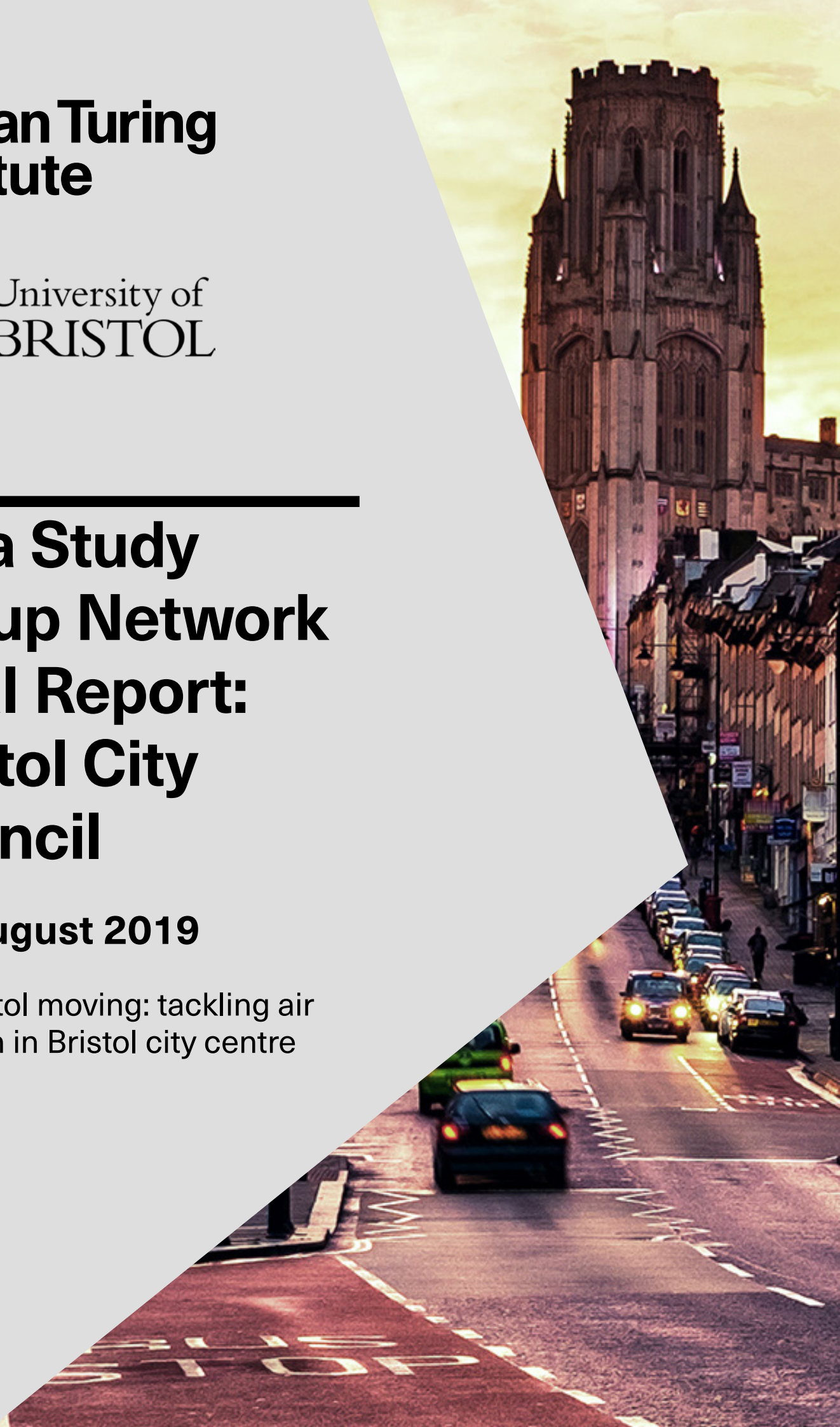
The Alan Turing Institute

University of BRISTOL

Data Study Group Network Final Report: Bristol City Council

5–9 August 2019

Get Bristol moving: tackling air pollution in Bristol city centre



https://doi.org/10.5281/zenodo.3775497

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, 'AI for Science and Government' programme at The Alan Turing Institute

UK Research and Innovation

The Alan Turing Institute

Jean Golding Institute - University of Bristol

Get Bristol moving: tackling air pollution in Bristol city centre

Authors

Carlos Mougan
Siddharth Dixit
Caitlyn Robinson
Hyesop Shin
Qian Fu
Ella M. Gale
Jojeena Kolath
Laurens Geffert
Huan Tong
Ahmad Abd Rabuh

Contents

List of Figures

1 Executive Summary

1.1 Challenge overview

Air quality is an increasing public health concern in UK cities, and Bristol is no exception breaching annual targets for nitrogen dioxide[1]. In Bristol, five people die each week as a result of poor air quality[2]. Like most UK cities, in Bristol the main cause of air pollution is traffic. To reduce the negative impacts of air pollution, and meet increasingly stringent European Union pollutant standards, Bristol is currently consulting on a Clean Air Strategy[3]. A range of policies is proposed including increased public transport, clean air zones and road closures. The city council currently collects a wealth of data (about air quality, traffic flows and weather) for analyzing the distribution of air quality and the relationship with various drivers. Such understanding has the potential to feed into the design and evaluation of air quality policies. However, the council has limited capacity for analyzing and interpreting such large and complex datasets. This project mines this data to understand both the spatial and temporal distribution of air quality. the driving factors of air pollution in different parts of the city, with a particular focus upon traffic.

1.2 Background

- There is ample evidence of inverse relationships between NO_x and children's risk of health e.g. exposure, asthma, headache, and mental health.
  - Roosbroeck et al.(2007)[4] measured personal exposure to traffic-related air pollution in Utrecht, the Netherlands. From 54 young participants, the personal exposure to NO_x was 37% lower on near-road schoolers compared to background schoolers.
  - Kim et al. (2004)[5] conducted 10 school-based surveys from students who walk to school to compare NO_x exposure levels based on their residential location in San Fransisco, USA. Using a logistic model, children who lived within 300m to major roads and at a downwind location have higher exposure levels(OR: 1.05 ± 0.1) compared to those who live far and upwind of major roads.
  - Robert et al. (2019)[6] analyzed a longitudinal cohort study of London-based teenagers and discovered that the 18-year olds who have had depression was associated with high pollution exposure when they were 12 years old (Age-12 pollution exposure was not associated with age-12 mental health problems).
- School zones in Bristol that have a high volume of traffic during drop-off and pick-up times are facing a serious risk of pollution exposure, and children whose schools are closer to the road will be more susceptible to these disease symptoms than the schools distant from the road.
- In line with the Bristol clean air zone proposals[7], the city council is in consideration of closing roads outside schools (Bristol local news[8]).

[1] <https://www.cleanairforbristol.org/what-is-air-pollution/what-is-air-quality-like-in-bristol/>
[2] <https://www.claircity.eu/bristol/city-shockers/air-pollution-map-of-bristol/>
[3] <https://www.cleanairforbristol.org/>
[4] <https://www.sciencedirect.com/science/article/pii/S1352231006012878>
[5] <https://www.atsjournals.org/doi/full/10.1164/rccm.200403-281OC>
[6] <https://www.sciencedirect.com/science/article/pii/S016517811830800X>
[7] <https://airqualitynews.com/2019/07/01/bristol-launch-clean-air-zone-consultation/>
[8] <https://www.bristolpost.co.uk/news/bristol-news/roads-outside-schools-could-closed-3001742>

- However, we argue that the proposed plans can only take into further consideration once a critical exploration of pollution, weather, and traffic is done.

1.3 Data overview

We build upon a wide range of exciting and open datasets provided by Bristol City Council. This includes continuous air quality datasets, point observation, and route observation traffic datasets, weather observations and other spatial datasets.

1.4 Main objectives

The projects aims were twofold:

- A1. To understand the spatial and temporal distribution of pollutants and drivers
- A2. To examine the relationship between air quality and its drivers (e.g. traffic and weather)

In working with the council's data to understand this relationship, the project also feeds into the Our Data Bristol[9], an exciting initiative that allows local people and organizations to access, use and benefit from a wide range of open-source datasets and technology. Experience of working with the datasets during the project will allow recommendations to be made about how datasets could be more accessible and comprehensive, and how to transport datasets could be made available in the future.

1.5 Approach

The project was divided into three work packages (WP). WP1 and WP2 address our first aim. WP3 addresses our second aim.

WP1. Analyzing the temporal distribution of air quality
WP2. Analyzing the spatial distribution of air quality
WP3. Estimating the importance of different drivers of air quality

1.6 Main conclusions

This study firstly (WP1) explored the temporal features of pollutants (NO_x), then examined Parson Street School as a case study. Our initial finding was that most stations had a daily and seasonal oscillation of NO_2 throughout the whole period ranging from less than $20g/m_3$ to over $1000g/m_3$ based on hourly measurement. Looking into an averaged hourly NO_2, there was a clear trough around $20g/m_3$ at 3-5 am but peaked at around $60g/m_3$ after 10 am. However, the concentration was always higher in the city center sites.

Our second finding (WP2) was that the NO_2 concentration between holidays and non-holidays (school holidays and bank holidays 2018-2019) was on average less than $5g/m_3$, however, varied by locations.

We also were able to find what are the most important drivers in different geospatial locations (WP3). We saw that some meteorological conditions were important in some places and that they were less important in others. One of the most important drivers that we found was the standard

[9] <https://www.bristol.gov.uk/data-protection-foi/open-data>

deviation over time of temperature, we interpreted this as the change of temperature in any given hour. This goes along with the intuition that changes in the meteorology have an impact on the pollutant concentration.

1.7 Limitations

In this project we found several limitations that did not allow us to proceed further and find better results:

Data Limitations: Most of the data sets were not pre-processed enough or they didn't have enough information to answer accurately the questions proposed. For example the diffusion tubes limited the spatial understanding of the pollution. It was also hard to work with the traffic data, and to extract solid conclusions due to difficulties while combining the tables (not a common and index). In order to improve that, we suggest trying to aggregate all features in one tabular dataset, where data is easier to understand and manipulate. Even if this is not achievable due to the task requirements just orientating the dataset in this way allows is for faster pre-processing.

Approach Limitations: With the data provided, we could not find the association between the diffusion tubes and the NO_x. This blocked us from trying more algorithms and eventually getting better results. Some of the approaches we wanted to try but we couldn't either for lack of time or lack of quality data are statistical hypothesis testing, geographically weighted modeling, hierarchical modeling, and k-nearest neighbors for spatial classification of tubes. On the other hand when we tried to answer if there was any difference in holiday vs term but we were only able to do visualization and we were unable to proceed with further statistical analysis.

1.8 Recommendations and further work

In the last section, we provide different ideas of how we would have approached the problem given more time and resources. Some of them are more focused on the data and others on mathematical modeling.

- For the data: improving the datasets, gathering more data, expanding the work on traffic volumes.
- For the mathematical modeling: the first improvement would be to do statistical hypothesis testing. Other recommendations that we think that would be convenient is to do geographically weighted modeling, as this will allow us to see how the driver's behaviour changes in a spatial way. Once we have this, we suggest doing hierarchical modeling to see if there is a station that behaves similarly. Some further recommendations in the Data engineering techniques: winsorization[10], logarithmic transformation and adding meteorological features such as geopotential height, dew point ...
- For determining the difference between holiday and term-time, it would be necessary to do a statistical analysis called hypothesis testing, without these tests, we cannot say whether there was a statistically significant different or not. One idea that we were not able to experiment with was making a model that predicted if a certain date was a holiday or not.

[10] <https://en.wikipedia.org/wiki/Winsorizing>

2 Dataset overview

The following section provides information about the datasets used during the project, how these were processed and the dataset limitations.

2.1 Data description

A wide variety of datasets was available to the group which is explained in detail below, including information about how to access the processed datasets.

2.1.1 Air quality data

Bristol has been monitoring nitros oxides NO_x, which includes NO and NO_2) from a single site since January 1998, and has expanded it to multiple stations around the city center.

At present, the city has seven stations, in the city center, that monitors NO_x , NO_2 , NO, and PM_{10} (see Figure 1). Each station monitors a different type of environment: Colston avenue (detects exposure in city center), AURN St Pauls (detects background exposure in residential area), Brislington depot (freight-caused exposure), Fishponds road (residential and shopping), Parson street school (residential area and school zone), Wells road (continuous traffic in and out of city center).

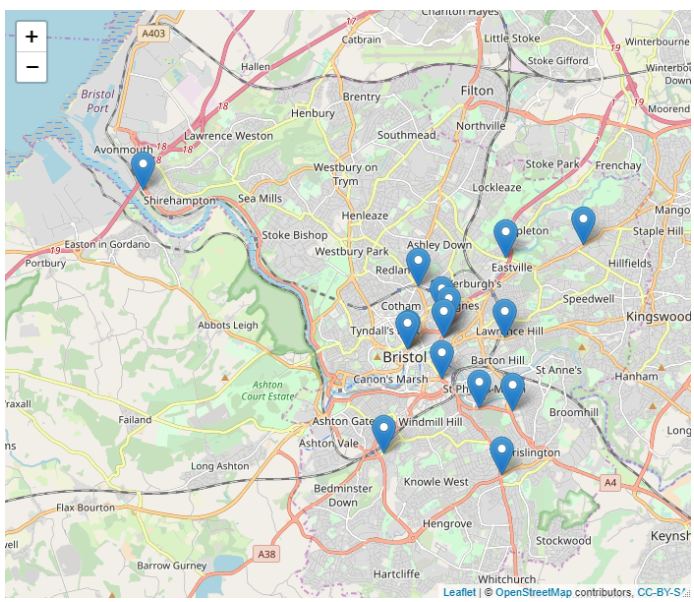


Figure 1: Location of Bristol area air pollution monitoring site

Exploring the continuous air quality data we can see that the stored value for many measurements of air pollutants was `NA`, indicating that the data was missing or it was not logged. We decided to retain these in the dataset for now though in case we wanted to use them in a later analysis.

Diffusion tube data measuring annual averages of NO_2 were also made available. Diffusion tubes/Diffusive samplers are widely used for indicative monitoring of ambient nitrogen dioxide (NO_2)[11]. Although much less accurate than the continuous air quality measurements, diffusion tube data was available at approximately 127 sites (varying depending upon the year) across the city making it useful for spatial analyses.

At the recommendation of the council, we primarily focused upon NO_x because the pollutant more accurately reflects traffic patterns. Some of the analyses also explored NO_2 which is useful for the council when considering regulatory compliance.

2.1.2 Weather data

There was weather data available from the challenge data and the Open Data Bristol[12] website. However, of the two stations available, one went offline in 2015[13]. The other was still up and running but was missing some essential variables[14] such as rainfall. We, therefore, obtained an external dataset from a metoffice weather station[15]. The data was downloaded from the website manually (one month at a time) and combined using the code in "products/01/*data*/clean/data/*reader*/weather.R". CSV formatting seemed to change as a function of the operating system of the downloading computer, which resulted in the necessity to use the two slightly different import functions that you see in the `R script`. It is worth noting that since there was only one weather data station available, we assume in our analysis that the weather conditions are the same across the entirety of Bristol.

2.1.3 Holiday data

To understand whether and how school traffic impacts air quality, we used school holidays and bank holidays as control group observations. Data on holiday days was not provided in the challenge data, so we created our own dataset. School term dates were taken from the local council website[16] and bank holiday dates were taken from the national government website[17]. From this information, we created Boolean variables for weekends, school holidays, and bank holidays for the school year 2018/19 (August to July).

2.1.4 Traffic data

Four traffic datasets were provided: flow detector data, link travel time data, camera plate data, and SCOOT link data. Flow detector data is generated by detectors measuring vehicles passing a certain point in the road. Camera plate data is generated by average speed cameras, which can monitor the time that has passed between two sightings of the same number plate, however, stops along the road are not taken into account. This means that the journey time recorded will be an upper bound but should be used with caution, we may see high-value outliers.

[11] `<https://laqm.defra.gov.uk/diffusion-tubes/diffusion-tubes.html>`
[12] `<https://opendata.bristol.gov.uk>`
[13] `<https://opendata.bristol.gov.uk/explore/dataset/meteorological-data-create/information/>`
[14] `<https://opendata.bristol.gov.uk/explore/dataset/met-data-bristol-lulsgate/table/?sort=date_time>`
[15] `<https://wow.metoffice.gov.uk/observations/details/20190806qbpnbopacce6ucrdyyb96sczue>`
[16] `<https://www.bristol.gov.uk/schools-learning-early-years/school-term-and-holiday-dates>`
[17] `<https://www.gov.uk/bank-holidays>`

2.1.5 Additional useful datasets

In addition to data about air quality, travel and weather, we also used several datasets to provide context to our analyses: area-based demographic datasets and boundaries from the Office for National Statistics[18] and Ordnance Survey Open Roads[19].

2.2 Data quality

Many of the datasets available to the group required pre-processing cleaning and aggregation. Here we give a list of issues identified when importing the data, explain how we addressed each issue, and provide cleaned versions of the datasets.

2.2.1 Metadata

Some of the metadata is not easily available. For example, geolocations of air quality measurement stations are only available in the full continuous measurement file, not just the coordinates in a separate file or table.

2.2.2 Data Integrity

Parts of the data contained unintended artifacts. For example, the traffic data contained repeated header rows pasted into the data rows of the CSV. Our data importing functions take care of this and create cleaned versions of these datasets. Another issue we encountered was inconsistent formatting of date fields but we were able to address this with the excellent conversion functions of the `lubridate` R package.

2.2.3 Data Schema

There seems to be no distinction between FACT (a table that containes[20] measurements, metrics or facts about a business process.) and DIM data (companion table to the FACT table that contains descriptive attributes to be used as query constraining.) in the data schema. It could be useful to use a star schema[21] and group small DIM tables around larger FACT tables. Examples for possible DIM tables could be locations, setup times, names, and types of traffic sensors or weather stations, which can then be referred to by ID in the larger FACT tables such as traffic measurement and air quality data measurement tables. This would both make DIM data more easily accessible and reduce the size of the FACT tables. Besides, this will ensure data consistency in the future when DIM values are updated centrally.

2.2.4 Naming Conventions

In several cases, we decided to rename columns from the names they had in the original datasets provided. We are listing some examples of our renaming policy in case this can inform a future database design:

[18] <https://www.ons.gov.uk/searchdata?q=neighbourhood%20statistics>
[19] <https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-roads.html>
[20] <https://www.guru99.com/fact-table-vs-dimension-table.html>
[21] <https://en.wikipedia.org/wiki/Star_schema>

Remove spaces and special characters - Spaces and special characters can make it difficult to access columns when working in python. We replaced all spaces with an underscore and removed special characters where possible.

Remove upper-case-letters - (unless from acronyms) Uppercase adds cognitive load for the user and can lead to inconsistencies in data processing pipelines. We converted all uppercase letters with lowercase letters, apart from acronyms or chemical formulas such as NO_2.

Use hierarchical semantic taxonomies - Informative short words with hierarchical group-ing sequences could be used to create a variable name hierarchy. We used this to make it eas-ier for the user to identify similar variables and to use command-line autocomplete for the de-sired result. For example, wind variables in the weather file were named`wind_direction_mean`,`wind_direction_sd`,`wind_speed_mean`,`wind_speed_sd`,`wind_gust_mean`, `wind_gust_sd` to keep similar variables together.

Ensure consistency across datasets - Consistency can be important when datasets are offered together. Examples for this that offer the potential for improvement are the coordinate variables which are sometimes called `lat`/`lon`, sometimes `Eastings`/`Northings`, and sometimes are misspelled (e.g. Longitude in traffic metadata). We renamed all of these, making it easier for the user to find variables that overlap and can be used to combine datasets.

2.2.5 Dataset information

Transport-related datasets utilised during the experiments can be located using the following:

- **A** - Weekly Cemara Plate - ("CSV CameraPlateWeekly")-
- **B** - Detector Flow - ("CSV DetectorFlow")-
- **C** - Link Travel Time -("CSV LinkTravelTime")-
- **D** - SCOOT Link Data -("CSV SCOOTLinkData")-

Understanding and cleaning the transport-related data is a significant challenge faced by the team.

- Geographical information was missing/unavailable in both the data sets of A *and* B, so that the group was unable to identify/locate where the information about traffic volumes was sourced.
- There was a lack of some common attribute(s) that might allow the team to merge these data sets.
- Identifiers of the `links` (i.e. "road sections" - to the understanding of the team) in both data sets C *and* D are different.
- In data set **D**, it happens frequently that traffic information (e.g. average speed) on any given link remains constant along the time.
- Finally, traffic datasets do not have a long history as the oldest goes back to 2018-12-05 as opposed to air pollution and weather data that are, in one station, go back to 1998-01-10.

3 Experiments and Results

Given the issues in the available data sets, only dataset **C** was further processed. The team defined an area, referred to as a "buffer zone", for each AQ monitoring station. For the sake of testing the idea, the buffer zone was designed as a circle for this task, with the location of the station being the center. The team examined different sizes of buffer zones, using the radii of 50, 100, 500 and 1000 meters, respectively, and extracted relevant road links within and/or intersecting the buffer zones. This can facilitate geo-coordinates transformation.

3.1 WP1: Analysing the temporal distribution of air quality

Addressing aim A1, WP1 explores the temporal features of NO_2 across Bristol monitoring sites. Besides, the temporal features of meteorological factors are examined. Due to the focus upon temporality, we chose to focus upon a case study of air pollution at Parson street primary school.

3.1.1 Experiments

There have been similar studies in other UK cities, for example, using air pollution measurements from a background site in Central London, Bigi and Harrison (2010)[22] analyze the seasonality of NO_2 recording minimum values in June/July and maximum values during Winter. Annual patterns of NO_2 were found to be similar to that of CO and NO. Meanwhile, the NO_2 weekly pattern is weak, with the mean concentration remaining steady during weekdays with only a small drop during weekends. NO_2 is a relatively stable pollutant.

3.1.2 Results and visualizations

In general, all stations have a daily and seasonal oscillation of NO_2. However, there was a variation by stations. The averaged NO_2 was the highest on Rupert Street (city center) at $93.1\mu g/m_3$, followed by Colston Avenue and Temple Meads station at 65.7, 63.2 respectively.

From all the stations, Parson Street school has been monitored since 2002 had an average of $47.7\mu g/m_3$, but the hourly concentration of NO_2 exceeded the legal limit on 154,189 data points. Although it has declined in recent years, the high counts are very much a concern. Overall, the averaged NO_2 tends to soar rapidly between 7am and 10am, which roughly peaks at twice the amount than the concentration at 5am. It remains the concentration until 6pm, then gradually decreases late at night. For example, Parson street school has the lowest NO_2 concentration just above $20\mu g/m_3$ at 4am, then rose up to $54\mu g/m_3$ at 10am.

The boxplot in figure 2 used the monthly average of NO_2 at 7 stations in 2018-2019. Overall, the weekday concentration was higher than on weekends. Amongst all stations, Colston avenue (city center) had the highest amount of NO_2 that ranged $50 - 70\mu g/m_3$ during weekdays and fallen 60 $\mu g/m_3$ on Saturdays, then decreased further on Sundays at around $50\mu g/m_3$. Parson Street school just managed to go below the national NO_2 limit of $40\mu g/m_3$ during weekdays.

[22] <https://www.sciencedirect.com/science/article/abs/pii/S135223101000155X>

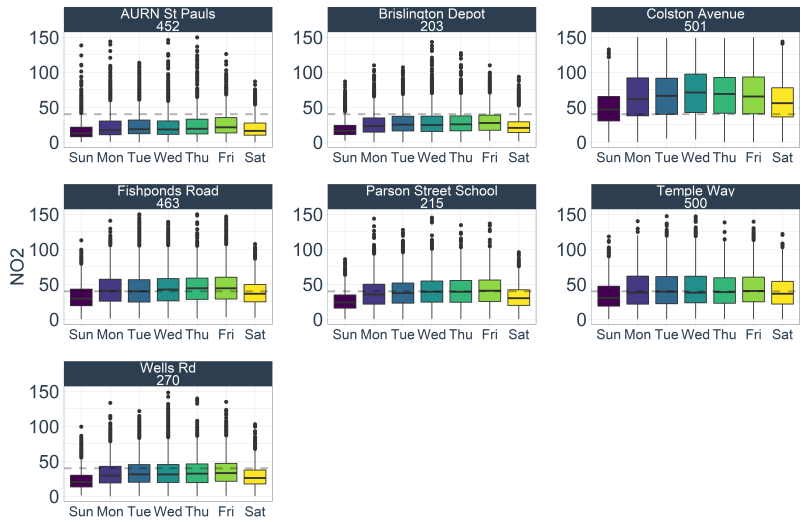


Figure 2: Daily aggregated concentrations of NO_2 over 2018-2019.

School Holiday

Overall, while there is a small difference between holidays and non-holidays ($5\mu g/m_3$), it has a distinction by locations.

- Colston Avenue a consistently high concentration of NO_2 above 50 regardless of holidays, but the concentration varied during holidays. Similar patterns were seen on Temple Way.
- During term time, children who were commuting to schools near Parson street primary school or passing Fishponds road might have experienced a high level of NO_2 since the concentration ranges up to $80\mu g/m_3$.

Figure 3: Holiday verus term box plot for NO_2.

Monthly evolution

The boxplots in figure 7 used the monthly average of NO_2 at 7 stations in 2018-2019. The visualizations of figure 4 show different time aggregations (hour, month and weekday) for an average of all the stations.

- Overall, the averaged NO_2 tends to soar rapidly between 7am and 10am, which roughly peaks at twice the amount than the concentration at 5am. It remains the concentration until 6pm, then gradually decreases late at night.
- For example, Parson street school has the lowest NO_2 concentration just above $20\mu g/m_3$ at 4am, then rose up to $54\mu g/m_3$ at 10am.
- Overall, the weekday concentration was higher than that of weekends, see figure 2. Amongst all stations, Colston avenue (city center) had the highest amount of NO_2 that ranged $50 - 70\mu g/m_3$ during weekdays and fell $60\mu g/m_3$ on Saturdays, then decreased further on Sundays at around $50\mu g/m_3$. Parson Street school just managed to go below the national NO_2 limit of $40\mu g/m_3$ during weekdays.

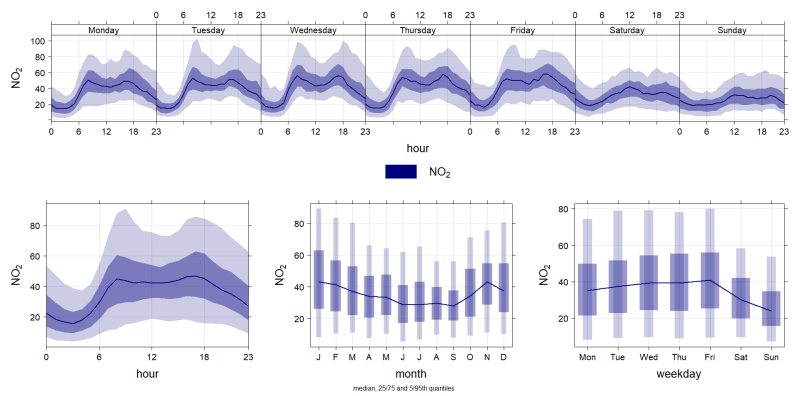


Figure 4: Monthly average temporal variation of NO_x concentration for all stations over 2018-2019.

Year Evolution

As can be seen in figure 5, the distribution for NO_2 concentrations in Bristol across 2010 to 2018 is not even. More ambient concentrations of NO_2 are clustered in the center of Bristol (around the University of Bristol and railway station) for each year. Changing from 2010 to 2018, the NO_2 spread from center to southwest in Bristol. From 2010 to 2012, the polluted area becomes larger, which is decreased from 2013 to 2018.

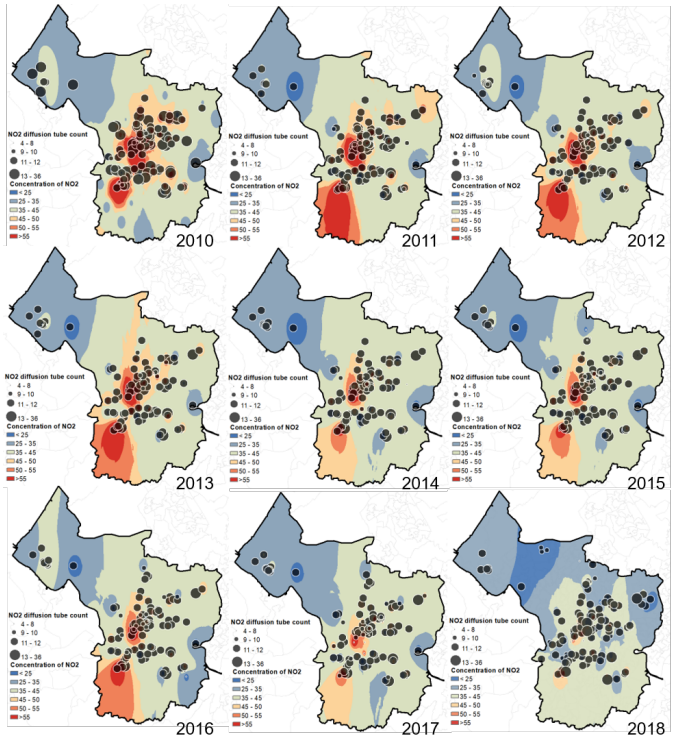


Figure 5: Visualization of the yearly evolution of the NO_2 concentration

3.2 WP2: Analysing the spatial distribution of air quality

3.2.1 Task description

This work package identified spatial 'hot spots' in the air pollution data provided by the city council. Mapping of spatial hot spots can inform future analyses about where air quality is of greatest concern, and where further exploration of the drivers of air quality more likely to be of greatest importance.

3.2.2 Experiments

In order to identify and understand the different hot spots of air pollution across Bristol several methods were used: Hot Spot Analysis[23], Kernel Density Estimation[24] and Inverse Distance Weighting[25]. Both the diffusion tube data and the continuous air pollution datasets were utilized, due to the small number of continuous air quality stations ($n = 7$) limiting the spatial understanding of pollution without complex modeling.

[23] <https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/hot-spot-analysis.htm>
[24] <https://en.wikipedia.org/wiki/Kernel_density_estimation>
[25] <https://en.wikipedia.org/wiki/Inverse_distance_weighting>

3.2.3 Results and visualizations

The method by which residents commute to work was extracted from census 2011 from National statistics in Lower Layer Super Output Area. In this section, the Spearman correlation was used for analysis because the variables were not normally distributed. There are two indicators in this section: car availability and commute transportation type. The commute transportation type was divided into two groups: public transportation and self-transportation.

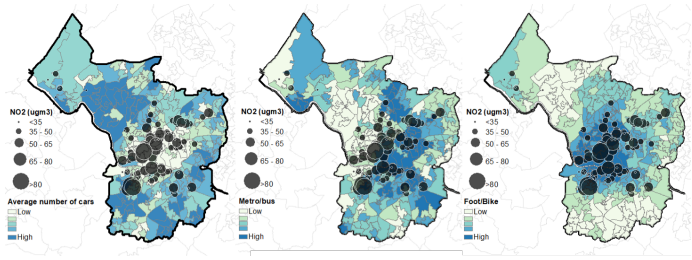


Figure 6: The spatial relationships between NO_2 distribution and transport type to work for residents

Figure 6 shows the results of commute analysis. The map shows that the areas where people are more likely to walk and cycle to work have a higher concentration of NO_2 and (using the Spearman correlation) this relation is statistically significant (at the 0.05 significance level, i.e. there is only a 5% chance that this relation is spurious). There were no statistically significant relationships was found between the NO_2 level and car availability or commuting via public transportation.

Temporal distribution by station

In general, all stations (except AQ Mesh Temple way site) have a daily and seasonal oscillation of NO_2, see figure 7. However, there was a variation by stations:

- The averaged NO_2 was the highest on Rupert Street (city center) at $93.1\mu g/m_3$, followed by Colston Avenue and Temple Meads station at 65.7, 63.2 respectively.
- Parson Street school which has been monitored since 2002 had an average of $47.7\mu g/m_3$, but the hourly concentration of NO_2 had exceeded the legal limit 154,189 times, and although the number of time the limit has been exceeded has declined in recent years, this is very much of a concern.

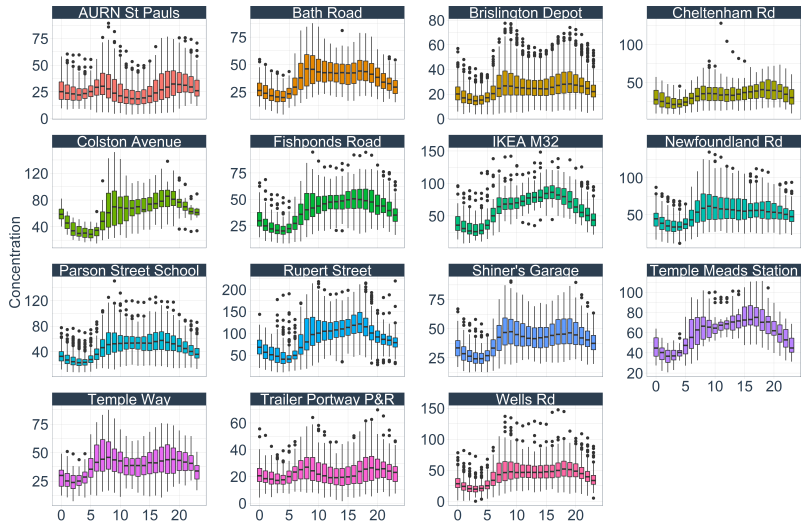


Figure 7: Concentrations of NO_2 in different measurement spots versus day hours.

3.2.4 Limitations

The main limitation for this method is that the diffusion tubes are clustered around the center, so the analyses at the Bristol boundary have lower accuracy.

3.3 WP3: Estimating the importance of different drivers of air quality

3.3.1 Task description

During this work package, we related air quality to the weather data (both recorded hourly). Our aim is to relate NO_x levels in the air to both meteorological and air pollution predictors. In the process, we also found the important factors which affect NO_x levels throughout Bristol. This is further extended to finding and comparing the important features location-wise to see if the importance of features varies with location.

3.3.2 Experiments

Since the data is large, we only consider the data for the last two years and split it into train and test set.

Model Building:- Since we had a lot of variables/features and wanted to find the least number of features that described the problem in the best possible way, we used a feature selection method called the *Best Subset Selection Method*[26] .The *regsubsets*() function (part of the *leaps* library) performs best *subregsubsets*() set selection by identifying the best model that contains a given number of predictors, where best is quantified using Residual Sum Squares (RSS). By default, it selects the best eight variables of the model, but that can be changed by specifying the *nvmax* parameter.

[26] `<https://www.stat.cmu.edu/~ryantibs/papers/bestsubset.pdf>`

Feature Engineering: Even after having good descriptive features we wanted to take into account the variation with time of these features. For this an easy way to calculate variation is when aggregating by features by the hour or day, to incorporate the aggregated standard deviation. The aggregated deviation can be grossly interpreted as the derivative of this feature respect to time. If we have, for example, "temperature standard deviation" chosen as a selected feature, this will mean that the variation of temperature is an important driver.

In this work package, we will fit different models to see what are the most descriptive features in different ways and how do they describe the prediction of NO_x.

3.3.3 Results and Visualizations

After training the model and identifying the best set of features using the best subset selection algorithm, we identifies the following eight features as important: NO_2, NO, temperature mean, wind direction mean, rainfall mean, temperature sd, wind direction sd and humidity sd (where sd is standard deviation).

Even though this method is a good technique to select a solid set of predictors, it does not give us an idea about the relative importance of those features. For this, we will have to use to evaluate the obtained coefficients.

Also, an important point to note is that the NO_x levels are a linear combination of NO and NO_2, so counting them as important predictors won't give us much information. Thus we drop them and make use of Lasso Regression[27] to compare the importance of features amongst themselves (relative importance) and get a more insightful outcome. The idea of using Lasso Regression is the following the lasso shrinkage causes the estimates of the non-zero coefficients to be biased towards zero and in general they are not consistent[28], this means that this shrinkage method will drop all features that are not relevant and no consistent in the dataset. In order to evaluate the accuracy of the model we used the Root Mean Squared Error (RMSE) for this Lasso Regression Model that turns out to be $rmse = 75.50$.

We first tried to identify overall feature importance across Bristol. Figure 8 indicates that, in general, the variation in temperature with time (temperature sd) had the highest effect on the NO_x levels all across Bristol.

- The mean wind speed, variation in rainfall, variation in wind speed and the mean temperature have a moderate effect on the NO_x levels.
- Wind Direction, humidity and the variation in wind direction play almost no role in affecting the NO_x levels.

We compared the feature importance for different locations around Bristol regarding those affecting NO_x levels with the general case of the whole of Bristol (i.e. Feature Importance Overall). As an example, taking one location the following differences were observed:

- The variation in "Wind Speed" is more important than the variation in "Rainfall".
- The variation in humidity does not affect NO_x levels in Brislington Depot.

[27] <http://statweb.stanford.edu/~tibs/lasso/lasso.pdf>
[28] <https://web.stanford.edu/~hastie/ElemStatLearn/>

Doing this kind of spatial modeling for all the locations in our dataset allows us to see which drivers are important in different locations. This means that the problem that we are modeling varies as we move geospatially, and that is why it is a good idea to tackle it with Machine Learning (ML).

We wanted to better understand the interplay of weather, measurement station location, and traffic with air quality. To explore the interaction of these variables we built a machine learning model aimed at predicting NO_x levels from the independent variables (weather and traffic).

The flow detector traffic dataset was used for this. We aggregated the data to hourly averages of traffic flow and converted the data to a wide format (a single row for every data point with multiple columns), making measurements at each detector into a single predictive feature. We also included weather data from the Bishopston weather station and one-hot-encoded site id to allow the model to generate a bias term for individual sites accounting for differences in setup and positioning of the air quality measurement station.

Data from all three variable categories (traffic, weather, and air quality) was only available for a limited time frame ranging from the fifth of February 2019 to the twelfth of May 2019. Given this limited data availability, we wanted to make sure that our model would generalize well to conditions outside the training window, thus we created a true temporal holdout, using the period of `2019-02-05` to `2019-04-30` for training and `2019-05-01` to `2019-05-12` as a blind holdout.

We fitted two models:

- a generalized linear model with elastic net regularisation[29]
- a gradient boosting machine model[30]

Elastic Net Model

This is a simple learning algorithm that we used as a baseline. It trains fast but only fits simple functions and doesn't by default include interactions between variables. It does, however, avoid overfitting by excluding unpredictive variables through elastic net regularization (we used an alpha parameter of '0.3' here.)

The results of this model were not great, with a RMSE of `69.0` on the cross-validation-folds, and an R^2 value of `0.61` on the blind holdout.

In the feature importance figure, figure 8, for the Elastic Net model, we can see that the most important feature is the standard deviation of the temperature. As explained in the feature engineering part this can be interpreted as the variation of temperature over time.

Gradient Boosting Model

This is a more complex model that uses decision trees boosting to fit more complex decision boundaries to the data. We tested a range of hyper-parameters for tree complexity (allows for variable interaction) and shrinkage (avoids overfitting by reducing the contribution of each fitted iteration).

The results this time have a lower error, with a RMSE of `45.81` on the cross-validation-folds and a root mean square error of `29.43` and an R^2 value of `0.94` on the holdout data.

It seems very surprising that we are doing better in predicting the holdout data than the cross-validation folds. Our hypothesis for an explanation is the temporal variation of the data. The training data contains winter and early spring months while the holdout is two weeks of May. Air pollution, especially NO_x, tends to be higher in winter. We see this in our data (mean NO_x of the

[29] `<http://ww.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf>`
[30] `<http://uc-r.github.io/gbm_regression#gbm>`

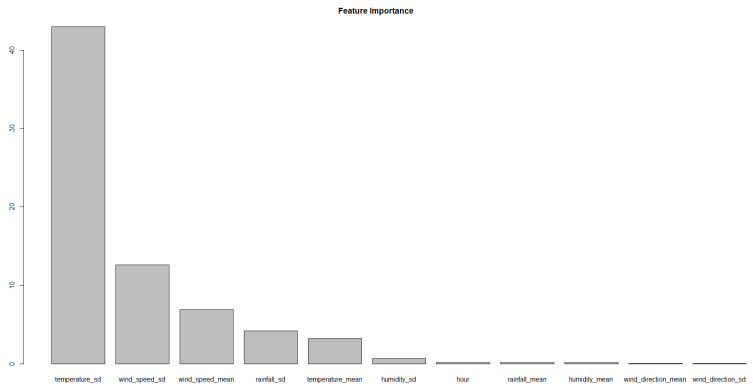


Figure 8: Feature importance in the ElasticNet model. In order of appearance (from left to right): temperature sd, wind speed sd, wind speed mean, rainfall sd, temperature mean, humidity sd, hour, rainfall mean, humidity means, wind direction means, wind direction sd

training set is `89.95` (sd +-`109.45`), mean NO_x of the test set `74.11` (sd +- `81.92`)). If we had year-round data then this could be accounted for by re-scaling our data. Since we don't have a whole year of training data available though, including the slightly different holdout data in our training setup would constitute data leakage.

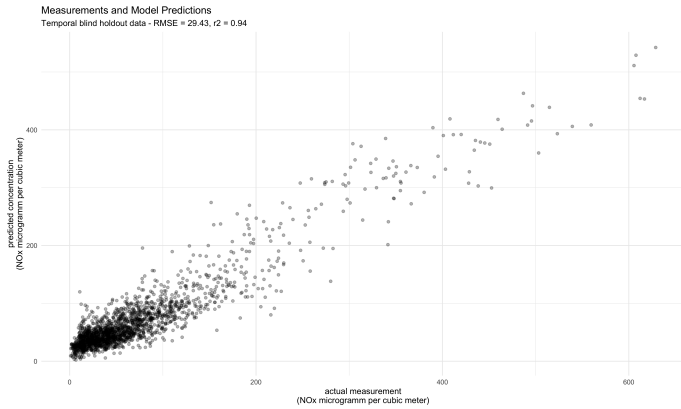


Figure 9: Model predictions on holdout

4 Conclusions

- Our initial finding was that most stations had a daily and seasonal oscillation of NO_2 throughout the whole period ranging from less than $20\mu g/m_3$ to over $1000\mu g/m_3$ based on hourly measurement.
- Looking into an averaged hourly NO_2, there was a clear trough around $20\mu g/m_3$ at 3-5 am but peaked at around $60\mu g/m_3$ after 10 am.
- However, the concentration was always higher in the city center sites (Parson Street School, Rupert Street). Although the average NO_2 levels of Parson street schoolers were just below the UK legal limit ($40\mu g/m_3$), the children might have consistently encountered instantaneous NO_2 exposure during drop-off and pick up times.
- The NO_2 concentration between holidays and non-holidays (school holidays and bank holidays 2018-2019) was on average less than $5\mu g/m_3$, however, varied by locations.
- Different stations have different pollutant drivers. This means that if we decide a policy to reduce contaminants (reduce traffic, block traffic in weekdays . . .), the effect won't be the same for all stations. This geospatial variation is an important factor to take into consideration when deciding how to reduce air contamination.

5 Future work and research avenues

5.1 Statistical Hypothesis Testing

For further assessment of some questions such us is there a difference between **non school days and school days** or **bank holiday versus non-bank holiday**, our proposal is just not to make data visualization but to do solid statistical hypothesis testing. There are several ways to do hypothesis testing:

- **Classical Statistics** Applying some of the traditional statistical test such us Kolmogorov–Smirnov, T of the student, Pearson ...
- **Machine Learning** Creating a target variable e.g. is it holiday day or not. Then fitting a machine learning algorithm such as Lasso[31], or gradient boosting[32]. With these modern statistical methods, we are able to determine whether is there any statistical significance or not between the created target variable and the rest of the features.

5.2 Future datasets and data collection

From the traffic data provided by Bristol City Council, it is currently not possible to differentiate between vehicle types. To provide useful conclusions that could meaningfully inform the proposed Clean Air Zone Framework, it would be useful to understand what proportion of vehicles are diesel, and whether the age of the vehicle means that it complies with European emission standards.

The closest weather station that forms part of the Met Office and continues to produce up to date weather information is outside the city. To improve modeling of air quality, Bristol city centre would benefit from its own weather station that can produce up-to-date meteorological information including wind speed and temperature. This would particularly benefit spatial modeling approaches.

5.3 Expanding work on traffic volumes

To expand upon the work carried out relating air quality and traffic (WP3) the extracted traffic information and air quality data should be integrated (and any other relevant information, e.g. meteorological observations) in both the temporal and spatial context. It would then be possible to develop and test a prototype data model capable of making a prediction of air quality given the traffic conditions considered in the model. In this DSG challenge, however, this task of model testing was not completed due to the data issues.

5.4 Geographically weighted modeling

The format of the air quality data lends itself in particular to temporal analyses of how air pollution levels varied throughout the course of the day, and also seasonally. However, air pollution is also a highly spatially variable and locally contingent phenomenon, negatively disproportionately affecting particular roads, neighborhoods, and communities. Building on the findings of WP3 which tries to understand the relative contribution of different variables to poor air quality, it would be useful to understand the spatial variance in these relationships.

Although we experimented with geographically weighted regression models, it was not possible to specify the models correctly. This was due to the small number of air pollution stations ($n = 7$) that also had weather and transport data associated with them. GW models could be an interesting avenue of further research for the city council, using the R package GW Model (Lu et al. 2015). As GWR works best with over 100 spatial features, this would be best applied to the NO_2 diffusion tubes which are more widely spatially distributed across the city.

[31] <http://statweb.stanford.edu/~tibs/lasso/lasso.pdf>
[32] <https://arxiv.org/abs/1908.06951>

5.5 Log Transformation

The training data contains many NO_x measurements between 20 and 200 while only a few observations are beyond the threshold of 200 or even 400, up to a maximum of 2164.25 micrograms per square meter. It may be worth a try fitting a model on the log-transformed NO_x values instead of the raw measurements to get a more evenly distributed response.

5.6 Winsorization

There were several (85) observations for which NO_x was measured as a negative value. This seems to be a measurement error. One solution would be to exclude these records. Another approach could the winsorization, i.e. confine extreme and nonsensical values to the marginal reasonable value, in this case, zero.

5.7 More Data

The flow detector data that we used to build this model was only available for a period of fewer than four months. With a longer time frame of available data, the model would most likely perform better and generate more accurate predictions. This would help especially when it comes to the prediction of different seasonal conditions!

5.8 Hierarchical/Conditional Modelling

We included a site-specific term into our model by creating one-hot-encoded site features. However, this only allows for limited differences between the sites as the general model term is unlikely to change much between sites. We think that a hierarchical modeling approach, site-specific modeling, or a mixed effect model may all be promising approaches to explore to both get maximum information from the data while allowing for local variability between the model terms in each site.

6 Participant profiles:

Carlos Mougan Facilitator

Carlos is currently a research student in two research institutes in Barcelona, Spain: Artificial Intelligence Institute of the Spanish National Research Council (IIIA-CSIC) and Barcelona Supercomputing Center.

Qian Fu Post-doctorial researcher

Qian is currently a post-doc research fellow at the Birmingham Centre for Railway Research and Education of the University of Birmingham. His work focuses on the development and use of advanced methods for data integration and analytics within the rail industry as well as the wider multimodal transport.

Jojeena Kolath MSc Big Data

Jojeena is an MSc. student of Big Data at the University of Stirling, with a background in Electronics and Communication Engineering. She is currently completing her MSc. dissertation in Natural Language Processing of online hate speech with Trilateral Research Ltd., London.

Huan Tong PhD Candidate Huan Tong is a PhD student from Institute for Environment Design and Engineering, The Bartlett, University College London. The aim of her PhD research is to investigate the impact of urban sound environment on human health and behaviours, based on crowd-sourced data. Currently, her research examines relationships between noise complaints and urban morphology.

Siddharth Dixit Participant

Sid is an undergraduate researcher currently majoring in Mathematics and specializing in AI at SNU, India. He has a passion for applying Machine Learning to diverse areas and its underlying mathematics. Collaborating with the University of Luxembourg in the past, he has to lead the ML part for projects involving the discovery of new materials for Thermoelectric applications using AI.

Laurens Geffert Participant

Laurens is a Data Science Manager at the Nielsen Marketing Cloud where he and his team build predictive machine learning models for applications in programmatic advertising. Originally studying Biology in Germany, he moved to the UK for a position in bioinformatics and conservation at UNEP-WCMC. He went on to complete a Ph.D. in Geography at the University of Cambridge and then transitioned to the Data Science industry in London through the "Science to Data Science" program.

Hyesop Shin Ph.D. candidate

Hyesop is a Ph.D. candidate in Geography, University of Cambridge. His Ph.D. research examines the potential health vulnerability of Seoul citizen's exposure to air pollution based on agent-based modeling (ABM).

Ahmad Abd Rabuh PhD candidate

Ahmad is a PhD candidate in his last year researching on applying machine learning models on financial time series data. Also, he is an intern at Experian Data Labs UK where he is currently working on a project that predicts life events from financial transactions and the patterns associated with those.

Caitlin Robinson Post-doc

Cait is a post-doc at Newcastle University in the Centre for Urban and Regional Development Studies (CURDS) and the Spatial Analytics and Modelling research group (SAM). She is currently working on the Turing Institute-funded project Spatial Inequality in the Smart City. Cait is a

quantitative human geographer, interested in using spatial analysis to understand the geographies of inequality.

Ella M. Gale Principal Investigator

Ella is a data scientist and machine learning expert in the school of Chemistry at the University of Bristol.

turing.ac.uk
@turinginst