Air Quality Monitoring with IoT Big Data

A Technical Guide for Data Processing and Analytics

July 2018

TABLE OF CONTENTS

1. INTRODUCTION

Today, the shift to becoming a data driven business is driving massive transformation across all industries. The telecommunication industry, by its nature being a massive producer of data which will only increase with the coming explosion of the IoT (Internet of Things) and 5G networks, is at the critical stage of transformation. Mobile operators are evolving from being connectivity providers to being intelligence service providers through the use of advanced analytics and big data technologies on IoT and other sources of data.

For decades, mobile operators have been capturing large volumes of data from both the Radio Access Network (RAN) and Core Network (CN) such as network statistics, customer calling patterns, data usage, subscriber profile and geographic information. Many operators have been applying data analytics tools and platforms to mobile data in areas such as network self - optimisation, network anomaly detection[1], user behaviour and mobility understanding[2], as well as transportation planning and management[3].

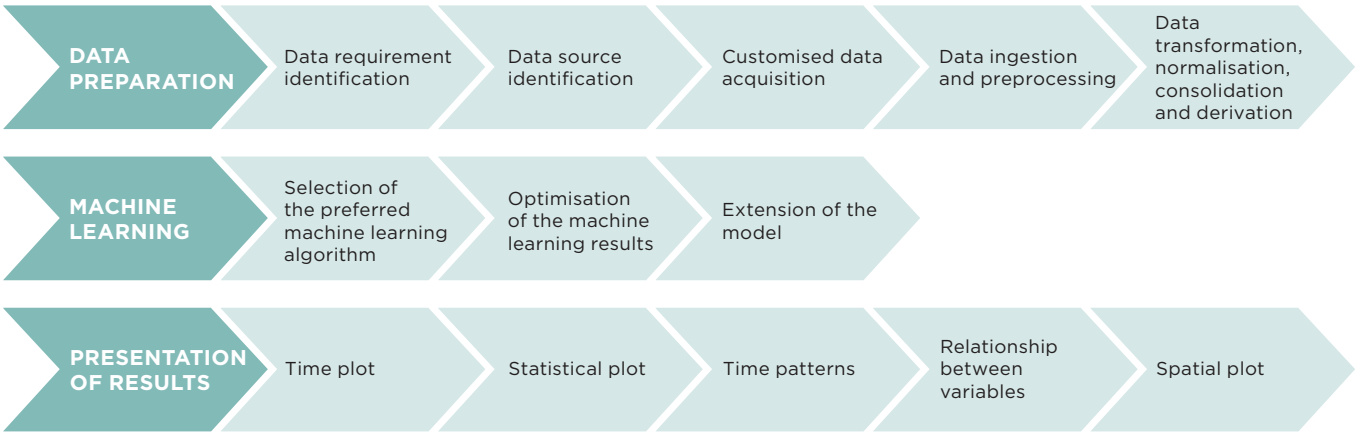
However, when entering into the big data era, the inherent characteristics of big data - volume, velocity, variety (heterogeneity), and veracity (uncertainty) - pose big challenges to traditional database and software technologies. Especially, with the scaling of IoT networks and the upcoming IoT data ocean inflowing from different verticals, new architectures and platforms which can help extract value from huge volumes of disparate data from multiple sources and enable real time analysis and decision making will be the crucial part in the whole transformation journey of the mobile industry.

According to recent forecasts from GSMAi[4], by 2025, total IoT connections will reach 25 billion and the total potential revenue generated from the IoT will reach $1.1 trillion globally, with more than two-thirds coming from platforms, applications, and intelligent services and only three per cent coming from connectivity services. With such significant market opportunities, mobile operators are already transitioning from being only connectivity providers to being end to end intelligence service providers. The advances in Artificial Intelligence (AI), Machine Learning, Cloud Computing and Big Data Technology, and collaboration with the wider ecosystem, make it possible for operators to take the advantage of their data assets and enable new business models to gain a big share from the potential IoT market[5].

This document, shares experiences and common data analytical techniques from the experience of developing analytics services in the air quality field based on joint projects firstly with Royal Borough of Greenwich in the UK and secondly with Far EasTone Telecommunications (FET) in Taiwan. The successful use case from FET is a good proof that mobile operators have the capability to not only deliver intelligence as a service but also add value by enriching the analysis

[1] Ilyas Aplper Karatepe and Engin Zeydan, “Anomaly Detection in Cellular Network Data Using Big Data Analytics”.
[2] Eyuphan Bulut and Boleslaw K.Szymanski, “Proc. IEEE ICC Workshops, Dynamic Social Networks, DYSON, London, June 8, 2015, pp. 1548-1553”.
[3] http://www.citilogik.com/uploads/1/3/7/9/13799627/utilising_mobile_network_data_for_transport_modelling_dft_dec2107.pdf
[4] https://www.gsmaintelligence.com/research/2018/05/iot-the-1-trillion-revenue-opportunity/670/
[5] https://www.gsma.com/iot/wp-content/uploads/2016/06/IoT.01-v1.0.pdf

with their unique mobile network data. Monetising the data from the IoT and delivering an integrated service are feasible and actionable for operators. The diagram below shows the approach taken from beginning to end in developing the air quality prediction models, and this document expands on each of these topics in turn. Logically the development process can be structured into three main phases, which allows for teams with quite different skills to work together to produce the results:



2. DATA PREPARATION

Data is the cornerstone for intelligence services - every big data analysis and machine learning project starts from data acquisition and collection. It is estimated that in the big data era, the rate of data generation is roughly 2.5 quintillion bytes a day[6]. To find the right data source for a specific application area is becoming more and more challenging because of the growth in data volumes and channels. This section describes the process of identifying useful data sources for air quality as well as the generalised process used by data analysts and scientists to prepare the data ready for analysis / machine learning.

2.1 Data requirement identification

It is useful as a first step to identify the data that may be relevant to the problem area. This step is to help direct the later search for suitable data sources. For some problems, such as air quality, there is a good body of published academic research which considers air quality outcomes along with the environmental and other factors that influence the outcomes.

[6] http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

With big data and machine learning techniques it isn't necessary to narrow the range of input parameters to a few high correlation input factors asthe algorithms can find those, and more subtle connections. However, this desk research step helps in identifying the types of inputs useful to the domain being studied so that the appropriate data sources can be identified.

For example, air quality inputs:

- There is an inter-relationship between certain pollutants such as Nitrogen Dioxide and Ozone gases, so measurements for these could form useful inputs as well as outputs;
- Weather factors including the intensity of sunlight, air temperature, wind speed, wind direction, air pressure and precipitation have an impact on the various pollutants;
- There is an element of hysteresis in the environment where pollutants build-up and disseminate over time;
- Human factors also influence the pollutant levels through factors such as the days forming the core of the working week, and the hours of the day when people travel to work, remain at work and return home after work.

Academic research can be identified using search engines[7,8,9], and references between academic papers are useful to identify the factors that might be useful to include in the eventual models. It is also useful to identify and engage with academics working in the related discipline who can help identify useful input factors and outcomes.

The figure below shows a general overview of the correlations between different variables of wind speed, wind direction, temperature, NO2, O3, PM10, and PM2.5 to identify which factors could become parameters for building a model. This 'pairwise' correlation is useful to identify any apparent correlations between inputs and outputs. Note this analysis is for a specific site in Greenwich.

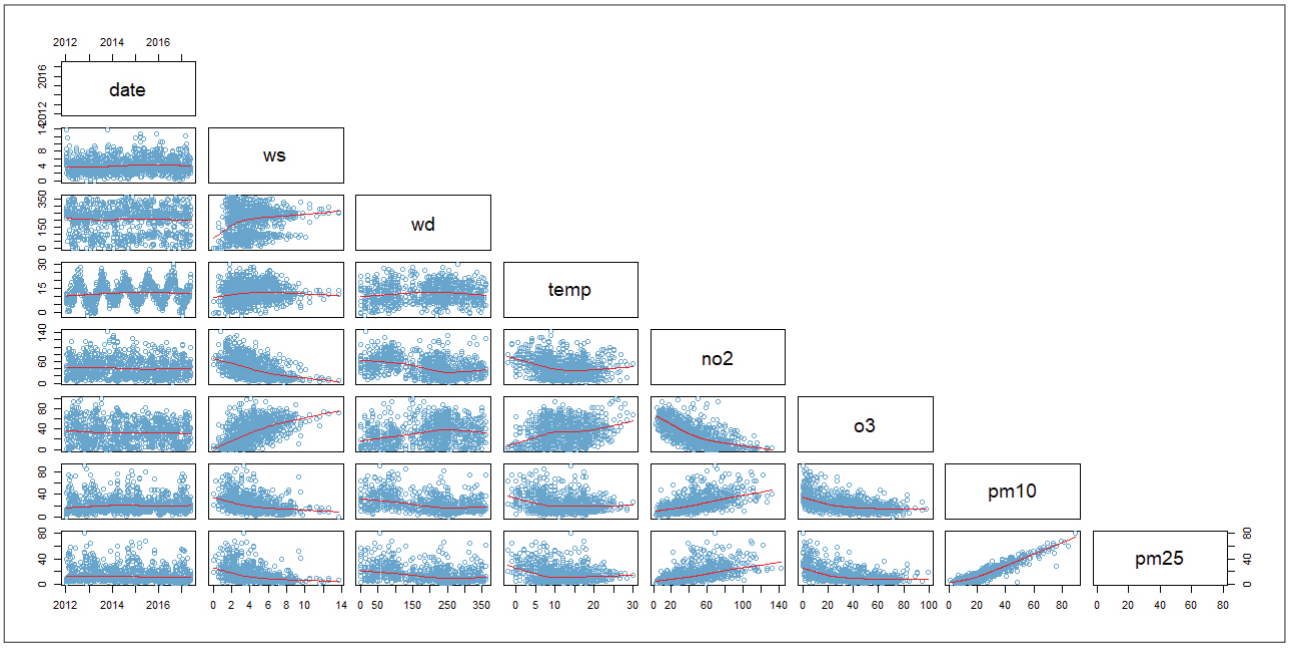


Figure 2.1 Pair plot of different variables

[3] https://www.london.gov.uk/sites/default/files/old_oak_and_park_royal_air_quality_report_draft_final_issued_new_cover.pdf
[8] https://www.diva-portal.org/smash/get/diva2:984443/FULLTEXT01.pdf
[9] https://www.nice.org.uk/guidance/ng70/resources/air-pollution-outdoor-air-quality-and-health-pdf-1837627509445

For example, Figure 2.1 shows that concentrations of NO2 generally gradually decrease with the increase of wind speed. Concentrations of NO2 show a slight curve shape with respect to temperature. This means the concentration first decreases with the increase of temperature above zero degrees Celsius, subsequently increases slightly and stabilises once the temperature reaches around ten degrees Celsius.

Not much information can be determined relating to the correlation of wind direction (compass degrees relative to north) to NO2 in Figure 2.1, however the polar plot in Figure 2.2 can provide further insights.

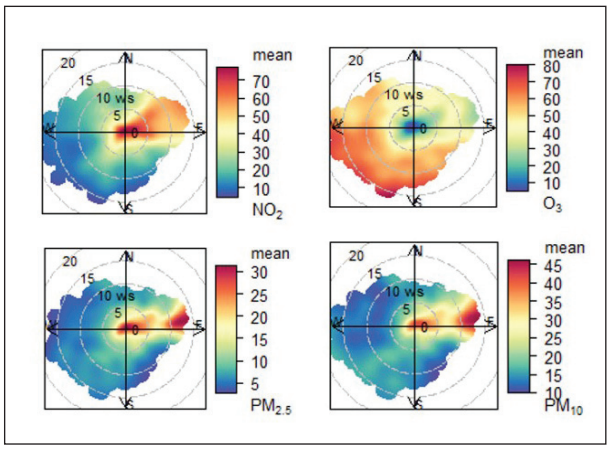


Figure 2.2 Polar plot of wind speed and wind direction

For NO2, the highest levels are recorded when there is little or no wind. As the wind speeds increase the NO2 levels remain relatively high when the wind direction is from the east (E) to east-north-east (ENE). NO2 levels are generally lower when the wind speeds increase and particularly when coming from directions from west (W) to south (S).

As mentioned earlier, this is an analysis for a specific monitoring site in a specific country and relates to both local conditions, such as the placement of the site with respect to pollution sources, as well as climate conditions. It is important to perform such correlations and analysis for areas of interest because the correlations for one site in a specific country are likely to be different to another site, especially if that is in a different country. Therefore, from this analysis we can see that weather factors such as temperature, wind speed and wind direction affect the air quality. This provides guidance that data sets providing these inputs should be useful to the analytics to be performed downstream.

2.2 Data source identification

Once there is a reasonable understanding of the data types needed to address the problem, the next step is to identify usable data sources for that data.

2.2.1 IoT sources

There is an ever-increasing number of IoT devices that can be used to provide a wide variety of data useful to solve various problems. For example, as part of the air quality study, the GSMA used data from IoT air quality devices to gain a better understanding of the relationship between air quality in detailed locations across Greenwich.

It is expected that in the future there will be substantial amounts of IoT data made available by companies and even individuals which can provide data useful to topics such as air quality prediction. For example, parking bay sensors, traffic junction and flow measuring devices, smart car or even CCTV image analysis could be used to provide additional context information that could improve air quality models in the future.

2.2.2 Public database/ data sets

A good starting point is often a search of public data sources. Many countries are maintaining a commitment to 'open' their data for use by for-profit and non-profit applications. The typical way to start is by the use of a search engine, and often assisted by country or region-specific directories of open data. Public data can be further grouped into three categories as:

- Government and official organisations

Each year, many government departments and official statistics agencies publish sets of

authoritative data covering domains such as national economy, social development, industry, energy, agriculture, entertainment, education, health, environment, climate. Some examples are:

- World Bank: https://data.worldbank.org
- World Health Organization: http://www.who.int/
- United Nations: http://data.un.org/
- US Government's open data: https://www.data.gov/
- Weatherbase: http://www.weatherbase.com/
- UK Data Portal: https://data.gov.uk/
- European Data Portal: https://www.europeandataportal.eu/en

In the case of the air quality studies for London and Taiwan, data sets relating to air quality and weather were obtained from official sources published by national environmental agencies or departments. The table below summarises the data sets and sources:

#	SOURCES	DATA SETS
1	http://www.londonair.org.uk	Air Quality Data for London based on the UK Air Quality monitoring network
2	https://mesonet.agron.iastate.edu/	UK Weather Data based on airport METAR reports from Iowa State University* (a substitute data set)
3	https://taqm.epa.gov.tw/taqm/en/	Taiwan Air Quality and Weather Data

Table 2.1 Summary of data sets and sources

Academic and open sources

There are also a number of academic or open source data sets which are worth identifying for particular problems. Some of these data sets are more focused on specific AI and machine learning problems, but it is also possible that data sets are available that are useful to specific problems such as air quality. Some popular sources include:

- UCI: http://archive.ics.uci.edu/ml/datasets.html
- Berkely Stat Lab: https://www.stat.berkeley.edu/users/statlabs/labs.html
- Figshare: https://figshare.com
- Github: https://github.com/openimages/dataset

Data contests

Data contests are popular with many statisticians and data scientists at different levels to compete and try to produce the best prediction model for solving a certain problem. The data sets are normally contributed both by companies and individuals, and those data sets are usually cleansed (see later) and of a high quality. The top ranked data contests include, but are not limited to, the following:

- Kaggle: the leading platform for data prediction competition, https://www.kaggle.com/
- DrivenData: Data Science competitions to save the world, https://www.drivendata.org/
- AlibabaCloud: https://tianchi.aliyun.com/competition/gameList.htm
- DataFoundation: http://www.datafoundation.org/

One general downside of 'open data' is that as a consumer of that data there is little support or influence over that data. E.g.

- The consumer is responsible for understanding what information is in that data set; what values each parameter represents; what represents valid versus invalid values; what the accuracy is;
- Data update frequency, intervals, and the delay in receiving updated data needs to be checked for suitability;
- The format or method by which the data is obtained can be difficult e.g. often there is not a good API to retrieve the data and it's therefore necessary to build 'custom adapters' (as described below) to read the data set.

The experience of the GSMA in acquiring data for the air quality study is useful to reference at this point as it shows that there can be considerable effort required to source data when there is not a good, reliable source for such data. For the air quality study in the UK the obvious source of weather data was the UK Government's Met Office. However, there was not an available API service for this data, and though a web-scraping process was developed this was extremely slow, required quite considerable bespoke software development, and ultimately stopped working when the source web feed was discontinued. The GSMA had then to establish an alternate data feed, ultimately using METAR data archived by Iowa State University, but having to rework some of the downstream processes due to differences in the available data. These difficulties were in contrast to the ease of sourcing of air quality data from the London Air website due to the availability of an easy to use and performant API. This emphasises the benefit of using supported APIs to obtain data rather than unofficial web scraping methods.

Code snippets of the process for reading air quality data from London Air and weather data from Iowa State University can be found here:

https://gist.github.com/GSMADeveloper/c0a8cc94603fa5444efa4eaf48a16200

https://gist.github.com/GSMADeveloper/2312e3d6b97d94afbb873d41f07479a5

2.2.3 Paid databases/data sets

Traditionally paid databases are offered mostly by intelligence agencies, consultancy companies and industry associations such as Bloomberg, Gartner, Experian and KMPG as well as new data market places such as Dawex and the IoT / Distributed Ledger enabled marketplace developed by the IOTA foundation.

In recent years, due to the increased demand for data, there has been growth in the number of data brokers and data transaction marketplaces, attracting participation even from some universities and research institutes. Note that there is also a possibility for mobile operators to participate as a data broker by cleansing, harmonising, aggregating and anonymising external as well as their own data assets and make available over APIs.

A significant advantage of paid databases is that the supplier can be expected to produce a higher quality, curated, and more frequently updated data set compared with open source data. In addition, the supplier will typically provide support both for obtaining the right data using a suitable distribution mechanism (API/ data feed), help with understanding the data provided, and assistance in the case that there are issues with the data.

2.2.4 Internal database/data set

In the Taiwan Air Quality study there was use of an internal mobile network data set processed to provide aggregate user mobility data as a representation of traffic movements and therefore an indication of vehicle generated pollution. Mobile operators have various internal databases and systems such as ERP, CRM and various transactional systems which might also be relevant to some problems. The mobile network comprises many different types of network element such as MME, PGW, SGW, GGSN, SGSN, etc. each of which will

produce vast amounts of data on a daily basis including attributes such as Cell ID, IMSI code, and timestamped spatial information comprised of longitude and latitude. In general, data from the mobile systems can be grouped into four categories:

DATA CATEGORIES	DESCRIPTION
User Profile Data	Basic user context info, such as IMSI, MSISSDN, subscriber profile, subscription plan, billing info etc.
User Behaviour Data	Data extracted and mined mostly from users' communication session records, including communication Behaviour, Mobility Behaviour, Social Connection Behaviour, Web Behavior etc.
Mobile Network Data	Network performance data aggregated over a certain period (e.g., 5mins) mainly reflected as Key Performance Indicator (KPI) and Measurement Report (MR) to monitor and evaluate the network.
Terminal Data	Information read from connected things such as smartphones and IoT devices which are comprised of device information, service characteristics, network parameters etc.

Table 2.2 Categories of data from mobile system

In the Taiwan Air Quality study, the mobile operator successfully extracted aggregated population density and population change information from mobile network data. Details can be referenced in section three of this report.

2.3 Customised data acquisition

Sometimes, for special projects or green field projects, it may be necessary to obtain data that does not exist anywhere else – by running user surveys or by tailored designed experiments. Usually the design needs to balance between comprehensiveness and the associated time, effort and cost.

In the case of the air quality study in Greenwich it was desirable to demonstrate the benefit of finer-granularity temporal and spatial data to evaluate the potential value of mobile IoT in the application of air quality monitoring. The GSMA commissioned an eight-day data collection exercise using an electrically powered mobile 'laboratory' vehicle called the 'Smogmobile'. In addition to the instruments from the Smogmobile itself, two IoT air quality sensors were also installed into the Smogmobile, one on the inside and one on the outside, to allow comparisons to be drawn between the internal and external air quality.

The Smogmobile data collected for each pollutant shown in the figure on page ten.. By applying such a purpose designed experiment, it could help us gain a highly detailed understanding of the pollution across the Greenwich area.

Air Quality Relative Pollution Level

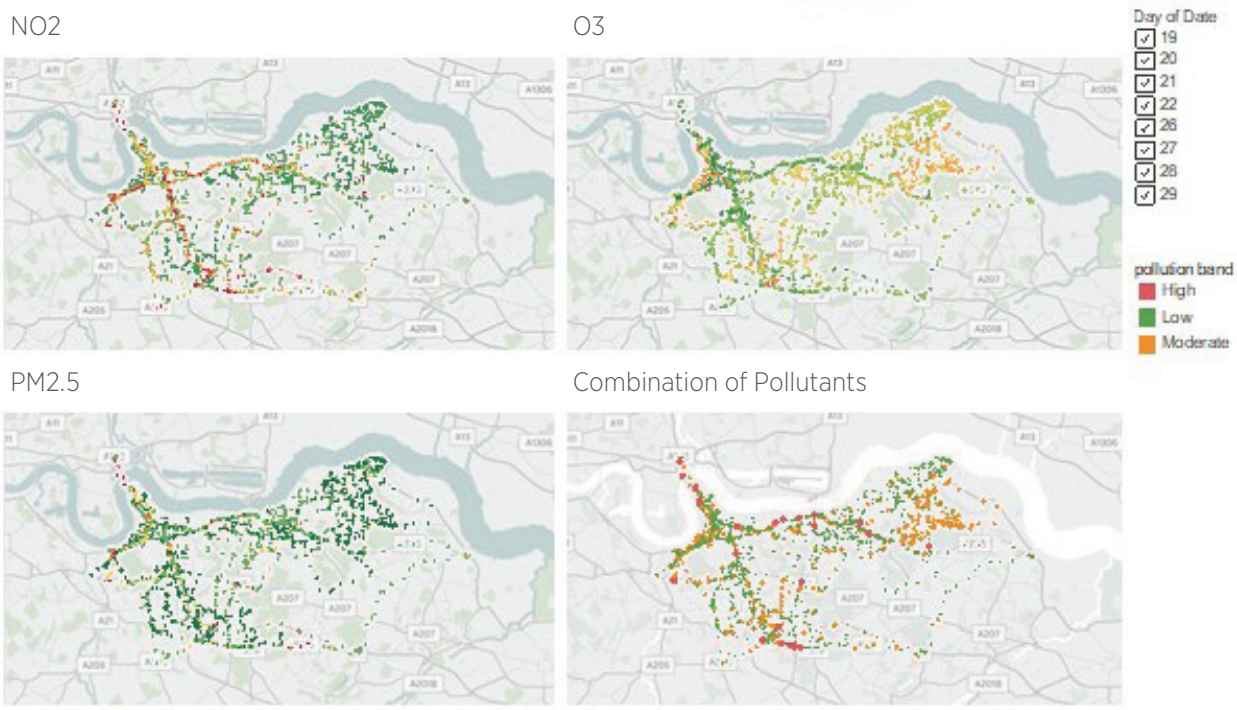


Figure 2.3 Air quality relative quality level

Some sample data rows from the Smogmobile capture can be found below:

TIMESTAMP (DD/MM/YYYY HH:MM)	LATITUDE_ AVG	LONGITUDE_ AVG	SPEED_AVG m/s	O3_AVG (ug/m3)	NO2_AVG (ug/m3)	PM2.5_AVG (ug/m3)	BATTV (volts)
19/07/2017 07:38	5130.1	0.06	8.05	39.71	35.87	13.42	13.39
19/07/2017 07:39	5129.92	0.2	20.62	35.27	58.79	13.52	13.37
19/07/2017 07:40	5129.66	0.53	14.71	33.02	52.3	13.75	13.38
19/07/2017 07:41	5129.59	0.59	1.65	37.97	54.64	16.54	13.36
19/07/2017 07:42	5129.53	0.56	5.78	34.65	51.68	13.25	13.36
19/07/2017 07:43	5129.36	0.5	14.83	29.81	58.18	16.68	13.36

Table 2.3 Sample data rows from Smogmobile

As can be seen from the table above, the Smogmobile gives per minute measurements of main pollutants of NO2, O3 and PM2.5 together with the detailed geo location information (GPS Latitude/ Longitude) as well as the remaining battery level of the vehicle. Battery level was not used in any air quality analysis as there was no expectation of any correlation with air quality outcomes.

For the visual representation of the pollution band the guidance from the World Health Organization (WHO)[10] was used, as listed below:

POLLUTANT	YEARLY AVG (ug/m3)	24 HOUR AVG (ug/m3)	8 HOUR AVG (ug/m3)	1 HOUR AVG(ug/m3)	10 MIN AVG (ug/m3)
NO2	40			200	
O3	50		100		
PM2.5	10	25			

Table 2.4 Guidance from WHO

Each pollutant was categorised into three levels according to one hour or eight hour or 24 hour or annual mean depending on the type of the pollutant. For example, level of NO2 below 40 ug/m^3 was categorized as green and level above 200 ug/m^3 was categorised as red while the level between 40 ug/m^3 and 200 ug/m^3 was categorised as yellow, while level of PM2.5 below 10 ug/m^3 was categorised as green and level above 25 was classified as red and level between was marked as yellow. This broad categorisation is useful for output purposes though when predicting pollution levels this is not very useful for input parameters – so in general it's better to use a 'real' value for parameters such as NO2 rather than a categorised value.

2.4 Data ingestion and pre-processing

2.4.1 Data ingestion

As mentioned above, the GSMA found that the ideal source of data is via an official API from the data provider. Typically, there will be a need to develop a bespoke software application to read the data from the source and form a local dataset.

When obtaining data from IoT devices it is often the case that the manufacturer provides suitable mechanisms (typically in the form of a well-structured API) to access to the data. Simpler IoT devices will often send their data to the manufacturer's cloud storage system which often provides dedicated APIs to allow the data to be read along with consent management solutions allowing the device owner to agree to the data being shared. More complex IoT devices may allow data to be stored directly to a database or using a cloud service API so that the data can be pushed directly to analysis platforms.

In the case that an API is not available it may be necessary to develop a custom web crawler, a piece of code or automated script, to crawl selected web pages on the internet and store the data content locally. Meaningful data can be extracted by parsing the page content and mined by using certain statistical algorithms. However, since the design of

[10] http://www.who.int/airpollution/en/

the page labels of each website is different, it usually requires data engineers to have special knowledge and skills to write tailored scripts. In Python, the most popular framework for web scraping is called 'BeautifulSoup' which can be used to parse page content. However, as noted earlier the web scraping technique is not usually as efficient as an API, generally requires much more development time and effort to get working, and requires much higher maintenance as any change to the underlying web site operation or layout can cause the scraping code to stop working.

As part of the project, the GSMA obtained air quality data from two different types of IoT devices. The first of these being a portable air quality IoT sensor unit 'SMAQ' from Urban Clouds[11] and the second being a fixed air quality IoT solution from Libelium.[12]

Urban Clouds SMAQ

The Urban Clouds 'Spatial Mapping Air Quality' unit is a battery powered sensor unit measuring various pollutants. This has a GSM/GPRS modem sending data to a cloud storage platform provided by Urban Clouds. This cloud storage platform offers API services allowing the recorded pollutant data to be read by customer systems for further analysis or ingestion into machine learning based applications.

Figure 2.4: Urban Clouds SMAQ Unit

Urban Clouds provide a customer dashboard which allows data to be visualised or downloaded in a CSV format.

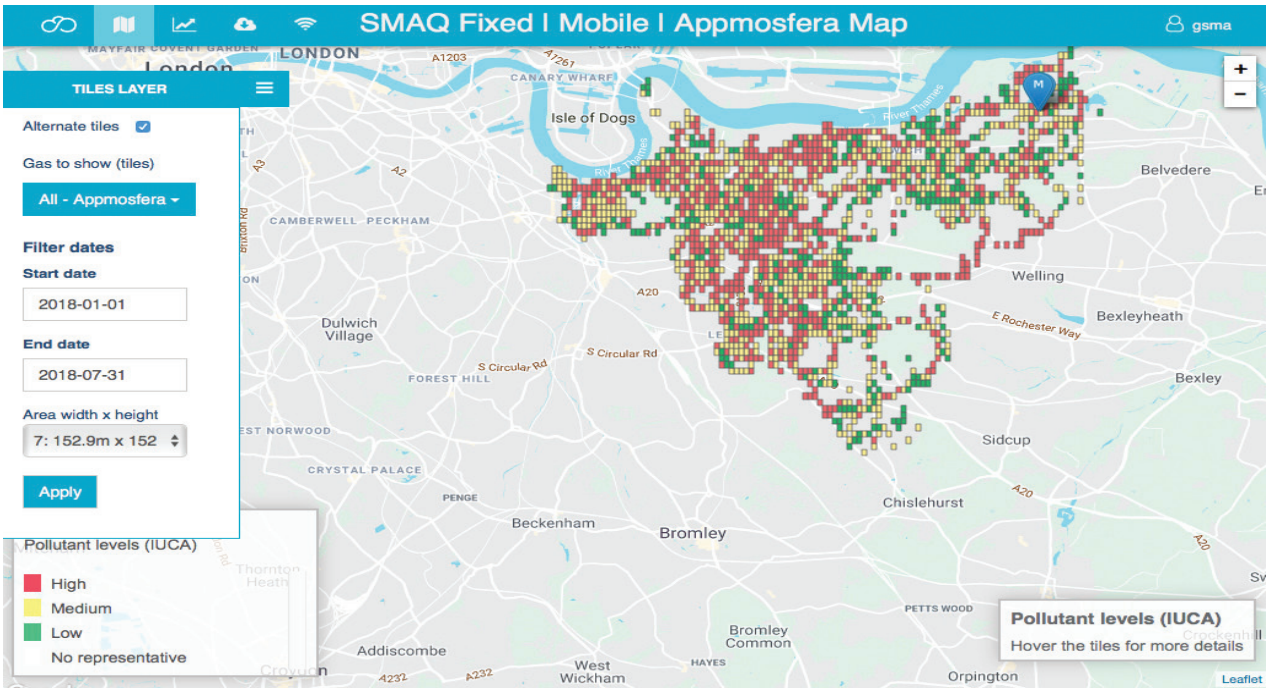


Figure 2.4.1 Spatial mapping of air pollution using 152.9m x 152.9m tiles

[11] https://urbanclouds.city/outdoor-air-quality/
[12] https://www.the-iot-marketplace.com/libelium-air-quality-index-iot-vertical-kit

An example Python script to download data using the APIs provided by the Urban Clouds platform is available here: https://gist.github.com/GSMADeveloper/a0eeb9c3fe7dd1fed7566c13429d8991

The SMAQ data downloaded then looks like the following

CO	DeviceID	DeviceName	IAQ	Latitude	Longitude	NO2	O3	Observed	PM1	PM10		Resp
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0338	2	0	19/06/2018 08:49	1	2	2	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	3.8	0	19/06/2018 08:48	0	3	1	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:48	1	3	2	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:47	1	3	3	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:47	0	3	1	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:47	1	1	1	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	12.6	0	19/06/2018 08:46	1	2	2	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451604	0.0339	6.9	0	19/06/2018 08:46	1	1	1	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451604	0.0339	17.1	0	19/06/2018 08:45	1	2	2	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451604	0.0339	74.4	0	19/06/2018 08:45	1	3	3	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451604	0.0339	34.4	0	19/06/2018 08:45	1	3	3	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451604	0.0339	57.8	0	19/06/2018 08:44	2	5	4	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	159.5	0	19/06/2018 08:44	2	6	4	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:43	2	3	3	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:43	1	3	3	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:43	1	2	2	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	7.3	0	19/06/2018 08:42	1	2	2	
0	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	47.7	0	19/06/2018 08:42	2	3	3	
88.5	5931530171c03a5219a7b929	SMAQ_0004		51.451504	0.0339	0	0	19/06/2018 08:41	2	4	4	
309.7	5931530171c03a5219a7b929	SMAQ_0004		51.451096	0.0329	0	6.7	19/06/2018 08:41	2	8	4	
675.1	5931530171c03a5219a7b929	SMAQ_0004		51.450504	0.0309	0	2	19/06/2018 08:41	2	5	5	
1186	5931530171c03a5219a7b929	SMAQ_0004		51.451404	0.0333	2.8	6.3	19/06/2018 08:40	7	11	11	
282.3	5931530171c03a5219a7b929	SMAQ_0004		51.452704	0.035899	9.1	6	19/06/2018 08:40	6	10	9	
73.2	5931530171c03a5219a7b929	SMAQ_0004		51.453	0.0354	22.5	9.6	19/06/2018 08:40	2	3	3	

Libelium Air Quality IoT Kit

The Libelium unit is a sophisticated IoT solution comprising a gateway unit which relays information over a 3G/4G connection to a cloud service or database, the gateway unit then connects locally to one or more remote monitoring sensor units.

Figure 2.5: Libelium – Air Quality IoT Kit

In our deployment at the GSMA offices in London, we used the database connectivity option of the Libelium gateway unit to store data directly to a cloud hosted MySQL database. Data rows were stored as in the following example, note that several database records (rows) form a single timed group.

id	id_wasp	id_s	frame _type	frame	sensor	value	timestamp	sync	raw	parser _type	MeshliumI
340422	Sep-01	426	6	77	TC	25.79	31/07/2018 13:04	0	noraw	1	meshliumfb40
340421	Sep-01	426	6	77	HUM	32.84375	31/07/2018 13:04	0	noraw	1	meshliumfb40
340420	Sep-01	426	6	77	PRES	100668.72	31/07/2018 13:04	0	noraw	1	meshliumfb40
340419	Sep-01	426	6	77	CO	0.20719433	31/07/2018 13:04	0	noraw	1	meshliumfb40
340418	Sep-01	426	6	77	NO2	0.05636322	31/07/2018 13:04	0	noraw	1	meshliumfb40
340417	Sep-01	426	6	77	O3	0.00705144	31/07/2018 13:04	0	noraw	1	meshliumfb40
340416	Sep-01	426	6	77	SO2	0	31/07/2018 13:04	0	noraw	1	meshliumfb40
340415	Sep-01	426	6	77	PM1	1.36	31/07/2018 13:04	0	noraw	1	meshliumfb40
340414	Sep-01	426	6	77	PM2_5	2.03	31/07/2018 13:04	0	noraw	1	meshliumfb40
340413	Sep-01	426	6	77	PM10	2.12	31/07/2018 13:04	0	noraw	1	meshliumfb40
340412	Sep-01	426	6	76	TC	24.49	31/07/2018 12:49	0	noraw	1	meshliumfb40
340411	Sep-01	426	6	76	HUM	37.464844	31/07/2018 12:49	0	noraw	1	meshliumfb40
340410	Sep-01	426	6	76	PRES	100665.11	31/07/2018 12:49	0	noraw	1	meshliumfb40
340409	Sep-01	426	6	76	CO	0.21957237	31/07/2018 12:49	0	noraw	1	meshliumfb40
340408	Sep-01	426	6	76	NO2	0.1346853	31/07/2018 12:49	0	noraw	1	meshliumfb40
340407	Sep-01	426	6	76	O3	0	31/07/2018 12:49	0	noraw	1	meshliumfb40
340406	Sep-01	426	6	76	SO2	0	31/07/2018 12:49	0	noraw	1	meshliumfb40
340405	Sep-01	426	6	76	PM1	1.1999999	31/07/2018 12:49	0	noraw	1	meshliumfb40
340404	Sep-01	426	6	76	PM2_5	1.8	31/07/2018 12:49	0	noraw	1	meshliumfb40
340403	Sep-01	426	6	76	PM10	2.06	31/07/2018 12:49	0	noraw	1	meshliumfb40
340402	Sep-01	426	6	75	TC	24.79	31/07/2018 12:34	0	noraw	1	meshliumfb40
340401	Sep-01	426	6	75	HUM	34.171875	31/07/2018 12:34	0	noraw	1	meshliumfb40
340400	Sep-01	426	6	75	PRES	100659.11	31/07/2018 12:34	0	noraw	1	meshliumfb40
340399	Sep-01	426	6	75	CO	0.18457577	31/07/2018 12:34	0	noraw	1	meshliumfb40
340398	Sep-01	426	6	75	NO2	0.2356804	31/07/2018 12:34	0	noraw	1	meshliumfb40
340397	Sep-01	426	6	75	O3	0	31/07/2018 12:34	0	noraw	1	meshliumfb40
340396	Sep-01	426	6	75	SO2	0	31/07/2018 12:34	0	noraw	1	meshliumfb40
340395	Sep-01	426	6	75	PM1	0.87	31/07/2018 12:34	0	noraw	1	meshliumfb40
340394	Sep-01	426	6	75	PM2_5	1.4699999	31/07/2018 12:34	0	noraw	1	meshliumfb40
340393	Sep-01	426	6	75	PM10	1.65	31/07/2018 12:34	0	noraw	1	meshliumfb40

2.4.2 Data pre-processing

Often when a data source is read it will be realised that there are various 'quality' issues with that data such as

- Certain individual attributes might not be reported at an expected interval;
- Certain whole records might not be reported at an expected interval;
- Certain sub-feeds produce different combinations of attributes e.g. in London not all air quality monitoring stations measure the same types of pollutants;
- The data feed might represent 'missing' attributes in different ways e.g. '0' might not mean a zero reading but instead a missed reading, or there might be an odd value such as '-99' that means the reading should not be used.

Incomplete or incorrect data records may produce significant bias in generating training models and some algorithms are particularly sensitive to the presence of outliers in data which may result in poor prediction results. As a rule of thumb, in any data science project at least 80 per cent of the time and effort will be spent in wrestling the data into a correct and usable format. Less than 20 per cent of the time is typically spent applying the fully prepared data to a process such as machine learning.
The pre-processing step is often highly bespoke to the data source and typically involves bespoke software development for automated processing and/or dedicated manual pre-processing steps before subsequent steps (data cleaning etc.) can be applied.

2.4.2.1 Data cleaning

This step involves removing invalid records, filling in (when practical) missing values, resolving data inconsistencies, identifying or removing outliers, smoothing noisy data, and correcting erroneous data.

- Incomplete data

Data is not always available due to it being missing at the source of collection or deletion on purpose caused by inconsistency with other recorded data.

Missing data is usually left blank in the record, or sometimes noted as specific characters such as 'NA' or '-99'/'999' in the raw data set. Normally, the simplest method to deal with missing data is just deleting the whole of the associated record or records. This is a practical method when there are only a very small portion of observations with missing values and the missing values are distributed completely at random. Sometimes, however, it might be useful to fill in the missing data rather than totally deleting the records to avoid introducing bias.
Often missing values can be replaced with the mean value or the most frequent value of the attribute for the dataset assuming each attribute follows some distribution. One sophisticated method, interpolation, tries to run a regression model based on available data and fill the missing values by some predicted results. In general, methods for handling incomplete data depend on the types of the missing pattern and usually need data scientist to understand the reason behind and analysing the distribution of the missing values.

It is recommended that data scientists closely inspect every data source to check for the quality, completeness and range of all input parameters so that decisions can be made regarding the strategies for dealing with quality issues in the data set.

In the case of the air quality study in Taiwan conducted by the GSMA and FET, it was seen that there was a significant nine per cent improvement in the prediction accuracy by applying the combination of both removing missing values from the air quality data and the population data plus linearly interpolating the weather data.

FET applied linear interpolation techniques to weather station data for Taiwan, for temperature, pressure and wind speed. Linear interpolation was not, however, used for wind direction since the variability of wind direction was considered likely to introduce ambiguity into the learning data.
As an example of linear interpolation where there was a missing value for temperature at the hour t+1 but data was available for hour t and hour t+2, a simple averaging method was used where the temperature readings for hour t and hour t+2 hours were summed and then the sum divided by 2 to provide an estimated reading for hour t+1.

- Noisy data

Noisy data generally refers to data containing errors or outliers. A simple quick visual check of the dataset may help detect some apparent errors e.g. a negative -10 figure for wind speed. Most of the time, data checking involving the range checking and validity or legitimacy checking which can be realized by using a basic visual checking first but ideally an automated method to identify anomalies would be used with input from domain knowledge.

For example, in the open source weather data set used in the air quality study, a temperature below -50 degree and wind speed below 0 are detected as errors and treated like missing values. In addition, in the open source air quality data set, all pollutant values below 0 are identified as errors. We have also seen some NO2 values as extreme as 2036 ug/m3 from some mobile IoT devices which would be accepted into a machine learning model with the risk of skewing the accuracy of its model. This shows the importance of domain knowledge in verifying the range of input parameters.

2.5 Data transformation, normalisation, consolidation and derivation

Data transformation, normalisation, consolidation and derivation is an important process for preparing the raw data into a specific format to meet the input requirements of the machine learning algorithms. This process correlates the input data sets and the final data analysis you want to achieve, typically handling issues such as where the raw data is usually not in the correct format for feeding into machine learning algorithms. Therefore, tailored scripts or code are usually needed to develop to support the bespoke requirements of the transformation process for a particular problem and data sets.

In the use case of the air quality study, several transformation, consolidation and derivation techniques were made on the raw air quality and weather data to prepare for the prediction algorithms. Here we will share some useful techniques when dealing with an air quality or weather data set.

2.5.1 Data transformation

The various transformations applied to the air quality and weather data included:

- Unit conversion

In order to compare the level of the results, it is advised to unify the unit of each measured variable. For example, in the air quality dataset, the unified unit chosen is micrograms per cubic metre (ug/m^3) for each pollutant so that comparison with the European Commission or World Health Organization air quality standards is much more convenient. Not all data sources provide inputs using this unit, instead providing measures as Parts Per Billion (PPB) which was found to be less useful in machine learning models than micrograms per cubic metre. A python code snippet for conversion of NO2 from PPB to ug/m3 is shared below:

https://gist.github.com/GSMADeveloper/84d0adb7b5d88d0e32dc512299573272

Certain machine learning algorithms work best on normalised data (see below) and it is important to implement unit conversion before normalisation.

- Coding transformation

For some machine learning algorithms, it is required to transform non-numeric variables to numeric values. For example, in the weather dataset, if you have a qualitative attribute similar to 'weathertype', it can be coded into numeric values. The code snippet can be found here:

https://gist.github.com/GSMADeveloper/84d0adb7b5d88d0e32dc512299573272

- Other ad hoc transformations

In the input data set for weather data, the raw data from the open database source for wind is reported as a wind speed in meters/second and wind direction either in cardinal or degree representation. We theorised that this would not be useful to the machine learning models. Also, if the wind direction is reported in cardinal direction, for example N, NNE, NE, etc., we theorised it would be advisable to convert it from cardinal to degrees first and then combine with the wind speed parameter to generate two wind vectors resolved to the northern and eastern direction respectively. The code for this conversion is shared below. By using this function, wind speed and wind direction will be transformed into two new 'real' numeric variables which can be fed into the prediction algorithms easily for future modelling. The code snippet can be found here:

https://gist.github.com/GSMADeveloper/05cfd6ac7d2ab06ae6f8b1c40ea9f204

2.5.2 Data normalisation

When the original data is comprised of attributes with different units or scales, normalisation may be helpful to rescale the attributes to fall within a small, specified range. Many machine learning algorithms as mentioned below can benefit from the normalising process. The common normalisation methods are divided into three categories:

- Min-max normalisation:

  Attributes are often scaled into the range between [0, 1]. This technique is often used in neural networks and algorithms based on Euclidean distance such as KNN.

$$new\ value = \frac{old\ value - min}{max - min}$$

- Z-score normalisation:

  When an attribute is assumed to follow a Gaussian distribution, it is also possible to transform the attribute with mean of 0 and standard deviation (sd) of 1 by using the formula:

$$new\ value = \frac{old\ value - mean}{sd}$$

  Z-score normalisation technique is often been found useful in ridge regression, logistic regression and linear discriminant analysis.

- Normalisation by decimal scaling:

  This technique tries to move the decimal point of values of certain attribute.

$$new\ value = \frac{old\ value}{10^{j}}$$

  Where j is obtained by finding the largest number in the range of the attribute and then counting the number of digits in the largest number.

In the use case of air quality study, normalisation methods were trialed both on air quality data and weather data e.g. 'Barometric Pressure' in millibars was divided by 1000.0. However, no improvements on prediction results by implementing normalisation technique were seen for the decision tree-based algorithms which were ultimately selected for the prediction models due to their leading performance.

2.5.3 Data consolidation and derivation

The air quality data retrieved from the London Air website is provided as structured JSON data as an array of records where each record has multiple key value pairs. Data in this structure cannot be fed into the machine learning algorithms directly. Therefore, the next step is to convert the cleansed and optionally normalised data into a 2D matrix with row dimension as the number of observations and column dimension as the number of features.

Since the data for the air quality and weather normally come from different sources and weather is an important factor for air quality prediction, a

necessary step is joining the two data sets into one. Both the air quality data and weather data pulled from the open source for Greenwich, London are on an hourly basis from year 2012 to 2017. In this case, the joining process is pretty straight forward. The air quality data set and the weather data set were joined together by the timestamp value. Geographically, as there was data available from four weather stations across the London area, for each air quality station, all of data from the four weather stations were joined with the air quality data along the date/time dimension.

When joining data sets which have different time dimensions or geographical levels, they need to be aggregated into a common time measurement period or into different geographical administrative levels for the purpose of query, reporting, data visualisation or further intelligent data analysis. Particularly, when cross combing data from mobile IoT devices which generally give 'real-time' readings at a granularity of seconds or minutes, it is essential to firstly pre-process and aggregate the raw data before joining with the other context data sets. For example, in the experiment with the SMOG Mobile for the air quality study, the data obtained from the mobile for NO2, O3, PM10/PM2.5 are as detailed as minute level. Whereas, the air quality data published by the UK government is usually reported on an hourly basis which is widely accepted by industry as a common reporting interval. To associate the two datasets for comparison or for complementary data analysis purpose, aggregation is necessary.

The final step before feeding the joined data into the selected machine learning algorithm is derivation. In the case of air quality the initial analysis had determined the buildup of poor air quality over preceding days and therefore it was decided to feed near term historical data into the air quality model. Specifically it was decided to train the machine learning model to predict the air quality at the current hour (t) given the air quality measurement and weather factors at the previous 1 day (t-24) and 2 days (t-48). A simple example of the data can be seen below:

Data Records Num	NO2D1 (t-24)	NO2D2 (t-48)	Current Temperature at t	TemperatureD1 (t-24)	TemperatureD2 (t-48)	NO2 - result (for prediction at t)
1	56.4	50.1	8.6	4	5.7	39.5
2	38.5	23.7	24.7	20.7	19.2	18.7
3	15.8	8.8	14.8	14	13.8	6.2
4	27.5	11.7	7.1	2.6	8.7	22.3
5	63.7	92.8	9.2	1.8	5.6	48.6

Table 2.5 Sample illustration of the input data format

So here is a simple demo of the result after derivation of the data matrix. To predict the NO2 result at time t and compare with the actual NO2 result on 7th column, day-1 and day-2 inputs of NO2 as well as day-1 and day-2 temperature for the respective hour of day need to be shifted on the same row. The detailed input parameters of the final data matrix fed into the ML learning algorithms are generalized as below and a sample input data file is share on github for a detailed look:

https://gist.github.com/GSMADeveloper/fc8bdbeb7f38a59bd460f81819db5f85#file-inputdata_gr9-txt

- Current temperature at time (t), temperature at previous 24 hours (t-24), temperature at previous 48 hours (t-48), average temperature over the past 48 hours
- Current barometric pressure at time (t), barometric pressure at previous 24 hours (t-24), barometric pressure at previous 48 hours (t-48), average barometric pressure over the past 48 hours
- Current wind speed at time (t), wind speed at previous 24 hours (t-24), wind speed at previous 48 hours (t-48)
- Current Eastern and Northern wind vectors at time (t), Eastern and Northern wind vectors at previous 24 hours (t-24), Eastern and Northern wind vectors at previous 48 hours (t-48). (methods to get the two vectors were already discussed in Section 2.3) for each of the nearby weather stations
- Current weather type at time (t)
- Pollutant level at previous 24 and 48 hours respectively, (t-24), (t-48)
- The difference (delta) between the NO2 level at previous 24 and 48 hours
- Day of week at time (t), coding into 0, 1, 2, 3, 4, 5, 6
- Hour of day at time (t), coding into 0, 1, 2, ..., 23
- Is holiday or not, coding into 0 or 1
- Is Saturday or not, coding into 0 or 1
- Is Sunday or not, coding into 0 or 1
- Is weekday or not, coding into 0 or 1

(Note that date and time were not included in the input data because it was felt this could lead to 'overfitting' i.e. the machine learning model learning that the NO2 measure at a specific date/time would be a specific value rather than creating a more generalised model. Instead date/time were used to derive other parameters e.g. day of week, hour of day.)

The general process for development of machine learning is to take the input data ('features') and result targets ('labels') and divide into a training data set and a test data set. This allows the performance of the machine learning algorithm to be evaluated using data in the testing data set that has not been seen before. The aim of this is to evaluate if the machine learning algorithm has learned a 'generalised' method, or has simply learned the results it needs to generate for a specific set of input data. In python, there is a function named **"train_test_split"** which can be called to automatically split the input data into training and testing data sets. A code snippet for this can be found here:

https://gist.github.com/GSMADeveloper/0aafb43791c460ad6e2b18f4c1d32804

3. MACHINE LEARNING

The complexity in traditional computer programming is in the code (programs that people write). In machine learning, algorithms (programs) are in principle simple and the complexity (structure) is in the data. Is there a way that we can automatically learn that structure? That is what is at the heart of machine learning.

---- Andrew Ng

As stated by Andrew Ng, in machine learning, algorithms are in principle simple and the complexity exists in the data. So, this section focuses on the methods applied to the use case of air quality study in Greenwich, London and Taiwan area on applying and optimising machine learning technique to the air quality and weather data sets.

For the initial Greenwich study, the GSMA had determined that machine learning could be applied to the task of predicting air quality for the near future based on historical air quality and weather data and a short-term weather forecast. The GSMA also hoped to include mobile network analytics into the machine learning models at a later date, as eventually was done with FarEasTone in Taiwan.

Machine learning finds connections between data that are often not known by the people training the models, and there are various processes that need to be applied as part of getting the best result.

Measurement of accuracy of algorithms

Having prepared the data ready for machine learning the next step is to select and optimise the machine learning algorithm best suited to the air quality problem. There are many algorithms to choose from, and for any new problem there is a need to evaluate the optimal choice of algorithm and the settings for that algorithm to produce the best result.

An important part of the process of maximising performance of machine learning is 'objectively' evaluating the performance of the algorithm and settings. R-squared is a frequently used statistical index to measure how well different machine learning algorithms perform.

R-squared measures the proportion of variability in the output variable Y that can be explained using the model built from the input data X. The best prediction result is when R-squared is equal to 1. When R-squared is equal to 0, it means the machine learning model makes a constant prediction with the expected value of the output variable Y (i.e. this is a poor result as the output is effectively independent of input values). Additionally, R-squared can be negative when the model performs worse than a random prediction.

R-squared is defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where RSS is the residual sum of squares:

$$\sum (y\_true - y\_predict)^2$$

And TSS is the total sum of squares:

$$\sum (y_{true} - mean\ of\ \_y\_true)^2$$

The goal therefore of choosing and tuning a machine learning algorithm is to get as close as possible to an R^2 value of 1.0.

3.1 Selection of the preferred machine learning algorithm

There are many machine learning algorithms available, the result of decades of academic research and development. There is, however, no guidance as to what machine learning algorithm suits a particular new application area and so the GSMA first, and then Far Eastone, subsequently sought to objectively compare algorithm performance in order to select the best for the air quality prediction.

An initial screening was run feeding the same data set through a range of machine learning algorithms to determine the R2 result. The results below compare different algorithms and show that so called 'ensemble-based' methods highlighted * below performed far better than other methods for this particular problem. Sample code snippets for calling different algorithms can be found here:

https://gist.github.com/GSMADeveloper/168e90ccdf1589836978144d60eabcf6

Trialled Methods	R-SQUARED RESULT
Linear Regression	0.343
Neural Net	0.4822
Support Vector Regression (Linear)	0.3384
Decision Tree	0.3368
Bagged Decision Tree[13]	0.6546
Random Forest[14]	0.7683
Extra Tree[15]	0.8078
Gradient Boosting[16]	0.8498

Table 3.1 Prediction results with different ML algorithms

Note in the table 3.1 the highest performance algorithms are variations of 'Decision Trees'. A decision tree is essentially a set of simple decisions which can be employed sequentially on input parameters – such as 'is today a Sunday' or 'Is the air temperature over five degrees'. Additional performance is obtained by the machine learning algorithm creating many such decision trees on subsets of input parameters and then combining the results. This is known as 'ensemble learning' and was found to exhibit good performance in the air quality model. The main idea of such ensemble learning is to average the output from complex models to reduce the variance so as to improve the overall prediction performance.

- Bagged Decision Tree

This is the simplest form of the ensemble learning method. To apply bagging technique to decision trees, we normally construct B regression trees by generating B bootstrapped training sets and then average the B results to reduce the high variance of each tree. For a detailed explanation refer to "Bagging", Ryan Tibshirani[17].

- Random forest

Random forest is based on the bagging of decision tree mentioned above, but when building each individual decision tree, it will randomly choose a subset of attributes at each split instead of using the whole attribute sets. So by doing this, in each split, weak predictors will have more chance to be considered rather than always queuing after those strong predictors. Plus, the correlation between the trees built in the ensemble is decreased which will further reduce the total variance. A detailed algorithm explanation could be found in the "Random Forest" algorithm[18].

[13] http://www.stat.cmu.edu/~ryantibs/datamining/lectures/24-bag.pdf
[14] http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf
[15] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.7485&rep=rep1&type=pdf
[16] http://www.cse.chalmers.se/~richajo/dit865/files/gb_explainer.pdf
[17] http://www.stat.cmu.edu/~ryantibs/datamining/lectures/24-bag.pdf
[18] http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf

- Extra Tree

Extra tree is short for 'Extremely Randomized Trees', which is, as the name implies, a further randomness on top of Random Forest. The Extra-Trees algorithm builds an ensemble of regression trees by not only randomly selecting attributes for each split but also randomly selecting the cut point for each attribute instead of choosing the best threshold. The other main difference is it uses the whole learning sample rather than a bootstrap to grow the trees. The extra randomness introduced on top of Random Forest can help reduce more variance in building a complicated forecast model.

- Gradient boosting

As listed in the table 3.1, among all of the ensemble methods, gradient boosting gives the best prediction result in terms of 24 hour air quality forecast capability. The main difference between gradient booting and bagged methods is that unlike fitting a single large tree to the data, it learns slowly and at each iteration adds only a bit of the new tree into the prediction rule in order to improve the residuals. To avoid overfitting, gradient boosting usually build small trees determined by depth of tree. So in this way, the model can be built incrementally and it has a chance to improve the areas where it does not perform well during the building process.

3.2 Optimisation of the machine learning results (Greenwich)

As discussed above, the gradient boosting algorithm was selected as the best performing algorithm for air quality prediction.

In Python the Gradient Boosting Regressor is part of the 'scikit-learn' library. Code snippets for invoking the gradient boosting algorithm can be found here:

https://gist.github.com/GSMADeveloper/dbe8fd6839b41bf15fce510a0a1ed832#file-gradientboost-py

The above function call shows the parameters used to optimise the performance of the machine learning algorithm. Whilst there is no programming as such required for execution of machine learning algorithms there are various factors that affect the performance and a key step in obtaining the best performance is choosing good values for the various parameters that can be tuned for the particular machine learning algorithm.

The process to choose the parameters is essentially iterative and involves choosing different parameter values, one parameter at a time, and checking if the R^2 value improves or decreases. A methodical approach was used first by the GSMA and subsequently by Far EasTone, varying the parameter values tried and ending up for the GSMA with the above choices.

Explanations about some these parameters for the gradient boost algorithm:

- n_estimators: number of iterations to perform. The default value is 100, but since the air quality data set is quite large across 5 years, a larger value such as 10000 was found to yield a better prediction accuracy.
- max_features: the maximum number of features at each split (of the tree). When set to 'auto', it equals to the number of features we feed into the model.
- max_depth: The maximum depth of each individual tree built. As mentioned in Section 3.3, a simple tree normally gives better result. For the air quality case a recommendation was made to choose a value between 1 and 10 according to the input data, and a depth of 4 was found to be good for the London air quality data. The parameter can be varied for other situations (different cities with different factors affecting air quality) to achieve the best performance.
- learning_rate: this parameter controls the pace of the learning and is usually a small number between 0.1 and 0.001. A very small learning rate can require a very large number of iterations which means a longer

time to train so there is a balance to be chosen between training speed and accuracy. For the London air quality case the value of 0.0875 was found to provide good accuracy with a reasonably fast learning rate.

- Loss: The loss function describes how far the model's prediction result is from the expected. 'ls', referring to least square, was chosen for the air quality study though other loss functions were tried.
- Criterion: the function to measure the quality of a split. The default "fried man_mse", which is the mean squared error with improvement score by Friedman, was used in the model and found to perform well against other criterion functions.
- Detailed explanation of other parameters can be found in Gradient Boosting for regression[19].

3.2.1 Results for 24 hour NO2 prediction in Greenwich, London

There are nine air quality stations in Greenwich area. By using Gradient Boosting, the 24 hour prediction accuracy for each station are listed below:

Code of Air Quality Station	R-SQUARED RESULT
GR8	0.8299
GR9	0.8558
GR4	0.8479
GN0	0.8311
GN2	0.8282
GN3	0.8411
GN4	0.8369
GR5	0.8360
GR7	0.8431

Table 3.2 Prediction results of NO2 level in Greenwich

In general, all of the R-squared results of the nine stations are above 0.8 which means the average error rate represented by RMSE(Rooted Mean Square Error)/Y_mean(Mean of Y value) is around 25 per cent. The results show that gradient boosting is a robust prediction method giving very good 24 hour NO2 forecasting capabilities in advance with accuracy rate of 75 per cent. It is also noted that a further experiment extended to 75 more air quality monitoring stations covering the whole Greater London area beyond the nine stations in Greenwich area shows a good average R^2 result of 0.82, which is significant to show the robustness of the prediction method.

An example scatter plot of the predicted NO2 vs the actual NO2 value is shown below for the monitoring station GR9 and shows a good cluster of actual and predicted readings with good proportionality:

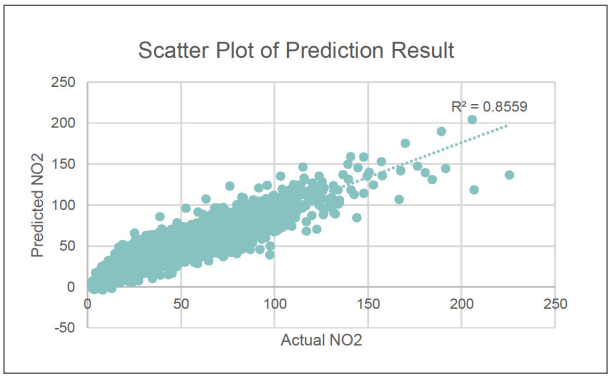


Figure 3.1 Predicted NO2 result vs Actual NO2 level

The visualisation for the prediction result at air quality station GR9 is plotted below:

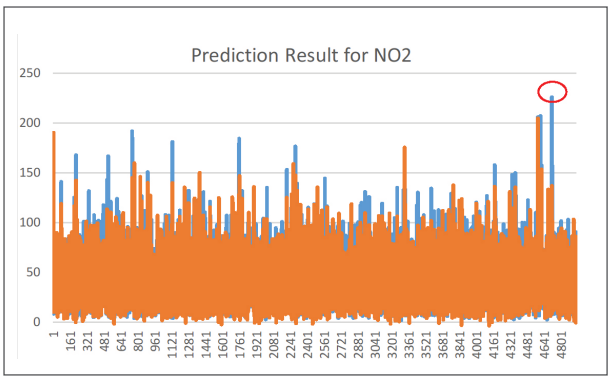


Figure 3.2 Prediction result at GR9 site

[19] http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html#sklearn.ensemble.GradientBoostingRegressor

By comparing the actual NO2 level plotted in blue in figure 3.2 and the predicted NO2 level plotted in yellow, we can see that there is a high coincidence between the yellow and blue curves. It shows that the prediction result by using this method can generally replicate the actual daily NO2 level of changes and especially can give the best result during the periods when the changes of NO2 level are smooth. Some sudden NO2 changes are still difficult to capture especially those high spike cases. For example the blue spike circled in red is much higher than its predicted NO2 level.

3.2.2 Results for 24 hour PM2.5 prediction in Greenwich, London

There are nine air quality stations in Greenwich area. By using Gradient Boosting, the 24 hour prediction accuracy for each station are listed below:

Code of Air Quality Station	R-SQUARED RESULT
GR9	0.7889
GR4	0.8293
GN0	0.7833
GN2	0.8365
GN3	0.8158

Table 3.3 Prediction results of PM2.5 in Greenwich

Compared with NO2, the R-squared value for PM2.5 is slightly lower on average than the previous results. This is in line with our expectations since we identified in the data analysis report [7], the nature of particulate matter formation is more complex than the gas NO2 and it has a wider variety of factors that influence this (not just combustion that is largely responsible for NO2). As well as weather effects and road traffic and other combustion sources, PM2.5 can also be affected by local industrial pollutions and longer-range transport from mainland Europe.

3.3 Extension of the model with Far EasTone (Taiwan)

In a follow-up collaboration with Far EastTone in Taiwan, the joint work, led by Dr Hau Chen Mike Lee, Executive Vice President of FET and supported by the GSMA, showed that all of the methods discussed in this document can be readily replicated to geographies with different environmental challenges. In addition, Far EasTone were able to demonstrate that mobile operators have the opportunity to improve the air quality forecast by adding people and traffic predictors determined from their mobile network mobility dataset.

This section will talk through the initial results obtained by replicating the GSMA Greenwich Machine Learning Model in three cities of Taiwan, the optimization process on top of the basic model, methods to extract population density from mobile network, as well as the added value from mobile data.

3.3.1 Initial results in Taiwan area

Twenty air quality stations in three cities in Taiwan area were selected for the study. The same gradient boosting model chosen for Greenwich, London was applied, like-for-like, on a Taiwan data set that broadly had the same available air quality and weather parameters as for London. The initial prediction results for the three cities are listed below:

Cities in Taiwan	Number of Air Quality Stations	Range of Rsquared result
Taichung	5	0.55-0.67
Kaohsiung	9	0.58-0.69
Taipei	6	0.67-0.8

Table 3.4 Prediction results for three cities in Taiwan

We can see from the table above that R-squared results for Taichung and Kaohsiung were initially lower, at a level of 0.6, compared with Greenwich results which were generally over 0.8. In terms of demographic distribution, economic activity and geographic and geomorphic conditions, Taipei is considered most similar to London, both are low level areas without mountains and both are metropolitan area without major industrial pollution sources (such as power plant). The result in Table 3.4 for Taipei already reflects this analysis.

However, Far EasTone were keen to optimise the prediction performance for all three cities in Taiwan.

3.3.2 Optimisation for Taiwan

The location type of the nine air quality stations in Greenwich, London are shown in the figure below:

Figure 3.3 Geolocation of air quality stations in Greenwich

Across Greenwich, only one site, (GN2) is an industrial site while the others are either suburban site or roadside sites. Reviewing the results of table 3.2, GN2 has the lowest R-squared value among the nine stations which reflects the similar results in Taiwan. Therefore, inspired from this analysis, four power plants were identified in the Taichung and Kaohsiung areas and hourly power plant emission data were added as inputs to the machine learning process using data obtained from the Taiwan government's public website. It was added as a supplementary predictor to the air quality prediction models for Taichung and Kaohsiung.

In addition, other optimisation methods described in the previous sections were also implemented, such as deleting 'NA' (not available) values in the air quality data set, using interpolation methods to fill empty cells in the weather data set, and also tuning of the gradient boosting algorithm parameters as described previously.

By applying all of these methods, the overall prediction performance was successfully improved by around 15% for the R-squared measure to a level of 0.8 and above – so broadly matching the results for Greenwich and demonstrating the possibility to replicate the same methods in other cities of the world.

3.3.3 Application of mobile network data

A key goal when starting this work was to identify the potential use of mobile data analytics to improve the prediction accuracy for air quality. Whilst the public data sets clearly provided a good amount of macro level input on air quality this starts from an 'effect' and we wanted to see if mobile network data could model a 'cause' i.e. population density, the presence of people in their homes/ workplaces and/ or the travel modes of people to workplaces and their later return home. Mobile operators clearly have a unique mobile network utilisation dataset to mine, and some operators are applying their analytics teams to various government and commercial challenges. We wanted to see if it was possible for operators to provide a specific intelligence service such as the air quality forecasting service using mobile network data.

From the previous air quality data analysis shared in[20], we determined that the primary NO2 emissions in big cities are from a combination of road transportation and other sources of combustion (e.g. commonly used gas fired heating in London). Our aim was to have another predictor

[20] GSMA, IoT Big Data, Greenwich Air Quality Proof of Concept Data Analysis v1.0

approximating the road traffic activity which could be added into the prediction model. Mobile network data was expected to be the right candidate to realise this.

Far EastTone successfully extracted population density and population change information from mobile network data aiming to imitate the presence and movement of the people and road traffic.

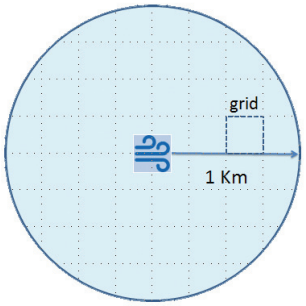
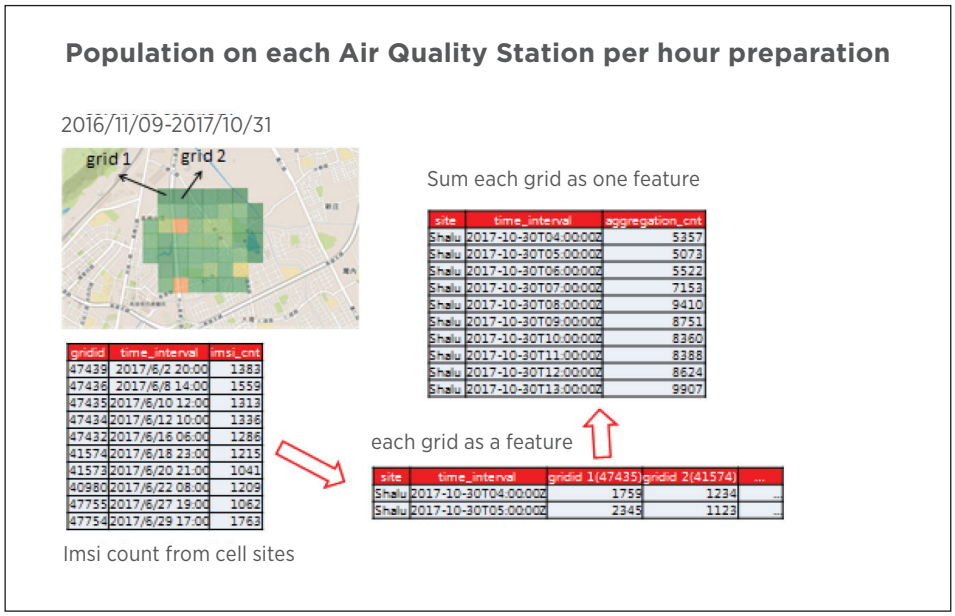


Figure 3.4 Demo of the grid of squares

The basic process assessed, within a one-kilometre radius of each air quality station, an aggregated count of distinct mobile users (essentially, IMSI count) occupying each 250*250 meters square within the circular area. This was used to calculate the total active, distinct, users for each grid during each hour of the day. It is important to mention that the information extracted from the mobile network is aggregated and anonymised information about users in general rather than identifying individual subscriber data, therefore the whole process is totally anonymous and maintains the privacy of individuals in line with regulations protecting use of personal data. An example of how FET aggregated population estimate near an air quality monitor site is illustrated as follows:



The process estimated users in each 250mx250m grid square from network registration data.
Using the Shalu air quality monitoring site, there is an association of corresponding grid ids forming a list. Then instead of using the active user count in each (250mx250m) grid square as an individual feature, the active user count for the list of all grid squares within 1km radius of the Shalu air quality station is summed and forms a single feature. In this example, the total sum of registered users in 51 grid squares was used as a feature for the machine learning algorithm.

Example population density data derived from the FET mobile network can be seen in the table below. Mobile network user counts are aggregated over hourly periods rather than identifying individuals.

site_eng _name	time_interval	Grid 1	Grid 2	Grid 3	Grid 4	Grid 5	Grid 6
S*	09/11/2016 07:00	37	117	122	188	388	27
S*	09/11/2016 08:00	102	289	335	405	818	43
S*	09/11/2016	100	305	346	440	874	62
S*	09/11/2016 10:00	102	297	406	472	932	60

Table 3.5 Sample data from FET mobile network

4 GRAPHS AND VISUALISATIONS OF ANALYTICS RESULTS

Visual exploration methods were a useful tool in analysing air quality and weather data and presenting the final prediction results. There is a detailed exploratory data analysis report shared in The Greenwich Air Quality Proof of Concept Data Analysis[21] with various visual plots generated from the open source air quality and weather data with Tableau, R or Python. This section is a summary of the useful types of graphs used in this air quality proof of concept project.

4.2 Time plot

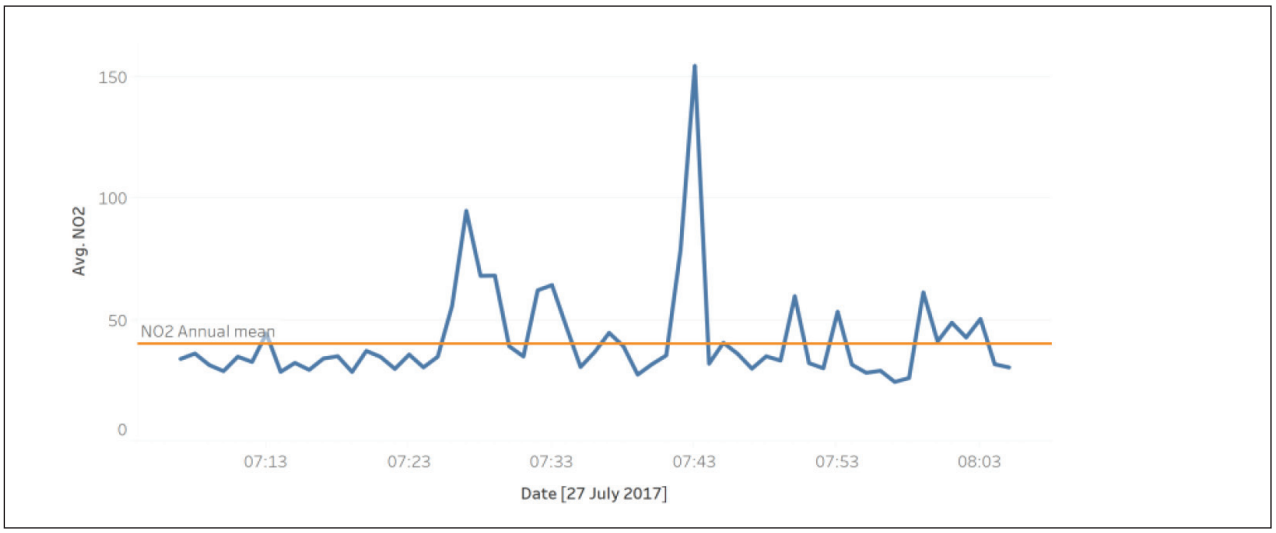


Figure 4.1 time plot of NO2 value

Both air quality and weather data are time-series data. Therefore, conventional plots such as line plots are usually used for a general preview of the air quality value along a time line.

[21] GSMA, IoT Big Data, Greenwich Air Quality Proof of Concept Data Analysis v1.0

4.2 Statistical plot

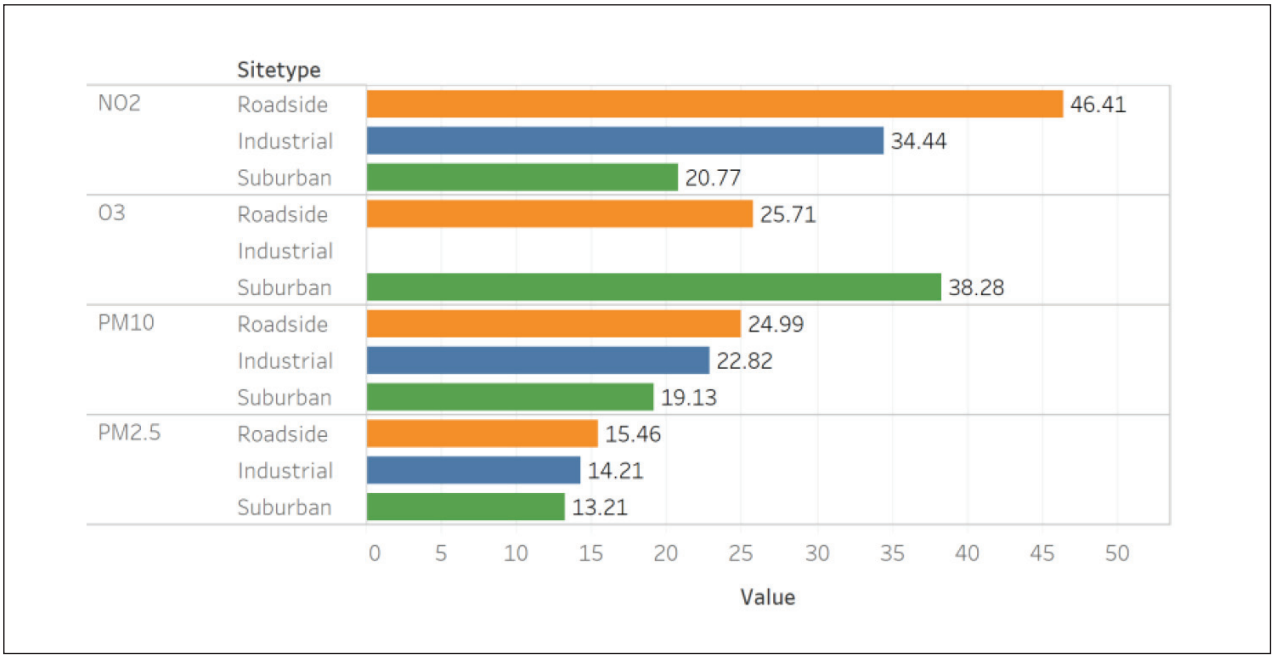


Figure 4.2 Detailed statistical comparison of site type

Bar plot marked with a numerical number gives a clear contrast when making a statistical comparison among different categories of data. It can be read easily from this figure that the average concentrations of different pollutants of NO2, O3, PM10, and PM2.5 across the roadside sites are ranked highest while those at suburban site are ranked the lowest with those at the industrial site in the middle.

4.3 Plots of time patterns

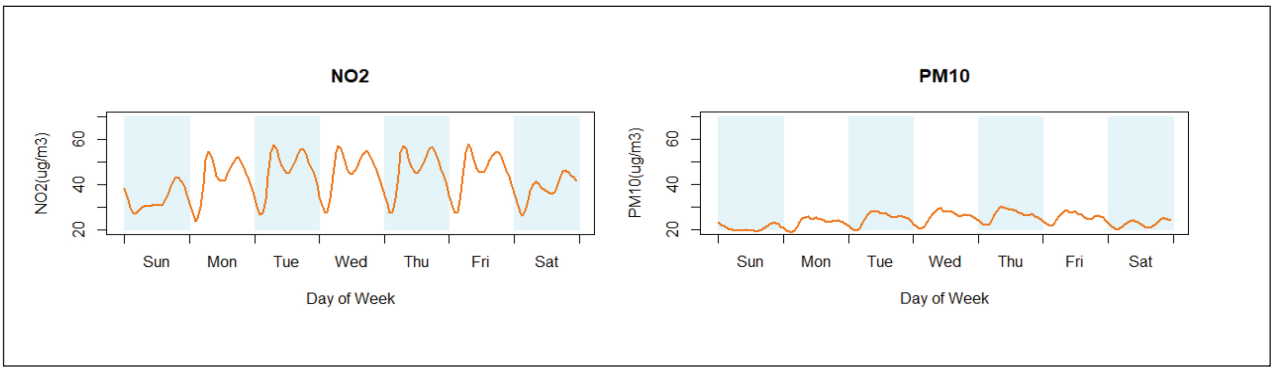


Figure 4.3 NO2 and PM10 weekly-hourly Pattern

- Plots of relationship among variables

The relationship between variables can be visualised through scatter plots like below:

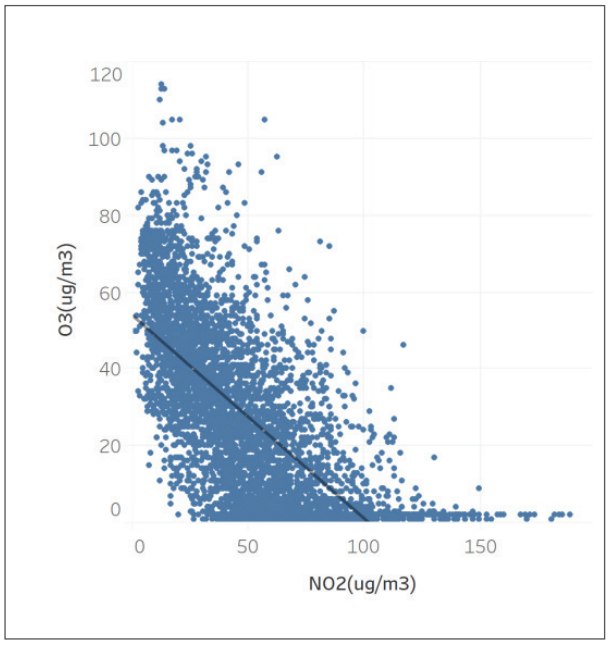


Figure 4.4 NO2 vs O3

By examining the scatter plot above between NO2 and O3, it has already been identified which was also scientific proved later by theory study there is an anti-correlated relationship between NO2 and O3. The findings of relationship identified through scatter plot can be used as reference for further statistical study.

4.4 Spatial plot

Geovisualisation of the pollutant value on a map can give an intuitive representation of the distribution of the pollution which can help identify the hotspot of the most polluted area. For example, the figure below plots the NO2 measurements from the Smogmobile concentrating on a map of Greenwich area.

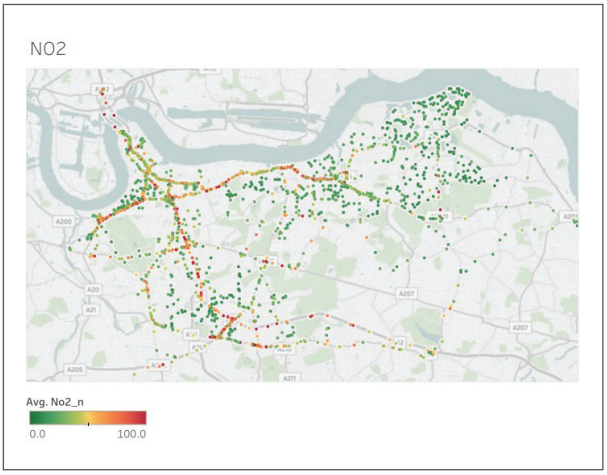


Figure 4.5 Map illustration of the mobile route and NO2 values

It can be seen clearly identified that NO2 values across the major roads experiencing significant pollutant.

The figure above gives good snapshot geographical representation of the air pollution level. When having vast volume of data collected over long period, another way to visualise the pollutant data is by tiling the map into squares like 150meter by 150 meter below. The figure below is a representation of the Smogmobile data collected over 14 weeks with 17.6 million measurement points. Each square represents the average value aggregated from all the data points falling into that square.

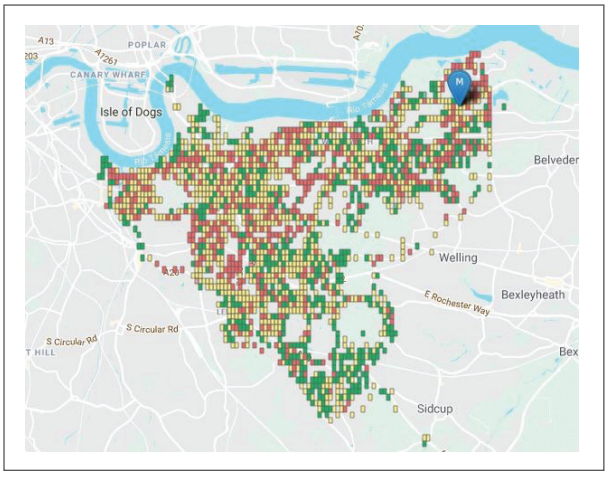


Figure 4.6 Grid level Map illustration of pollutants

A heat map is also a common tool in geovisualising air quality data. For example, the heat map below is generated from a satellite data set which gives an overall view of the distribution of NO2 across whole London area.

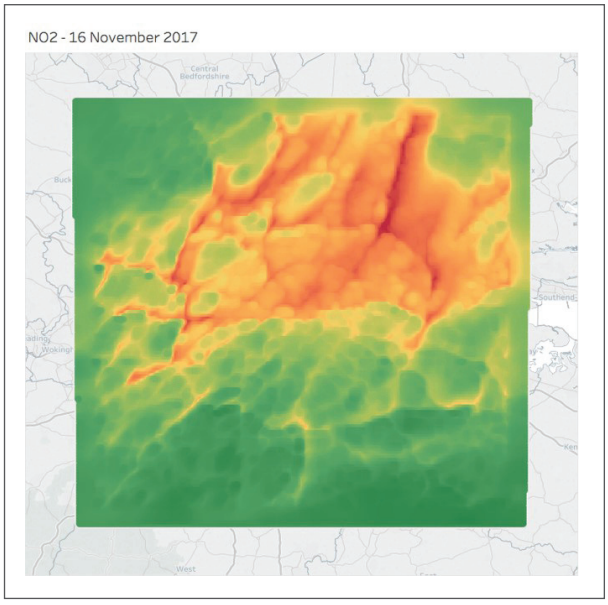


Figure 4.7 Heat map of NO2 satellite measurement

The heat map can help visualise the trend and movement of the pollutants across a large geographical area. Reading from the figure above, the major road network such as the M25 can be clearly identified and it also shows an elevated levels of NO2 heading eastwards from central London out into the suburban and even rural areas of Kent and Essex. The impact of pollution is not limited in local area but propagated large scale.

For more information please visit:
www.gsma.com/IoT

GSMA HEAD OFFICE
Floor 2
The Walbrook Building
25 Walbrook
London EC4N 8AF
United Kingdom
Tel: +44 (0)20 7356 0600
Fax: +44 (0)20 7356 0601