A Geo-Spatial Study of Bikeshare Station Locations

With the massive rise in popularity of bikeshare services in recent years, there is a natural concern of whether these programs are operating equitably. Many efforts to address this concern thus far have utilized either survey-based methods or performed analyses similar to ours on substantially smaller samples. We make use of a standardized information release protocol used by bikeshare programs across the United States to scrape geographic coordinates of over 5000 bikeshare stations, and, joining this data with census tract level demographic data, perform Poisson regression to understand the influences of demographics on the extent to which communities are served by bikeshare programs.

Introduction

A central concern around bikeshare programs in the U.S., especially those funded by public money, is whether such programs are effectively serving traditionally underserved communities. Previous research on this topic has been largely centered on community surveys, allowing for understandings of the motivations of possible bikeshare users at the level of observation of individual people (McNeil). However, by design, these studies must be limited in scope, focusing on a small set of bikeshare programs and randomly sampling a small portion of the relevant population. Shifting the level of observation from people to communities, we can explore how the demographic characteristics of U.S. communities inform our knowledge of the number of bikeshare stations in them. Several studies have taken this approach, though are limited to a small number of programs (Smith; Ursaki). In this study, we construct a dataset containing demographic predictors at the census tract level of observation, along with the number of bikeshare stations in that community for over 5000 bikeshare stations in 51 programs. Modeling this dataset using Poisson regression, we find that the demographic characteristics of a community are deeply intertwined with the extent to which that community is served by bikeshare services.

Data

The dataset used in analysis contains demographic information by census tract as well as the number of bikeshare stations in that census tract. This dataset was constructed by a join between two different datasets. The first was scraped using the statistical software *gbfs*, which makes use of a standardized information release protocol by the same name used by bikeshare programs nationwide, and contains the latitudes and longitudes of 5451 bikeshare station docks from 51 bikeshare programs throughout the United States (Couch; Chan-Norris). The other dataset was obtained from Harvard University's Opportunity Insights, a research institute providing data including demographic information at the census tract level of observation. The dataset contains 74,123 observations of 38 variables, 13 of which contain demographic measures from 2010 or later. For every observation in the first dataset, we reverse geocoded the geographic coordinates to their respective census tracts to produce a table of counts of bikeshare stations per census tract. This table was then “matched up” with the second dataset such that, in addition to demographic measures, each row had an associated number of bikeshare stations in that tract recorded, to produce the “master” dataset.

We first performed forward subset selection on a multiple Poisson regression of number of bikeshare stations in a census tract, finding the most effective explanatory subset of size n explanatory variables for values of n from 1 to 12, using a Bayesian information criterion. Poisson regression is a technique applicable when the output is a natural number (e.g. count data). Since we would like to predict the *number of* bikeshare stations in a census tract, this model makes the most sense, given that negative bikeshare stations has little meaning. The forward subset selection algorithm operates as follows:

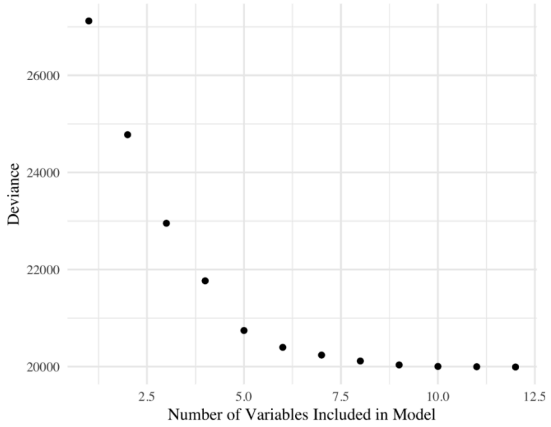
For i from 1 to n:

- *Fit n-i models of size i by adding 1 variable into the present model at a time.*
- *Use an information criterion to select the "best" model.*
- *Fix the newly chosen variable in the model.*

The forward subset selection algorithm stops adding in variables at 1 less than the number of possible explanatory variables, since the choice of the last variable is trivial—we can add the last one in manually. See the discussion section for a consideration of the limitations of this algorithm.

We then performed cross-validation on the size of the subset of variables, using deviance as an information criterion to choose the optimum number of variables to include in the model in order to explain a significant amount of the variation in the data while also restricting the number of included variables to

make model interpretation more straightforward.



This curve is highly smooth, so the choice of an "elbow" is largely arbitrary. We will explore the full model in more depth.

Results

We will first discuss the variables that best explain the relative presence of bikeshare stations in a given census tract. In this model, variables are scaled to a standard normal such that the exponent of their coefficients is proportional to their relative explanatory value, with coefficients a_i on the right-hand side exponentiated such that a unit increase in the relevant predictor (scaled to a standard normal) corresponds to multiplicative scaling in the predicted number of bikeshare stations by a_i. Thus, a coefficient below 1 indicates a *decrease* in the predicted value of the output. We emphasize, though, that coefficients from the Poisson are only additive on the log-odds scale. Thus, if one wanted to predict the number of stations using several predictors, they must construct a linear combination of predictors using the coefficients on the left-hand side, and *then* exponentiate.

Predictor	Coefficient	exp(Coefficient)	P-Value
Intercept	-5.4587111	0.0042590	0.0000000
Rent	0.0009435	1.0009439	0.0000000
% in Poverty	2.9867864	19.8218810	0.0000000
% College-Educated	5.1989059	181.0740163	0.0000000
% Single Parents	1.1880556	3.2806959	0.0000000
# Jobs w/in 5mi	0.0000033	1.0000033	0.0000000
Public School Quality	-0.2837444	0.7529591	0.0000000
Median Income	-0.0000115	0.9999885	0.0000000
Population Density	-0.0000104	0.9999896	0.0000000
% Nonwhite	0.7190488	2.0524800	0.0000000
# High-Paying Jobs w/in 5mi	-0.0000037	0.9999963	0.0000000
% Foreign-Born	-0.3614245	0.6966832	0.0108625
Job Density	-0.0000003	0.9999997	0.0503055

Interestingly, the first two variables initially chosen to be added to the model were purely economic measures, and the next two are known to be highly correlated with socioeconomic class: the cost of rent, the percentage of residents living below the poverty line, a measure of educational attainment, and the proportion of single parents in the tract, respectively (each with a scaling factor above 1). As apparent in the coefficients for educational attainment, the level of education in the census tract (as measured by the percentage of residents with a 4-year college degree or greater) is an especially valuable predictor for the number of bikeshare stations in the census tract. This is particularly interesting because of the nature of the forward subset selection algorithm. As these economic measures are added into the model one by one, their variability is controlled for in subsequent searches for the next best explanatory variable. Thus, we find that these different measures provide unique explanatory value in understanding the factors that influence where bikeshare stations are located. In general, the finding that increases in measures of economic wellness in a community are highly correlated with an increase in the number of bikeshare stations is unsurprising; the majority of bikeshare programs are privately-owned, paid services, and often are more expensive and unpredictable services than public transportation options.

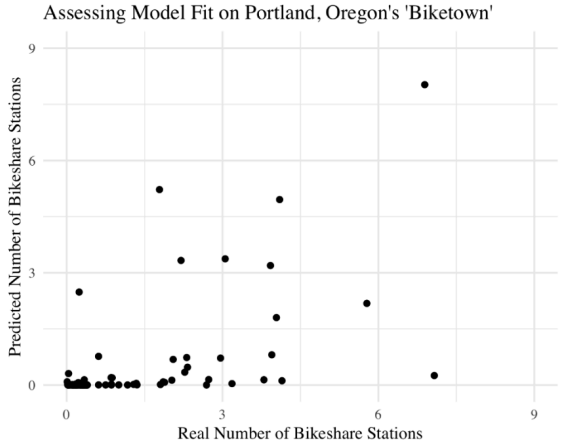
The next most important predictor is the number of jobs within a 5 mile radius of the census tract. We can treat this predictor as a sort of interaction effect between concentration of jobs and population density, both of which are statistically significant predictors themselves when included in larger subsets.

Other predictors statistically significant at the .05 significance level include a measure of school quality, median household income, population density, percentage of non-white residents, the number of high-paying jobs within 5 miles, and the percentage of foreign-born residents. We note that the coefficient for the percentage of

foreign-born residents is the only one substantially below 1; communities with high percentages of residents born outside of the U.S. are less effectively served by bikeshare services.

A Case Study Cross-Validation

In fitting our model, we excluded Portland, Oregon from our training dataset. A bikeshare program known as Biketown operates in Portland, with a total of 146 bikeshare stations distributed across 171 census tracts. We use Biketown as a case study to cross-validate our model.



For one, we note that the majority of census tracts have no bikeshare stations in them, and our model correctly predicts this outcome in the majority of cases. However, in census tracts with 1 or 2 bikeshare stations in them, which is a majority of cases when there *is* a bikeshare station in a given census tract, the model consistently predicts that there are no bikeshare stations. In cases of census tracts with 3 or more bikeshare stations, though, the model often outputs predictions well above zero. We can conclude from this case study that the model has modest predictive power.

Discussion

There are several limitations to the data that could affect our results:

- Though every census tract in the US is included in the dataset, only bikeshare stations from programs that release valid *.json* in the generalized bikeshare feed specification format are included in the dataset. Thus, for some number of census tracts that actually have a nonzero number of bikeshare stations, the recorded number of bikeshare stations will be zero. Since *gbfs* is only a technical specification, though, we can make the assumption that bikeshare programs that do not release *gbfs* feeds do not differ systematically in the geographic and demographic distribution of their bikeshare stations from programs that do. Thus, model coefficients are uniformly closer to zero, and therefore less likely to be significant due to this insufficiency.
- Due to computational limitations, we could only perform forward subset selection as opposed to best subset selection. By fixing previous sources of variation accounted for when considering new variables to add to the model, the forward subset selection algorithm could have chosen a number of variables early on that were highly collinear with other, more effectively explanatory variables.

Nonetheless, we find that the demographic characteristics of a community are deeply intertwined with the extent to which that community is served by bikeshare services. Namely, several measures of economic wellness are significant predictors of the number of bikeshare stations, indicating that each contributes unique explanatory value on our output. This finding poses a problem for publicly funded bikeshare programs, in that they serve communities based on projected economic demand more so than they are a solution to the logistical problem of moving citizens around the city.

Future work on this topic could benefit from simultaneous usage of both the traditional survey-based, individual level of observation, and this demography-based approach at the community level of observation. Such work could synthesize the social and narrative clarity offered by the former with the objective integrity of the latter, making a compelling argument to a wider variety of audiences and offering actionable insights for both the direction of currently existing bikeshare programs and the development of future ones.

Bibliography

Chan-Norris, Jesse. “General Bikeshare Feed Specification.” *North American Bikeshare Association*, 29 Nov. 2018, github.com/NABSA/gbfs.

Couch, Simon, and Kaelyn Rosenberg. “gbfs.”, 1.1.0, *Comprehensive R Archive Network*, 21 Sept. 2018.

Harvard University. “Neighborhood Characteristics by Census Tract.” *Opportunity Insights*, 2018, opportunityinsights.org/data/.

McNeil, Nathan et. al. Evaluating Efforts to Improve the Equity of Bike Share Systems. *National Institute for Transportation and Communities*, 2016, pp. 1–20.

Smith, C. Scott, Jun-Seok Oh, and Cheyenne Lei. "Exploring the Equity Dimensions of US Bicycle Sharing Systems." *Transportation Research Center for Livable Communities*. Western Michigan University Kalamazoo, MI, 2015.

Ursaki, Julia, and Lisa Aultman-Hall. "Quantifying the equity of bikeshare access in US cities." *95th Annual Meeting of the Transportation Research Board, Washington, DC*. 2016.