How Greenspace Generates the Green

Andrew Hill, Rohun Iyer, Angel Liu, Sarah Sachs

New York City is a city renowned for its parks. From the vast globally-known Central Park to the Community Gardens in the East Village, each park brings its own influence and social good. The introduction of a park to a community generally sees a positive impact in social health, property values and tax revenue (Lin, 2016). Assessing the relationship between parks and their surrounding communities can provide valuable insight into the equity of park benefits and influence future affordable housing and zoning projects.

In this paper we examine one aspect of a park's influence on the neighboring communities: property value. A well-documented effect of a park or greenspace is the increased value of proximate properties, such as Central Park raising values 800 percent more than expected over an 18 year period (Lin, 2016). Following past research, we hypothesize a decline in Airbnb listing prices as the walking distance increases while controlling for the park's amenities, the neighborhood demographics, and Airbnb listing amenities.

In order to formulate our analysis, we gathered data from three separate datasets. We chose Airbnb listing data instead of general real estate prices because the former is more granular, the data is unified, and geo-located. We also selected green spaces data from Department of Information Technology & Telecommunications and Park Amenity data from the Department of Parks and Recreation NYC Open data. Demographic data came from American Community Survey (ACS) for 2016.

Integral to our analysis was gathering a large amount of Airbnb data. We culled through Airbnb's publicly available data and pulled New York City listings within the date range of November 2017 to October 2018. This selection yielded 582,308 listings with associated data points on the number of guests the listing accommodates, the square footage of the listing, and many other features which could prove useful in our final analysis. From that large set of listings, we removed outliers such as listings with more bedrooms than beds, listings with prices above $1,000, unique property types (e.g. 'Igloo'), listings where the location (in longitude and latitude) was not entirely exact, and other statistical and structural outliers. The final clean dataset had less than 50% of the original observations.

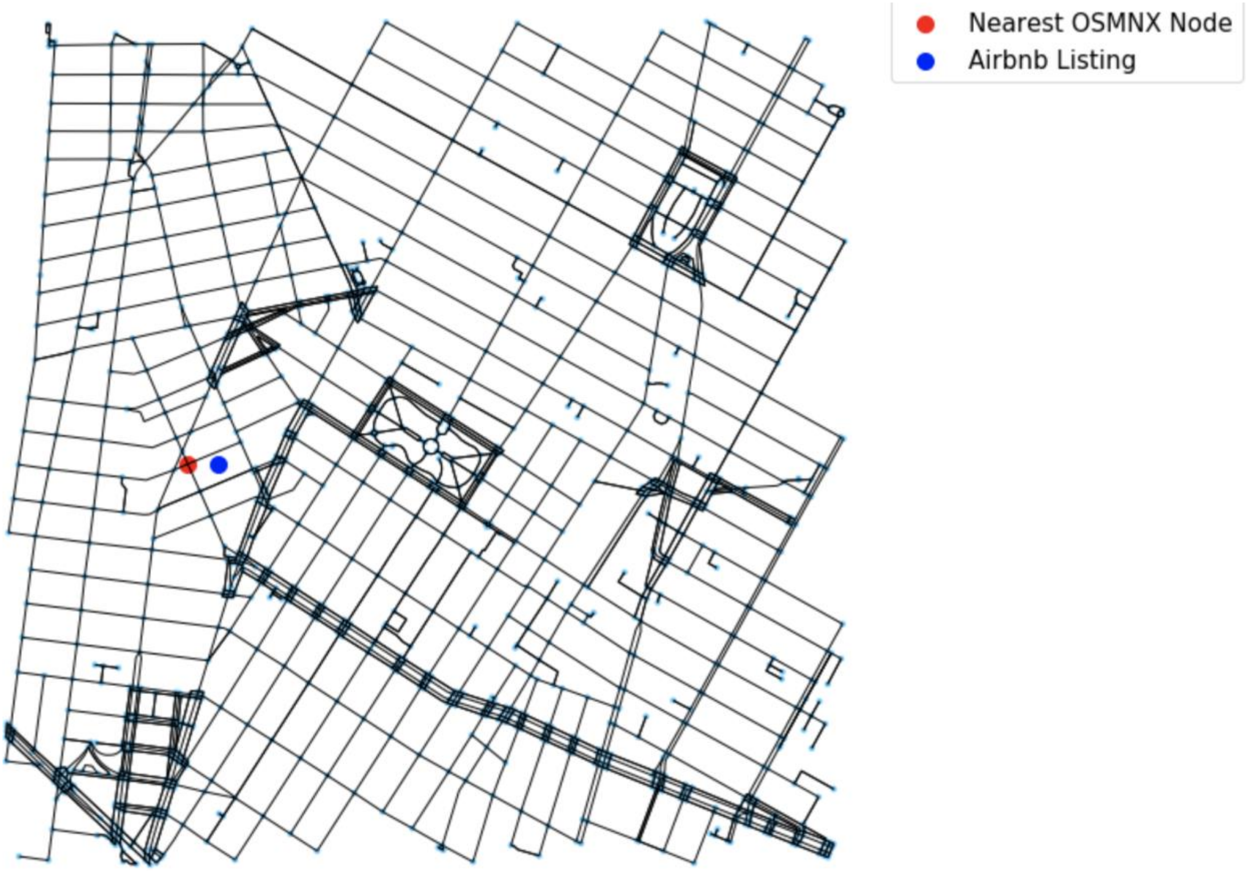
We worked with park zones and greenspace data, to find relevant spaces across NYC and add amenities for future analysis. The amenities under study include barbecues, basketball courts, eateries, playgrounds, tennis courts, ice-skating rinks, and dog runs. We were unable to choose amenities that could affect the desirability of a park like Wi-Fi data which is not included in NYC Park Open data. Parks can be partitioned into two categories: Active and Passive. Active parks generally include recreational amenities such as basketball and tennis courts and baseball diamonds whereas Passive parks tend to have more open green spaces, dog-friendly areas and eateries. Literature suggests Active parks more greatly contributes to a neighborhood's health than Passive parks (Bolitzer, 2000).

Finally, we selected the 2016 ACS data as our 'Demographic' dataset. Although we could have achieved a higher level of geographic granularity with the 2010 decennial census, we believed both the granularity of census tracts and the more contemporary data would better serve our analysis. From the data we specified the counties of New York City (i.e. the Manhattan borough), Kings (Brooklyn), Richmond (Staten Island), Queens and Bronx. From this dataset, we created four features, bound to census tracts, for our analysis: Total Population, Median Age, Percentage with a Bachelors Degree, and Median Household Income. Our next step was to cluster these demographic statistics into four groups, using the K-Means Algorithm, so we could analytically pinpoint geographic groupings of census tracts when selecting which parks will enter our final analysis.

This clustering allowed us to pare down the overwhelmingly large dataset to fewer observations surrounding certain parks. We chose some parks because of the homogeneity of the census tracts around it (e.g. Washington Square Park) and complemented those choices with more diverse parks (e.g. Fort Greene Park). Along with pinpointing our choices based on diversity, or lack thereof, we focused on choosing parks so our total dataset would include equal amounts of each cluster grouping.

Through the Open Street Map Network (OSMNX) library on python, we ran a network path analysis between different points across New York City. We chose to use OSMNX because of its ability to measure walking distance versus a straight-line distance. Realizing there exists significant differences between walking distance and direct distance, and that this disparity would skew our final analyses and regressions. Any proximity tests we carried out were solely done with OSMNX. Our reasoning for paring down our dataset was largely a result of the computational complexity of the walking distance function.

While OSMNX provided an incredible opportunity to calculate proximity through walking distances, it comes with its own limitations. OSMNX is primarily used for analyzing street networks and relies exclusively on a network of nodes and edges. Using the geographic coordinate system, there are theoretically infinite locations and a finite number of nodes within OSMNX. Unfortunately, our Airbnb listing locations did not necessarily map to a node. As such, in order to calculate the walking distance from an Airbnb listing to a specified park, we found the nearest node to that listing and then carried out the shortest path analysis. In most cases, this did not prove to be an issue, however as we got to listings in the outer boroughs, there was no guarantee an Airbnb would be within a negligible distance to an OSMNX node. Below is an OSMNX map of the area surrounding Washington Square Park and an example of an Airbnb listing location and the nearest OSMNX node:



In order to test our hypothesis, we decided linear models would provide the best blend of interpretability, low computational complexity, and analytical power. But before the analysis, we further cleaned our final merged dataset. This cleansing included renaming column headers, transforming features to achieve a normal or near-normal distribution, creating dummy variables from categorical ones, compositing groups of categorical variables, and outlier removal. After checking for collinearity between features, we selected our final independent variables which had Airbnb characteristics, demographic data, and park characteristics. We then standardized these variables, as they were not cohesively scaled. Finally, we conducted analysis using six types of models. Using Ordinary Least Squares (OLS), Ridge, and Lasso regression, we conducted tests on our data with two types of independent variables: Airbnb listing price and Airbnb listing price with cleaning fees. We hypothesized the cleaning fees were a part of the consumer's overall price estimation of the listing. This hypothesis proved to be true as the latter type of independent variable models performed far better than the former, in terms of R-squared. We decided to interpret the results of only the OLS model rather than the Lasso and Ridge models, despite their performing slightly better than the OLS model. This decision was a result of prioritizing interpretability, as we could not sufficiently isolate and quantify the effect of an Airbnb listing's distance to a park to its listing price. However, if one desires to predict a listing's price with our methodology, we recommend the Ridge model, which had an R-squared commensurate to the Lasso model but was far less computationally expensive in comparison. Our OLS model showed that we could reject the null hypothesis at a 5% level and that a listing's distance to the park has a statistically significant negative effect on a listings price. The full results are shown in the table below.

OLS Model Results		
Variable Type	**Variable**	**Relationship/ Significance**
	Super Host (Binary)	+***
	Accommodation Capacity	+***

Airbnb	Number of Guests Included	+**
	Maximum Number of Nights	+***
	Review Rating	-***
	Reviews Per Month	-***
	Instant Bookable (Binary)	-*
Demographic	Census Tract Population	+***
	C.T. Median Age	+***
	C.T. Median Household Income	-***
Park	Log (Distance to the Park)	-***
	Number of Eateries	-***
	Number of Playgrounds	-***
	Active Park (Binary)	-***
The Dependent Variable is Airbnb Listing Price with Cleaning Fee '+' and '-' signify a statistically significant positive and negative relationship, respectively * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001		

Our analysis substantiates our initial hypothesis, as an Airbnb listing's distance from a park increases the listing's price decreases, holding the Airbnb amenities, demographic data, and park amenities constant. This result leads us to believe more should be done to make access to a vital public resource more open to less well-off people, as they are being priced out of neighborhoods near parks. More research can be done by analyzing the density, along with the proximity, of multiple parks near an Airbnb listing. Also, one could expand the dataset to include more parks and even more cities.

Model	R-Squared	Alpha
OLS	0.545	N/A
OLS with Cleaning Fee	0.579	N/A
Ridge	0.560	192
Ridge with Cleaning Fee	0.585	63
Lasso	0.560	0.226
Lasso With Cleaning Fee	0.584	0.312

References

1. Boeing, G. 2017. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." *Computers, Environment and Urban Systems* 65, 126-139. doi:10.1016/j.compenvurbsys.2017.05.004

2. Bolitzer, B., & Netusil, N. R. (2000). The impact of open spaces on property values in Portland, Oregon. *Journal of Environmental Management,* 59, 185–193.

3. Espey, M., & Owusu-Edusei, K. (2001). Neighborhood parks and residential property values in Greenville, South Carolina. *Journal of Agricultural and Applied Economics*, 33(3), 487–492.

4. Inside Airbnb. "Inside Airbnb: New York City October 2017-November 2018". Airbnb, 2018. Web. 20 November 2018.

5. Lin, I-Hui, "Assessing the Effect of Parks on Surrounding Property Values Using Hedonic Models and Multilevel Models" (2016). *Theses and Dissertations*. Paper 1291.

6. NYC Department of Information and Technology & Telecommunications. "Open Space (Parks)". NYC Open Data, 2018. Web. 2 December 2018.

7. United States Census Bureau / American FactFinder. *2015 American Community Survey*. U.S. Census Bureau's American Community Survey Office, 2018. Web. 2 December 2018