Inferring Latent Patterns in Air Quality from Urban Big Data

Suparna De, University of Winchester (s.de@ieee.org)
Usamah Jassat, University of Loughborough
Wei Wang, Xi'an Jiaotong Liverpool University, China
Charith Perera, Cardiff University
Klaus Moessner, Chemnitz University of Technology, Germany; Institute for Communication Systems, University of Surrey, UK

ABSTRACT
The emerging paradigm of urban computing aims to infer latent patterns from various aspects of a city's environment and possibly identify their hidden correlations by analyzing urban big data. This paper provides a fine-grained analysis of air quality from diverse sensor data streams retrieved from regions in the city of London. The analysis derives spatio-temporal patterns, i.e. across different location categories and time spans, and also reveals the interplay between urban phenomena such as human commuting behavior and the built environment, with the observed air quality patterns. The findings have important implications for the health of ordinary citizens and for city authorities who may formulate policies for a better environment.

INTRODUCTION
As an ever-increasing percentage of the population resides in cities (83.7% of the UK's population is urbanized, as of 2019 [1]), air pollution has become a growing major concern, in terms of both the environment and citizen's health. Since the industrial revolution, air quality in urban areas has been a huge problem for city dwellers. In the twentieth century, urban practices such as burning of coal in factories and houses resulted in heavy emission of pollutants, i.e. sulphur dioxide and black smoke (soot). More recently, reactive gases such as Nitrogen Dioxide (NO2) and fine particulate matters have emerged as the major concerns in cities. The main sources of NO2 are high temperature combustion processes, predominantly produced by vehicles such as cars and trains. Particulate matters refer to particles of diameter less than 10 µm (PM10) and 2.5 µm (PM2.5) suspended in the air and include both natural (volcanic ash and sea salt) and manmade (fuel combustion) ones.

Studies have shown that PM2.5 has strong adverse health effects, specifically, short term exposure can trigger acute cardiovascular events whilst long term exposure can increase the risk of cardiovascular mortality [2]. This is particularly relevant during the current Covid-19 disease pandemic, where a number of initial studies have explored the correlation between Covid-19 mortality and chronic exposure to air pollution. A study conducted in northern Italy [3] has noted the correlation between the high prevalence and lethality of SARS-CoV-2, the pathogenic agent of Covid-19, and the high pollution levels in the Lombardy and Emilia Romagna regions. A Harvard study [4] has concluded that a $1\mu g/m^3$ increase in PM2.5 is linked to a 15% increase in Covid-19 mortality. In the UK, a Medical Research Council (MRC) study [5] has showed that regional nitrogen oxide levels have significant associations to Covid-19 cases and deaths in England, as well as increased individual infectivity due to PM10 and PM2.5.

In this scenario, identifying long-term trends in outdoor pollution in cities can serve as a useful marker in epidemiological studies. Alongside this, surfacing temporal and spatial patterns in the observed pollutant measurements can help drive efforts to target the main polluting sources and control their production in order to improve the general air quality. In addition to identifying spatio-temporal trends, studies also need to take into account the distinct natural and built environment characteristics of the urban area, which might influence local pollution patterns, since pollution is highly location dependent [6].

Since the establishment of the world's first national air pollution monitoring network in the UK in 1961 and the subsequent setting up of networks globally, studies have looked at the temporal trends in different pollutants and suspended particulates. Air quality data has been examined on a daily basis to represent trends over and across different days and seasons in [7, 8]. However, these studies were only based upon data from single days rather than derived from a large amount of data. Long term changes in black smoke concentrations in the UK were modelled using an approximate Bayesian inference model in [9], where the researchers also investigated the effects of preferential sampling in the estimation method. Diurnal and seasonal variations in air pollutants in the Veneto region of Italy are explored in [10], by plotting the mean daily and monthly concentrations, respectively. Diurnal patterns showed average levels increasing from Mondays to Thursdays and dropping on weekends. However, the method was not able to reveal more nuanced variations in the measurements and eliminate the effect of high values in certain pollution episodes dominating the analysis.

In urban computing, inferring air quality levels has received growing research attention. A number of studies [6, 11] have applied various machine learning algorithms for short-term forecasting of pollutant levels, such as a non-linear autoregressive neural network [12] with a feedback loop featuring past pollutant values for the prediction. Inferring air quality at an arbitrary location that doesn't have a monitoring station close by is studied in [13], which uses two separate classifiers for temporal and spatial features along with co-training to infer the air quality. The main limitation of these methods is the computational complexity: when applied to large spatial datasets, processing large matrices within each iteration and convergence of parameters are extremely time consuming.

We need an appropriate technique that can handle the requirements of manipulating large matrices, resulting from large datasets derived from a number of monitoring stations covering possibly long time spans. Nonnegative Matrix Factorization (NMF) is a powerful technique that can produce clear and interpretable representations of latent patterns. Non-negativity of factors is built in NMF, allowing the decomposition of a large matrix as a positive combination of multiple, smaller ones. It is naturally applicable to the air quality dataset as it doesn't have negative pollution values; thus, we apply NMF to uncover the latent spatio-temporal patterns across different areas of an urban region and time spans, i.e. diurnal/weekly/seasonal. The analysis focusses on the city of London, which is characterized in terms of roadside, urban, suburban, and industrial areas. The derived results reveal some prominent and consistent patterns in these different areas.

CHALLENGES

Long-term monitoring and trend analysis of outdoor pollution encompasses a number of challenges, arising from the specific characteristics of the monitoring networks and the resultant datasets:

Strong locality concerns – decisions regarding the siting of monitoring stations consider not only a primary location that is representative of the pollution experienced by the local community, but also a second location sited far from pollution sources in order to obtain background pollution levels in the larger area [9]. Therefore, the spatial effects of the monitoring station location on the recorded observations cannot be assumed to be stationary spatial processes, with the actual location also determining the air quality trends.

Preferential sampling – as the monitoring network evolves with the closure or opening of some sites, the possibility of selection bias arises, with the remaining stations maintained in order to adhere to policies or directives. This can lead to preferential sampling, i.e. dependency between the monitoring site location and the recorded pollutant concentrations. This needs to be taken into account in the analysis.

Data sparsity – the high installation and maintenance costs have led to shrinking numbers of monitoring stations. Moreover, with the variance in the monitoring methods, (e.g. hourly data communicated through web platforms, or, diffusion tubes left at a site and analyzed at monthly intervals to give the average monthly pollutant concentration), and locality characteristics of the site (e.g. residential or industrial location, roadside or background monitoring station) also mean that stations vary in terms of pollutants measured, the frequency of measurements and the reporting of observations. This could result in a sparse and inconsistent dataset with missing labels [12], requiring data collection and normalization over an extended time period for inferring trends.

Diurnal/weekly/seasonal variations – inferring latent patterns that are specific to the natural and built environment of a city requires that the sampling method covers a number of sites from different locality types and extends over a large enough time span to obtain enough data on the spatial and temporal variations of pollutant concentrations.

METHODOLOGY

We present here the characteristics of the collected dataset and the NMF algorithm to infer latent patterns in pollutant concentrations. We also discuss how the challenges identified above are addressed.

Dataset

Figure 1 shows the monitoring sites on a map. Sites are identified by their initials and their category. They were selected according to the following specification: i) data availability to cover the three-year period of 2015-2018, to address the data sparsity challenge, ii) sites to be representative of varying pollution climate scenarios, i.e. city-scale pollution, regional background, traffic or industrial hotspots, to address locality and preferential sampling concerns.

In particular, data was collected from 4 types of sites: roadside, kerbside, urban and suburban. Roadside and kerbside monitoring sites are situated close to roads and

representative of traffic hotspots or congestions, with the kerbside station within 1m and roadside stations being between 1–5 m away from the road. Urban sites represent citywide levels of pollution. Suburban monitoring sites are located away from main roads and usually close to parks, i.e. not influenced by road traffic, and thus representative of background pollution. Two common pollutant species, NO_2 and PM_{10} are consistently recorded across the various sites and are used in this analysis.

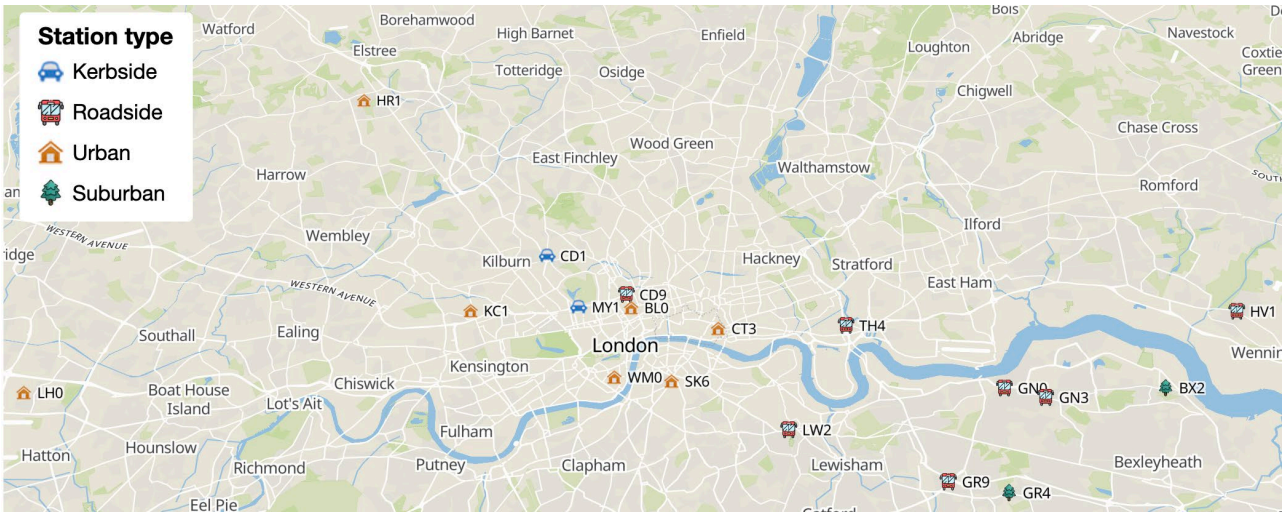


Figure 1. Map showing the location of monitoring sites in the London air quality dataset

Nonnegative Matrix Factorization

The difference between NMF and other techniques, such as vector quantization and principal component analysis (PCA), is that NMF only works on non-negative matrices, with non-negativity of both loading matrices and scores built-in, and thus, being intuitively applicable to datasets with all positive values. In contrast to factor analysis, where insufficient information is derived from the covariance matrix and PCA, which relies on information from the correlation matrix, NMF employs a point-by-point least squares minimization method. This enables direct comparison to the input matrix without transformation.

NMF decomposes an input matrix *V* of dimension n rows and m columns, into two non-negative matrices, a feature matrix *W* ($n \times k$) and a coefficient matrix *H* ($k \times m$) [14], which can be used to approximate V, i.e. $V \approx WH$, where k is the number of factors extracted.

The product of *W* and *H* can explain the systematic variations in *V*. The decomposition of *V* as the product of two matrices is similar to that in PCA, which however, focusses on explaining the non-weighted sum of the squares of the residuals. High concentrations, corresponding to anomalous pollution levels, can dominate such analysis, while ignoring low values. In contrast, NMF places emphasis on the information from all samples, by weighting the residual squares with the reciprocals of the squares of the standard deviations of the data values [15]. As pollutants with high values will have larger absolute standard deviations, their weight coefficients will be smaller than in unweighted models such as PCA. Finding the values of the matrices W and H is done by solving the following optimization problem that minimizes the Frobenius norm of the difference between V and WH, i.e.

$0.5 \times \|V - WH\|_F^2$, while requiring that all components of W (or H) should be non-negative and where the Frobenius norm is the sum of element-wise squared errors.

1. Input: filename.csv
2. Clip time series values < 0;
3. *numDays* = length (time series) / 24;
4. Initialize *V* with all Zeros with size: 24 × *numDays* ;
5. for *day* from 0 to *numDays* do
6. $V_{1,2...24,day}$ = timeseries[day*24 to (day + 1)*24] ;
7. if $V_{1,2...24,day}$ contains NaN value then
8. Delete $V_{1,2...24,day}$;
9. end
10. end
11. *W*, *H* = NMF(*V*);
12. plot *W* × mean(*H*);

Algorithm 1: Generation of pollutant trend plots.

Data was retrieved from the London Air Quality Network (LAQN) HTTP API (http://api.erg.kcl.ac.uk/AirQuality/help), with the selected interface accepting the monitoring station, species, start date, and end date as inputs and returning the data in a csv file format.

The retrieved csv file has hourly entries for the entire duration selected for the location, with separate series for each location and subsequently each pollutant. As shown in Algorithm 1, the csv file is read in using the Pandas Python library (line 1). Pre-processing of the dataset includes checking for anomalous and missing values. For this, all negative data values (below 0) were marked as invalid and removed (line 2). The data is then arranged into a data matrix *V* of dimension $24 \times M$ where *M* is the total number of days in the dataset (corresponding to the 3 years of data collection: 2015-2018) (lines 3-4). The input matrix V is then arranged into a time series format and checked for missing values (lines 5-9). The csv file may have missing entries as empty cells, which become NaN entries when it is opened in Python Pandas (line 7). The feature W and coefficient H matrices are obtained by applying the python sklearn library NMF implementation on the V matrix (line 11). The feature matrix contains the daily pattern for a particular location and is scaled by the average of the coefficient matrix to compare the amount of pollution across locations. This is then plotted into a chart, using the matplotlib library (line 12).

To extend this analysis to other temporal patterns, multiple matrices are created with the data of each subset, for instance, to analyze the difference between weekends and weekdays at a particular location, two matrices are created, one with data for weekdays and the other with the weekends. These matrices are then decomposed separately, with the plots drawn on the same graph to analyze differences in patterns across the days.

RESULTS AND ANALYSIS

Diurnal Patterns

Suburban Monitoring Sites

Figure 2a shows the diurnal NO_2 patterns at two suburban sites. The diurnal NO_2 trends are very similar in both locations, and higher concentrations of NO_2 at Bexley Belverde (BX2) is observed. This is most likely due to the fact that the Greenwich Eltham (GR4) station is situated in a location that has a lot more parks and fields, resulting in a lower level of general pollution in the area. There are two peaks that occur at 07:00 and 19:00, these correspond to commuting times for the general working population in London. The dip in between these times corresponds to a time when people are least likely to be travelling in suburban locations.

The diurnal pattern of PM_{10} levels (Fig. 2c) is quite different from the NO_2 trend. Although there are two peaks in the morning and evening, corresponding to commute times, these occur at different times, not only when compared to the NO_2 peaks, but also to each other. These peaks occur at 08:00 and 19:00 at BX2 monitoring station and at 09:00 and 20:00 at GR4. The difference between the NO2 and PM10 peaks indicates that PM10 pollution takes significantly longer to travel through the air than NO2 pollution.

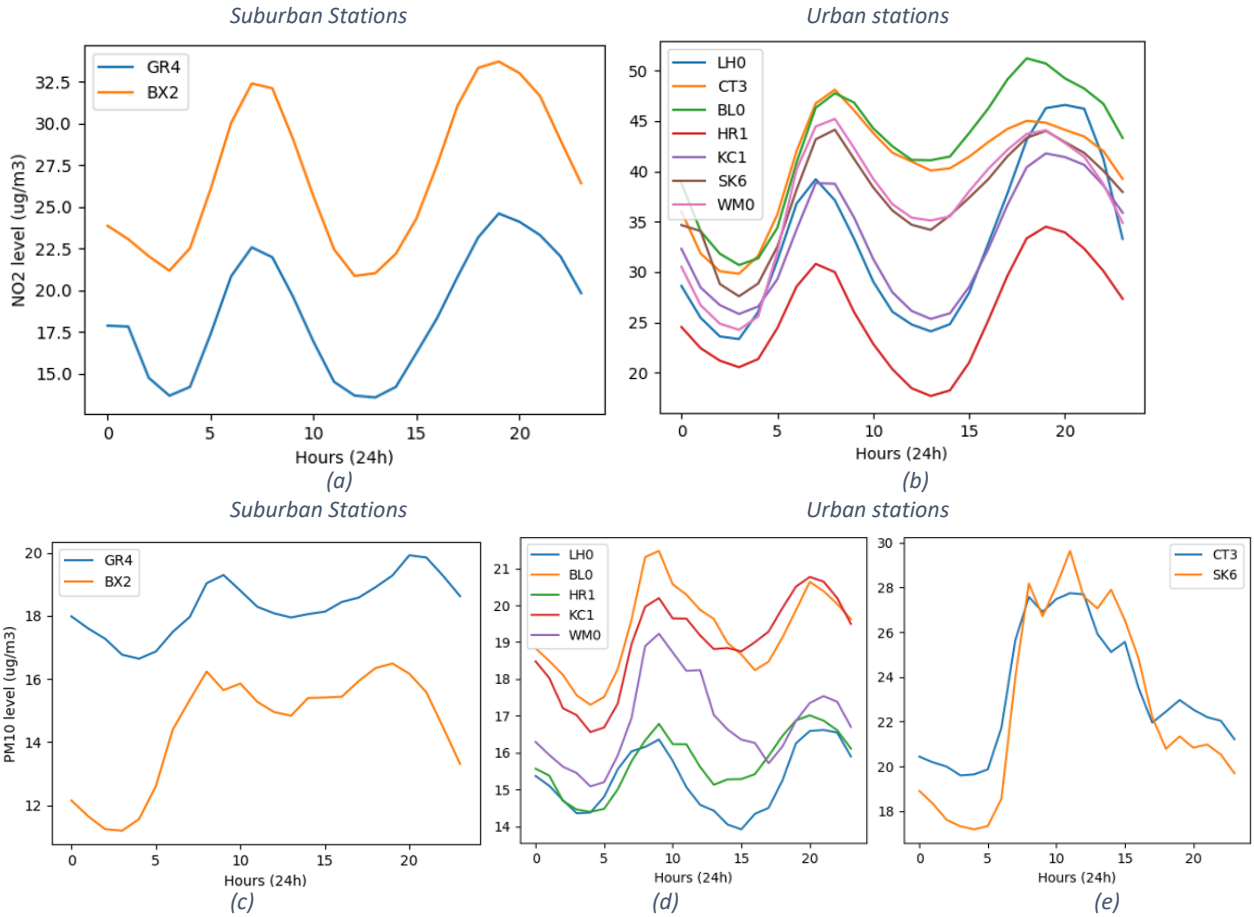


Figure 2. Diurnal variations in levels of measured pollutants at suburban and urban monitoring stations, computed over the 3-year sampling period

This is further supported when the distance from the monitoring sites to the closest main source of pollution for both stations is considered (main roads and railroads). Looking at the map of the two locations, the closest main source of pollution from the GR4 station is

0.57km away and 0.36km away for the BX2 station. These distances are reflected in the different peak times with the GR4 station showing peaks later than the BX2 station. The dips in the PM10 levels corresponding to midday and nighttime also show this same time lag. The fact that the PM10 dip at midday is smaller relative to the night time dip when compared to the NO2 pattern suggests that trains running along the nearby railroads contribute a significant amount of PM10 pollution in comparison to NO2; although the rate of trains is lower during midday, they still run on a steady schedule.

Urban Monitoring Sites

Fig. 2b shows that the daily trends of NO2 levels across all the 7 urban monitoring stations are very similar to that exhibited by suburban stations, with a similar two peak pattern, and dips during midday and at night. However, the daily trends can be distinguished as two slightly different groups: group 1 consisting of locations CT3, BL0, WM0 and SK6 that exhibit NO2 peaks later in the morning but earlier in the afternoon, and group 2 consisting of LH0, HR1, and KC1. The stations in group 1 are all located very close to or in central London, meaning people commuting from or through these locations would be arriving at them later than people would be leaving or travelling through areas further away from central London. Similar reasoning can be used for the earlier peaks in NO2 during the evening, people would be arriving at these locations earlier. This can explain the slight difference in the peak NO2 timings during the day. Another difference in the daily trends for these groups is that the fall during the day is substantially higher than the fall at night. This is likely due to these sites being in central London, where there are a lot of businesses and tourist destinations, where although there isn't as much pollution as during commuting times, there still is a lot of traffic, resulting in the values staying relatively high.

Figure 2d shows the diurnal PM10 patterns in 5 urban monitoring stations, which are similar to the diurnal suburban pattern. The patterns exhibited at the other two urban stations, SK6 and CT3, (Fig. 2e) are remarkably different (also from the NO2 trend at these locations), suggesting that there is another heavy source of PM10 pollution near the stations.

Roadside and Kerbside Monitoring Sites

Fig. 3a shows the diurnal NO_2 patterns at kerbside monitoring stations. The trends show a sharp rise in the morning and a decline in the evening, however the in-between NO_2 concentrations is different at the two stations, with one showing a steady increase and the other dipping slightly at midday, similar to the diurnal patterns seen at other site types. The MY1 Marylebone site in central London, is notorious as one of London's busiest roads, located near a main train station as well as Madame Tussauds, one of the biggest tourist attractions in London. This would explain the constant traffic across the whole day. The NO2 peak in the afternoon occurs a lot earlier than the peaks at urban and suburban stations, suggesting traffic starts to subside earlier in the center of London than areas further out.

Fig. 3c shows that the diurnal PM10 patterns at the same locations are different from the NO2 patterns, with the morning peak being higher than the afternoon peak and presence of a dip between peaks for both stations. The peaks also occur later than the NO2 peaks; all this suggests that there are different PM10 and NO2 pollution sources at these locations.

Fig. 3b and 3d show the diurnal NO2 and PM10 patterns at roadside monitoring stations. The NO2 patterns seen here are very similar to the urban and suburban ones, with the major difference being that the dip between the two peaks isn't as large; this is likely due to the site locations being close to roads with constant traffic, even during midday.

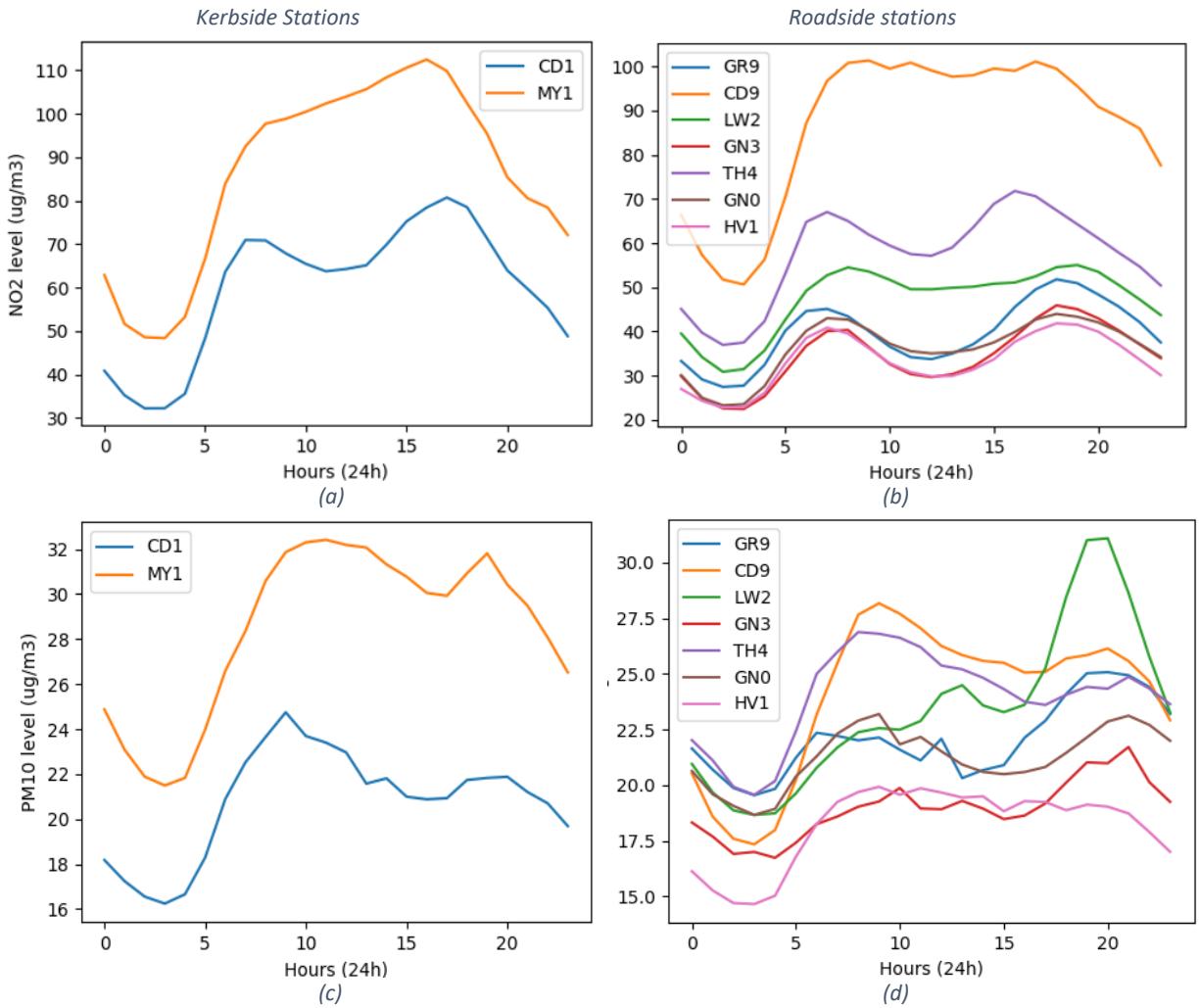


Figure 3. Diurnal variations in levels of measured pollutants at kerbside and roadside monitoring stations, computed over the 3-year sampling period

The PM10 patterns vary widely but show a similar two-peak pattern with a dip in the middle. The biggest outlier seen is at the LW2 station which exhibits an extremely high evening peak. Delving further shows that this site is located right next to a train station as well as a bus station, suggesting that either busses or trains are a major source of PM10 pollutants.

Weekly Patterns

Fig. 4a shows the difference between the weekday and weekend NO2 patterns at the two suburban locations. There are two apparent distinctions at both locations: (1) lower NO2 concentrations in general across the entire day during weekends, and (2), the difference between the morning and evening peaks is a lot larger during weekends, with the morning peak being significantly lower, suggesting less morning activity than the evenings.

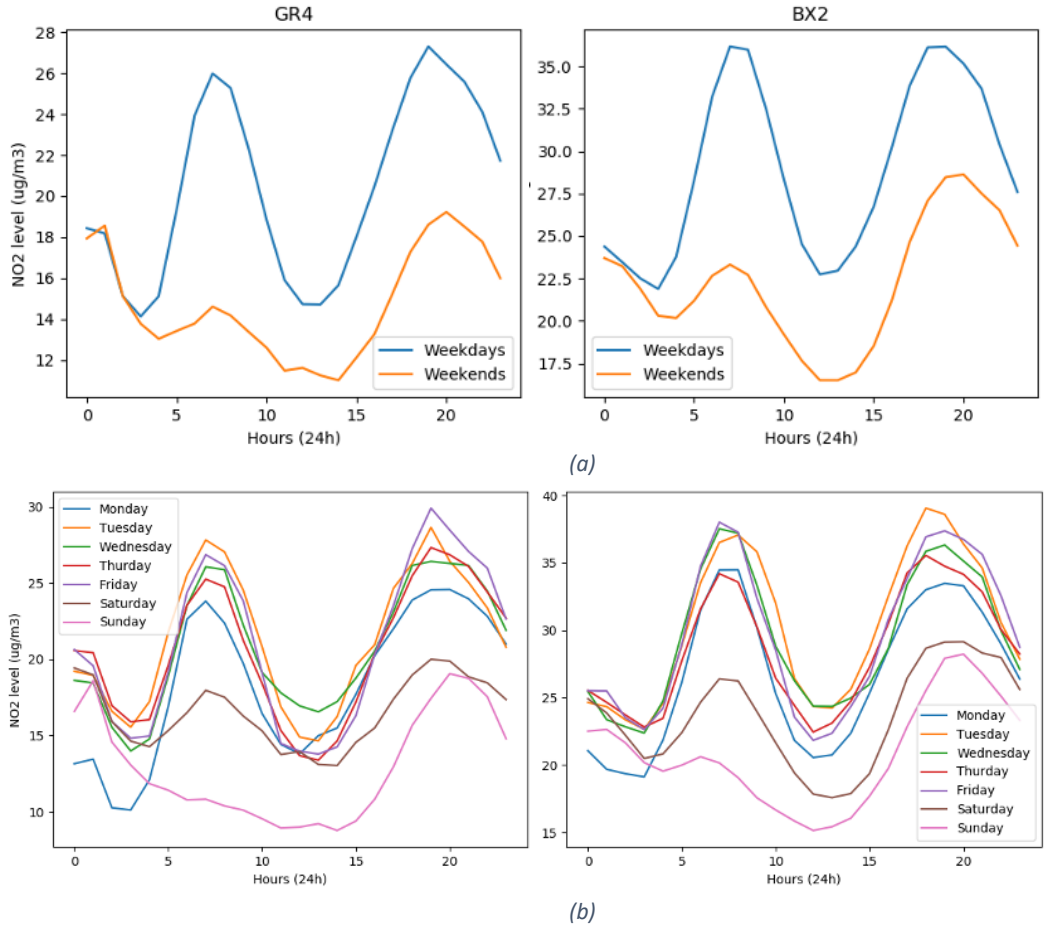


Figure 4. Difference in daily patterns of NO2 levels for a) weekends and weekdays, and b) different days of the week, at two suburban locations

The individual patterns for each day of the week (Fig. 4b) show that the difference in the two peaks mainly comes from the Sunday pattern, with the peaks being of a similar size on Saturdays. The trends also show that the NO2 levels seem to be the lowest during the early hours of Monday and Sunday evening, suggesting that this is the time that the roads are the quietest. Similar trends can be seen in the PM10 patterns.

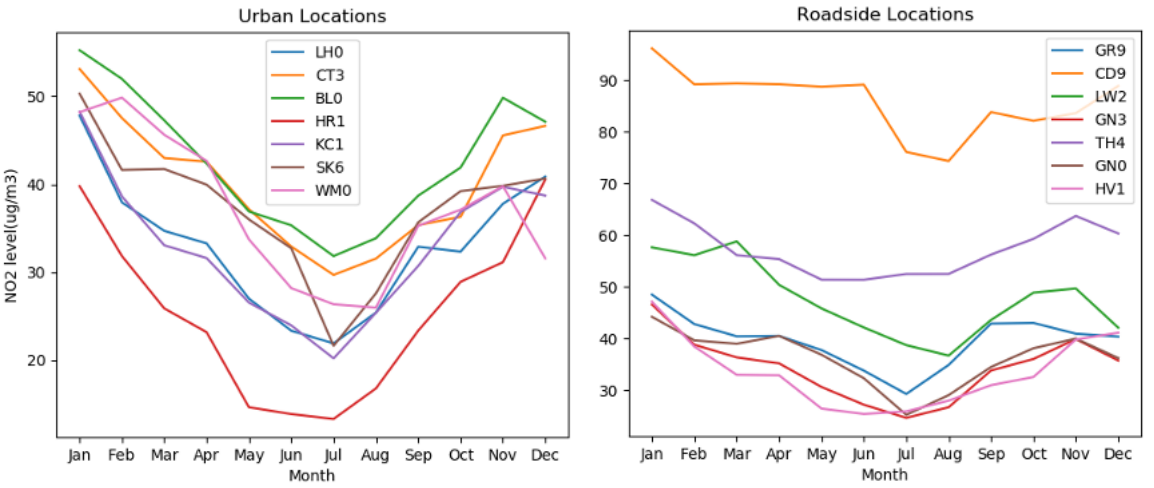
The same weekend – weekday pattern repeats across all monitoring stations and pollutants.

Seasonal Pattern

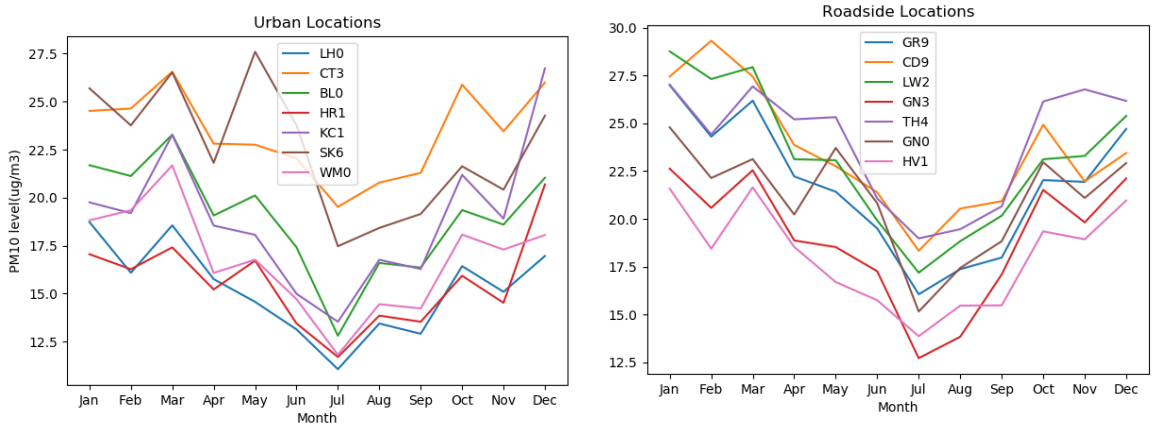
Fig. 5a shows the annual NO2 patterns for urban and roadside monitoring stations. The trends are very similar at both monitoring station types across the year, with the NO2 level peaking in the middle of winter (January) and reaching a low point in the middle of summer (July). This suggests that NO2 concentrations in the air are related to the weather at the time. Fig. 5b shows the corresponding annual PM10 patterns, which exhibit the same annual trend as NO2.

The patterns indicate that meteorological conditions associated with seasonal changes, such as temperature and wind speed, as well as the lower mixing layer heights in winter, which

may limit the dispersion of pollutants generated locally and hence play a part in the level of pollutants in the air. Increased pollutant concentrations in winter are also attributable to increased domestic heating demands. This seasonal trend agrees with patterns reported in other research works [10].



(a)



(b)

Figure 5. Change in average a) NO2 and b) PM_{10} values across the year for urban and roadside monitoring stations

Discussion of Results

Despite the range of different monitoring station types and pollutants, two main diurnal patterns were identified. The first pattern that was the most prominent was the 'two-peak' pattern seen at nearly all stations, with the only exception being a number of roadside sites. This pattern stands out with two distinct peaks, one occurring in the early morning, usually between 7-9 am and the other in the evening between 5-7 pm. These relate to the times of commuter travel within and around London. This is reinforced with how the peaks occur at slightly different times at locations depending on how close they are to the center of London. Stations that are further away from central London have morning peaks that occur earlier and evening peaks that occur later in comparison to stations that are situated in the center. This is intuitive as most commuting within London is from the more residential areas on the outskirts of London to the business and industrial center. The relationship of this pattern with commuting times and the prominent nature of stations that exhibit this

pattern suggest that transport modes are a heavy contributor to pollution within London. This has implications for those with specific air quality sensitive conditions, who should avoid travelling during these commuter times to minimize the exposure to these pollutants.

There are slight variations to the main 'two-peak' patterns, including differences in the times at which the two peaks in pollution levels occur and the size of the dip between the peaks. Some stations exhibit a large dip during the middle of the day, with pollutant levels dropping to levels similar to those seen during the middle of the night, whilst other stations and pollutant combinations saw a much smaller reduction in the pollutant levels during midday. Large midday dips occurred in the NO2 levels at suburban stations as well as at a number of urban stations, whilst the roadside stations displayed very small dips in the midday relative to the drop exhibited at night. This would suggest that vehicles and particularly cars are a significant contributor to the daily fluctuations in NO2 levels around London.

The second pattern detected features no dip in diurnal pollution, the pollutant (NO2) level stays constant or even increases over the course of the day, as seen in some of the roadside stations. The stations that exhibit this pattern are located in busy areas of London, which would explain the constantly high levels of NO2. The same is seen with the daily PM10 level for some other stations.

There was an interesting finding when comparing the patterns of different pollutants at the same station. The patterns for different pollutants can vary quite widely, this suggests that the levels of these different pollutants have different controlling factors, either in terms of their sources or influence of meteorological factors. This implies that methods for tackling different pollutants require different approaches.

The key features discovered from the analysis show that pollution levels are similar across all 5 weekdays with slight variations in the overall magnitude of the pollutant levels across each day. The weekend pattern is different, with the general daily pattern on Saturday being similar to the weekday pattern but with significantly lower overall magnitude in pollutant levels. The daily pollution pattern seen on Sundays stands out very distinctly across all analyzed stations, displaying a much lower overall magnitude than all other days and also a significantly lower pollutant levels in the morning. This resulted in a significant difference being seen between peak pollutant levels on Sundays with a morning peak being absent at some stations.

CONCLUSIONS

We applied the NMF technique to find patterns from the London air quality data, inferring prominent and consistent patterns from 4 different types of monitoring stations, namely, roadside, kerbside, urban, and suburban. Some interesting conclusions about human mobility can be drawn from the inferred patterns. The variations in the general magnitude of pollutant level for NO2 is most likely indicative of the amount of travel with cars. The reduction in general pollutant levels on the weekend in comparison to weekdays suggests that roads in London are quieter on the weekend with less people traveling. This is intuitive and understandable as many office jobs which are based in the center of London have days off on the weekend, meaning that the number of people travelling for these types of jobs

are greatly reduced. The unique pattern observed on Sundays implies that people are less active during Sunday mornings and start their day later. This agrees with the general known behavior of individuals as well as businesses, with many businesses opening later on Sundays when compared to other days.

The results from this study can have wider societal impact by triggering behavioral change, for instance, to motivate people to use public transport or car-share schemes in the push for clean air, driven by the discovery of patterns in pollution levels during peak-hours versus non-peak hours in different regions, seasonal variations and weekend/weekday patterns. Citizens with specific air quality sensitive conditions (e.g. asthma, heart and lung conditions), vulnerable groups (people with weakened immune systems, babies/toddlers and their carers) can plan their daily activities accordingly. Such insights can also form a quantitative evidence base to motivate city authorities to implement programs for cleaner air quality that are tailored to different districts within an urban region. For instance, the 'ultra-low emissions zone' (ULEZ) road pricing scheme introduced in central London in April 2019, that charges vehicles differently, based on not only their emissions but also the time of day and the road being used.

The technique proposed in this paper is easily applicable to other urban areas which have monitoring stations equipped with automatic analyzers for collecting the gaseous and particulate pollutants at hourly intervals. Such continuous monitoring sites measure pollutants in near real-time at high accuracy. Other sensing options include passive diffusion tubes for NO2 and beta gauge monitors for particulate matter.

The future study aims to extend this research by developing mechanisms to infer the environmental impact of social and cultural events (involving large-scale traffic and human movement) in public spaces in a city, as well as unpredictable events, e.g. the Covid-19 lockdown in London (March-June 2020), which saw unprecedented low human and traffic volumes. The separation between holidays/lockdown, regular weekends and weekdays can provide further insights on how the pollution develops in accordance with the change in human activities. Cross-domain data fusion techniques will be investigated to combine social network data (to detect real world city events) with pollution patterns to detect and identify correlations between events and pollution in public spaces.

REFERENCES

[1] "The World Factbook — Central Intelligence Agency." https://www.cia.gov/library/publications/the-world-factbook/fields/2212.html (accessed 03-Jan, 2019).

[2] R. D. Brook and S. Rajagopalan, "Particulate Matter Air Pollution and Atherosclerosis," *Curr. Atheroscler. Rep.,* vol. 12, no. 5, pp. 291–300, Sep. 2010.

[3] E. Conticini, B. Frediani, and D. Caro, "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?," *Environmental Pollution,* vol. 261, p. 114465, 2020/06/01/ 2020, doi: https://doi.org/10.1016/j.envpol.2020.114465.

[4] X. Wu, R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici, "Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study," *medRxiv,* p. 2020.04.05.20054502, 2020, doi: 10.1101/2020.04.05.20054502.

[5] M. Travaglio, Y. Yu, R. Popovic, L. Selley, N. S. Leal, and L. M. Martins, "Links between air pollution and COVID-19 in England," MRC Toxicology Unit, University of Cambridge, 2020.

[6] J. Y. Zhu, C. Sun, and V. O. K. Li, "An Extended Spatio-Temporal Granger Causality Model for Air Quality Estimation with Heterogeneous Urban Big Data," *IEEE Transactions on Big Data,* vol. 3, no. 3, pp. 307-319, 2017, doi: 10.1109/TBDATA.2017.2651898.

[7] F. Famoso, R. Lanzafame, P. Monforte, C. Oliveri, and P. F. Scandura, "Air Quality Data for Catania: Analysis and Investigation Casestudy 2012-2013," *Energy Procedia,* vol. 81, pp. 644–654, Dec. 2015.

[8] H. Li, H. Fan, and F. Mao, "A Visualization Approach to Air Pollution Data Exploration—A Case Study of Air Quality Index (PM2.5) in Beijing, China," *Atmosphere,* vol. 7, no. 3, 2016, Art no. 35.

[9] G. Shaddick and J. V. Zidek, "Preferential sampling in long term monitoring of air pollution: a case study," in "Technical Report #267," University of British Columbia, Department of Statistics, Technical Report #267 2012.

[10] M. Masiol, S. Squizzato, G. Formenton, R. M. Harrison, and C. Agostinelli, "Air quality across a European hotspot: Spatial gradients, seasonality, diurnal cycles and trends in the Veneto region, NE Italy," (in eng), *Sci Total Environ,* vol. 576, pp. 210-224, 2017/01// 2017, doi: 10.1016/j.scitotenv.2016.10.042.

[11] Z. Qi, T. Wang, G. Song, W. Hu, X. Li, and Z. M. Zhang, "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality," *IEEE Transactions on Knowledge and Data Engineering,* pp. 1-1, 2018, doi: 10.1109/TKDE.2018.2823740.

[12] Yuchao Zhou, Suparna De, Gideon Ewa, Charith Perera, and Klaus Moessner, "Data-Driven Air Quality Characterization for Urban Environments: A Case Study," *IEEE Access,* vol. 6, pp. 77996-78006, 2018-01-01 2018, doi: 10.1109/access.2018.2884647.

[13] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: when urban air quality inference meets big data," presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, Illinois, USA, 2013.

[14] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2000, pp. 535–541.

[15] E. Lee, C. K. Chan, and P. Paatero, "Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong," *Atmospheric Environment,* vol. 33, no. 19, pp. 3201-3212, 1999/08/01/ 1999, doi: https://doi.org/10.1016/S1352-2310(99)00113-2.