Machine Learning Prediction of Asthma Rates in Pennsylvania

Min Zhong

Pennsylvania Department of Environmental Protection (Employer Project)

Introduction

Asthma is a chronic lung disease that causes coughing, wheezing, chest tightness or shortness of breath [1]. Asthma can be deadly and cannot be cured. It is one of the most common and costly diseases in the United States. Approximately 25 million people in the U.S. have asthma, which equals about 1 in 13 people. From 2008-2013, the annual economic cost of asthma was more than $81.9 billion – including medical costs and loss of work and school days [2]. From 2012 - 2020, Pennsylvania had consistently higher asthma rates (9.6-10.9%) compared to the rest of the United States (8.9-9.6%) [3]. Two cities in Pennsylvania, Allentown and Philadelphia, were ranked 1st and 7th, respectively, among the top 100 most challenging places to live with asthma according to the latest 2021 asthma Capitals report published by the Asthma and Allergy Foundation of America [4].

Distribution of asthma rates is the driver for researchers and policymakers to discover asthma clusters, understand the link between environment and asthma, develop control policies and adapt healthcare capacity. The Centers for Disease Control and Prevention's (CDC') Behavioral Risk Factor Surveillance System (BRFSS) collects and publishes asthma data annually based on near 400,000 adult surveys each year. CDC's PLACES project combines this BRFSS county-level asthma rates with socioeconomic and demographical data the American Community Survey to estimate asthma rates or other health outcomes at the census tract level [5]. This census tract level data lags 2 years with 2019 as the latest year due to the resource and time consuming of conducting surveys. In addition, annual monitoring of asthma and socioeconomic status using surveys is prohibitively expensive. More update-to-date asthma rates are needed to help the public health policy makers for policy implementation and adjustment. Our first objective is to predict census tract level asthma rates in Pennsylvania using machine learning modeling. Specifically, our focus is to present asthma rates at the census tract level which provides much higher spatial resolution than the county level.

There are several factors associated with the development and exacerbation of asthma: exposure to allergies, respiratory infections, poor air quality, and weather [6]. Since an important part of these factors is related to environment, it is important to identify environmental factors in reducing its effects. Our second objective is to investigate the role of environmental factors in predicting asthma rates.

Data Sources

Asthma rates: The health outcome or dependent variable is current adult asthma rate. We collected the only available 2018 and 2019 annual adult asthma rate data at the census tract level from CDC's PLACES project . In Pennsylvania, there are 3,218 census tracts.

***Air pollution**:* Annual mean ozone, $PM_{2.5}$, diesel particulate matter, and air toxics were collected from EPA's environmental justice project. We use cancer risks of all 52 hazardous air pollutants to represent air toxics. Nitrogen dioxide (NO_2) column levels were obtained from the Tropospheric Monitoring Instrument (TROPOMI) satellite data with a 1-km resolution.

***Natural environment**:* To measure exposure to overall greenness, we used the Normalized Difference Vegetation Index (NDVI), which is a greenness index bounded by -1 and 1 and typically derived from satellite imagery. Specifically, we used maximum NDVI derived from the U.S. Geological Survey's Earth Resources Observation and Science (EROS) Archive. Urban land cover data was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite MCD12Q1 V6 product on a 500-meter grid.

Temperature: Maximum daily surface temperature on a 1-km grid was obtained from the Daymet V4 Product of the Oak Ridge National Laboratory.

Night light: The monthly radiance of night light data was obtained from Visible Infrared Imaging Radiometer Suite (VIIRS) Day-Night Band (DNB) with resolution of 464 meters.

Data Preprocess

All satellite products, including TROPOMI NO_2, NDVI, Urban land cover, Daymet temperature, and Night light, were accessed and processed using Google Earth Engine. Daily or monthly data were averaged to obtain annual average data. Census tract level data from raster satellite images were aggregated in the 2010 census tract boundary. Data in both 2018 and 2019 were used for training and testing. We randomly split data into two parts: 75% for training and building models and 25% for testing model performance. Since our feature variables vary in units, magnitudes, and ranges, Min-Max scaling was applied to all features.

Modeling Methods

We built seven models using Sklearn packages and used 5-fold cross validation to optimize parameters with training dataset. We calculated four error metrics: root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), and mean absolute percentage error (MAPE) using predicted and observed asthma rates. In addition to these error metrics, we also reported R-squared (R^2) score for model evaluation. We wanted to select a model that could achieve the lowest errors and highest R^2 score. We considered error metrics for both training and testing datasets. However, we selected the model with better performance in the testing dataset.

Since the asthma rate is a continuous variable, we employed multiple machine learning regression models: k-nearest neighbors (KNN), random forest, neural network, gradient boosting regression, and support vector regression (SVR). We also included the traditional statistical models such as linear regression and generalized additive model (GAM) as refences for comparison. The workflow of the project is shown in Fig. 1. Once the best model was created, we predicted 2020 asthmas rates in Pennsylvania at the census tract level.

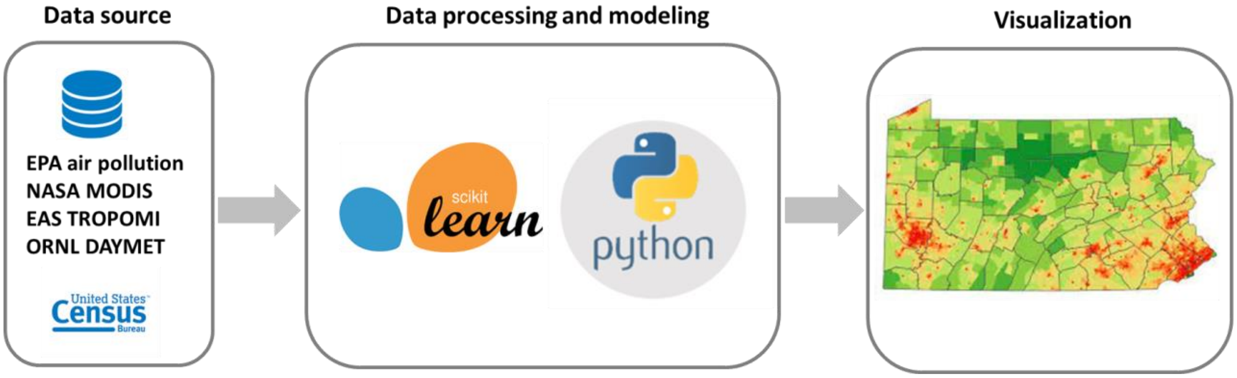

Figure 1: Workflow of the project.

Results and Discussion

Statistic Distribution of Variables

The CDC PLACES project contains asthma rates for 3,196 out of 3,218 census tracts. Table 1 provides descriptive statistics for both outcome and predicting variables. The statistical distribution was calculated by averaging the annual data in 2018 and 2019.

Table 1. Statistic distribution of asthma rates and predicting variables for 2018 and 2019 across 3,196 census tracts in Pennsylvania.

Variable	Mean	Minimum	Maximum	Median	25th percentile	75th percentile
Asthma rate (%)	10.53	7.20	17.70	10.20	9.60	11.00
$PM_{2.5}$ (μg/m^3)	10.14	6.11	12.91	10.24	9.62	10.96
Diesel PM (μg/m^3)	0.70	0.08	6.32	0.56	0.33	0.88
Ozone (ppb)	44.71	36.76	47.85	44.78	43.53	46.23
Traffic proximity (count per day)	456.55	0.00	22701.23	191.92	48.57	514.21

Air toxics (incident per million)	37.33	15.93	596.46	34.51	27.74	43.06
TROPOMI NO$_2$ (e^{15} molecules/cm^2)	3.00	1.26	6.36	2.80	2.16	3.53
Temperature (degrees C)	16.56	12.37	18.80	16.81	15.74	17.61
NDVI	0.55	0.10	0.78	0.57	0.46	0.67
Urban landcover (%)	51.17	0.00	100.00	54.27	2.78	99.29
Night light (nanoWatts/cm^2/sr)	19.14	0.27	216.50	12.56	3.22	25.31

Model Comparison

We first provided the general pros and cons of each model used in this project before analyzing the results. Linear regression is the simplest statistical model. It has a closed form of solution, and it is easy to interpret the regression coefficients; however, it requires linear correlation between features and response. GAM can fit a non-linear relationship and examine the effect of each feature, but it is restricted to be additive. KNN uses the neighboring points for prediction and does not require distribution assumptions for the data. The biggest challenge of KNN is calculating distance with high dimensional data. Random forest is an ensemble of decision trees which ensures reduction of the overall error. It can handle a very large amount of data with higher dimensions; however, it appears as a black box providing little control over what it does. Neural network is good for nonlinear data with many inputs, but it is computationally expensive, depends a great deal on training data, and has the risk of overfitting. Gradient boosting is an ensemble of weak trees with each tree learning and improving on the previous. It often provides excellent predictive accuracy, but it is computationally expensive and has the potential of overfitting. SVR is more accurate in high dimensional space, but it may not perform well when the dataset is noisy.

Table 2 shows the error metrics of different models for both the training and testing datasets. The desired errors and R^2 score are highlighted in green. Among the seven models, gradient boosting yields the lowest RMSE, MAE and MAPE for both the testing and training datasets. However, gradient boosting has the risk of overfitting. KNN obtains the lowest MPE. Random forest performs slightly poorer than gradient boosting, however the differences of error metrics between the testing and training datasets are less than those of gradient boosting. Linear regression consistently underperforms with highest errors and lowest R^2 for testing dataset.

To avoid potential overfitting of gradient boosting and improve model performance, we select top three models with low RMSE and high R^2 scores: gradient boosting, random forest, and KNN, to create a simple ensemble model. We average the prediction of these three modeling results to produce improved results. As shown in Table 2, the ensemble model does yield the lowest RMSE and MAE for the testing datasets.

Table 2. Error metrics of different models for training and testing datasets.

Models	Training					Testing				
	RMSE	MAE	MAPE	MPE	R^2	RMSE	MAE	MAPE	MPE	R^2
Linear Regression	1.225	0.894	0.083	-0.012	0.26	1.168	0.853	0.079	-0.006	0.32
GAM	1.039	0.742	0.069	-0.009	0.47	0.976	0.701	0.065	-0.004	0.52
KNN	0.878	0.604	0.056	-0.003	0.62	0.929	0.647	0.059	-0.002	0.57
Random Forest	0.691	0.501	0.047	-0.006	0.77	0.879	0.626	0.058	-0.004	0.61
Neural Network	1.074	0.780	0.074	-0.021	0.43	0.997	0.723	0.068	-0.016	0.5
Gradient Boosting	0.520	0.401	0.038	-0.004	0.87	0.866	0.607	0.056	-0.003	0.63
SVM	1.063	0.750	0.069	-0.008	0.44	0.989	0.699	0.064	-0.003	0.51
Ensemble	0.659	0.482	0.045	-0.004	0.79	0.855	0.601	0.056	-0.003	0.63

Feature Importance

The importance of each variable is ranked in Fig. 2 based on the permutation importance technique for feature evaluation using the ensemble method. Night light is the most important variable in modeling asthma rates, followed by NDVI, TROPOMI NO_2, temperature, and urban landcover. The other variables, such as air toxics, ozone, $PM_{2.5}$, diesel particulate matter, and traffic proximity, are less important. Night light data which reflects human activities has been used as a proxy of socioeconomic variables and these socioeconomic variables such as race, education, and income are correlated with asthma rate estimates provided by CDC's PLACES project.

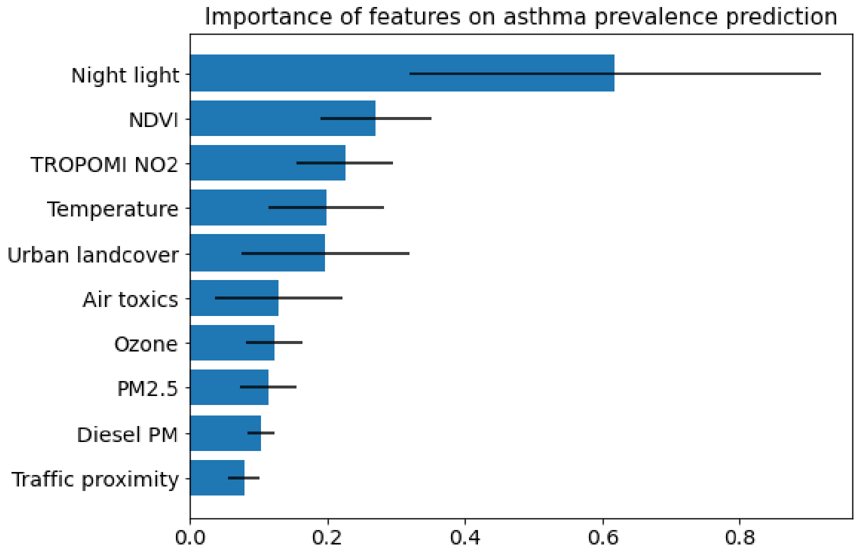

Figure 2: Ranking of importance of features

Prediction of Asthma Rates for 2020

Using the ensemble model and predicting variables of 2020, we predicted census tract asthma rates in Pennsylvania for 2020. The mean and median asthma rates across all census tracts are 10.76% and 10.50%, respectively. The spatial distribution of predicted asthma rates in 2020 is shown in Fig. 3. The asthma rates in most rural census tracts are in the range of 10-11%. The relative lower asthma rates (less than 9.5%) are mainly located in suburban areas near the city of Pittsburgh. High asthma rates (larger than 12%) are more likely concentrated in urban areas, especially the city of Philadelphia. Although Allentown was ranked as 1st place among asthma capitals, only one census tract has an asthma rate above 12% in Allentown.

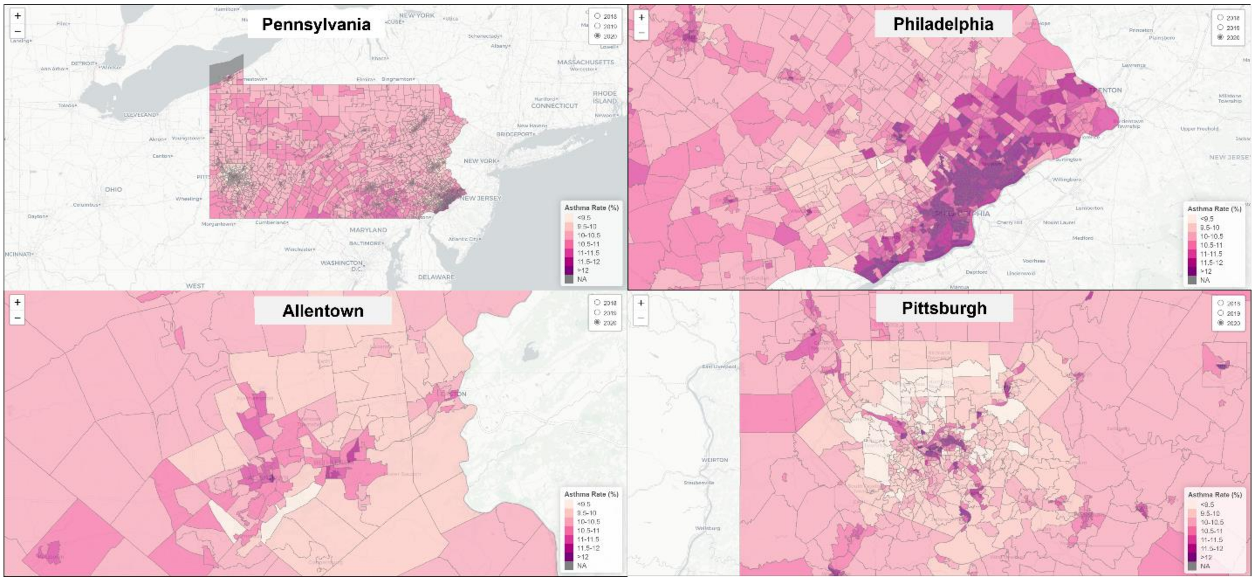

Figure 3: Predicted asthma rates at census tract level in Pennsylvania 2020.

Limitations

Our study has several limitations. First, we relied on asthma data at census tract level from the CDC PLACES project that only provides estimated adult asthma rates for the year of 2018 and 2019 in Pennsylvania. We do not have estimates for asthma in children, which is a focus of asthma study. Secondly, since CDC's census tract asthma estimates were calculated from a multilevel logistic regressions model that uses socio-demographic variables from the American Community Survey, we cannot use these same socio-demographic features as predictors to compare the performances of our model. Thirdly, we were not able to validate the robustness of the models over time because the data from the CDC PLACES project is only available for 2018 and 2019.

Conclusion

We predicted asthma rates in Pennsylvania at the census tract level using multiple regression models including five machine learning regressors and two traditional statistical models. The best-performing ensemble model was created using three machine learning regressors: KNN, random forest, and gradient boosting. From the training results of this model, we identified night light is the most important feature in predicting asthma rates, followed by NDVI, TROPOMI NO_2, temperature, and urban landcover. Using the ensemble model, we predicted census tract level asthma rates for the year of 2020. This method could potentially be used to assess other health outcomes that are correlated with the environment for epidemiological studies.

References:

1. https://www.cdc.gov/asthma/default.htm
2. https://www.thoracic.org/about/newsroom/press-releases/resources/asthma-costs-in-us.pdf
3. https://www.cdc.gov/asthma/brfss/default.htm
4. https://www.aafa.org/media/3040/aafa-2021-asthma-capitals-report.pdf
5. https://www.cdc.gov/places/about/index.html
6. https://www.epa.gov/asthma/asthma-triggers-gain-control