Predicting Bike Usage for New York City's Bike Sharing System

Divya Singhvi,[1] Somya Singhvi,[1] Peter I. Frazier,[1] Shane G. Henderson,[1]
Eoin O' Mahony,[2] David B. Shmoys,[1] Dawn B. Woodard[1]
[1]School of Operations Research and Information Engineering, [2]Department of Computer Science
Cornell University, Ithaca, NY 14850, USA
{ds576@, ss989@, pf98@, sgh9@, eoin@cs., david.shmoys@, woodard@}cornell.edu

Abstract

Bike sharing systems consist of a fleet of bikes placed in a network of docking stations. These bikes can then be rented and returned to any of the docking stations after usage. Predicting unrealized bike demand at locations currently without bike stations is important for effectively designing and expanding bike sharing systems. We predict pairwise bike demand for New York City's Citi Bike system. Since the system is driven by daily commuters we focus only on the morning rush hours between 7:00 AM to 11:00 AM during weekdays. We use taxi usage, weather and spatial variables as covariates to predict bike demand, and further analyze the influence of precipitation and day of week. We show that aggregating stations in neighborhoods can substantially improve predictions. The presented model can assist planners by predicting bike demand at a macroscopic level, between pairs of neighborhoods.

Introduction

Bike-sharing systems are in place in several cities in the world, and are an increasingly important support for multimodal transport systems (Shaheen, Guzman, and Zhang 2010). The objective of our research is to develop an accurate prediction model that estimates demand for bike trips between pairs of locations. Our model can be used to make predictions even when one or both of these locations do not currently have bike stations in place, and can be used as a planning tool when deciding how to expand a city's bike sharing system. By providing pairwise predictions of the demand for trips between each origin-destination pair, we provide not just an estimate of how much incoming and outgoing demand will be realized if a new bike station were built, but also where the incoming/outgoing demand will originate/terminate, predicting this new station's effect on the existing network.

We predict bike demands by running regression models with covariates that include population, weather and taxi usage. Since bike usage is affected by temporal and weather characteristics (Imani et al. 2014), we focus only on the morning rush hours of 7:00 AM-11:00 AM and test our model in two different scenarios: dry weekdays and rainy weekdays. We show that, although accurately predicting flows at the station pair level is difficult, a simple aggregation at the neighborhood level can substantially improve flow predictions. Neighborhood definitions (geographical boundaries) are obtained from an external source and are based on demographic and economic variables. Pairwise neighborhood predictions can then assist decision-makers in deciding in which neighborhoods to expand their network. We validate our prediction model by predicting demands for the existing network of New York City's Citi Bike system.

Over the past few years, several studies have analyzed factors affecting bike flow and usage. (Rixey 2013) analyzes the impact of demographics and built environment characteristics on bike usage and concludes that population density critically affects bike usage. (Buck and Buehler 2012) explores the influence of population, bicycle lanes and retail destinations on bike usage in Washington DC.(Imani et al. 2014) analyzes the BIXI system in Montreal using meteorological data, temporal characteristics and built environment attributes. (Etinne and Latifa 2012) uses model based clustering to explore the usage of bike sharing systems. (Shu et al. 2010) uses train ridership data as demand estimates and develops a network flow model to analyze bike sharing systems. While these other studies have analyzed the effects of various factors on bike demands, no studies to our knowledge have used taxi data to predict bike trip volume. We use taxi usage as a co-variate in bike usage predictions, finding it particularly useful for predicting pairwise demand, and propose a neighborhood approach in analyzing flows between stations.

Data

We obtained bike usage statistics for April, May, June and July 2014 from Citi Bike's website[1]. This dataset contains start station id, end station id, station latitude, station longitude and trip time for each bike trip. 332 bike stations have one or more originating bike trips. 253 of these are in Manhattan while 79 are in Brooklyn (left panel of Figure 1). We processed this raw data to get the number of bike trips between each station pair during morning rush hours.
We obtained publicly available taxi usage data from New York City's Taxi and Limousine Commission (TLC) for

[1]https://www.citibikenyc.com/system-data

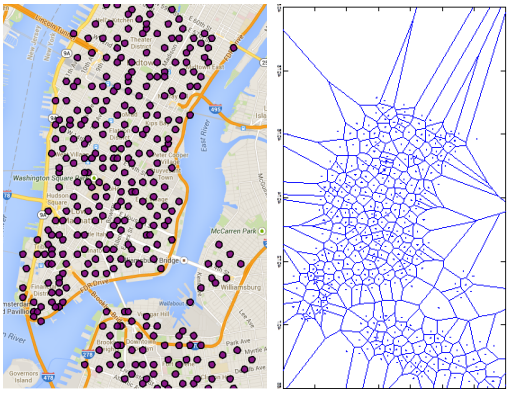

Figure 1: Bike Station Locations in Manhattan and Brooklyn (left), and corresponding Voronoi regions (right). Taxi trip origins and destinations were assigned to bike stations when making predictions according to these Voronoi regions.

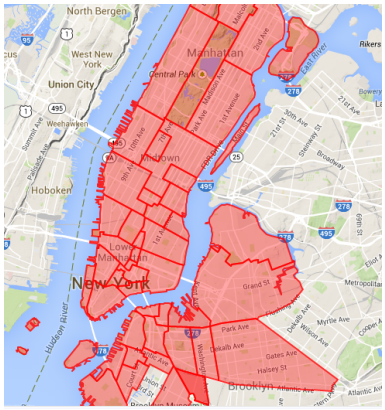

Figure 2: Neighborhood Boundaries in Manhattan and Brooklyn. We aggregate bike stations within neighborhoods to improve predictive performance.

April, May, June and July 2013[2]. Each record in the data set contains pickup date and time, drop-off date and time, passenger count, trip time, trip distance, pickup latitude/longitude and drop-off latitude/longitude for a taxi trip. We processed this raw data to obtain all trips during morning rush hours.

To every processed taxi trip, we then assign a pickup and drop-off bike station ID using the taxi trip's pickup and drop-off locations. To do this, we create a Voronoi diagram with bike station locations as Voronoi centers (right panel of Figure 1) and find, for each taxi trip, the Voronoi regions in which the taxi pickup and drop-off occurred. We assign the trip's origin and destination to the corresponding bike stations. To avoid bias favoring stations with large Voronoi areas, and more accurately predict bike trips, we include only those taxi trips for which both pickup and drop-off location are within a quarter mile of a bike station. The retained trips are then grouped by pickup and drop-off bike station id to get a count of taxi trips for each station pair.

We obtained New York City's daily precipitation data for the months of April, May, June and July in 2013 and 2014 from the National Climatic Data Center[3]. The average number of daily bike trips on weekdays with rain less than 1 mm (dry days) is 26% more than on weekdays with rain greater than 1mm (rainy days), and so we focus on dry days in our analysis.

We obtained population and housing data from the 2010 US Census[4]. Each record in this data set contains geographical ID, display label and population count for a census block group (a small geographic area). Each bike station is assigned the population and housing units in the census block group to which it belongs.

We use neighborhood boundaries defined with economic and demographic variables to classify different neighborhoods in Manhattan and Brooklyn[5]. There are 38 neighborhoods in Manhattan and 18 in Brooklyn (Figure 2). Out of these, 27 neighborhoods in Manhattan and 12 neighborhoods in Brooklyn have existing bike stations. Each of the existing bike stations is assigned to a neighborhood based on its geographic location. Neighborhood population and housing units are calculated by aggregating values for bike stations that lie in different census block groups within that neighborhood. This ensures that we do not double-count population in neighborhoods with multiple stations in the same block group.

Regression Models

We use regression analysis to predict bike trips during morning rush hours for station and neighborhood pairs. We use log-log regression models and test the model in different scenarios based on weather and temporal characteristics. To satisfy normality assumptions made by linear regression, we use base-10 log transformations of bike trips, taxi trips, population and housing units. We add 1 to both bike and taxi trips before the transformation to avoid infinite values for station pairs with 0 bike or taxi trips.

Analysis at the Station Level

A scatter plot of log-transformed bike and taxi trips between station pairs (Figure 3) suggests that although there is some positive correlation between the two, there is also a great deal of unexplained variation. We fit a linear regression model using log-transformations of taxi trips, population, and housing units along with Euclidean and Manhattan distance between station pairs as covariates. Manhattan distance (in miles) is calculated by taking the absolute difference of latitudes and longitude for pickup and drop-off stations and multiplying by 69.1. To exploit the difference in usage patterns between Manhattan and Brooklyn, we also introduce pairwise indicator variables. $I_{M,B}$ is an indicator for bike trips that start in Manhattan and end in Brooklyn.

[2]http://www.andresmh.com/nyctaxitrips

[3]http://www.ncdc.noaa.gov/cdo-web/

[4]http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml

[5]http://nyc.pediacities.com/New_York_City_Neighborhoods

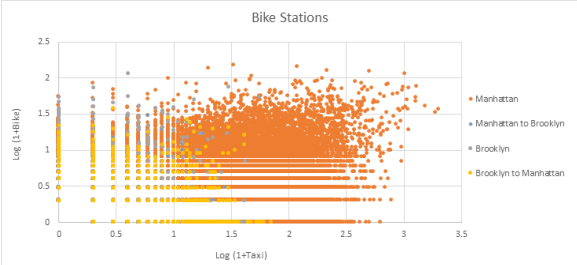

Figure 3: Taxi vs. bike usage for individual stations. Each point corresponds to a pair of bike stations. The x-axis shows taxi trips between Voronoi regions centered at those bike stations. The y-axis shows bike trips between those stations. Although there is some correlation, the correlation is much stronger at the neighborhood level (see Figure 5).

We define $I_{M,M}, I_{B,M}$ and $I_{B,B}$ similarly. We used stepwise backward selection to select significant independent variables. We trained the model on dry weekdays in April 2014 and tested on dry weekdays in May 2014.
The regression at the station pair level does not yield the desired level of accuracy in predictions for the test set, as the model consistently underpredicts log(1+bike trips) relative to actual values, and a substantial amount of variation remains unpredicted (Figure 4). Moreover, the adjusted R-squared value for the model on the training set is 24% which implies that the model captures only 24% of the variability in the training set. Below, to improve our predictions we run the regression models after aggregating stations based on neighborhood definitions.

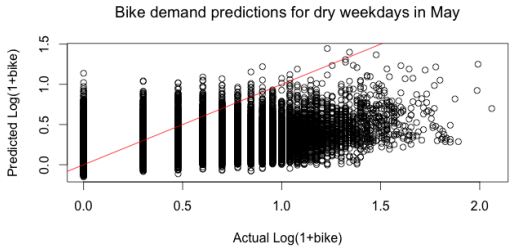

Figure 4: Actual vs predicted bike demand, for demand between station pairs. The red line shows actual=predicted. Compare to Figure 6, which shows substantially better predictive performance for demand between neighborhoods.

Analysis at the Neighborhood Level

A scatter plot of log-transformed bike and taxi trips between neighborhood pairs shows that there is a substantial positive correlation between taxi and bike trips during morning rush hours (Figure 5). We fit a regression model to predict pairwise bike demand between neighborhoods using log-transformations of taxi trips, population at pickup and drop-off stations and indicator variables as described in the previous section. The covariates are selected using stepwise back-

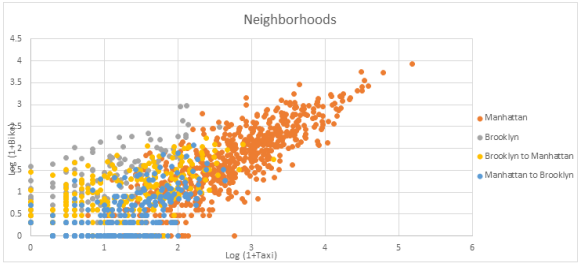

Figure 5: Taxi vs. bike usage for neighborhoods. Each point corresponds to a pair of neighborhoods. The x and y-axis show taxi and bike trips respectively between that pair of neighborhoods.

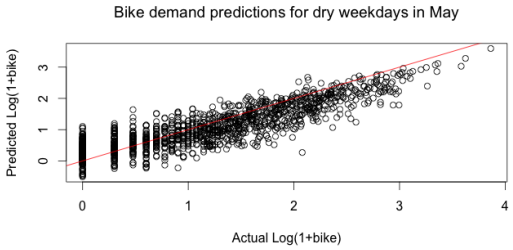

Figure 6: Actual vs predicted bike demand, for demand between neighborhoods. The red line shows actual=predicted. This plot shows substantially better predictive performance than when predicting demand at the station level.

ward selection and we include only those covariates that are significant. The adjusted R squared value of the new model with dry weekdays from April as the training set increases to 74% on the training set. While this is a significant improvement from the individual station model, it is important to note that the two values cannot be compared directly since the two models are predicting trips at different aggregation levels(Robinson 1950). Nevertheless, accurate predictions at the neighborhood level is important because it can provide important and interpretable insights to planners about flows of bikes between neighborhoods.
Table 1 shows the estimated coefficients of each covariate in our model. A positive coefficient value implies that log(1+bike trips) increases with this covariate. We choose $I_{B,M}$ as the base for our indicator variables, setting its effect to 0 without loss of generality. Thus, a negative coefficient value for $I_{M,M}$ signifies that, for each neighborhood pair with origin and destination in Manhattan, there are fewer log(1+bike trips) on average than pairs with an origin in Brooklyn and destination in Manhattan. The actual vs predicted values for May 2014 (Figure 6) suggest that the predictions are accurate. We do a similar analysis for rainy weekdays and observe that the predictive power of the models as measured by adjusted R squared is 68.5%.

Model Validation Linear regression (in our case, applied to log-transformed outcome variables and covariates) as-

	Dependent variable: log_bike	
	est. coefficient	stderr
log_taxi	0.378***	(0.040)
log(pick_population)	0.206***	(0.021)
log(drop_population)	0.092***	(0.019)
$I_{M,B}$	−0.478***	(0.079)
$I_{B,B}$	0.330***	(0.078)
$I_{M,M}$	−0.850***	(0.077)
log_taxi:$I_{M,B}$	−0.021	(0.063)
log_taxi:$I_{B,B}$	0.155**	(0.071)
log_taxi:$I_{M,M}$	0.379***	(0.044)
Constant	−0.700***	(0.122)
Observations		1,277
R^2		0.746
Adjusted R^2		0.745
Residual Std. Error		0.419 (df = 1267)
F Statistic		413.704*** (df = 9; 1267)
		*p<0.1; **p<0.05; ***p<0.01

Table 1: Regression Output. The first section shows, for each independent variable, its estimated effect on the dependent variable log_bike=log10(bike trips + 1), the statistical significance of this effect, and the standard error associated with this estimate. The second section shows summary statistics on the training set. This table was created using the R package (Hlavac 2014).

Training Set	Test Set	RMSE	St.Dev.of Errors
April	May	0.431	0.412
May	June	0.429	0.416
June	July	0.410	0.408

Table 2: Prediction performance on test sets, using three different training sets.

sumes that residuals are normally distributed. We test this assumption by creating a quantile-quantile (QQ) plot (Figure 7) on the residuals. The straight line in the QQ plot validates the normality assumption. To check the model predictions, we use a test-train framework. We run 3 iterations of the model on dry weekdays with different training and test sets and report the Root Mean Squared Error (RMSE) and standard deviation of the errors in each case (Table 2). We observe that the RMSE on the log-transformed data is approximately 0.42, which corresponds to predictions that are larger or smaller than the actual value by a factor of $10^{0.42} \approx 2.6$. While these errors are perhaps too large for many operational uses, observed demand values span 5 orders of magnitude (from 10^0 up to 10^4, as seen in Figure 6), making a prediction with this level of accuracy quite useful for strategic decisions about network expansion. When calculated on an absolute scale, rather than a logarithmic scale, the average RMSE across the three runs was 163.

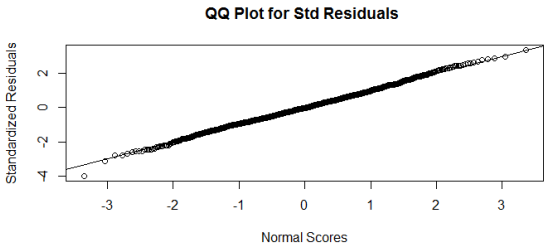

Figure 7: Q-Q plot of the residuals from fit of demand between neighborhood pairs. This plot validates linear regression's assumption of normal residuals.

Conclusion and Future Work

This study predicts the bike usage pattern of New York City's Citi Bike system during morning rush hours of 7:00 AM-11:00 AM. We use taxi usage in addition to temporal, demographic and weather factors as covariates in predicting pairwise trips. We observe that analyzing pairwise trips at the neighborhood level instead of looking at individual stations in bike sharing systems can substantially improve the predictions.

In the future, we will analyze the effects of other weather and demographic covariates on bike usage patterns. A comparative study of a similar analysis for evening rush hours will provide important insights about temporal effects on bike usage patterns. We will use these pairwise demand estimates to predict the required number of bikes and racks at different stations that would maximize the total number of bike trips.

Acknowledgments

Peter Frazier was partially supported by NSF CAREER CMMI-1254298, NSF IIS-1247696, AFOSR FA9550-12-1-0200, and the Atkinson Center for a Sustainable Future. Shane Henderson was partially supported by NSF grant CMMI-1200315. David Shmoys and Eoin O' Mahony were partially supported by NSF grants CCF-0832782 and CCF-1017688. Dawn Woodard was partially supported by NSF grants DMS-1209103 and DMS-1406599.

References

Buck, D., and Buehler, R. 2012. Bike lanes and other determinants of capital bikeshare trips. In *91st Transportation Research Board Annual Meeting*.

Etinne, C., and Latifa, O. 2012. Model-based count series clustering for bike-sharing system usage mining, a case study with the vélib'system of paris.

Hlavac, M. 2014. *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*. Harvard University, Cambridge, USA. R package version 5.1.

Imani, A. F.; Eluru, N.; El-Geneidy, A. M.; Rabbat, M.; and Haq, U. 2014. How land-use and urban form impact bicycle

flows: Evidence from the bicycle-sharing system (bixi) in montreal.

Rixey, R. A. 2013. Station-level forecasting of bikesharing ridership. *Transportation Research Record: Journal of the Transportation Research Board* 2387(1):46–55.

Robinson, W. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review* 15(3):351–357.

Shaheen, S. A.; Guzman, S.; and Zhang, H. 2010. Bikesharing in europe, the americas, and asia. *Transportation Research Record: Journal of the Transportation Research Board* 2143(1):159–167.

Shu, J.; Chou, M.; Liu, Q.; Teo, C.-P.; and Wang, I.-L. 2010. Bicycle-sharing system: deployment, utilization and the value of re-distribution.