Mining and Analysis of Air Quality Data to Aid Climate Change

Lakshmi Babu Saheer(✉), Mohamed Shahawy, and Javad Zarrin

Anglia Ruskin University, Cambridge, UK
{lakshmi.babu-saheer,javad.zarrin}@aru.ac.uk,
mohamed.shahawy@student.aru.ac.uk

Abstract. The data science and AI community has gathered around the world to support tackling the climate change problem in different domains. This research aims to work on the air quality through emissions and pollutant concentration data along with vegetation information. Authorities especially in urban cities like London have been very vigilant in monitoring these different aspects of air quality and reliable sources of big data are available in this domain. This study aims to mine and collate this information spread all over the place in different formats into usable knowledge base on which further data analysis and powerful Machine Learning approaches can be built to extract strong evidences useful in building better policies around climate change.

Keywords: Data mining and analysis · Data pre-processing · Air quality · Climate change · Urban planning and machine learning · Geographic information systems

1 Introduction

Climate change is the main challenge that humanity is facing today, threatening the existence of life on earth. Awareness campaigns and drastic steps towards bringing the situation under control have been initiated in every nook and corner of the world. Both developed and under developed countries are working hard to tackle this problem. Data Analysis, Data Science and Artificial Intelligence have big potential to help mankind where ever (big) data is available to help build models and make predictions or provide prescriptive solutions. Such models for emissions, resources, energy consumption, etc., have already been statistically worked out by specialist groups around the world to tackle climate change [10]. ClimateChangeAI [8] is a classic example of such an initiative.

1.1 Motivation

There are sources of big data available in the fields of energy consumption, transport, building & cities, carbon footprint, farming, climate change and prediction etc [21]. It is really hard to mine all this data in a reasonable time to get useful

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020 Workshops, IFIP AICT 585, pp. 232–243, 2020.
https://doi.org/10.1007/978-3-030-49190-1_21

resources with mathematical models. The approach for our research would be to first identify the most useful and informative data for the identified topics of climate change and start to build machine learning models and methodologies to extract useful and relevant information in the form of predictions or prescriptions. Even though a lot of data is available for every field, it is very difficult to gather the majority of these information in a structured useful format.

1.2 Research Objective

The research presented in this work is to look at the parallel data on pollutant concentrations, road traffic emissions and vegetation specifically around city like London. "Parallel" in this scenario refers to different datasets collected in close timelines and physical locations. London is chosen as the city to look at mainly because there is a lot of publicly available data for this city from various sources. This can directly help us plan our cities and traffic routes or even come up with laws to keep our carbon footprint under control. Further, our hypothesis (evidenced by initial experiments and literature) is that one of the major factors affecting the concentration of pollutants in atmosphere is the vegetation. This research aims to find traffic and air quality or emissions monitoring datasets along with the information of vegetation around the area to analyse the importance and effectiveness of vegetation in urban planning.

2 Background

There have been several studies to understand the potential benefits of vegetation in mitigating urban air pollution problems [4,11,19]. Bealey et al. [3] states that trees have been widely quoted as effective scavengers of both gaseous and particulate pollutants from the atmosphere. Further, effect of tree planting strategies on dispersion and deposition of airborne urban aerosol concentrations onto woodland is suggested to be considered in the planning process. Several studies have suggested the importance of including the vegetation in urban road planning [1,2]. It is even suggested that vegetation in urban settings can provide benefits beyond improvements in air quality—these include carbon sequestration, temperature and storm water regulation, noise reduction, aesthetic improvements, and opportunities for physical exercise and the experience of nature.

Most of these research use expert knowledge and postulations to predict these impacts. In this modern age, scientists should work on big data and deductions from the proofs evidenced from data. To this end, this research will collate data from different sources dispersed all over internet regarding emissions (like $CO_2, NO_x, PM_{2.5}, PM_{10}$), concentration of gaseous and particulate matter (like $NO_2, NO_x, PM_{2.5}, PM_{10}$), road transport, and vegetation details into useful parallel information base. This pool of information could help us model different Machine learning systems to understand the relations between and impact of each of these factors on the overall quality of air. The main aim of this research is finding the links between vegetation, transport and other factors on pollutant

concentration. This should help us validate the earlier suggestions on urban vegetation planning. Further, this project should help in building a proposal system to the local authorities on the type and form of vegetation to be planted around our cities to help control the effects of emissions.

There has been other efforts in the area of data analysis of air quality using advanced Machine Learning techniques for data imputation [14]. Junninena et al. [14] looks at only recreating missing pollutant values in a specific air quality dataset. The other studies mentioned earlier [3,4,11,19] approach the problem as a data collection exercise as a validation of their theoretical hypothesis. It seems like our research is the first systematic effort combining the different aspects like emissions, concentration and vegetation by collating existing big data collected from different sources to perform a detailed data analysis. This in turn leads to several challenges in data pre-processing as discussed in Sect. 3.

2.1 Data Mining Approach

The first steps in this work is to identify parallel data from different sources for road transport information including number and type of vehicles, emissions in the air with different types of pollutants, pollutant concentrations, vegetation type and extend. Parallel in this scenario, as mentioned earlier, refers to both same geographical location and similar or same time scales. The idea is to build such an open source parallel database with different levels of geographic and timescale granularity. The geographic information could be as coarse as starting from big cities, to different regions in a city (boroughs), to finally narrowing down to road level details. Similarly, time scales could vary from seasons (especially where the difference would be evident with leaf shedding in trees) to peak traffic timings in the day. Time scale granularity is left as a future exercise due to unavailability of reliable resources. For the sake of simplicity, the location granularity is also taken to be at a region or borough level. Hence, the task is to find the parallel data for air pollutants in terms of concentration and emission along with data on vegetation at borough level. Later on, this information could be made more granular in terms of location and time.

This huge knowledge base once ready is expected to provide a wealth of information for the research community to study the interactions between these factors and how these learning can be combined to form useful environmental policies. Some of the main challenges in this work is mainly around finding parallel compatible information including the formats of the information is discussed in detail in the next section. The main hypothesis in this work is that Vegetation has a positive effect in reducing the concentration of the pollutants despite the emission rates. The aim of this research is to prove if the data available can support this hypothesis.

3 Research Challenges

As mentioned earlier, the UK city of London is chosen as the first city to target our study merely due to the availability of good quality data from various reliable

sources. For this particular study, different public domain datasets available through government and public authorities are considered. The traffic monitoring data in UK is available with different authorities like: Cambridgeshire county council: [5], Highways England data: [13] and Traffic for London: [25].

Similarly, air quality datasets in the form of emissions and pollutant concentration monitoring data in London is available with authorities like: London Air: [16], UK Government Monitoring: [9,17], Traffic for London: [25]. Particularly for London, the local authorities have also kept track record of Vegetation information around London [18].

The challenge would be to map these different sources to come up with parallel data to extract useful information through machine learning approaches like SVMs or Neural Networks [12,15,20,26]. Traffic for London seems to have usable parallel data for traffic and emissions and is known to publicly share this information to support research. This again validates London to be a good starting point for this research, later on extending to other cities or more resources for further information.

There are several challenges even with these available datasets for London. The data available is only partially aligned with varied features between the measured concentrations and emissions. The Table 1 clearly shows the misalignment between the concentrations or emissions features, which limits the correlation analysis and the amount of useful information that could be inferred from the data. Next section presents how some of these challenges are dealt with in this study in order to perform further data analysis.

3.1 Technical Challenges

The London Atmospheric Emissions [17] dataset, collected by the Greater London (GLA) and Transport for London [24], is chosen as the primary source of data in the following analysis. While the dataset provides a vital and invaluable source of geographic information, it also has its drawbacks regarding formatting and data-types. The varied types of geographic references (as shown in the Table 1) pose an issue with both the granularity of the entire analysis (limited to London boroughs at the highest-level), as well as finding means of converting the reference types for uniformity purposes.

The first challenge was to find the data related to each borough in London. Due to the lack of support for Ordinance Survey National Grid coordinates converters, especially all-numeric grid references, a supplementary OSGB36 or WGS84 python library (PyBNG) was developed and open-sourced on GitHub and PyPi [23] as an outcome of this research. Using PyBNG, OSGB36 coordinates were converted to latitude and longitude, which can be used to find the corresponding London boroughs. The large volume of the concentration dataset, however, poses an issue with the London borough-search using APIs. Nevertheless, a GeoJSON dataset (open-sourced by Ordnance Survey) containing geographic polygonal boundaries of all the London boroughs could be used to find the encapsulated coordinates and ultimately generate the required missing data.

Table 1. Dataset format details

Data	File-type	Geographic reference	Available data	Unit
Transportation	XLSB (multi-sheet)	London borough-level, main motorways, ArcGIS project	Emissions/ vehicle type	Tonnes/ year
Concentration	CSV	All-numberic, OSGB36 coordinates	NO_2, NO_x, PM_{10} and $PM_{2.5}$	μ g/m^3
Emission	XLSB (multi-sheet)	London borough-level, ArcGIS project	CO_2, NO_x, PM_{10} and $PM_{2.5}$	Tonnes/ year
Trees	CSV	London borough-level, OSGB36 coordinates, WGS84 coordinates	Species name	N/A

Fig. 1. Greater London Boroughs' geo-spatial boundaries

In order to synthesize or validate the *boroughs* feature in some of the datasets mentioned, the ray-cast algorithm was used. This algorithm is used to determine whether or not the corresponding longitude & latitude coordinates are encapsulated within the boundaries of a geo-polygon of the corresponding London borough (Fig. 1).

Another challenge was to check for wrong or inconsistent data. Whilst verifying some of the trees dataset, it became apparent that the boroughs were incorrectly entered; the coordinates given for some rows do not correspond to the boroughs provided. To validate the entries, PyBNG was used to convert easting/northing to latitude and longitude coordinates, which were then used to infer and ultimately validate the borough for every entry. Another hurdle encountered was the varying invalid entries like the tree species names ("Failed Planting Site", "Unknown Conifer", "Zz Tree Missing", etc). Manually filtering the valid species names was a painstaking task given the fact that there are

thousands of different species in the dataset. Alternative techniques for filtering using crawlers may be pursued in future to expand this data.

Sample output dataset after pre-processing and collation is shown in Table 2. Given that the data pre-processing and aggregation proved to be a challenging task, the finalized dataset will be released publicly in the near future to assist with other potential climate change solutions.

Table 2. Processed dataset sample

Borough names	Tree count numbers	Road emissions (tonnes/yr)				Concentrations ($\mu g/m^3$)			
		CO_2	NO_x	$PM_{2.5}$	PM_{10}	NO_2	NO_x	$PM_{2.5}$	PM_{10}
City	1361	52817.85	220.46	9.33	15.57	54.04	126.96	16.02	27.41
Ealing	42082	285274.47	934.35	51.08	95.17	36.35	65.51	13.17	21.79
...	...	...	...	...	...	...	...	...	

4 Data Analysis

The data-analysis process heavily relied on the aggregation of the datasets given the high volume of features. The goal was to extract the useful trends out of thousands of features whilst minimizing the loss of data and maintaining uniformity of geographic level. This was achieved by pre-processing subsets of data individually and guaranteeing uniformity prior to collating all the subsets together (essentially a divide-and-conquer data pre-processing approach). A more abstract level of the data was ensured in the end to be consistent.

The pollutants monitored for air quality can be categorised into two classes - gaseous (CO_2, SO_2, NO_x, NO_2) and particulate matter ($PM_{10}, PM_{2.5}$). It could be postulated that the trees help absorbing only gaseous pollutants. But, there have been reports on certain types or species of trees that could help absorb the particulate matter as well [3,6]. Data analysis is performed separately on the effects on gaseous and particulate matter. Unfortunately, this study has not yet looked into the species of trees in the dataset. This may be included in the future extensions of this research. Again, the particulate matter may seem to have more effect on the general health rather than climate change. But, it is known that the particulate matter has fractions of elementary carbon [7] which results in global warming and hence, affecting climate change directly. These types of pollutants also need to be included in this study.

Another limitation with the current emissions and pollutant concentration datasets is that the type of pollutants monitored are not the same. The emissions data has CO_2, NO_x, $PM_{2.5}$ and PM_{10} and the concentration data has NO_x, NO_2, $PM_{2.5}$ and PM_{10}. It is unfortunate that there is only NO_x, $PM_{2.5}$ and PM_{10} that are aligned and could be studied in parallel. Efforts are underway to

gather more information from the authorities on other pollutants like CO_2 monitored as pollutant concentration. At this stage of the research, focus is only on these 3 pollutants to study the direct relations between emissions and concentration. It has to be also noted that there is difference in metrics of representation for concentration as $\mu g/m^3$ and emissions as tonnes/year. This time granularity for the data is annual values and hence concentrations are aggregated to mean concentrations per year as well.

Once the parallel data for emissions, pollutant concentration and vegetation (as number of trees) at borough level is obtained, the next step is to look for correlations between these features. The correlations between these three features are first plotted as graphs and heatmap. A negative correlation is expected between the number of trees and pollutant concentration. The values in the graph should be representing each borough in London. A closer look at the number of trees against mean values of concentrations of gaseous and particulate matter distributions for each borough could be plotted separately. Finally, number of trees, emissions and concentrations should be looked at separately for the three common pollutants viz., $NO_x, PM_{2.5}$ and PM_{10}. Next section visualizes and analyzes these plots in detail.

5 Results and Discussion

It is quite evident even from the initial analysis that the existence of trees has a negative correlation with the mean concentrations of both gaseous and particulate matter as shown in Fig. 2. Figure 3 shows a heatmap of correlation between the emissions, concentration and trees. As expected, there is a negative correlation between trees and concentrations. Also, there is a negative correlation between emissions and concentration which could be the effect of trees. This could be further investigated by looking at the boroughs with lower tree counts.

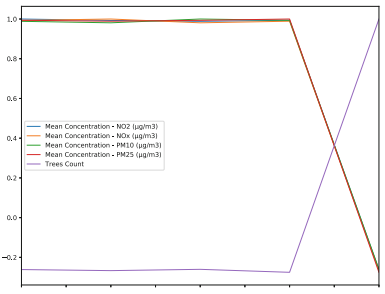

Fig. 2. Trees-concentration correlations

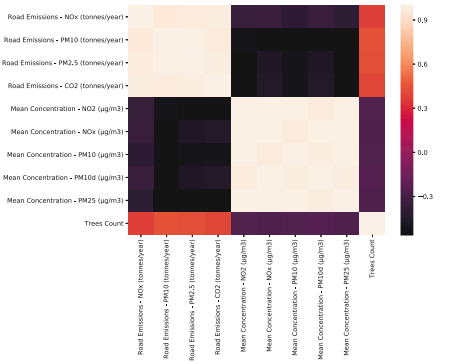

Fig. 3. Correlation: emissions, concentration and trees

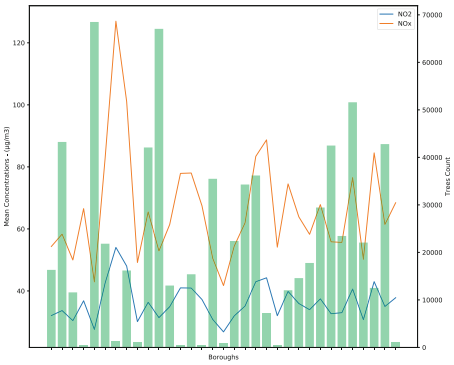

Fig. 4. Trees: effect on mean gas concentrations

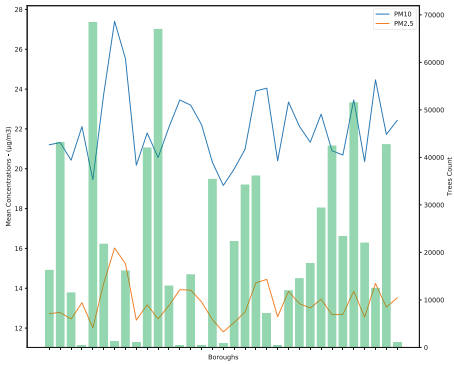

Fig. 5. Trees: effect on mean PM concentrations

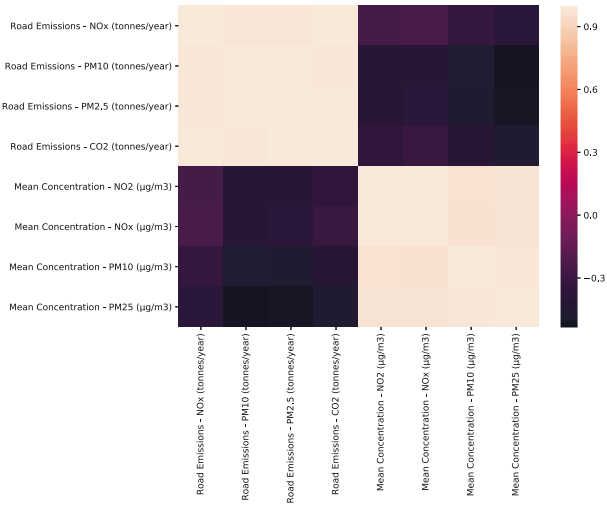

Fig. 6. Trees > 20000: correlation of emissions vs concentrations

Looking at each borough in detail, Fig. 5 shows the relation of number of trees with the particulate matter concentrations. Similarly, Fig. 4 shows the effect of trees on gaseous pollutants. It can be observed that most of the boroughs with a low volume of vegetation have relatively higher pollutant concentration means, while those with a higher volume of vegetation have low pollutant concentration. Each point or bar in the graph represents a borough. The difference in y-axis scales of these two graphs needs to be noted which suggests that there is lower variation in the particulate matter and more variation in the gaseous matter. There is no doubt that all types of trees would help in CO_2 absorption. Not all trees are particularly known to absorb nitrous oxide which could explain some discrepancies here, again, suggesting a deeper look into the species of the trees.

Fig. 7. Comparison: emission with concentration and trees for NO_x, $PM_{2.5}$ and PM_{10}

A statistical concern that arose was the potential collinearity (negative correlation) between the emissions and the number of trees per borough. The initial hypothesis can be contested by claiming that the concentrations measured are lower than expected due to low emissions in the area rather than it being the vegetation's effect. Separate plots for each of the three parallel pollutant in both datasets have been presented in Fig. 7. These figures further prove the hypothesis that concentration is low with even high emissions where there are more trees. It needs to be clarified that the emissions in the dataset includes only the transport or road emissions and hence in some cases, it is observed that concentrations are much higher than emissions. This again is a future work to look into. For the sake of sanity, if we look at boroughs with more than 20000 trees, it is seen that the there is a strong negative correlation between concentration and emissions as shown in Fig. 6. Hence the results of our data analysis supports and proves our hypothesis.

6 Conclusion

More and more climate data from different sources are being collected now than ever before. Analysing such big data can offer unprecedented opportunities to create innovative climate solutions and influence decision-making on a global level. This research initiates the study of influence of AI on tackling climate change with regards to air quality data. The various sources of big data that could contribute to a knowledge base facilitating a detailed data analysis (prior to applying Machine Learning algorithms) are being collated. The challenges faced during this process is briefly outlined to gauge the efforts needed in this research. The initial aim was to build borough level information for London city in order to perform data analysis. The initial analysis of this parallel data at borough level confirms the hypothesis that the vegetation is negatively correlated to the pollutant concentration. This motivates further detailed analysis on this topic to look at better granularity in time and location and impact of other factors like transport emissions, types of trees, weather etc.

7 Future Work

As mentioned in the earlier sections, there are a number of details to be figured out in the existing datasets extending the quality, quantity and granularity of information. This paves way for further data mining and data processing. Examples of these extensions include looking at tree species, more types of pollutants, more granularity of information in location and time, etc. Authorities like TFL will be contacted along with searching for other (online) resources. Finally, the research is intended to progress as a data gathering project to acquire good quality first hand information to implement the machine learning approaches derived from the initial studies.

To further eliminate any potential collinearity that might be biasing the measured concentrations and hence disproving the deductions made, other sources of

emissions must be taken into account. Seeing as road transportation constitute roughly 25% to 35% of the total emissions in the Greater London region [17] and the concentrations of pollutants are measured regardless of the source, the total emissions produced should be considered.

Using AI for climate change control through efforts like the traffic-emissions control could be a major step forward to build upon strategies for tackling climate change based on big data available in various fields to build a sustainable future. Once this big data knowledge base is available, different data analysis techniques could be worked on to understand the correlations, Machine learning techniques like standard regression models (linear and non-linear) have already been used to predict the air quality [27]. Another review [22] shows that air quality estimation problems tend to implement Ensemble Learning and Regressions, whereas forecasting makes use of Neural Networks and Support Vector Machines. Since different factors are at play, it might be worth identifying correlated information using unsupervised clustering algorithms and later on use expert systems with some Bayesian prior probabilities to take into consideration several related factors.

Due to the multi-aspect nature of the problem and its internal dynamism, Multi-Perspective Machine Learning (MPML) and Classifier Ensemble can also be considered to investigate various parallel aspects and co-factors involved in the scenario. Furthermore, it is possible to enrich weak datasets through leveraging open source data and using automated approaches for harvesting and integration of publicly available online data.

References

1. Al-Dabbous, A.N., Kumar, P.: The influence of roadside vegetation barriers on airborne nanoparticles and pedestrians exposure under varying wind conditions. Atmos. Environ. **90**, 113–124 (2014)
2. Baldauf, R., et al.: Integrating vegetation and green infrastructure into sustainable transportation planning. Transp. News **288**(5), 14–18 (2013)
3. Bealey, W., et al.: Estimating the reduction of urban PM10 concentrations by trees within an environmental information system for planners. J. Environ. Manag. **85**(1), 44–58 (2007)
4. Benjamin, M.T., Winer, A.M.: Estimating the ozone-forming potential of urban trees and shrubs. Atmos. Environ. **32**(1), 53–68 (1998). Conference on the Benefits of the Urban Forest
5. Cambridge County Council (2019). https://www.cambridgeshire.gov.uk/residents/travel-roads-and-parking/roads-and-pathways/road-traffic-data/
6. Chen, L., Liu, C., Zhang, L., Zou, R., Zhang, Z.: Variation in tree species ability to capture and retain airborne fine particulate matter. Sci. Rep. **7**(1), 3206 (2017)
7. Chernyshev, V., et al.: Morphological and chemical composition of particulate matter in buses exhaust. Toxicol. Rep. **6**, 120–125 (2019)
8. Climate Change AI (2019). https://www.climatechange.ai/
9. Defra (2019). https://uk-air.defra.gov.uk/data/
10. Energy models at the UCL Energy Institute (2019). https://www.ucl.ac.uk/energy-models/

11. Fares, S., et al.: Particle deposition in a peri-urban Mediterranean forest. Environ. Pollut. **218**, 1278–1286 (2016)
12. Gastaldi, M., Rossi, R., Gecchele, G., Della Lucia, L.: Annual average daily traffic estimation from seasonal traffic counts. Procedia-Soc. Behav. Sci. **87**, 279–291 (2013)
13. Highway England (2019). http://webtris.highwaysengland.co.uk/
14. Junninena, H., Niskaa, H., Tuppurainenc, K., Ruuskanena, J., Kolehmainena, M.: Methods for imputation of missing values in air quality data sets. Atmos. Environ. **38**(18), 2895–2907 (2004)
15. Krile, R., Todt, F., Schroeder, J., Jessberger, S.: Assessing roadway traffic count duration and frequency impacts on annual average daily traffic estimation: assessing accuracy issues related to annual factoring. Technical report, United States. Federal Highway Administration (2016)
16. London Air Quality Network (2019). http://www.londonair.org.uk/LondonAir/
17. London Atmospheric Emissions (LAEI) (2016). https://data.london.gov.uk/dataset/london-atmospheric-emissions-inventory-laei-2016
18. London Local Authority Maintained Trees (2019). https://data.london.gov.uk/dataset/local-authority-maintained-trees
19. Monks, P., Allan, J., Carruthers, D., Carslaw, D., Dore, C., Fuller, G.: Air quality expert group: impacts of vegetation on urban air pollution. UK Air Quality Reports (2018)
20. Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K.: A large contextual dataset for classification, detection and counting of cars with deep learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 785–800. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_48
21. Rolnick, D., et al.: Tackling climate change with machine learning. CoRR abs/1906.05433 (2019). http://arxiv.org/abs/1906.05433
22. Rybarczyk, Y., Zalakeviciute, R.: Machine learning approaches for outdoor air quality modelling: a systematic review. Appl. Sci. **8**, 2570 (2018)
23. Shahawy, M.: PyBNG (2019). https://pypi.org/project/PyBNG/
24. Traffic for London (2019). https://tfl.gov.uk/corporate/publications-and-reports/travel-in-london-reports
25. Transport for London, London Air Quality (2019). https://tfl.gov.uk/corporate/about-tfl/air-quality
26. Tsapakis, I., Schneider, W.H.: Use of support vector machines to assign short-term counts to seasonal adjustment factor groups. Transp. Res. Rec. **2527**(1), 8–17 (2015)
27. Zhu, D., Cai, C., Yang, T., Zhou, X.: A machine learning approach for air quality prediction: model regularization and optimization. Big Data Cogn. Comput. **2**, 5 (2018)