Predicting Asthma Rates from Geographic Features

Keya Sengupta

**Executive Summary**

This paper explores the possibility of using land cover data to predict asthma rates in a census tract. It uses data from the EnviroAtlas (land cover information), the CDC (census tract level asthma rates), and the American Community Survey (census tract population estimates), and runs linear regression, quadratic regression, and K-Nearest Neighbour regression to determine which yields the best model and potential future steps.

**Introduction**

Researchers have suggested that there exists a correlation between air quality and the likelihood of developing asthma (Dick, Doust, Cowie, et. al., 2021). With that in mind, factors that influence air quality, such as the percentage of trees in an area, may be useful for predicting asthma rates in that same area. If this is true, researchers may be able to use a model trained on the current data of a tract to predict the impact of a physical change to that tract, such as adding a new road, thereby removing tree cover or agricultural space, on health outcomes.

There are many existing approaches to predicting asthma. However, most such methodological approaches are on the individual level, with demographic information, comorbidities, and personal medical history being used to predict that individual’s level of asthma risk (Alharbi, Nadeem, and Cherif, 2021). As asthma development is partially genetic and partially associated with environmental factors, using demographic information and personal history is likely to yield stronger individual level predictions than using solely environmental factors. However, if the goal is to predict population level rates, rather than the specific likelihoods of any individual developing asthma, and environmental factors can be used to make such a prediction, it will be possible to determine general population outcomes without needing extremely detailed information about the individuals within the population.

Environmental impact assessments (EIAs) typically seek to answer these questions and determine how developmental projects will impact human health. However, such assessments are difficult, time consuming, and expensive to do correctly, and evidence suggests they are often extremely inaccurate. If existing datasets collected for different purposes can predict general health outcomes, EIAs can become more feasible.

This paper explores whether asthma rates in a census tract can be predicted from land cover information. Rather than using individual level data, I will instead use census tract data. If asthma rates can be predicted using this census tract level environmental data, policymakers could determine the health impact of a proposed project and be better equipped to understand the extent to which it will or will not harm the population in the area of the proposed project.

**Data**

The ideal dataset to use to for this analysis would include census tract level information about the environment and the population. Attributes such as asthma rates, percentage tree cover, total population, demographics breakdown of the population, geographical area, primary industry, and number of toxic facilities would be useful. A single record would exist for each tract. An approximation of this ideal dataset can be constructed using several existing public datasets.

I will retrieve information about my target variable from the CDC Places (CDC). The CDC Places dataset contains census tract level information about a number of health indicators, including asthma. Table 1 gives some basic summary statistics about the target variable. The median and mean rates appear to be approximately 1.6 standard deviations apart. This suggests there are not significant outliers in the dataset, as the values are only about 0.1 standard deviations apart, and outliers would lead to a large difference between the mean and median. Also, considering that the mean and the median are both closer to the minimum asthma rate than the maximum asthma rate, the summary statistics suggest that the data is not normally distributed, and lower rates occur more frequently than do higher rates. Examining this graphically through a histogram and a Q-Q plot will lead to a better understanding of the form of this variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MinimumAsthmaRate (%) | MaximumAsthmaRate (%) | MedianAsthmaRate (%) | MeanAsthmaRate (%) | StandardDeviationAsthmaRate (%) |
| 4.900 | 21.500 | 9.600 | 9.767 | 1.560 |

Table 1: Asthma Summary Statistics, US Census Tracts

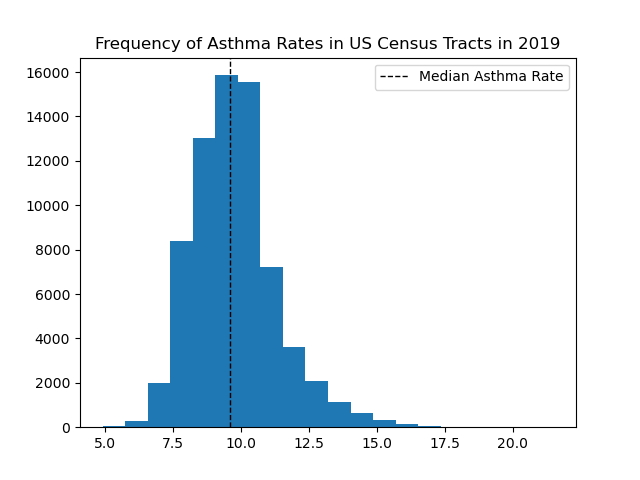


Figure 1: Frequency of Asthma Rates

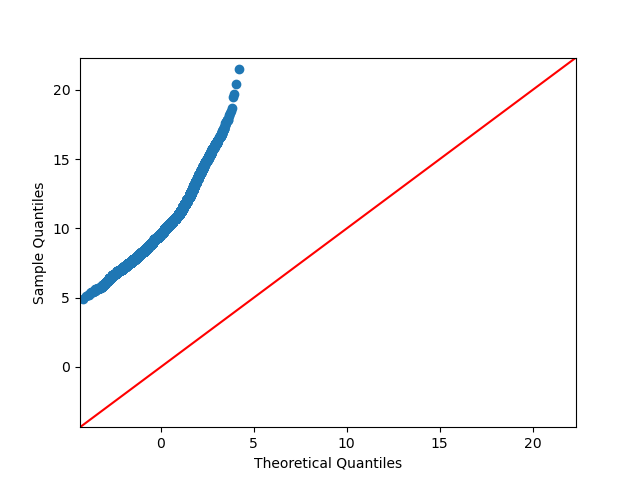


Figure 2: Q-Q Plot of Asthma Rates

Figure 1 suggests that asthma rates are not normally distributed, as values appear more frequently towards the right of the median, while the right side has a longer tail. Additionally, the Q-Q plot in Figure 2 is curved, rather than straight, supporting the idea of there being skew in the dataset, as was suggested both by the summary statistics and the Figure 1 histogram. Therefore, that the data is not normally distributed is a reasonable conclusion to draw. This has implications for the model selection. Models that assume a normal distribution of the variables are therefore not the best options for this scenario.

One resource I will use in order to retrieve my feature variables is the EPA’s EnviroAtlas (EPA). This provides detailed information about landcover, including not only trees, but water, wetland, grasses, and impervious areas. There may be challenges associated with this dataset. One such challenge is the resolution of the data. The majority of national data is aggregated by hydrologic unit code. This may make it difficult to join to other datasets, such as the CDC census tract or county level data. This concern is resolvable by using the EnviroAtlas’s high resolution data, which is on a census block group level, then aggregating to the level of the other datasets.

What may be more difficult to resolve is the locations of this information. By virtue of the fact it is high resolution data, it is more limited than other datasets. This high-resolution community data is limited to high population areas, i.e. major cities and the surrounding suburbs. These areas are likely to have a lot in common. Therefore, any model trained on this data, even if it performs well on a test subset, may not be useful for predicting asthma in rural areas. In addition to the potentially low variation between the locations with EnviroAtlas data, there may not be sufficiently detailed fields to have much predictive power. For example, one of the fields is area covered by impervious surfaces. This includes roads, sidewalks, and buildings, all of which are likely to have very different impacts on emissions and thus health outcomes. In order to expand this in later iterations such that whatever model is derived is more broadly applicable, I would need larger datasets with more specific features, representing a broader range of geographic locations.

From this dataset, I will use tree coverage, wetlands coverage, impervious area coverage, agricultural coverage, and population as features. Trees are associated with lower rates of asthma, potentially through improving air quality (Lovasi, Quinn, Neckerman, Perzanowski, & Rundle). Therefore, it is important to examine the tree cover feature to gain a better understanding of the distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MinimumTreeCover (%) | MaximumTreeCover (%) | MedianTreeCover (%) | MeanTreeCover (%) | StandardDeviationTreeCover (%) |
| 0.0 | 94.020 | 23.490 | 27.617 | 16.339 |

Table 2: Tree Cover Summary Statistics, US Census Block Groups

As can be seen in Table 2, there is a small difference between the mean and median of the percentage of tree cover in a census block group. There is an extremely wide range of values this variable takes on, and from the summary statistics, there is some indication of right skew, and the mean is higher than the median.

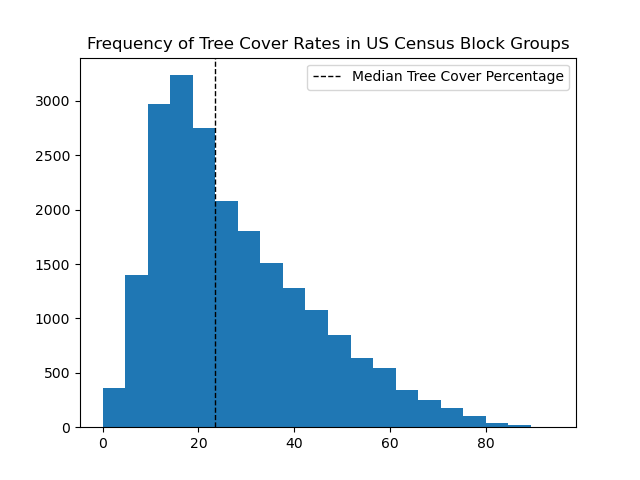


Figure 3: Frequency of Tree Cover Rates

From the histogram of the frequencies of tree cover rates in Figure 3, the distribution appears right skewed, supporting the initial understanding from the summary statistics.

Unlike the other land cover attributes, the PercentAgriculture attribute has values outside the expected range of 0 to 100: there are instances where both the magnitude and the direction are outside of the range of percentages, eg: -888888. These cases appear to be entirely from the LA metropolitan area land cover dataset. As can be seen in Figure 4, these values occur frequently. I will try running the models thrice, with the negative agricultural rates recoded to 0, with only observations with positive rates, and with the negative rates filled with predicted values from running a regression using PercentAgriculture as the target.

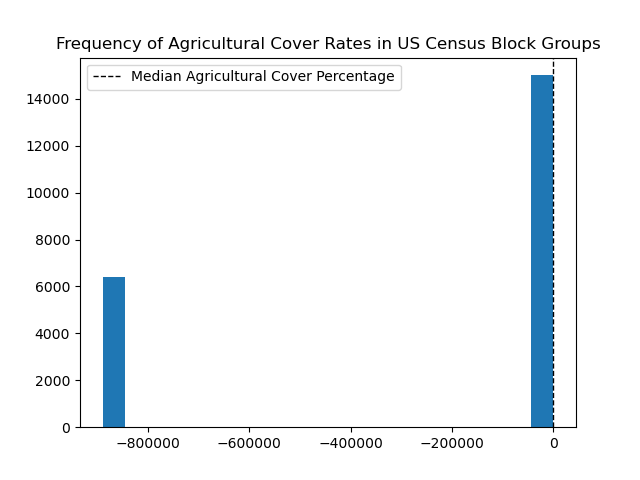


Figure 4: Frequency of Agricultural Cover Rates

The PercentAgriculture attribute is the only one with such outliers that make such a significant impact on the modeling approach. Refer to the Implementation Appendix for the frequency histograms for the other attributes.

**Methodology**

The target variable of this analysis will be asthma rates. This is a numeric, continuous variable and thus necessitates regression, rather than classification. This narrows down the potential models that can be used. Additionally, the non-normal distribution of the target variable has implications on the models that can be used.

As discussed in the data section, I will use tree coverage, wetlands coverage, impervious area coverage, agricultural coverage, green space coverage, and population as features in the model. These features are all numeric and continuous.

Interpretability is an important quality of a model that would be well suited to this topic. This is because a potential use case for the model would be as a preliminary step in an environmental impact analysis. Understanding what features the model is prioritizing is important for using the generated results to make recommendations to improve the environmental impact. In addition to this, as discovered in the preliminary data analysis, asthma rates are not normally distributed. This suggests that Gaussian Naïve Bayes will be an inappropriate algorithm to use, as it assumes a normal distribution in order to calculate probabilities.

Therefore, for my parametric model, I will use polynomial regression. Polynomial regression models the relationship between the feature variables and the target variables by fitting a curve to the observations. Linear regression, the form of polynomial regression where the degree is one and the curve is a straight line, is a valuable tool because it provides insights into the strength of the relationships between the variables and can be used for prediction of a numeric variable, such as my target variable. This means that it is well suited for achieving the goals of both interpretability and prediction. Even if other models perform better, the additional insight into what may be factoring into a high rate of asthma can guide future steps of environmental impact analysis, and so it is a strong choice for a preliminary model. It also does not depend on a normally distributed target variable.

An instance in which linear regression has been used for prediction of health outcomes can be found in Abolmaali and Shirzael’s 2021 paper, *A comparative study of SIR Model, Linear Regression, Logistic Function and ARIMA Model for forecasting COVID-19 cases.* This study used simple linear regression, in which the relationship is between a single regressor *x* and target variable *y.* This was found to not be sufficient for predicting the number of COVID cases *x* days after the first case was diagnosed, even though it has a low error rate in the short term, which highlights the limitations of linear regression for forecasting.

One of the dangers of linear regression is that it is inadequate for accurate predictions when the relationship between variables is fundamentally nonlinear. This is a disadvantage of all parametric models, due to their assumption of functional form. In this specific case, it is likely that the relationship between these attributes and the target variable is nonlinear – different geographic areas can look vastly different while still having similar health outcomes. For example, wetlands and forested lands both sequester carbon and absorb pollution. Therefore, health conditions that are caused or exacerbated by pollution may occur at similar rates in a census tract that is primarily wetlands and a census tract that is primarily forested. Because of this, I will also try higher order regression to determine what degree yields the best results. This comes with its own sets of challenges; primarily, using higher order polynomials runs the risk of overfitting. However, I have a sufficiently large dataset to be able to reserve points for testing, enabling me to iteratively improve the model and validate its functionality.

Because I am using continuous numeric variables for my features, a decision tree is likely not the best choice for my nonparametric model. This is because training may take a long time and that a small change in data may lead to a large change in the ideal tree – this situation is likely to involve overfitting. Therefore, I will use k nearest neighbours (KNN). This algorithm determines the approximate value of the target variable by averaging the targets of the most similar observations to the observation being predicted on. While KNN is not interpretable, the fact that it is a local regressor may be helpful for handling this use case. Since there is a lot of variability within the features, and the dataset contains very different census block groups in terms of their geographic composition (eg: primarily forested vs. primarily agriculture), comparing only to the most similar points may avoid relying on spurious correlations. An instance of KNN being used for regression can be seen in Tanuwijaya and Hansun’s 2019 paper, *LQ45 Stock Index Prediction using k-Nearest Neighbors Regression.* This study used KNN to predict the price of a given stock. One potential pitfall of KNN is that it is sensitive to outliers. This may be problematic in the situation being explored in this paper, due to the non-normally distributed variables, some of which are only rarely significantly higher than 0.

**Findings**

These models were built using a 70-30 train test split. The reason for this choice was out of a desire to have more data available for testing and desire to err on the side of caution such that I assume a higher error rate than I would likely encounter if I used a larger training set.

To select the number of neighbours for each run of my KNN models, I used a validation curve to estimate where the training and test estimates begin to approach each other. As such, I opted for an initial value of 20 for the dataset with invalid values recoded to 0 and 15 for the datasets with invalid values dropped and predicted. Refer to the appendix for the validation curves.

|  |  |  |  |
| --- | --- | --- | --- |
|  | R2 | Adjusted R2 | Mean Squared Error |
| Linear Regression | 0.088 | 0.138 | 3.706 |
| Quadratic Regression | 0.308 | 0.418 | 2.814 |
| K-Nearest Neighbour | 0.035 | 0.035 | 3.921 |

Table 3: Metrics for Recoded Data

Table 3 presents the metrics for all the models run with the negative agricultural rates recoded to 0. Linear regression yielded an R2 score of 0.088, which is to say that 8.8% of the variation can be explained by land cover information, and a mean squared error of 3.706, suggesting that the average squared difference between the actual asthma rate and the predicted asthma rate is 3.706%. Figure 7 suggests that the distribution of predictions peaks in the same ranges as the distribution of actual rates, while the more extreme values on the upper end of the spectrum are not predicted at all by the linear model. Overall, it appears the distributions do not align very well.

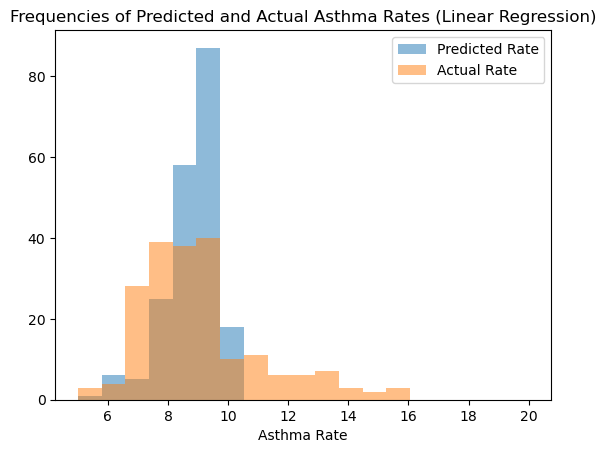


Figure 5: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with Linear Regression

Using polynomial features of degree 2 reduces the mean squared error relative to linear regression, bringing it down to 2.814. As can be seen in the histogram presented in Figure 8, increasing the degree improves the extent to which the distributions of predictions align. While this model still fails to predict values at the extreme end of the spectrum and overpredicts in the most frequent ranges, it does a slightly better job of capturing the potential ranges of values. In the most frequent ranges, quadratic regression seemingly does a better job of capturing the local minima and maxima. This suggests that a quadratic curve fits the data better than a straight line. However, this model still cannot capture the variation in the cases of high asthma rates.

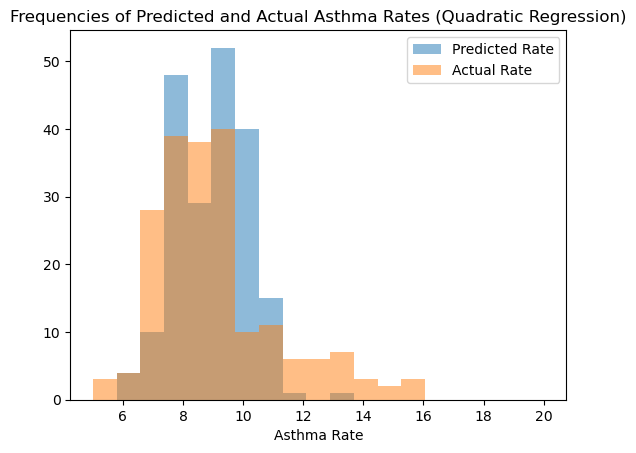


Figure 6: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with Quadratic Regression

Unlike I had suspected would be the case when selecting models, there does not appear to be a benefit to using local models over global – KNN using 20 neighbours yielded a lower R2 score than either linear regression or quadratic regression, with a value of approximately 0.035.

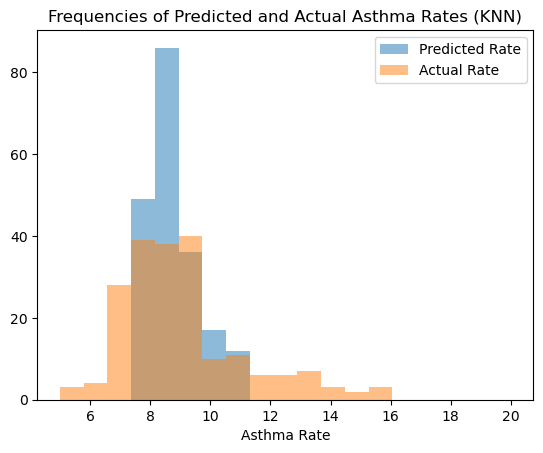


Figure 7: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with K Nearest Neighbours Regressor

As can be seen in Figure 7, KNN has similar issues with the distribution of predictions as linear and quadratic regression in that it does not predict values at the extreme end of the spectrum. However, this could be partially influenced by the choice in the number of neighbours. Looking at Figure 7, there are fewer than 10 items in all the bins above approximately 11.5. Using a lower k-value might reduce the overestimation around the modal range, while increasing the under or overestimation in the other ranges. Running the same model with the number of neighbours reduced to 8 results in a higher MSE overall, but a distribution that is visually slightly closer to the actual distribution, as can be seen in Figure 8.

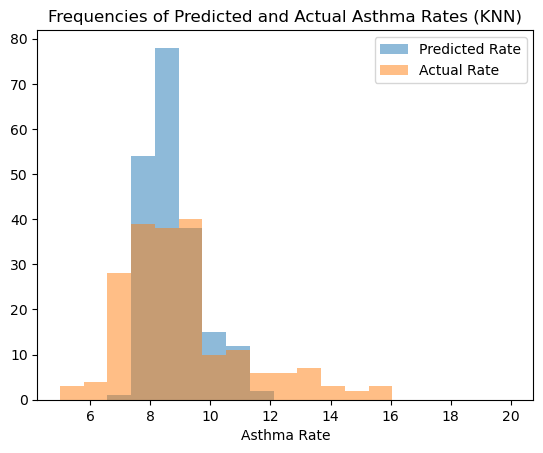


Figure 8: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with K Nearest Neighbours Regressor with a Lower K

When excluding invalid values from the dataset altogether, the performance of the linear and KNN models improves slightly, as can be seen by the decrease in mean squared error in Table 3 as compared to Table 2. The quadratic model, on the other hand, has a greater mean square error when invalid values are excluded, in comparison to when they are recoded to 0. This is expected behaviour for the quadratic model, as it is susceptible to overfitting and reducing the number of datapoints exacerbates that risk. It is also a reasonable change for KNN, as removing the records altogether results in more distinct points, such that all the neighbours are not clustered together. Similarly, linear regression with the records removed ensures less influence by points with assumed values.

|  |  |  |  |
| --- | --- | --- | --- |
|  | R2 | Adjusted R2 | Mean Squared Error |
| Linear Regression | 0.095 | 0.133 | 3.572 |
| Quadratic Regression | 0.060 | -0.232 | 3.713 |
| K-Nearest Neighbour | 0.122 | 0.183 | 3.467 |

Table 4: Metrics for Data With Dropped Values

Because PercentAgriculture is a linear continuous variable, it is possible to use estimated values for the invalid values, rather than either recoding them to 0 or excluding them. I chose linear regression to generate these predicted values from the other land cover values, based on the fact that there is likely to be a simple relationship between these values, as they are rates of cover for the same area. The metrics for the models to predict asthma rates based on land cover with the filled PercentAgriculture values can be seen in Table 5.

|  |  |  |  |
| --- | --- | --- | --- |
|  | R2 | Adjusted R2 | Mean Squared Error |
| Linear Regression | 0.084 | 0.131 | 3.721 |
| Quadratic Regression | 0.311 | 0.424 | 2.800 |
| K-Nearest Neighbour | 0.028 | 0.020 | 3.951 |

Table 5: Metrics for Data with Predicted Values

Using predicted values yielded results similar to the recoded values. As can be seen from the histogram shown in Figure 9, the predicted distribution for linear regression looks nearly identical to the one generated from the dataset with recoded values. There are minor differences in the predicted distribution for quadratic regression and KNN, as can be seen from the histogram showed in Figures 10 and 11 in comparison to Figures 6 and 7, but overall, the distributions appear similar.

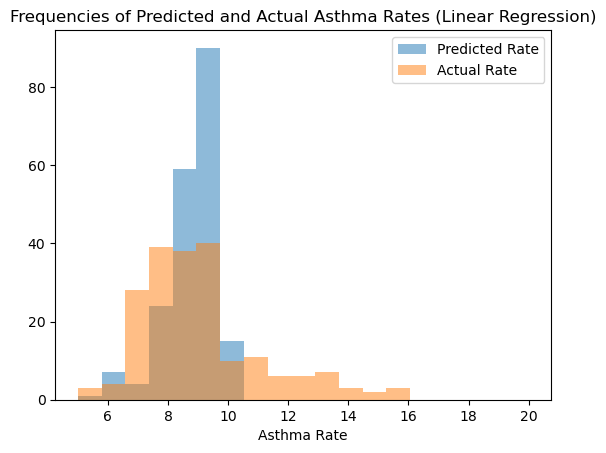


Figure 9: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with Linear Regression from Predicted Agriculture Rates

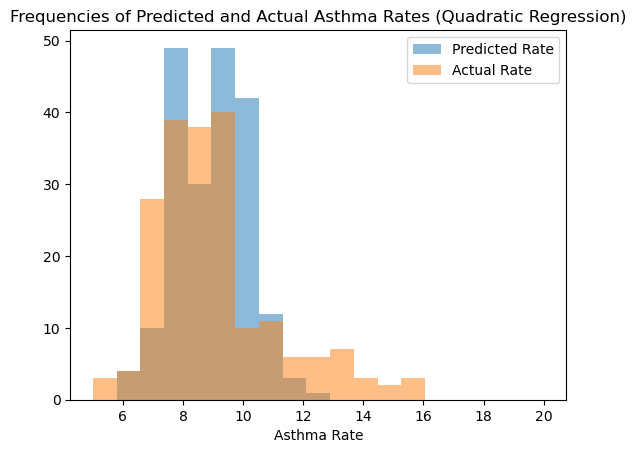


Figure 10: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with Quadratic Regression from Predicted Agriculture Rates

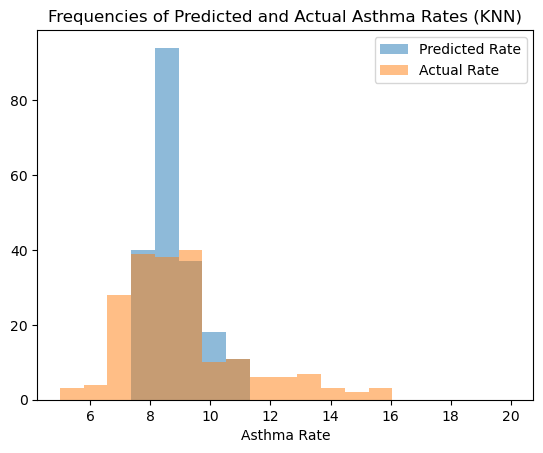


Figure 11: A Histogram of Frequencies of Actual Asthma Rates and Those Predicted with KNN from Predicted Agriculture Rates

Changing the method used to handle invalid values made a large difference to the results. Given the range of R2 scores across the different models and methods of handling invalid values, it is not clear how much of the variation in asthma rates can be explained by land cover information.

Linear regression yields the most consistent results. Across all methods, it yields an R2 score of between 8 and 10%, which is a narrower range than quadratic regression (6–31%) and the KNN range (3–12%). This is somewhat surprising, as one might expect KNN to be more consistent across methods involving either removing or recoding values, as it is a local regressor and such methods would only influence some, not all, predictions.

**Conclusion**

There were limitations with this data due to a lack of expertise on the EnviroAtlas. It is unclear what those invalid values truly represent, and as such, what is the best method of handling them. Furthermore, the existence of so many invalid values for one of the attributes raises questions about the overall quality of the EnviroAtlas data. As such, the results of this analysis are inconclusive.

In future iterations of this project, using a wider range of geographic features than just land cover and predicting asthma rates rounded to the nearest percent might yield more conclusive results than using such specific values. Similarly, information for a wider variety of census tracts would be valuable. This project was limited due to a low number of both features and observations. Expanding to every site the EnviroAtlas has high resolution data on, as well as integrating other datasets than simply land cover, could yield more consistent results. It would also be preferable to use cross validation, as that would result in more easily comparable metrics than a single run.

While the extent of the relationship is inconclusive, this project suggests that land cover information can provide some insight into the variation of asthma rates. This may also indicate that land cover information can explain some of the variation in other health outcomes, such as respiratory diseases other than asthma. By bringing in other features and more locations, researchers may be able to gain greater insight into the impact of geographical features on health outcomes.

Bibliography

Abolmaali, S., & Shirzaei, S. (2021). A comparative study of SIR Model, Linear Regression,

Logistic Function and ARIMA Model for forecasting COVID-19 cases. *AIMS public health*, *8*(4), 598–613. https://doi.org/10.3934/publichealth.2021048

CDC. 2022. PLACES: Local Data for Better Health, Census Tract Data 2021 release.

<https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health->Census-Tract-D/373s-ayzu

Environmental Protection Agency. (n.d.). EPA Enviroatlas. Retrieved March 3, 2023, from https://www.epa.gov/enviroatlas/about-data

Dick S, Doust E, Cowie H*, et al*. Associations between environmental exposures and asthma

control and exacerbations in young children: a systematic review. *BMJ Open*2014; **4:**e003827.doi: 10.1136/bmjopen-2013-003827

Lovasi, G. S., Quinn, J. W., Neckerman, K. M., Perzanowski, M. S., & Rundle, A. (2008).

Children living in areas with more street trees have lower prevalence of asthma. *Journal of epidemiology and community health*, *62*(7), 647–649. <https://doi.org/10.1136/jech.2007.071894>

Tanuwijaya, Julius & Hansun, Seng. (2019). LQ45 Stock Index Prediction using k-Nearest

Neighbors Regression. International Journal of Recent Technology and Engineering. 8.

10.35940/ijrte.C4663.098319.

Zein, J. G., Wu, C. P., Attaway, A. H., Zhang, P., & Nazha, A. (2021). Novel Machine Learning

Can Predict Acute Asthma Exacerbation. *Chest*, *159*(5), 1747–1757. <https://doi.org/10.1016/j.chest.2020.12.051>

**Implementation Appendix**

The datasets for this paper were retrieved through a combination of downloaded CSVs and API calls. The CDC Places dataset and EnviroAtlas datasets came in the form of CSVs, and the population data came from API calls to the Census API to retrieve estimates from the American Community Survey. I used the ACS, rather than the decennial census, as the EnviroAtlas and CDC data both came from 2019, and there was no census data for that year. This may be a source of error, as the ACS data is typically less accurate than the decennial Census.

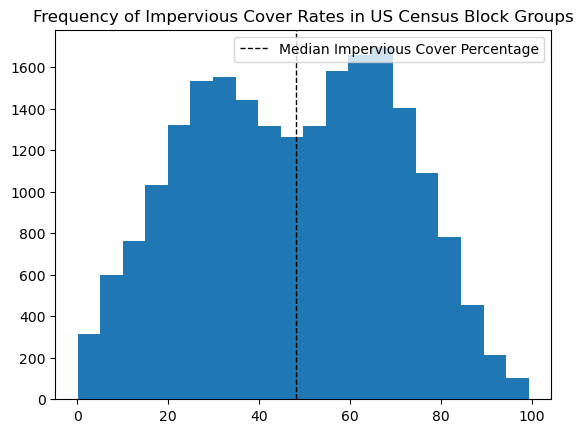


Figure A1: Frequency of Impervious Cover Rates

Figure A1 indicates that there is a bimodal distribution for the impervious land cover rate. Since all the values fell within the expected range, this required no further processing. Similarly, Figure A2 indicates a bimodal distribution for green space cover rates, with values all falling within the expected range and in no need of further processing.

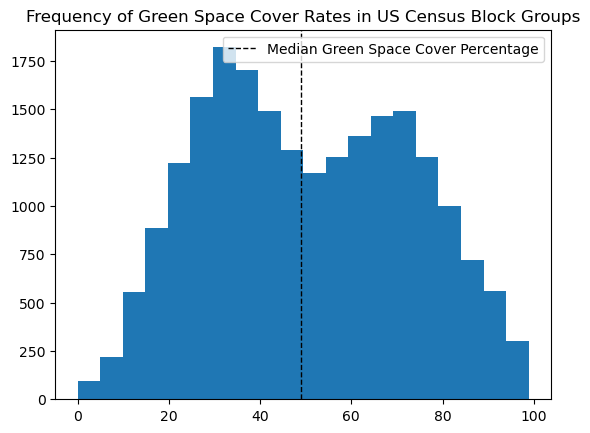


Figure A: Frequency of Green Space Cover Rates

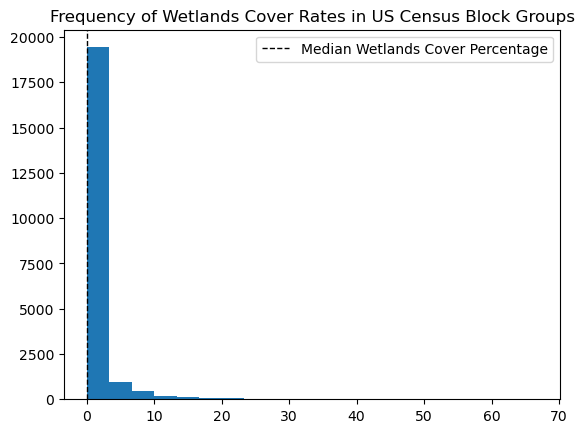


Figure A: Frequency of Wetlands Cover Rates

Like the tree cover rates, impervious land cover rates, and green space cover rates, the wetlands cover rates, as can be seen in Figure A3, required no further preprocessing. This variable is not normally distributed, with the majority of values close to 0. These histograms all supported my belief that a model that assumed a normal distribution would not be the best option.

Given that the EnviroAtlas data is on a census block group level, I first needed to aggregate to a census tract level in order to merge the EnviroAtlas and CDC Places datasets. However, the data provided was in percentages of the total area, rather than a unit of square area. Therefore, I needed to first convert to raw numbers, using the ShapeArea attribute, which provides the census block group area in square metres. That enabled me to sum all the attributes to reach the tract level data, and then convert back to percentages.

To determine the number of neighbours, I used validation curves. Figures A4, A5, and A6 depict the validation curves for the dataset with invalid values recoded to 0, the dataset with invalid values excluded, and the dataset with invalid values filled through a prediction with linear regression respectively. I selected k neighbours based upon a visual inspection of the curves to determine the point at which the disparity between them seemed to be narrowing or becoming more consistent.

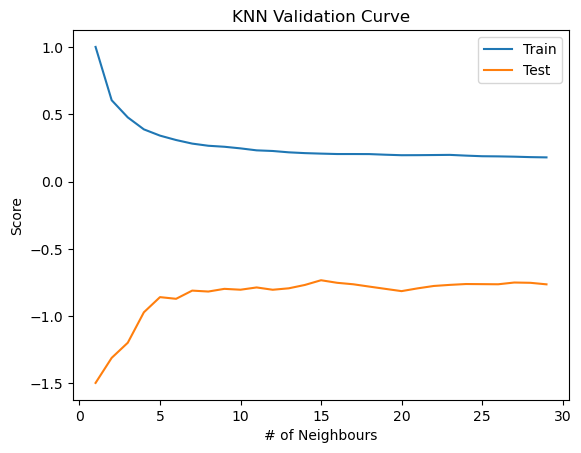


Figure A: Validation Curve for Recoded Values

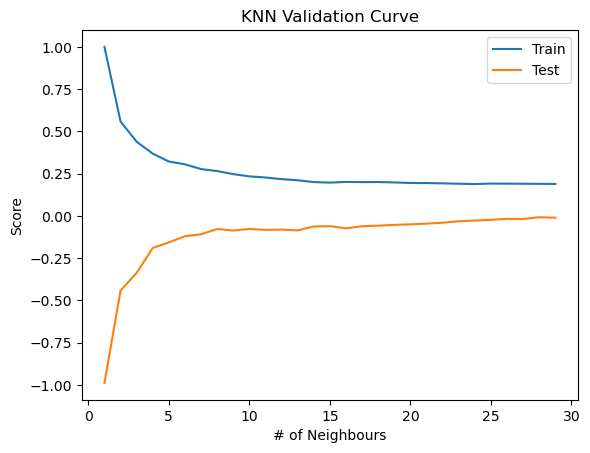


Figure A: Validation Curve for Limited Values

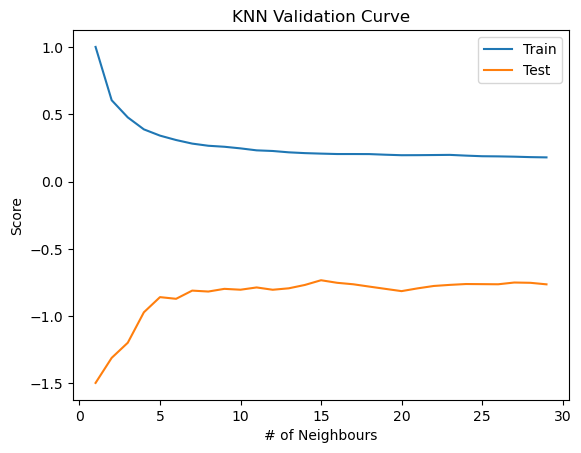


Figure A: Validation Curve for Estimated Values

Removing records containing negative PercentAgriculture values dropped the number of observations I had available to less than 500. This was the reason I decided to explore multiple methods of recoding the values, in hopes of larger datasets yielding more conclusive results.