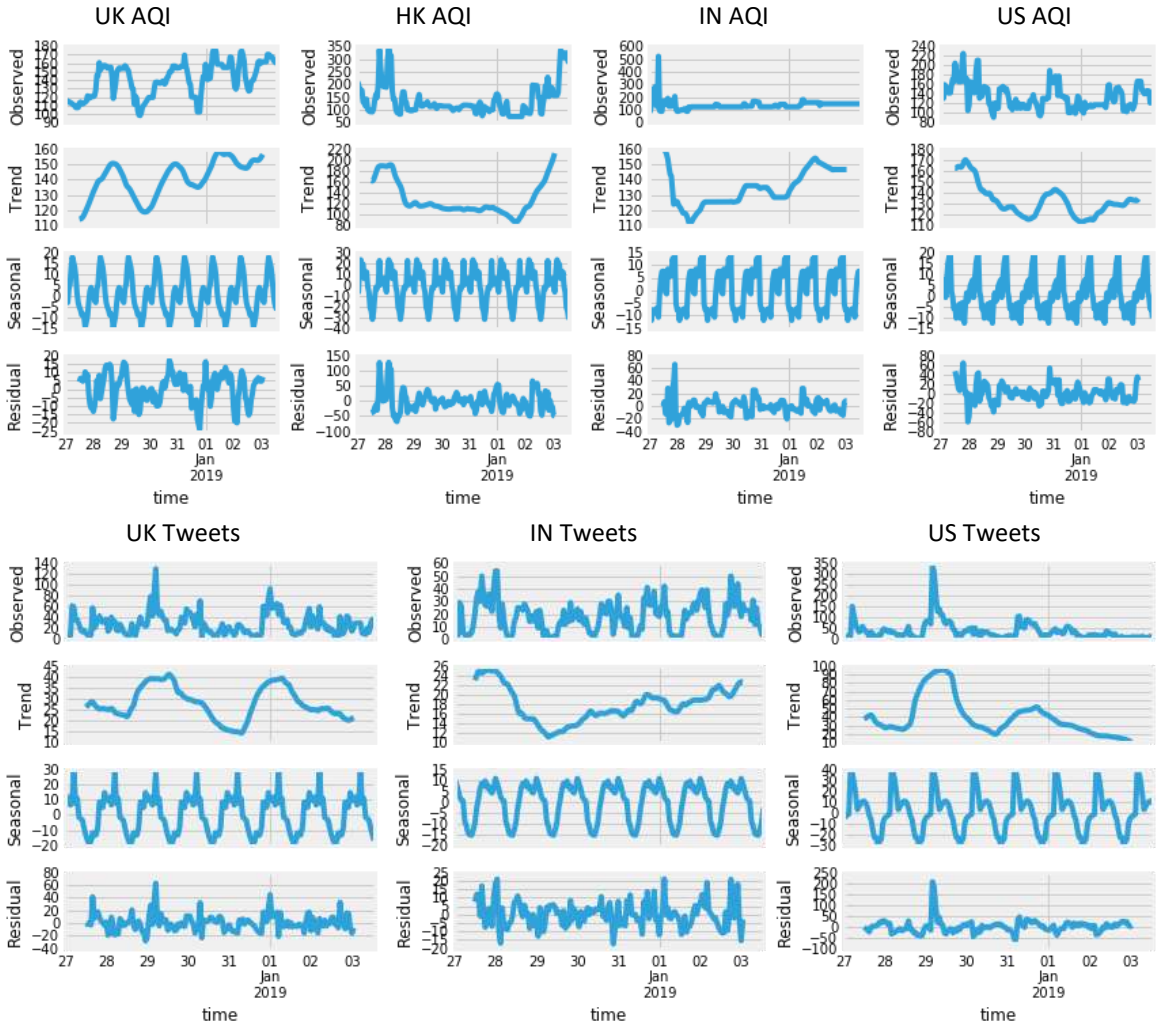


Fig.2: *Graphs showing decomposition of AQI and Tweets*

From the autocorrelation, and partial autocorrelation graphs for AQI and tweet rate, most lags are less than the confidence interval and are statistically insignificant. This means that there is a lack of correlation between a lagged function of itself, and hence a lack of seasonality for both variables.

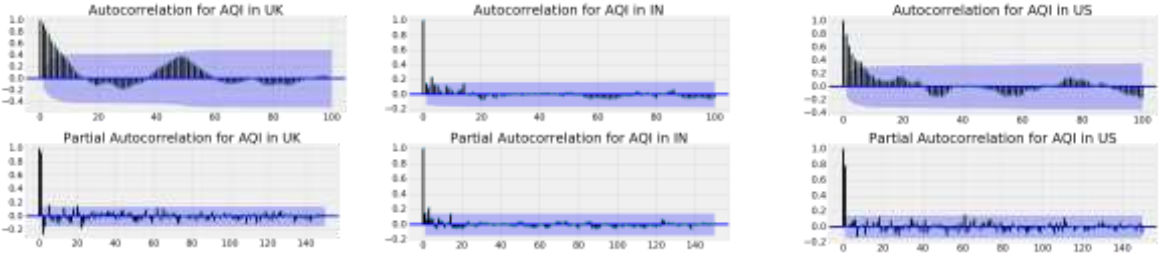


Fig.3: *Graphs showing autocorrelation and partial autocorrelation for AQI in different places*

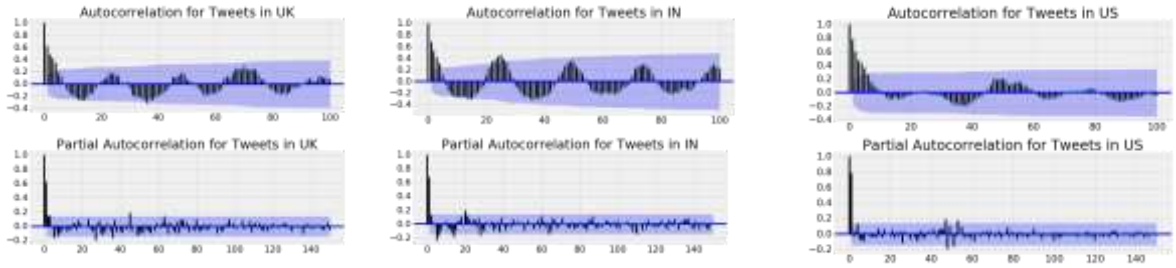
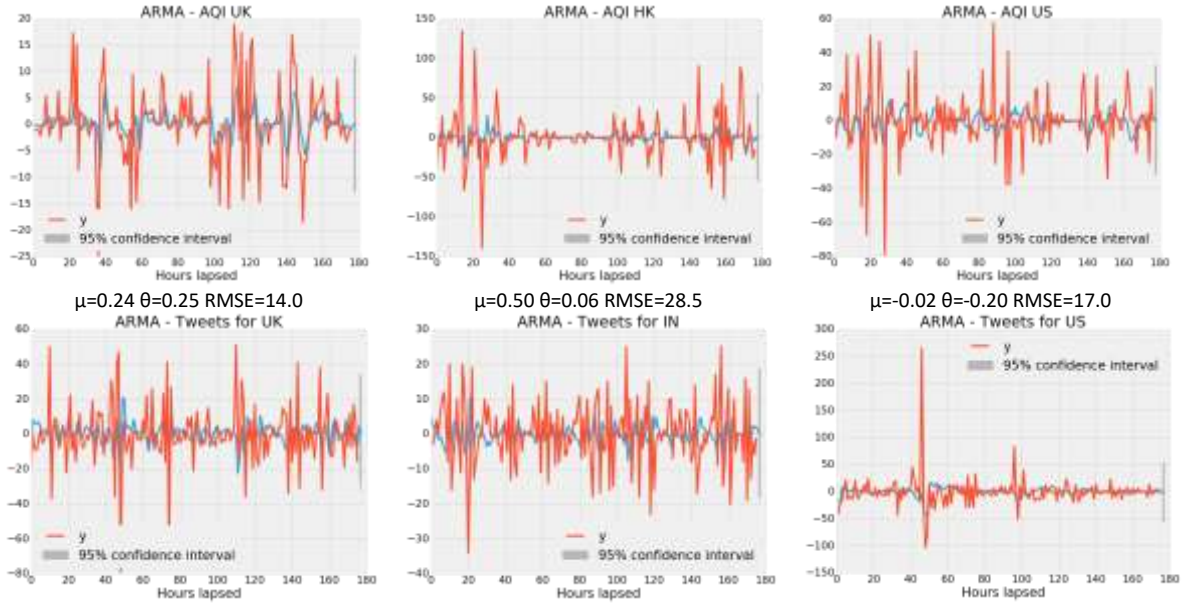


Fig.4: *Graphs showing autocorrelation and partial autocorrelation for tweet rate in different places*

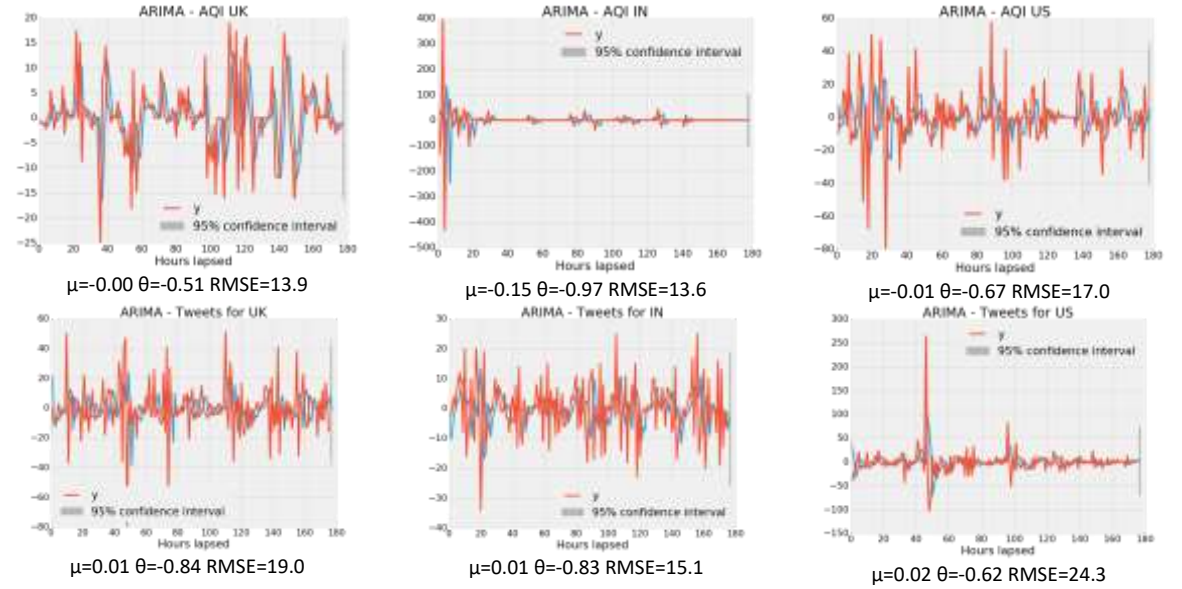
The graphs below show the fitting of linear time series models, ARMA (Autoregressive Moving Average) model and ARIMA (Autoregressive Integrated Moving Average) model, to help understand the time-series datasets and predict future points. Since the ARIMA model is a generalization of ARMA model to include cases of non-stationarity, it models the datasets better and hence the RMSE (root mean square error) is less or equal to that of ARMA's RMSE.



μ=0.24 θ=0.25 RMSE=14.0 μ=0.50 θ=0.06 RMSE=28.5 μ=-0.02 θ=-0.20 RMSE=17.0

μ=0.11 θ=-0.42 RMSE=19.0 μ=-0.13 θ=-0.31 RMSE=15.1 μ=-0.04 θ=-0.13 RMSE=24.3

Fig.5: *Graphs showing ARMA model of AQI and tweet rate in different places*



μ=-0.00 θ=-0.51 RMSE=13.9 μ=-0.15 θ=-0.97 RMSE=13.6 μ=-0.01 θ=-0.67 RMSE=17.0

μ=0.01 θ=-0.84 RMSE=19.0 μ=0.01 θ=-0.83 RMSE=15.1 μ=0.02 θ=-0.62 RMSE=24.3

Fig.6: *Graphs showing ARIMA model of AQI and tweet rate in different places*

Basic characteristics of the end-to-end systems set up and data

Below is a flowchart showing the data collection and processing set-up:

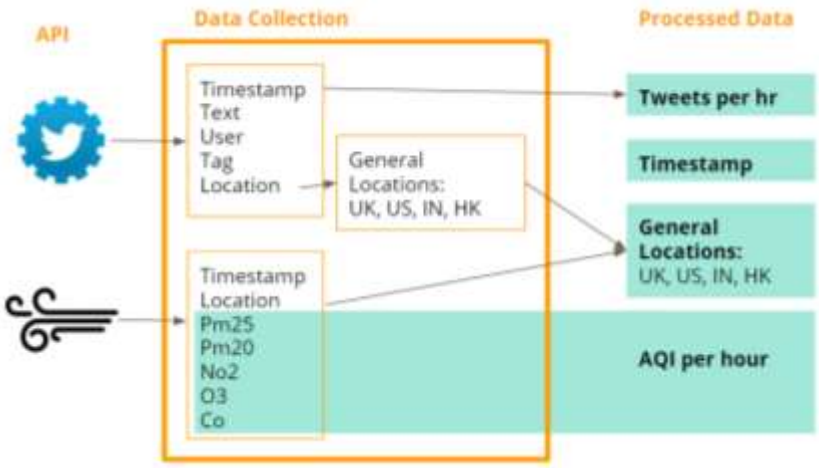


Fig.7: *Data collection and processing system set up*

Discussion

There is an initial assumption that the tweet rate about air pollution in a place would correlate to the care people would have to pollution in the area, however this may not be the case. As showcased by the anomaly, there are many other different reasons, such as internet virality, that may cause tweets about air pollution apart from the air quality of the area being poor. Filtering tweets with the keywords “air pollution” is not a rigorous method in filtering tweets, as there are other ways in saying air pollution, for example “air quality”, or comments on how clear the view is, or fog.

In addition to the assumptions, the data quality and quantity could be improved on. The labelling of the user’s location relies on a script which is very conservative and unthorough. Large time gaps and missing information were created due to not being able to collect data when physically travelling and when the internet connection was inconsistent. Although the time quantity was ten days, it was reduced to seven after removing the large time gaps. More data may help in detecting possible larger trends and periodicity, and hence change the results.

Conclusion

Based on the data collected, to conclude, the AQI and tweet rate about air pollution based on the data collected is stochastic and non-stationary. Unfortunately, no time-series patterns could be found.

Coursework 2: Internet of Things

Data analytics, inferences and insights

Using the data in Coursework 1, the relationship between the Air Quality Index (AQI), and Twitter tweets about pollution per hour in different places across time is explored. The cross-correlation below shows the similarity of AQI and the Twitter tweet rate for different values of displacement. The correlation is the highest when the functions displacement is zero, which means no displacement. Even though this is the case, the highest correlation is not significantly higher than the other displacement's correlations.

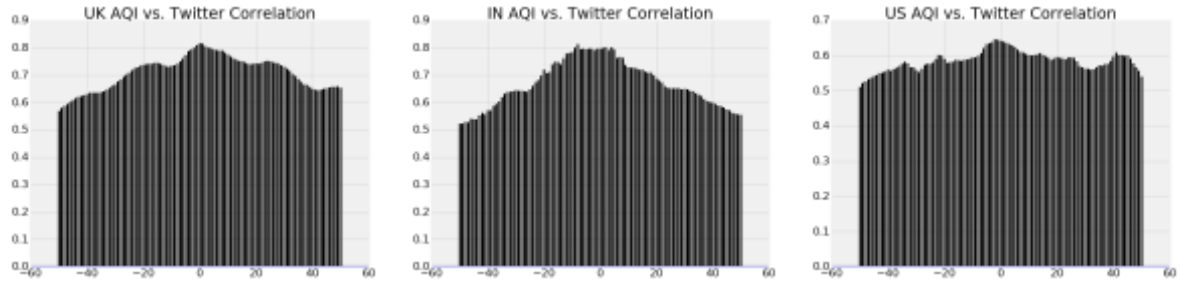


Fig.8: *Graphs showing cross-correlation between AQI and Twitter's tweet rate*

The respective Pearson product-moment correlation coefficients for the relationship between AQI and tweet rates in UK (0.13), IN (-0.03), and US (-0.01) suggest that there are insignificant correlations between the two variables for the different places. This can infer that there is no relationship between the pollution intensity of a locational area, and how much people in the area care about pollution.

Problem, Aim & Research

Ideally it would be beneficial for people to care about air pollution especially when there is high pollution intensity in the area. However, the data shows that this may not be the case. Although the problems with air pollution exist, some people are unaware of the impact. A London study (Taylor and Laville, 2017) showed that only one in ten British adults rated air quality as poor despite the country exceeding the legal limits of air pollution. An article that describes air pollution as a "silent killer" emphasizes the unfamiliarity of Air Quality Index (AQI) to some people in Vietnam (Dat, T., 2018). The effects of air pollution are difficult to comprehend due to it being invisible.

From the problem, the aim of the platform was established. It is to visualise the data found in an impactful way and to encourage people to care about air pollution. A website is chosen as the platform because of the ease of access to viewers. To achieve this, the different stages of design were used, and organised in a Gantt Chart, which can be viewed in this link.

Research on existing methods of making the air quality measurements more impactful were looked at for inspiration. Here are the two more creative ones: There is a website that animates the pollution at different months of the year as dots in a box. There has also been a shirt that visualises the pollution around you.

(Nieman, n.d) (Budds, 2016)

Fig.10: *Visualisations of* air *quality*

Data visualisation platform

Design Concept

Different ways of embodying pollution were considered, such as animating particles in a human body and blurring of a skyline of where the user is located. In the end, the chosen design concept was related to emphasizing the health impacts of pollution. This is to allow viewers to directly associate the pollution as impacting their own bodies, hopefully encouraging them to act upon the air quality around them as self-interest. The chosen idea was about animating particles in a lung, such that the state of the lung would vary depending on the pollution index. The number of grey particles would increase, and the lung would grow greyer as the pollution index increases. Pink was chosen as the background to resemble the human flesh. The lung changes were scaled following US-EPA (Aqicn, 2019).

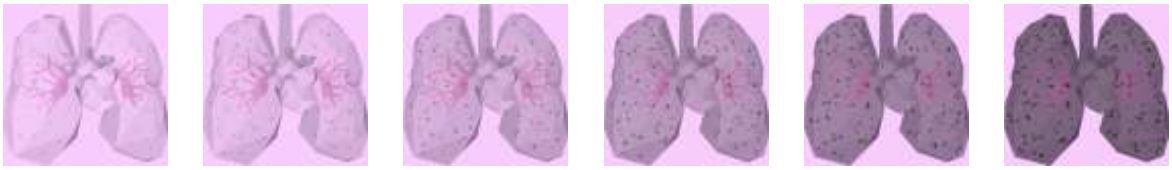


Fig.11: *Lungs reflecting air pollution index*

Data on the tweets per hour about air pollution was included through creating an index called the care index. The assumption that the number of tweets about air pollution per hour reflected how much people cared about air pollution was used. The care index is the current tweets per hour. To visualise the care index, similar lungs to the pollution index were animated such that the lungs had more pink particles and grew pinker when there were more tweets. A scale was created by comparing the current tweets per hour with the maximum recorded tweets per hour.

Fig.12: *Lungs reflecting care index*

To encourage people to care, a third lung was animated to predict how the affect of pollution on the lung changes as the care index and pollution index changes. The third lung reflected the pollution index that people were exposed to in consideration to the care index. This index was calculated as $pollution\ index - 30\%\ of\ pollution\ index * \frac{tweets\ per\ hour}{\max tweets\ per\ hour}$. It was found that a possible 30% of exposure to air pollution can be reduced through transportation decisions (Buechler, 2018.), hence to be conservative, a maximum of 30% reduction in pollution exposure was assumed and can be see in the equation above.

Prototyping and testing

The website was made as realistic as possible using HTML, CSS, and Javascript, and was hosted using Github Pages because it is free. The animations were created using gifs. Even though there is no cloud data storage, the data in the website was made as realistic as possible through collecting real-time data using API. Data for the pollution index was taken using the same API as the data collection. There was an attempt to use Twitter's API to generate the data for the Care Index, however, it was not possible due to the requirement of credentials to access the data. Github Pages does not support a dynamic site, which meant that the credential details cannot be hidden. Due to security reasons, the Twitter API was not used, and instead a random number generator was utilised to mimic the change number of Tweets per hour.

Fig.13: *Screen shot of front page of the website*

Features

Iterations were done through asking a user to see if they understood what the website was about and how to use it. To improve with the ease of use and access, the website was minimised and simplified. This prevented confusion. A help button was added upon user feedback for further explanation.

Discussions on the important aspects of the project

In reflection of the website created, there is room for improvements. Thorough testing and validation could have been carried out to see whether the website does positively affect the users and fulfil the aim. Although an API is used to generate realistic data, because of the requesting of data, the prototype can lag which reduces the user's experience. A preload animation may help lift some of the frustration of users, but perhaps better coding strategies around efficiencies may be utilised. Taking into consideration of user comments, although the minimalistic design was attractive, more information regarding what the website does, the significance of the tweets, and why the website is important could have been better explained.

In addition, as mentioned in the Discussion section of Coursework 1, the reliability of the data provided is also questionable due to many assumptions.

Avenues for future work and potential impact.

For future work, in extension of the website, an app that can visualise what the user's lung may look like depending on the amount of pollution the user has been exposed to is a possibility. This can make the system more personalised. Perhaps collecting data in different aspects of user's life, such as other social media and GPS, can be used to better estimate exposure of individuals and how much they care about air pollution.

For future avenues for future work, there are many applications for the types of data collected: how much pollution is in the area and how much people care about it. Gamification may be an option in terms of encouraging people to choose healthy options and act to the different pollution index. Social media can also be a channel in terms of targeting pollution information or posts towards places where there is a negative gap between the air quality and how much people care.

References for Coursework 1 & 2

Aqicn, 2019. Air Quality Index Scale and Color Legend. Available at: https://aqicn.org/scale/. [Accessed 10 January, 2019].

Budds, D., 2016. Forget Smog Alerts: These Shirts Visualize The Pollution Around You. *Fast Company*. Available at: https://www.fastcompany.com/3062129/these-pollution-sensitive-shirts-visualize-the-filthy-air-youre-breathing [Accessed 9 January 2019].

Buechler, J. 2018. Your Exposure to Air pollution could be much higher you're your neighbours-here's why. 22 June. *The conversation.* Available at: https://theconversation.com/your-exposure-to-air-pollution-could-be-much-higher-than-your-neighbours-heres-why-98486 [Accessed 9 January 2019].

Dat, T., 2018. Air pollution is Vietnam's silent killer. *Vietname Investment Review.* 26 February. [Viewed 26 December 2018] Available from: https://www.vir.com.vn/air-pollution-is-vietnams-silent-killer-56542.html [Accessed 8 January 2019].

EPA Victoria, 2015. Calculating a station air quality index. [online] Available at: https://www.epa.vic.gov.au/your-environment/air/air-pollution/air-quality-index/calculating-a-station-air-quality-index [Accessed 3 December 2018].

Excell, J., 2015. The lethal effects of London fog. *BBC.* 22 December. [Viewed 14 November 2018]. Available from: http://www.bbc.com/future/story/20151221-the-lethal-effects-of-london-fog

Nieman, A., n.d. A Breath of Fresh Air: Visualising Air Pollution. *Carbon Visuals.* Available at: http://www.carbonvisuals.com/blog/2016/2/3/a-breath-of-fresh-air-visualising-air-pollution [Accessed 9 January 2019].

Ogden, L.E., 2016. The tiny changes air pollution makes inside you. *BBC*. 11 February. [Viewed 14 November 2018]. Available from: http://www.bbc.com/future/story/20160210-the-tiny-changes-air-pollution-makes-inside-you.

OpenAQ, 2018. *Open AQ Platform API.* [online] Available at: https://docs.openaq.org/ [Accessed 23 December 2018].

Pandas, n.d. *Working with missing data.* [online] Available at: https://pandas.pydata.org/pandas-docs/stable/missing_data.html [Accessed 8 January 2019].

Simhan, R., 2018. Global logistics meet in Delhi on Sept 26. *The Hindu Business Line.* [online] Available at: https://www.thehindubusinessline.com/news/global-logistics-meet-in-delhi-on-sept-26/article24970412.ece [Accessed 8 January 2019].

Taylor, M., and Laville, S., 2017. British people unaware pollution levels air they breathe -study. [Viewed 14 November 2018]. Available from: https://www.theguardian.com/environment/2017/mar/01/british-people-unaware-pollution-levels-air-breathe-friends-earth [Accessed 8 January 2019].