Discernment of High-Impact Urban Air Pollutants Using Unsupervised and Supervised Machine Learning Methods

Katharine Voorhees (kv871@nyu.edu)
Nathan Caplan (nbc270@nyu.edu)
Rohun Iyer (rohun.iyer@nyu.edu)
Ursula Kaczmarek (uak211@nyu.edu)

May 9, 2019

Introduction

The link between air pollution exposure and cardiovascular illness is well established (Pope et al. 2004). However, urban air pollution has a complex dynamic that is not yet fully understood. Nearly 3000 individual pollutants are of anthropogenic origin, yet only a small fraction has been subject to scientific investigation (Fenger 1999). In an effort to address the complex problem of urban air pollution, national and international bodies have developed pollution indicators that condense and simplify into a smaller set of parameters the available pollution monitoring data to make them suitable for public reporting and decision makers (Wiederkehr and Yoon 1998).

Among the widely-used air quality indicators are two classes of pollutants containing aggregates of individual pollutants: hazardous air pollutants (HAPs) and major air pollutants (MAPs).[1] HAPs consist of a wide variety of organic chemicals, including heavy metals and volatile organic compounds, emitted from vehicles and industrial facilities. The MAP indicators include ozone, particulate pollution, and carbon monoxide and other byproducts of burning fossil fuels. This class is often used as a single indicator for the multitude of pollutants present in ambient air. Combustion from motor vehicle use and power generation is the dominant source of pollution in urban areas, with sources emitting the same pollutants in varying proportions (Fenger 1999). Thus, researchers interested in understanding the effects of urban air pollution on human health have assessed combustion-generated multipollutant exposure with a particular focus on the MAP pollutants oxides of nitrogen (NO_x), ozone (O_3), sulfur dioxide (SO_2), and particulate matter less than 2.5 μm in diameter (PM2.5) (Solomon et al. 2000; Guarnieri et al. 2014; Brauer et al. 2002; Trasande and Thurston, 2005).

This project aims to better understand the influence of a wider array of constituent pollutants within the HAP and MAP classes by applying a k-means cluster model and a random forest model to source-specific pollution emissions and examining the structure and feature importances to assess the effects of individual pollutants on health outcomes. We choose to

[1] The U.S. Environmental Protection Agency denotes this class as criteria air pollutants.

represent outcomes as the rates of pediatric asthma hospitalization per ten thousand children in each New York City neighborhood tabulation area (NTA) because children experience a unique vulnerability to prevalent outdoor air pollutants (Trasande and Thurston 2005).

Whereas our approach focuses on identifying high-impact individual pollutants within urban ambient pollution, it relies on total annual point source emission and does not account for meteorological, geographical, and other influences on pollution concentration variation, as previous studies have done. These studies have primarily relied on class pollution concentrations measured at monitoring stations or modeled from traffic intensity data (Pénard-Morand et al. 2010; Pershagen et al. 1995). Our approach of identifying individual pollutants with high levels of influence on health outcomes makes it possible to develop targeted public health policies and appropriate future planning. Armed with a deeper understanding of the roles constituent pollutants play in certain health outcomes, policymakers can enact evidence-based targeted emissions reductions, and residents of areas near sources of the identified high-impact pollutants can take steps to mitigate the effects of exposure to these emissions.

Data and Methods

Sources

The New York City Department of Health maintains the New York City Neighborhood Health Atlas[2], which provides data on approximately 100 measures related to health and social factors, including our pediatric asthma dependent variable, for 188 neighborhoods of each Neighborhood Tabulation Area (NTA).[3] Air pollutant data come from two federal government sources, the National Emissions Inventory (NEI) and the Toxics Release Inventory (TRI). Together, these datasets cover emissions of over 185 individual pollutants. NEI data is collected every 3 years and contains data on chosen pollutants in the National Ambient Air Quality Standards from facilities including "large industrial facilities and electric power plants, airports, and smaller industrial, non-industrial and commercial facilities."[4] The TRI program compiles data on the release into the air of certain toxic chemicals via stacks and fugitive release that have significant adverse acute human health and environmental effects.[5] Whereas the NEI dataset required geocoding using the ggmap package in R[6] to obtain source emission geographies, the native TRI dataset contained point geographies for source emissions. To determine which NTAs encompassed the source pollutant points, we performed spatial joins using the GeoPandas Python package.[7]

[2]New York City Department of Health and Mental Hygiene. New York City Neighborhood Health Atlas. 9 April 2019. <https://www1.nyc.gov/site/doh/health/neighborhood-health/nyc-neighborhood-health-atlas.page>

[3]New York City Department of City Planning. NTA Map. 19 March 2019. <https://data.cityofnewyork.us/City-Government/NTA-map/d3qk-pfyz/data>

[4]U.S. Environmental Protection Agency. *National Emissions Inventory (NEI)*. Web. 7 May 2019.

[5]U.S. Environmental Protection Agency. *https://www.epa.gov/toxics-release-inventory-tri-program*. Web. 9 April 2019.

[6]Kahle, D., and H. Wickham. "ggmap: Spatial visualization with ggplot2." *The R Journal*, 5(1): 144-161. <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>

[7]GeoPandas developers, GeoPandas 0.4.0.

The Housing and Transportation Affordability Index[8] provides neighborhood-level data on housing and transportation costs, including household vehicle CO_2 emissions by census tract. We aggregated household CO_2 emissions totals for census tracts alling within each NTA.

Methods

We initially sought to discern the natural groupings in our predictor pollutant variables based on NTA geography, as we expected to see locations with higher numbers of point source polluters experiencing higher rates of pediatric asthma hospitalization. We employed a k-means cluster model on the whole pollutant dataset (Pedregosa et al. 2011), a method well suited for simple and speedy exploration that optimally groups observations based on a chosen number of k cluster centroids (Jain 2010). We arrived at the optimal number of k clusters using the within-cluster sum of squared error (SSE) metric, the so-called 'elbow method'. In plotting the SSE against the number of clusters, we could not unambiguously identify the elbow. However, at $k = 6$, we detected a balance in the complexity of the model and the SSE cost function.

```
np.random.seed(999)
s = np.zeros(10)
for k in range(0, 10):
    est = KMeans(n_clusters = k+2, n_init = 100)
    est.fit(all_emissions_standard)
    s[k] = est.inertia_
```

```
km = KMeans(random_state=999,n_clusters=6, n_init=100)
res = km.fit(all_emissions_standard)
results = km.predict(all_emissions_standard)
```

Figure 1: Inertia Plot for Clustering on Emissions

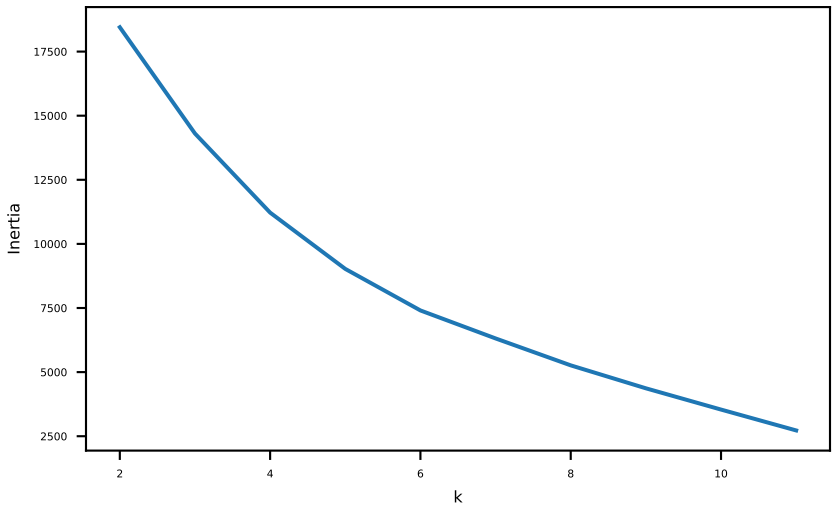


fig. 1: plot of in-cluster SSE against k cluster count shows no definitive optimal k value.

[8]Center for Neighborhood Technology. *The Housing and Transportation (H+T®) Affordability Index.* <https://htaindex.cnt.org/download/>

With the standard Scikit-learn *KMeans()* function, we divided New York City NTAs based on our six clusters of all pollutant emissions point source totals.

Figure 2: NTA Clusters Based on Pollutant Emittance

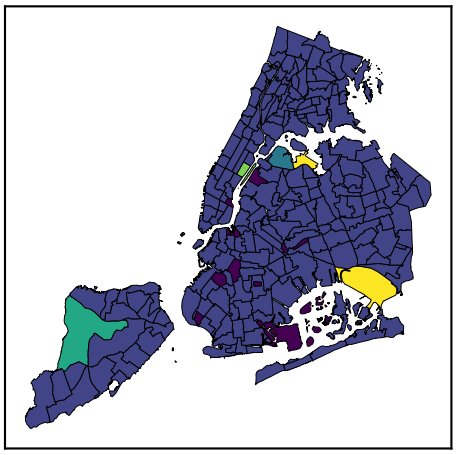


fig. 2: plot of New York City NTAs divided by six clusters of pollution emission totals shows a large number of NTAs falling into a single cluster, with airport locations and the former Fresh Kills landfill site constituting other notable clusters.

To test our expectation that NTAs housing higher numbers of point source polluters would experience higher rates of pediatric asthma hospitalization, we conducted an additional k-means clustering of the health outcome with an elbow method-generated optimum k of 3 clusters.

Figure 3: Asthma Hospitalization NTA Clusters

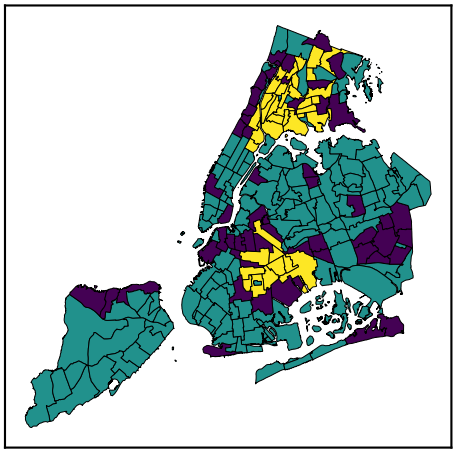


fig. 3: plot of New York City NTAs divided by clusters of pediatric asthma hospitalization rates show central Brooklyn and Bronx NTAs constituting a cluster representing high rates.

To examine the role of constituent pollutants within the MAP and HAP classes, we split our data into training and testing sets of two-thirds and one-third proportions, and employed a random forest regression. We then generated a feature importance metric by calculating the impurity reduction in each node in each tree across the forest. With 124 recorded pollutants, feature reduction was necessary to control for covariance while identifying key pollutants that lead to higher asthma rates. To implement the forest, we first performed a logarithmic transformation on our target variable to ensure it assumed a gaussian distribution. Next, to determine maximum tree depth, we looped through candidate parameters of two to 15 on our training data to identify the value resulting in the highest out-of-sample r-squared score. Our resulting forest consisted of 100 estimators at a maximum depth of 7. Using this method we identified 17 pollutants influential in predicting asthma hospitalization rates.

We relied on random forest regression over the alternative Ordinary Least Squares regression and Principal Component Analysis (PCA) methods in order to address interpretability and accuracy concerns. A key feature of this project was identifying the most significant pollutants, and PCA, while likely improving accuracy, does not lend itself to interpretability.

```
out_sample_score = 0
for i in range(2, 15):
    regr = RandomForestRegressor(max_depth=i, random_state=0,
                        n_estimators=100)
    regr.fit(x_valid, y_valid)
    print(regr.score(x_valid_test, y_valid_test))
    if regr.score(x_valid_test, y_valid_test) > out_sample_score:
        out_sample_score = regr.score(x_valid_test, y_valid_test)
        optimal_depth = i
```

```
regr = RandomForestRegressor(max_depth=optimal_depth, random_state=0, n_estimators=100)
regr.fit(x_train, y_train)
```

```
importances = regr.feature_importances_
indices = np.argsort(importances)[::-1]
features = {}
for i in indices:
    features[df_x.columns[i]] = regr.feature_importances_[i]
```

Results

Our K-means clustering yielded surprising results that suggest there is no tight association between total source point source pollution in an NTA and that NTA's rate of pediatric asthma hospitalization. The majority of NTAs (162 out of 192) comprised a single cluster based on emissions totals, yet asthma hospitalization rates noticeably varied within this single cluster. Emissions Cluster 0, which comprises of all cemeteries and parks within Brooklyn, Stuyvesant Town, and Astoria, was found to be the most perplexing in that it contains areas of both high point source concentration and low point source concentration. Several clusters with a single NTA are notable as well. There are three power plants with air stacks found within Steinway NTA (comprising all of Cluster 2), including the Con Edison Plant that had a transformer explode in December 2018. The New Springfield NTA (comprising all of Cluster 3) within Staten Island contains the Fresh Kills Landfill-turned-wetland area.

LaGuardia and John F. Kennedy airports (combining to comprise Cluster 4) had the highest levels of all NTAs.

Regarding clustering of asthma hospitalization rates, we see he clusters acting as a sort of heatmap for the asthma rates, with higher-rate NTAs, intermediate-rate NTAs, and lower-rate NTAs comprising different clusters. Moreover, NTAs clustered around high asthma hospitalization rates were not found to occupy the same NTAs featuring high emissions rates. We detected no pattern corresponding between the two cluster sets.

After running the random forest regression using our optimal hyperparameters, the regression feature importance metric returned 17 individual pollutants that have a noticeable impact on asthma rates. Our out-of-sample r^{2} came to 0.2, which does not instill much confidence in the model's ability to explain variations in pediatric asthma hospitalization rates. This low value is likely due to the large number of observations, i.e. NTAs, featuring point source emissions totals of zero. The most significant pollutants identified in the feature importance metric are NO_x boiler emissions, household CO_2, and propylene oxide, an industrial compound used in the manufacture of plastics. Data for boiler emissions and household CO_2 were most robust with values for every single NTA, and those pollutants may have wield an outsized effect on the feature importance as a result.

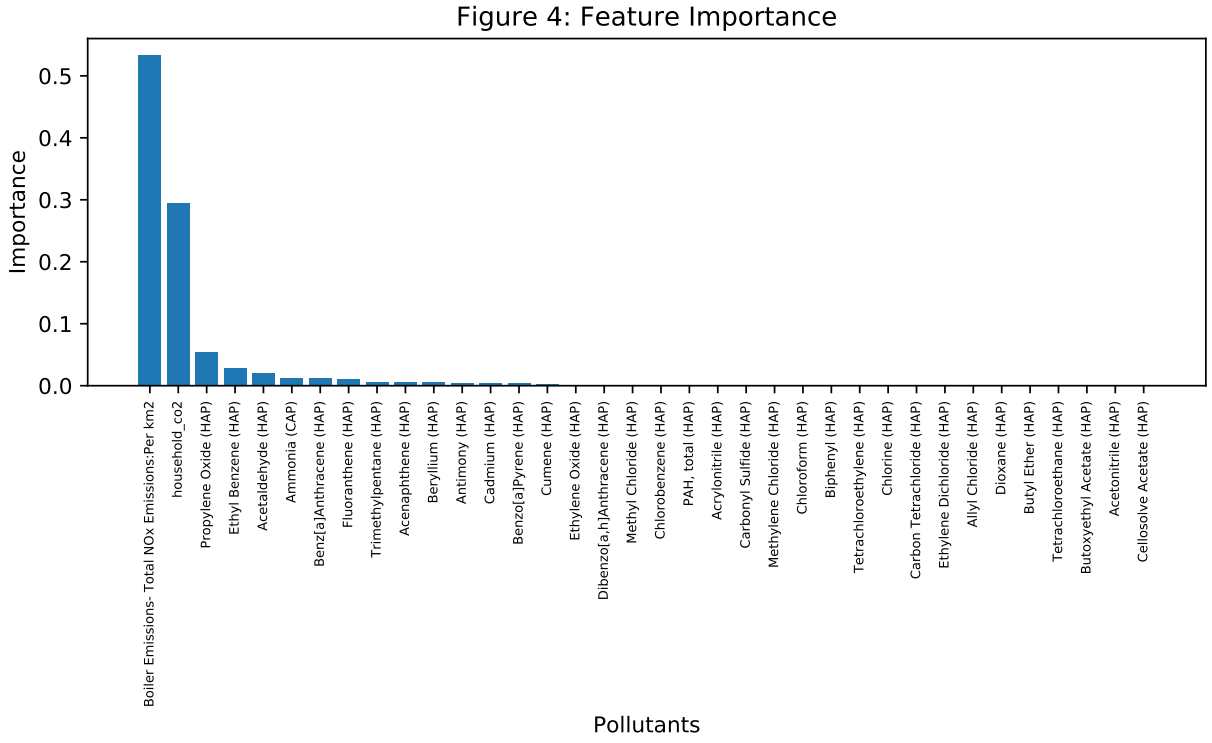


fig. 4: random forest model feature importance metric shows boiler emissions with the highest value

NEI and TRI values reflect emissions from a relatively small number of facilities within New York City, and thus emissions values for many NTAs and many individual pollutants are zero. Among significant pollutants, almost all are contained within the HAP class with ammonia being the lone MAP pollutant making an appearance as an important feature.

Additional Analytical Considerations and Limitations

We concentrated on pediatric asthma hospitalization rates across New York City because the most granular and robust dataset available to us was asthma-related hospitalizations, not overall rates of pediatric asthma incidence. Several factors are at play to differentiate the two rates. Sky-high health insurance rates often deter low-income individuals from buying insurance, leading to a lack of consistent medical attention and greater chance of serious health risks. (Antwi et al. 2015) Our results support this assertion, as many of the low-income areas of the Bronx, Bedford-Stuyvesant, East New York, East Harlem, and Flatbush exhibit higher than normal asthma-related hospitalizations.

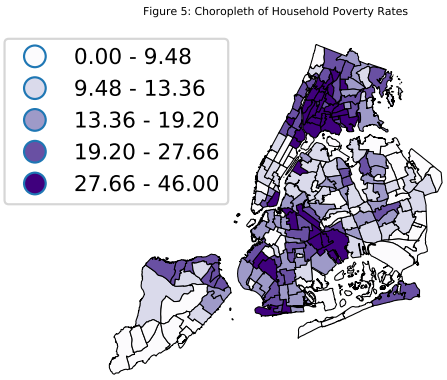
Additionally, the robustness of our underlying emissions data may have skewed our results. The NEI and TRI data ollection focuses on emissions from qualifying industrial and commercial pollutant generators, resulting in a number of neighborhoods containing no facilities generating pollutants. Conversely, as mentioned above, boiler emissions and household CO_2 data was very robust. Lastly, as modeling the geographic and meteorological dispersion of pollutants is outside the scope of our pollutant-centric analysis, our model relied only on pollutant generation totals and not actual ambient pollutant concentrations.

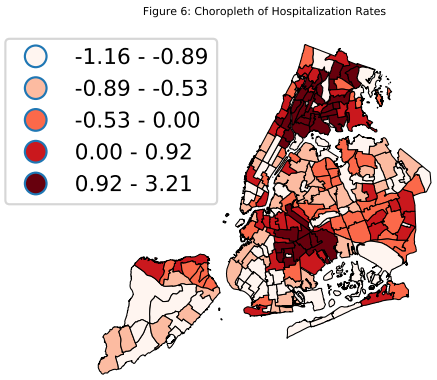
Conclusions

Random forest regression revealed a list of 17 industrial pollutants that influence health outcomes in New York City, and notably revealed the strong influence of boiler emissions. K-means clustering, however, did not uncover similar NTA relationships viz à viz pollutant point source emissions and rates of pediatric asthma hospitalization rates. Research on the relationship between asthma incidence and socioeconomic factors, particularly residence in low-income urban areas, suggests such factors covary with HAP and MAP pollution concentrations. We did not control for these possibly-confounding variables in our analysis and our results do not account for differences in hospital admission rates attributable to these factors. We may see poverty rates playing a particularly important role:ow-income families may forego hospitalization owing to limited access and/or unaffordability, yet paradoxically struggle with long-term asthma control and monitoring, which may precipitate acute events requiring hospitalization.

Research has also indicated low-income and minority families in New York City are more likely to live in areas near industrial and commercial polluters, which may contribute to our difficulty in discerning whether particular pollutant emissions, socioeconomic factors, healthcare-related challenges, or combinations of these factors are the greatest influence on asthma rates. We see below a striking similarity of geographic hotspots of both household poverty rates[9] and pediatric asthma hospitalization rates. Further research accounting for these healthcare access and geographical factors is necessary.

[9]U.S. Census Bureau; American Community Survey (ACS), 2017 Five-Year Census Tract, Table B17017; accessed via API ; (7 May 2019).





figs. 5, 6: choropleths of New York City NTA household poverty rates and normalized pediatric hospitalization rates show a high degree of geographic similarity

Statement of Work

All team members engaged in problem statement generation, model selection, and data collection.

Nathan was responsible for conducting the K-means clustering, determining number of clusters to use, and visualization of the clusters across the NTAs.

Rohun was responsible for determining that the random forest regression would be the best method for feature reduction. He then implemented the regression to determine the most important pollutants.

Ursula performed literature review on pollution and asthma studies, geocoded the NEI data, consolidated and performed spatial joins on the NEI and TRI datasets, and rendered the document in LaTeX.

Katie was responsible for aspects of cleaning Department of Health data, including spatially joining the neighborhood boundaries used by DOH to NTAs for purposes of aggregation with the other data. Katie was also responsible for research and data exploration regarding the possible relationship between poverty and child asthma rates, including collecting American Community Survey census data and aggregating at the NTA level.

References

Andrew Aligne, C., et al. "Risk factors for pediatric asthma: contributions of poverty, race, and urban residence." American journal of respiratory and critical care medicine 162.3 (2000): 873-877.

Antwi, Yaa Akosa, et al. "Changes in emergency department use among young adults after the Patient Protection and Affordable Care Act's dependent coverage provision." *Annals of emergency medicine* 65.6 (2015): 664-672.

El Din, Hamam Serag, et al. "Principles of urban quality of life for a neighborhood." *HBRC Journal* 9.1 (2013): 86-92.

Fenger, Jes. "Urban air quality." *Atmospheric Environment* 33.29 (1999): 4877-4900.

Guarnieri, Michael, and John R. Balmes. "Outdoor air pollution and asthma." *The Lancet* 383.9928 (2014): 1581-1592.

Jain, Anil K. "Data clustering: 50 years beyond K-means." *Pattern recognition letters* 31.8 (2010): 651-666.

Künzli, Nino, et al. "Public-health impact of outdoor and traffic-related air pollution: a European assessment" *The Lancet* 356.9232 (2000): 795-801.

Leikauf, George D., et al. "Evaluation of a possible association of urban air toxics and asthma." *Environmental Health Perspectives* 103.suppl 6 (1995): 253-271.

Pedregosa et al., "Scikit-learn: Machine Learning in Python", *JMLR* 12 (2011): 2825-2830.

Pénard-Morand, Céline, et al. "Long-term exposure to close-proximity air pollution and asthma and allergies in urban children." *European Respiratory Journal* 36.1 (2010): 33-40.

Pershagen, Göran, et al. "Air pollution involving nitrogen dioxide exposure and wheezing bronchitis in children." *International Journal of Epidemiology* 24.6 (1995): 1147-1153.

Pope III, C. Arden, et al. "Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease." *Circulation* 109.1 (2004): 71-77.

Shmool, J. L., Kubzansky, L. D., Newman, O. D., Spengler, J., Shepard, P., & Clougherty, J. E. "Social stressors and air pollution across New York City communities: a spatial approach for assessing correlations among multiple exposures." *Environmental Health : A Global Access Science Source*, 13, (2014): 91.

Solomon, C., et al. "Effect of serial-day exposure to nitrogen dioxide on airway and blood leukocytes and lymphocyte subsets." *European Respiratory Journal* 15.5 (2000): 922-928.

Trasande, Leonardo, and George D. Thurston. "The role of air pollution in asthma and other pediatric morbidities." *Journal of Allergy and Clinical Immunology* 115.4 (2005): 689-699.

Wiederkehr, Peter, and Seung-Joon Yoon. "Air quality indicators." *Urban Air Pollution—European Aspects*. Springer, Dordrecht, 1998. 403-418.