How could urban green space impact people's physical and mental health?

Anvay Vats, Benji Duan, Hao Zhu, Liam Lee Kitt

Executive Summary

This project aims to answer the following question: how could urban green space impact people's physical and mental health? We chose this question based on where we believed the data was most applicable and showed the most trends. Studying the relationship between health and tree diversity, nativity, condition and size, we find small but significant effects of tree diversity and tree health on improving the probability of having good mental or physical health.

Methodology

A logistic regression model is used for the analysis. For example, consider the task of investigating the relationship between green space and mental health. Let X_i be the mental health status for each person in a specific area.

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Suppose the mental health status of each person is independent and identically distributed according to the above distribution, then the number of people having good mental health in a specific area follows a Binomial(n, p) distribution. Since we have easy access to the number of people having good mental health in an area, it is natural to use a logistic model to answer the research question. We model the relationship between mental health and green space as follows:

$$\log\left(\frac{p}{1-p}\right) = X\beta$$

where X is the design matrix for the features we consider.

The regression results are summarized in Table 1 and Table 2 in the appendix. Having 1 more tree species leads to an approximately 0.03% increase in the odds of having good mental health and an approximately 0.07% increase in the odds of having good physical health. Having 1 more dead/dying tree leads to an approximately 0.03% decrease in the odds of having good mental health and an approximately 0.01% decrease in the odds of having good physical health. Having 1 more healthy tree leads to an approximately 0.02% increase in the odds of having good physical health.

Data

Dataset used

A total of 4 datasets were used in this project: a dataset of 5 million city trees from cities across the US was used to construct the features; data from the PLACES project provided estimates of health measures by zip code across the United States; data from simplemaps.com contained mappings from zip code to latitude and longitude; an

external dataset from the US census bureau included the adult population size for different zip codes in the United States and can be accessed from https://www.kaggle.com/datasets/census/us-population-by-zip-code. These data sets were analyzed at the zip code tabulation area level to better preserve the richness of the provided datasets while also limiting the granularity of the data as an attempt to reveal more subtle relationships between the response variable and the features while controlling for the noise. Matching the datasets by zip code is also a natural choice given that observations from all the data sets contained zip code information. We hope that the choice of conducting the analysis at the zip code level could help inform policy makers on how to improve the mental and physical health of their citizens through urban green space.

Data cleaning and feature construction

Community estimates of health measure across the United States

Data from the PLACES project contained estimated percent prevalence for adults who report 14 or more days during the past 30 days during which their mental or physical health was not good. We find that all the observations were collected in 2020, contained no missing data and were all recorded as crude prevalence, meaning there is no need to do any adjustments. The PLACES data set is then matched with the adult population data set to estimate counts of people who are mentally or physically healthy for each zip code.

Planted date and retired date

Before conducting any further analysis, any trees planted after 2020 or retired before 2021 are dropped from the data set since all observations in the health measures data set were recorded in 2020. The time frame of after 2020 and before 2021 is chosen to be conservative. On the other hand, it is reasonable to hypothesize that trees might have an effect on people's mental and physical health over time and include some aggregate measure for the number of years that trees have been planted for each zip code. However, the planted date variable included a large amount of missing data. With no clear way to do accurate data imputation, we deemed it not worth the trade-off to include planted date as a feature in the model.

Zip code

We note that the tree data set contained much more coordinate information than zip code information. Since our analysis relies on producing aggregated data for each zip code, it is beneficial to have more available data to work with. Coordinates are therefore converted to zip codes by finding its nearest latitude longitude centroid. A K-dimensional tree is constructed from the dataset of zip codes and coordinates. The tree is then searched to find the nearest neighbor. Compared to other methods like using the geopy package to find exact zip codes for a given coordinate, this method is much faster and does not need to consider issues like API rate limits, but the conversion from coordinates to zip codes may not be exactly accurate. It is nonetheless reasonable: the implied zip code will be at least close to the real zip code, and the proximity means that the condition of green space in one area could affect the response which may be in a different area. Using coordinate implied zip codes resulted in about a 4 times increase in the amount of available zip codes. All feature constructions in the project take advantage of the implied zip codes.

Tree diversity

For each zip code, the number of unique tree species is counted. Zip codes that contain no data for tree species were not considered. The resulting data frame is 1605 rows $\times$ 2 columns where the columns are zip code and count of unique tree species.

Tree condition

For each zip code, the number of trees having different health conditions is counted. There are 7 categories for tree health: excellent, fair, good, poor, dead/dying, dead, n/a. The dead/dying category is merged with the dead category due to their similarity. The n/a category is dropped to help with interpretability and deal with collinearity: the total number of trees for each zip code can be inferred from other features that will be included in the model. The resulting data frame is 1021 rows $\times$ 6 columns.

Tree nativity

For each zip code, the number of trees having different nativity is counted. There are 3 categories in the data set: introduced, naturally occurring and no info. The no info category is dropped to help with interpretability and deal with collinearity. The resulting data frame is 1844 rows $\times$ 3 columns.

Tree size

For each zip code, the number of trees having different sizes is counted. Size is defined as one of six groups of trunk diameter at breast height: 0 to 15.24 cm; 15.24 to 30.48 cm; 30.48 to 45.72 cm; 45.72 to 60.96 cm; 60.96 to 76.2 cm; and more than 76.2 cm. We hypothesize that trees that have different sizes could have different effects on people's mental or physical health. The resulting data frame is 1562 rows $\times$ 6 columns.

Appendix

Table 1: Regression results for mental health

	coef	std err	z	P>|z|
const	1.7128	0.001	1221.378	0.000
common_name	0.0003	1.65e-05	18.660	0.000
dead/dying	-0.0003	1.58e-05	-20.732	0.000
excellent	8.209e-05	1.8e-06	45.630	0.000
fair	-8.035e-06	1.37e-06	-5.879	0.000
good	4.368e-06	6.74e-07	6.486	0.000
poor	-2.279e-05	6.05e-06	-3.763	0.000
introduced	-2.405e-06	6.93e-07	-3.470	0.001
naturally_occurring	-2.411e-05	9.81e-07	-24.587	0.000
0 to 15.24 cm	-4.412e-05	1.1e-06	-40.077	0.000
15.24 to 30.48 cm	9.074e-05	1.93e-06	47.069	0.000
30.48 to 45.72 cm	-2.724e-05	3.93e-06	-6.932	0.000
45.72 to 60.96 cm	7.528e-06	7.08e-06	1.064	0.287
60.96 to 76.2 cm	5.124e-05	1.01e-05	5.094	0.000
more than 76.2 cm	0.0001	9.29e-06	15.086	0.000

Table 2: Regression results for physical health

	coef	std err	z	P>|z|
const	2.1651	0.002	1265.337	0.000
common_name	0.0007	2.07e-05	33.700	0.000
dead/dying	-0.0001	1.95e-05	-6.077	0.000
excellent	0.0002	2.59e-06	77.003	0.000
fair	1.258e-05	1.7e-06	7.408	0.000
good	8.337e-06	8.3e-07	10.045	0.000
poor	4.781e-05	7.48e-06	6.390	0.000
introduced	-3.054e-06	8.59e-07	-3.554	0.000
naturally_occurring	-5.899e-05	1.22e-06	-48.322	0.000
0 to 15.24 cm	-8.422e-05	1.33e-06	-63.182	0.000
15.24 to 30.48 cm	0.0001	2.38e-06	59.509	0.000
30.48 to 45.72 cm	-3.783e-05	4.79e-06	-7.890	0.000
45.72 to 60.96 cm	-2.734e-05	8.66e-06	-3.157	0.002
60.96 to 76.2 cm	2.03e-05	1.23e-05	1.652	0.098
more than 76.2 cm	0.0003	1.15e-05	23.361	0.000