Bike Sharing Ridership Forecast using Structural Equation Modeling

Fatemeh Ranaiefar
Transportation Planner
Fehr & Peers
600 Wilshire Boulevard, Suite 1050
Los Angeles, CA 90017
Tel: +1 (213) 261-3055
Fax: +1 (310) 394-7663
f.ranaiefar@fehrandpeers.com

R. Alexander Rixey
Transportation Planner
Fehr & Peers
600 Wilshire Boulevard, Suite 1050
Los Angeles, CA 90017
Tel: +1 (213) 261-3055
Fax: +1 (310) 394-7663
a.rixey@fehrandpeers.com

Submission Date: August 1, 2015
Word Count: 5,900
Figure Count: 4
Table Count: 2

ABSTRACT

This study investigates the effects of demographic and built environment characteristics near bike sharing stations on bike sharing ridership levels in four U.S. bike share systems: California Bay Area, Chicago, New York, and Minneapolis/St. Paul Bike Share. While previous studies have focused on the analysis of origin stations in isolation, this project considers both origin and destination station characteristics as well as the tripmaking patterns between station pairs. We developed BikeSEM, A Structural Equation Model (SEM) to forecast the ridership between every pair of stations in the bike share system. Using SEM methodology instead of conventional linear regressions we were able to make best use of available ridership data from three demotions: total rentals from each stations, total returns to each stations and ridership between each pair of stations. Forecasting bike share ridership at Origin-Destination level can help to improve future active transportation planning. For example we can prioritize bike path projects knowing popular origin and destination bike station pairs. This project also expands on previous studies by including the network effects of the size and spatial distribution of the bike sharing station network and by comparing experiences across systems; particular attention is paid to data quality and consistency issues raised by a multi-city analysis. Relative to previous models, this model will be more widely applicable to a diverse range of communities and help those interested in adopting bike sharing systems to predict potential levels of ridership and identify station locations that will serve the greatest number of riders.

Key Words:

Bike Sharing Forecast, Clustering, SEM

INTRODUCTION

Public bike sharing systems, which provide users with short-term access to bicycles through an automated kiosk or mobile or on-bike interface, are becoming an increasingly popular part of multimodal transportation systems. In the United States as well as around the world, the bicycle is increasingly considered an effective mode to serve short trips, improve first-/last-mile transit access, and decrease congestion and air pollution.

By December 2014 over 50 cities in the United States offered more than 22,000 public bikes to visitors and commuters.[1] As more jurisdictions, nonprofits, and private companies plan to implement bike sharing systems, the question of feasibility is at the forefront. In order to assess feasibility, planners need to determine an appropriate service area and the number, size, and location of bike sharing stations. With this information they can develop reasonable estimates of capital and operating costs. Accurate estimates of ridership at potential station locations can help planners to locate stations to maximize system-wide ridership.

Previous researchers have developed models to estimate the performance of bike sharing stations. The performance of each station is often defined as the number of bikes rented each day from that station. Researchers have investigated the effects of demographic, built environment, and bike sharing network characteristics at each station on ridership levels at that station. However, bike share stations are not independent, but part of a connected network in which many trips start at a different station from where they end, connecting complementary origins and destinations such as home to work, transit station to work, or work to lunch. In this study we use historical ridership data from four major bike sharing systems to estimate a direct demand ridership model using Structural Equation Modeling. The goal is to estimate the number of bike trips between each pair of stations in the system. Data for the analyzed systems come from the California Bay Area (including San Francisco, Mountain View, Palo Alto, Redwood City, and San Jose), New York, Chicago, and Minneapolis.

There is substantial variation both within and among the analyzed cities in demographic characteristics, such as population density; high-, low-, and medium-income employment density;

[1] http://bike-sharing.blogspot.com/2015/01/the-bike-sharing-world-2014-year-end.html

age distribution; and education level. They also vary by public transportation accessibility, bicycle infrastructure, and automobile traffic volumes. Even within a single system the performance of different stations is very different. Given the heterogeneity of bike rental at each station, a single model is not able to provide accurate ridership forecast for all stations. We used hierarchical clustering methods to classify stations based on demographic, socio-economic, land use and bike sharing network characteristics. Our statistical analysis shows that there are at least five distinct patterns among the stations. We show that using origin-destination (OD) data instead of using only bike rentals or returns at each station will lead to more accurate estimates as well as providing information to calculate valuable performance measures such as trip length distribution and popular origin destination pairs that requires bike ways.

BIKE SHARING RIDERSHIP FORECASTING MODELS

With the spread of bike sharing systems and the growth of bike sharing ridership, a number of studies have attempted to assess the feasibility of bike sharing systems, identify appropriate service areas, and forecast station-level and system-wide ridership. These studies identify a variety of relationships between demographic and built environment variables and bike sharing ridership.

Rixey (2013) reviewed several empirical and theoretic studies that provide methodologies for bike share ridership forecasting. He classified these methodologies in two categories: The first category is based on weighted sum raster analysis such as the Krykewycz et al. (*3*) study for Philadelphia, Gregerson et al. for Seattle, Olson et al. for Providence, RI and Maurer (*6*) for Vancouver and New York City. These studies aggregate weighted demographic and spatial variables for a hypothetical grid in the service area to generate a suitability score for each grid cell. These variables include total population and job densities, population density of 20-49 year olds, companies with commute trip reduction programs, tourist attractions, parks and recreation areas, topography, local and regional transit stations, bicycle friendly streets, streets with bicycle lanes, and proximity to colleges, libraries and historic places. These heatmaps forecast relative levels of ridership potential for different locations in the study area. Furthermore, Krykewycz et al (*3*) and Gregerson et al (*4*) also estimate system ridership by applying diversion rates from other modes to bike sharing based on data from Lyon, Paris, and Barcelona.

The second category of studies focused on more explanatory analysis of bike sharing ridership rather than forecasting. Buck and Buehler (xx) and Daddio (*8*) separately investigated Washington, D.C.'s station level ridership data. Total population, the supply of bike lanes, and the number of liquor license holders as a measure of retail destinations were positively correlated with ridership. However, the percentage of households without access to a car, in contrast with theory, has a negative correlation with ridership. The distance from the ridership-weighted average center of the bike sharing system, as a measure of the effect of the bike sharing station network is also negatively correlated with station ridership, so that the farther from the system center, the lower ridership levels are. Daddio showed that the structure of the network affects ridership at each station. Two stations with similar demographic characteristics may have different ridership due to their relative location in the bike sharing system. Hampshire and Marla (*9)* also found that in Barcelona and Seville the number of bike stations within a district affects bike trip generation and attraction.

The above-mentioned studies were focused on explaining the factors that affect the ridership of different stations in a system. Some tried to go a step further and forecast the ridership of new potential stations in the system. Another advancement would be to examine the geographic transferability of these models. In other words, does population density have the same effect on station level bike share ridership in Minneapolis and San Francisco? To examine this, Maurer (*6*) combines empirical analysis of existing bike sharing ridership in Minneapolis/St. Paul, MN with the weighted sum raster approach applied to Sacramento, CA. She developed a linear regression model with 16 covariates, collected at the station level. Her goal was to maximize total model R^2 without considering the significance or sign of the coefficients for individual variables. As a result some variables such as total jobs, total population and presence of bikeways had negative coefficients, in contrast with intuition and theory. This could be due to interdependency of covariates and a relatively small sample size (n=65). For example, because the number of total jobs and retail jobs in an area are usually highly correlated, including both variables in a linear regression model would increase the R^2, but may not provide valid estimates for covariates. Maurer did not consider network effects in her model.

Rixey (2013) developed a comprehensive regression model using data from operational U.S. systems. He identified a number of variables as having statistically significant correlations with station-level bike sharing ridership: population density; retail job density; bike, walk, and transit

commuters; median income; education; presence of bikeways; non-white population (negative correlation); days of precipitation (negative correlation); and proximity to a network of other bike sharing stations. The last variable exhibits a strong positive correlation with ridership in a variety of model specifications suggesting that access to a comprehensive network of stations is a critical factor supporting ridership.

In this study we combine the lessons learned from literature and build a bike share ridership forecast model using direct demand methodology. We address some limitations of the previous studies by several new contributions such as: 1) Using Origin-Destination data instead of station-level data to maximize the use of available information; 2) introducing a new set of network connectivity variables to measure the effect of network structure on station ridership; 3) using Structural Equation Modeling instead of linear regression to enable inclusion of causation and correlation relationships and improve model fitness measures without facing the multicollinearity issue; 4) analyzing multiple systems to increase sample size, diversity of station characteristics, and geographic transferability of the model; and 5) using a clustering approach to respect the diversity of stations' characteristics and estimate a more accurate model for each specific cluster instead of a single model that fits all stations poorly.

METHODOLOGY

Direct Demand Model

In regional science theory, spatial interaction models have been developed to model transaction flows or migration of population between regions. The goal is to predict the "flow" directly based on demographic or economic parameters. In other words, these are comparable to no-constraint gravity models. Working with spatial econometric models requires special data preparation due to 1) spatial dependency between the observations and 2) spatial heterogeneity in the relationships. Spatial dependency means that observations at one location depend on observations at other locations. Spatial heterogeneity refers to variation in relationships over space. In the most general case we might expect a different relationship to hold for every point in space. A basic spatial interaction model is defined in equation (1), where T_{ij} is any transaction between region i and region j such as dollar value or tonnages of goods, number of people migrated, or amount of information transferred; $f(O_i)$ is a function based on parameters at the origin such as population or wage, $f(D_j)$ is a function based on measures of attractiveness at the destination such as number of

jobs, and $f(c_{ij})$ shows relative accessibility or cost of flow or transaction between origin and destination.

$$T_{ij} = f(O_i).f(D_j).f(c_{ij}) \qquad \text{(Eq.1)}$$

In the transportation literature, this model is known as a direct demand distribution model (13). The equation is rewritten in a log-linear form for ease of computation. Direct demand models have been used in passenger and freight studies such as (15,16). However, no previous uses of direct demand distribution modeling could be found in the bike share ridership literature. Given recent advances in IT-based bike sharing system and availability of origin-destination (OD) data it is possible to explore and evaluate direct demand methodology for bike sharing forecasting relative to conventional models. Direct demand models work well for network settings with sparse OD pairs. The model in Equation 1 can be transformed to be in linear format for ease of estimatin.

Structural Equation Model

We developed a direct demand model with a structural equation modeling (SEM) framework. SEM is a flexible linear-in-parameters multivariate statistical modeling technique that has gained acceptance in the travel behavior research community (17). SEM is a more generalized form of linear regression that allows endogenous variables to serve as causal variables for other endogenous variables, and can identify unobservable factors called latent variables. There are different methods of estimating the parameters of these models, such as full information maximum likelihood estimation or three-stage least squares estimation. SEM allows for both confirmatory and exploratory modeling, such that hypothesized causal relationships and correlations can be tested (18). In this study SEM is initially used as a confirmatory approach: the structural design is hypothesized and the sample data are evaluated to confirm whether they fit the hypothesized design. Even if a well-fitting structure naturally exists, it needs to be identified prior to the estimation.

The hindrance of the direct demand method is lack of control over total inflow and outflow of each zone. We overcome this issue by relating the total inflow and outflow to the flow between each pair of stations using a doubly constrained Fratar method (13) to balance the OD matrix. We call this model BikeSEM . We compare the estimated OD table with the result of conventional direct demand formulation. The OD matrix using the BikeSEM process is more similar to the original

OD matrix using likelihood ratio test statistic (Box M-test) for comparing two metrices(11) . This process will be explained below in further detail.

Figure 1.a and 1.b shows the simplified path diagram for BikeSEM and direct demand models. A Path Diagram is a visualization method to represent a SEM model. The rectangles in the model are observed variables. The ovals/circles are latent or unobserved variables. The straight, one-way arrows are causal relationships and the curved two sided arrows show correlation between variables.

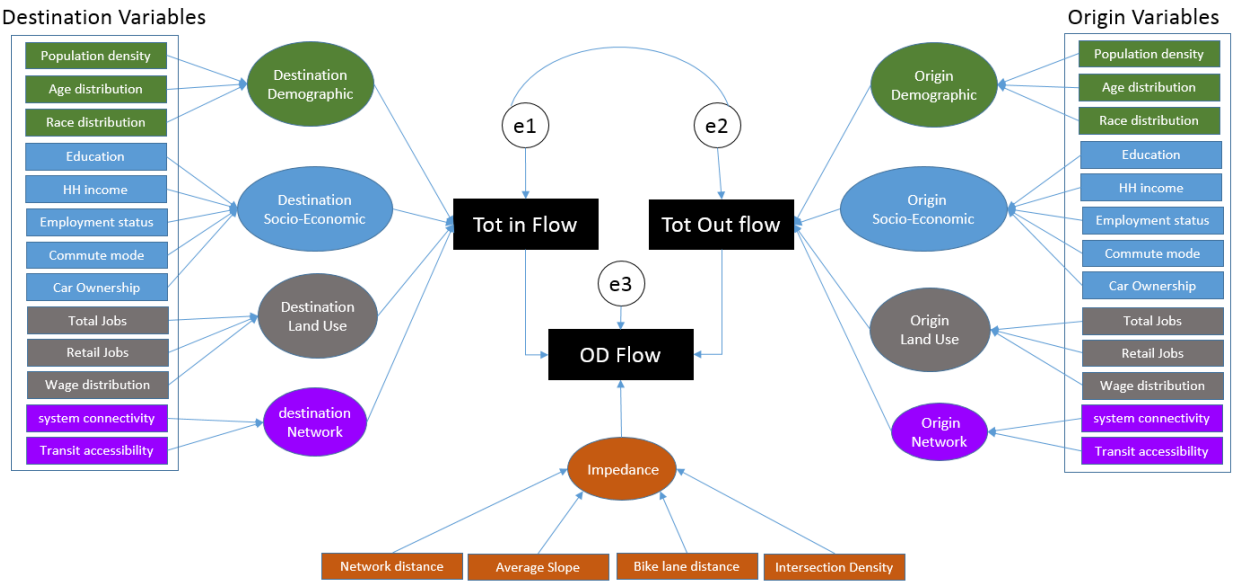

Figure 1.a BikeSEM Model

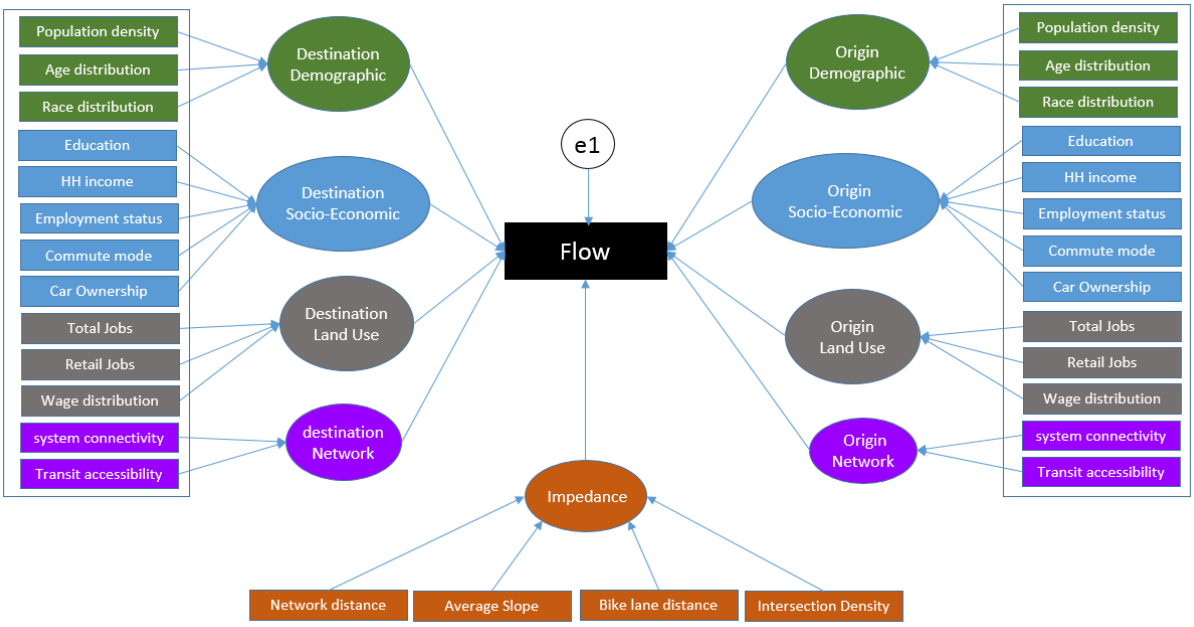

Figure 1.b Direct Demand model

Analysis was performed using stations in the Divvy bike share system in Chicago, the Citi Bike system in New York, Nice Ride Minnesota and California's Bay Area Bike Share system in San Francisco, Mountain View, Palo Alto, Redwood City, and San Jose as observations. In total, there

are n=819 stations in the sample. We used the average monthly rentals by station and the average monthly OD flow as the dependent variables.

A consistent dataset of independent variables was collected across all four systems and compiled; variables selected are widely available so that this analysis can be expanded as additional ridership data become available. Stations are clustered into four distinct categories. Each category encompasses a set of stations that have similar characteristics using hierarchical k-mean methodology. To group the variables and determine relationships among them, we conducted Confirmatory Factor Analysis to identify the best set of variables for the model. Grouping variables helps to understand and explain the relationships when there are many independent variables. Finally, we used AMOS to estimate the SEM Model. We focused on maintaining the intuitive direction and statistical significance of the independent variables used in the model, rather than only maximizing SEM fitness measures such as Chi-Square by adding insignificant relationships.

Clustering

Using meta-data from different cities with a variety of geographic and demographic characteristics increases the possibility of heterogeneity. In statistics, heterogeneity raises questions about the validity of the often-convenient assumption that the statistical properties of any one part of an overall dataset are the same as any other part (14). Ignoring heterogeneity in the data will lead to inconsistency of estimates. There are three formal tests to assess heterogeneity: Cochran *Q* (Chi-square, X^2), I^2 and Tau^2 (14). For our study we used I^2 test which is a robust test to show the presence and magnitude of heterogeneity in Meta data. $I^2 = 100\% \times (Q - df)/Q$ where Q is Cochran's heterogeneity statistic and df the degrees of freedom. Negative values of I^2 are put equal to zero so that I^2 lies between 0% and 100%. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity. Since I^2 was greater than 50% (I^2 =76%), heterogeneity severely exist in our meta data.

One treatment for heterogeneity is to divide the sample into subgroups of observations with similar characteristics and estimate a model for each group. Our clustering analysis shows there are at least three distinct categories of stations in the meta data set. Therefore, we used three different sets of origin and destination variables in the model. The number of clusters used in the model is constrained by sample size and available degrees of freedom. Not all variables appear significant

for all clusters; insignificant variables can be removed from the model. Figure xx shows the structure for two clusters.

We identify all stations with their cluster ID using a k-mean clustering method and demographic, socio-economic and land use variables as clustering variables. Then we used "multi-group analysis" in AMOS-SPSS to estimate BikeSEM for all clusters simultaneously. Figure 3 shows the simplified path diagram. To avoid confusion all correlational relationships and error terms are removed from the diagram.

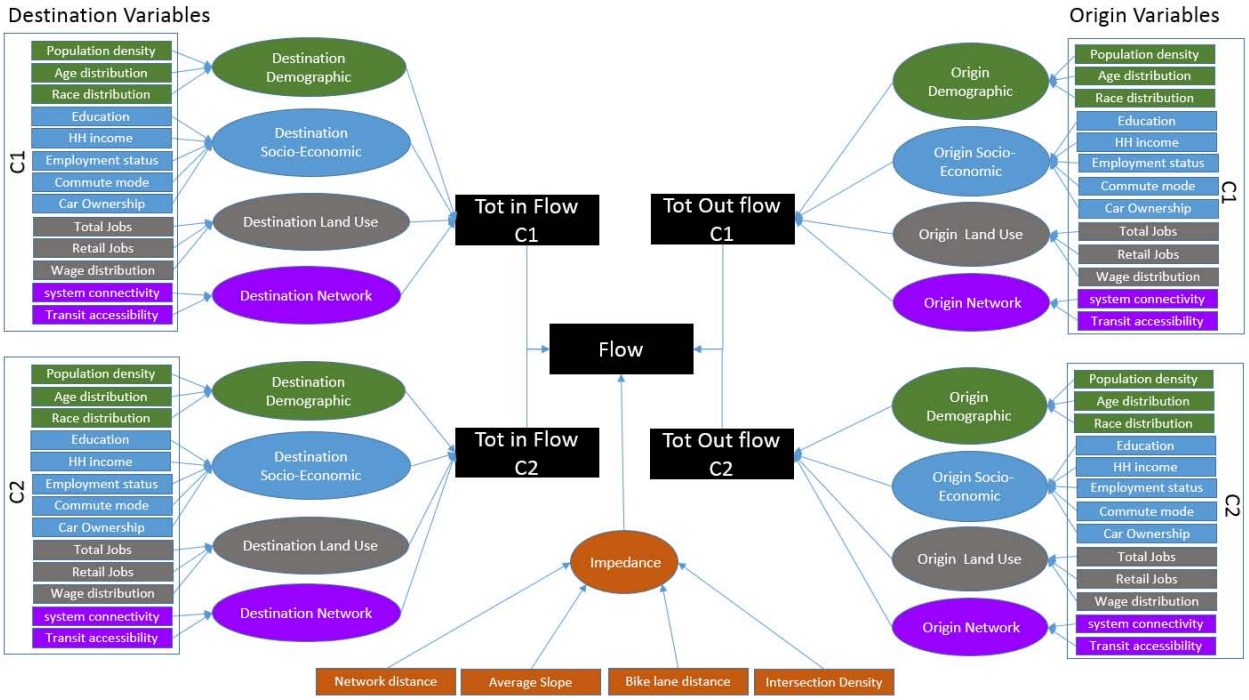

Figure 3. Simplified Multi-Group Path Diagram for BikeSEM

Using above methodology we can measure the effect of each factor based on spatial characteristic of each station. The other benefit of this approach is the Geographic transferability of the model to estimate bike share ridership for a new bike share system in a different city. We can identify the respective cluster for every station in new system and apply the appropriate model.

DATA

This section defines the variables used in the analysis and the final model, discusses the process of compiling the dataset, addresses data quality, consistency and limitations, and presents descriptive statistics of the data.

We have two sets of dependent variables: The monthly average rentals, by station, and the monthly average OD flow. The independent variables address a variety of demographic, socio-economic, built environment, and transportation network factors, collected for all four bike share systems so that a consistent dataset could be created across the sample data set.

We tested several system-specific factors such as number of sunny days, average temperature, annual membership, and hourly rental cost. We assumed they are the same for all stations within a given system. However none of these variables appear to be significant in the model.

Table 1 presents definitions of all variables considered for the BikeSEM model. Unless otherwise specified, variables are based on a half-mile buffer around each bike sharing station to account for a catchment area of users likely to walk to the station.

Data Compilation, Quality, Consistency, and Limitations

Developing a multi-city dataset presented several challenges in gathering comparable variables across all systems. This section discusses 1) the approach used in preparing each group of variables for dataset, 2) concerns regarding the quality and consistency of the data, and 3) potential implications of the data concerns for the model. We used the same methodology as (Rixey, 2013) (12) to compile the input data. The master meta-data should accommodate a clustering structure for the stations. The database includes three sets of variables, one set for each cluster. For each station, only the set of variables that represent the station cluster are populated with values.

Bike Sharing Rentals

Bike sharing station rental data were collected from the bike sharing operators. Information technology connected to the docks themselves ensures consistent records of the number of times bikes have been checked out from each station, and provides time-stamped origin and destination information for each trip. Because the data for different systems spanned different seasons, we used the number of average monthly rentals across approximately an entire operating system of each system for consistency. Data from the Divvy system spanned June 2013 through December 2013, Nice Ride MN data spanned April 2011 through November 2011, Bay Area Bike Share data spanned August 2013 through April 2014, and Citi Bike data spanned July 2013 through May 2014. The total number of checkouts for each station or number of trips between each OD pair over the course of the season was divided by the number of operating days in the season and

reported on a monthly basis; where data were available regarding the opening date of a station, that information was taken into account in the monthly average as well. After system launch, ridership of bike sharing systems tends to increase over time as awareness of the system grows and more users are able to become long-term members. Thus we excluded the first few months of their operations, when ridership was relatively low and would thereby overstate average monthly ridership in that system.

Table.1 Variable Definition

Variable	Definition	Source
Dependent		
Monthly Rental	Number of rentals by station; normalized by number of months in the data set	Bike sharing system operators
Monthly Return	Number of bikes returned by station; normalized by number of months in the data set	Bike sharing system operators
OD Ridership	Number of trips between each pair of origin-Destination; normalized by number of months in the data set	Bike sharing system operators
Independent		
Demographic Factors		
Population[1]	Total population (in 100s of persons)	U.S. Census Bureau, 2010
Jobs[1]	Total jobs (in 100s), by work area	Longitudinal Employer-Household Dynamics, 2010
High-wage Jobs[1]	Number of jobs (in 100s) paying more than $3,333 per month, by work area	Longitudinal Employer-Household Dynamics, 2010
Full Time Employment	Number of full time jobs (in 100s)	Longitudinal Employer-Household Dynamics, 2010
Retail Jobs[1]	Total retail workers (in 100s)	Longitudinal Employer-Household Dynamics, 2010
Alternative Commuters[2]	Proportion of workers who commuted by bicycle, walking, or public transportation (100s of workers)	U.S. Census Bureau, 2010
Median Income[2]	Median household income (in 1,000s of dollars)	U.S. Census Bureau, 2010
Non-White Population[2]	Proportion of population that is of a race other than "white alone"	U.S. Census Bureau, 2010

Low-Vehicle Households[2]	Proportion of workers who commuted by bicycle, walking, or public transportation (100s of workers); weighted average by 2010 Census Tract	U.S. Census Bureau, 2010
Transportation Network Factors		
Stations Within (X) mile	Number of bike sharing stations within (X) mile	Bike sharing system operators

[1] Summed proportionally by area intersecting 2010 Census Blocks.

[2] Weighted average by area of buffer intersecting 2010 Census Tracts.

Census Block-Level and Tract level Data

The Population, Jobs, Jobs by wage distribution, and Retail Jobs variables were collected from 2010 Census and Longitudinal Employer-Household Dynamic data at the Census Block level. The data were aggregated for the half-mile buffer surrounding each station.

The Commuter Mode Distribution, Median Income, Race distribution, Household Vehicle Ownership, and Education Level variables were collected from 2010 Census data at the Census Tract level. Like the Census Block-level data, the data were aggregated to the half mile buffer surrounding each station with a mean weighted by the proportion of the area of the intersection of the buffer and each Census Tract to the sum of the areas of each intersected Census Block.

The 2010 Census data are the best consistent data available to reflect conditions during the time period represented by the station ridership data for all systems. Using a single data source helps to ensure consistency across all cities, and makes this analysis readily scalable to include other cities and bike sharing systems.

Network Effects

Rixey, 2012 proposed to use stepwise buffers to assess the number of stations in vicinity of each station. Our empirical results show that using true network-based distance provides better estimates than "as-the-crow-flies" distance. The Stations Within (X) miles variables were created using shapefiles of the bike sharing stations in each of the systems and their roadway networks. ArcGIS Network Analyst was used to calculate the shortest network distance between each pair of stations. The number of accessible stations within 0.5, 1, 1.5, 2, 2.5 and 3 miles of each station was calculated for each station. Freeways and Highways are heavily penalized and where bikeway

network is available it is prioritized in the network skimming process. These variables provide a way to assess the availability of destination bike stations from a given origin bike station at a variety of scales considering the real biking distance.

The Transit Accessibility Score measures the impact of availability of transit options from each station on bike share rentals. Aggregate frequency of transit service per hour during the evening peak period within 0.25 miles of block group boundary from the EPA's Smart Location Dataset is used as a proxy for transit accessibility. We averaged the aggregate transit frequency of all blocks that intersect with the half-mile buffer of each station to assign a Transit Accessibility Score to each Station.

Impedance Factor

The major contribution of BikeSEM is using the origin-destination ridership combined with station-level rental data to provide a better forecast of demand at each station and other performance measures such as trip length distribution. The impedance factor is an important input to the model. It is a proxy for users' relative preference between similar destination choices. We considered several variables to create the impedance factor including network distance between two stations, average slope of the shortest route between two stations, the portion of the shortest route between two stations that utilizes bikeways, and the number of intersections on the shortest route between bikeways. Our results could be significantly improved by having more information about bikers' travel behavior and route choice parameters and their perception of travel cost by bike.

Descriptive Statistics of Ridership Data

The stations in the Bay Area Bike Share (n=71), DIVVY (n=300), Citi Bike (n=332), and Nice Ride MN (n=116) systems provided a total of 819 observations with comparable data. The model also considers combinations of origins and destinations as inputs. Although the 819 stations provide a total of 218,721 potential origin-destination pairs (assuming all stations within a system are connected and stations are not connected across systems), only 152,529 (70%) of these station pairs recorded trip activity. This percentage varies across systems, with dense, well-connected systems like Citi Bike reporting trips among 90% of its potential OD pairs, while more dispersed

systems like Bay Area Bike Share, which is spread among five different cities, reported trips among only 34% of potential OD pairs. In other words, almost every station in the Citi Bike system is well-connected to every other station, whereas in the Bay Area Bike Share system, there are many station pairs between which users do not ride. Another measure of the concentration or spread of tripmaking within each system is a cumulative plot of the percent of total system trips against the percent of total available OD pairs, sorted in descending order by tripmaking (Figure 4). In the Divvy and Nice Ride MN systems, some OD pairs constitute a disproportionate share of total system tripmaking; just two percent of OD pairs constitute nearly 40% of all trips taken within those systems. By comparison, trips in the Citi Bike system are slightly more evenly distributed throughout the system, with two percent of OD pairs serving only 20% of system trips.

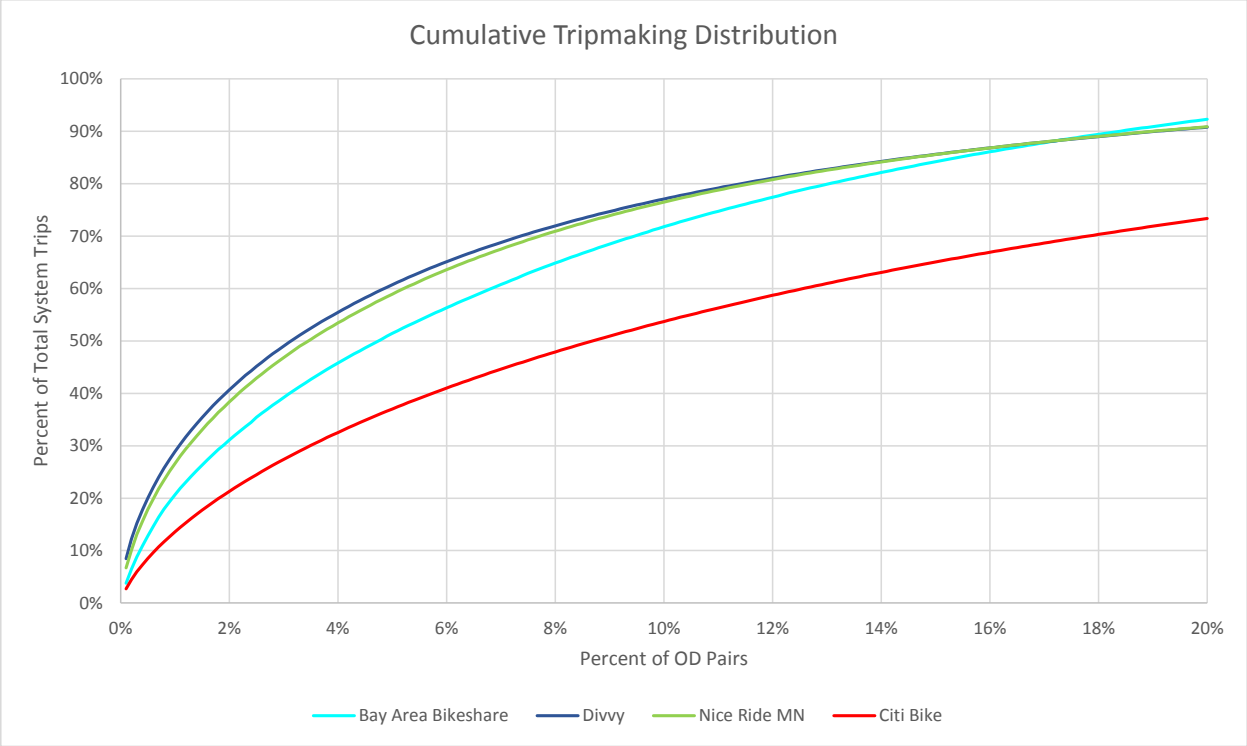

Figure 4. Total System Trips against the Percent of Total Available OD Pairs

Additional descriptive statistics of the dependent variables are provided in Table xx. With the highest number of stations and the longest timespan of available data, the Citi Bike system reported the highest level of total trips, over 7.5 million, which was nearly ten times the next most-utilized system, Divvy; even adjusting for the longer dataset timespan and larger number of stations, Citi

Bike ridership approaches 70 rides per station per day, more than four times Divvy's ridership level.

Table 2. Dependent Variable Descriptive Statistics

Attribute	Bay Area Bike Share	Divvy	Citi Bike	Nice Ride MN	Total or Mean
Stations	71	300	332	116	819
Potential Station Pairs*	5,041	90,000	110,224	13,456	218,721
Station Pairs with Trips	1,727	44,422	99,130	7,250	152,529
% of potential Pairs Used	34%	49%	90%	54%	70%
Start of Ridership Data	8/29/2013	6/27/2013	7/1/2013	4/8/2011	n/a
End of Ridership Data	4/30/2014	12/31/2013	5/31/2014	11/6/2011	n/a
Count of Days	244	187	334	212	n/a
Total Trips	194,827	759,788	7,538,335	217,530	8,710,480
Trips per Day	798	4,063	22,570	1,026	n/a
Trips per Month**	23,954	121,891	677,096	30,783	n/a
Trips per Station per Day (average)	11.25	13.54	67.98	8.85	n/a
Trips per Station per Month (average)	337	406	2,039	265	n/a

*= n^2, assuming each station can serve as an origin for each other station, including itself.
** assuming 30-day months.

The four analyze systems also varied in terms of seasonality, or the distribution of tripmaking across the year. Figure 5 shows the portion of trips made in each month for the four systems. Bay Area Bikeshare experienced the most consistent level of ridership throughout the year. Disregarding a short August of only three operating days, the remaining months through the end of the dataset in April each account for between 10 and 15 percent of total annual tripmaking. Divvy, on the other hand, experienced substantial peaking: in total, August, September, and October accounted for over 70 percent of tripmaking in the dataset, with substantially lower ridership in July, November, and December. Nice Ride MN and Citi Bike exhibit more subdued seasonality, with ridership peaks in the seasonable late summer and autumn months and lower ridership in the colder winter months.

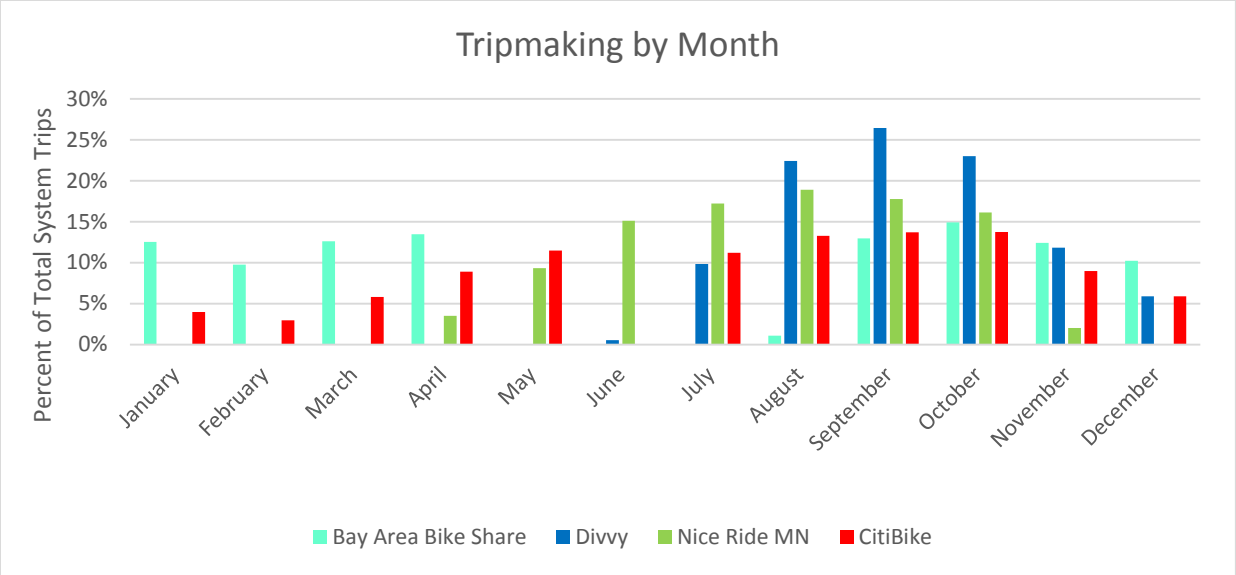

Figure 5. Summary of Number of Rentals for Five Bike Share Systems in the Study

BikeSEM MODEL ESTIMATION

Identification of BikeSEM Variables

Dependent variables for each structural model were selected based on a review of bike sharing ridership estimation literature (*3*),(*4*),(*5*),(*6*),(*7*),(*8*),(*9*) and *(12)* and intuition regarding relationships to ridership, and availability of consistent data sources across the four cities selected for analysis. The station network variables ("Stations Within (X) Meters") were included to test the effects of bike sharing station network density, distribution and size on ridership. SEM software provide "Modification Indices (MI)" as part of their general output (18). This index measure the change in overall fitness measure by adding a new causal or correlational relationship to the model. However, it is researcher's responsibility to verify theoretical validity of each relationship and the model as an integrated system. We tested different sets of variables for different clusters. Factor analysis was also used to reduce the dimension of the model shows. The station network variables were all significantly correlated with ridership at the 1% level, as were the majority of the other independent variables tested. Only the Race variables did not show significant relationship for any of the clusters.

BikeSEM Fitness Metrics

The conceptual framework presented in figure 3 was refined for each cluster in order to 1) maximize the predictive power of the model as a whole, and model fitness as measured by the model Chi-Square[2] and RMSEA, GFI (16) 2) incorporate a variety of independent variables, and 3) maintain statistical significance and intuitive direction of the included variables. As an advantage of SEM methodology the correlation between Independent variables with a high degree of multicollinearity, such as Alternative Commuters and Low-Vehicle Households, or the multiple Jobs or Stations Within (X) mile variables can be defined in the model and used as an information to estimate variance covariance matrix between all variables. In the ideal situation all the non zero (and significant) correlation relationships will be identified in MI report and can be added to the model to improve the overall model fitness.

There are different fitness measures in the SEM literature, as summarized by Hooper et al. (18). Chi-Square2 and RMSEA, GFI are the most cited metrics. If the data is not multivariate normal The Chi-Square test can fail, even though the model itself is properly specified. Since the data is indeed non-normal in this case, the Chi-Square test is disregarded. The RMSEA is 0.08 . RMSEA less than 0.05 is good fitness and less than 0.1 is acceptable. GFI id 0.78. GFI greater than 0.9 defines a good model. Overall BikeSEM is a fair model given limited available data.

For brevity and respecting TRB word limit, detail results and coefficients and estimates are not listed here. Please contact authors to receive complete results of BikeSEM.

Our Future plan is to validate BikeSEM using a different bike sharing system such as Capital in Washington D.C. We are also investigating models to estimate weekday and weekend average tripmaking as we identify significant temporal patterns between weekdays' and weekends' tripmaking

ACKNOWLEDGMENTS

The author would like to express his gratitude to the operators of Citi Bike share, Bay Area Bike share, DEVVY, and Nice Ride MN for making their ridership data available, and to Fehr & Peers for the opportunity to explore this interesting topic. Many thanks also to Molie Polen for GIS data analysis.

REFERENCES

1. Shaheen, S.A., S. Guzman, and H. Zhang. Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2143, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 159-167.
2. Shaheen, S.A., E.W. Martin, A.P. Cohen, and R.S. Finson. *Public Bikesharing in North America: Early Operator and User Understanding.* MTI Report 11-26. Mineta Transportation Institute, 2012.
3. Krykewycz, G.R., Puchalsky, C.M., Rocks, J., Bonnette, B., and F. Jaskiewicz. Defining a Primary Market Area and Estimating Demand for a Large-Scale Bicycle Sharing Program in Philadelphia. In *Transportation Research Record: Journal of the Transportation Research* Board, No. 2143, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 117-124.
4. Gregerson, J., M. Hepp-Buchanan, D. Rowe, J. Vander Sluis, E. Wygonik, M. Xenakis, and E. McCormack. Seattle Bicycle Share Feasibility Study. University of Washington College of Built Environment Department of Urban Design and Planning, 2011.
5. Olson, J., P. Goff, and S. Simms. City of Providence Bike Share Feasibility Study Final Report. Prepared by Alta Planning + Design for City of Providence. May 2011.
6. Maurer, L.K. Feasibility Study for a Bicycle Sharing Program in Sacramento, California. Presented at the 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.
7. Buck, D. and R. Buehler. Bike Lanes and Other Determinants of Capital Bike share Trips. Presented at the 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.
8. Hampshire, R.C. and L. Marla. An Analysis of Bike Sharing Usage: Explaining Trip Generation and Attraction from Observed Demand. Presented at the 91st Annual Meeting of the Transportation Research Board, Washington, D.C., 2012.
9. Daddio, D.W. Maximizing Bicycle Sharing: An Empirical Analysis of Capital Bike share Usage. University of North Carolina at Chapel Hill, 2012.
10. Osborne, J. Notes on the Use of Data Transformations. Practical Assessment, Research & Evaluation, 8(6), 2002. http://PAREonline.net/getvn.asp?v=8&n=6. Accessed Apr. 14, 2008.

11. Rencher Alvin C. Multivariate Analysis, Wiley Series in Probability and Statistics , 2002
12. Rixey, R. Alexander, Station-Level Forecasting of Bike Sharing Ridership: Station Network Effects in Three U.S. Systems, In *Transportation Research Record: Journal of the Transportation Research* Board, No. 2387, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 46-55.
13. Ortuzar Juan de Dios, Willumsen Luis G, MODELLING TRANSPORT Fourth Edition, A John Wiley and Sons, Ltd., Publication, 2011
14. Huedo-Medina, Tania; Sanchez-Meca, Julio; Marin-Martinez, Fulgencio; and Botella, Juan, "Assessing heterogeneity in meta-analysis: Q statistic or I2 index?" (2006). CHIP Documents. Paper 19.
15. Ranaiefar Fatemeh, Chow Joseph, Rodriguez-Roman Daniel, Camargo Pedro, Ritchie Stephen Structural Commodity Generation Model That Uses Public Data *Transportation Research Record: Journal of the Transportation Research Board* Dec 2013, Vol. 2378, pp. 73-83.
16. Ranaiefar Fatemeh, Chow Joseph, McNally Michael, Ritchie Stephen, Structural Direct Demand Model for Interregional Commodity Flow Forecasting, Presented at the 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
17. Kline, R.B. Principles and Practice of Structural Equation Modeling. 2nd ed., The Guilford Press, New York, 2005.
18. Bollen, K. A., and J. S. Long. Introduction. In Testing Structural Equation Models, Sage Publications, Newbury Park, California, 1993, pp. 1– 15.
19. Hooper, D., J. Coughlan, J., and M. Mullen, 2008. Structural equation modelling: guidelines for determining model fit. Electronic Journal of Business Research Methods, Vol. 6, No. 1, 2008, pp. 53-60.