EXAMINING THE IMPACT OF SAMPLE SIZE IN THE ANALYSIS OF BICYCLE SHARING SYSTEMS

Ahmadreza Faghih-Imani, McGill University
Naveen Eluru, University of Central Florida

Introduction

In recent years, there has been growing attention on bicycle sharing systems (BSS) as an alternative and complementary mode of transportation (Shaheen et al. 2010; Faghih-Imani et al. 2014). A bicycle-sharing system provides increased flexibility to ride a bicycle without the costs and responsibilities associated with owning a bicycle (such as the need to secure their bicycles or perform regular maintenance). With the growing installation of BSS infrastructure across the world there is a substantial interest in understanding how these systems impact the urban transportation system. Research efforts examining BSS employed a wide range of sample sizes depending on the temporal or spatial aggregation. While it is beneficial to use large sample sizes for analysis, increase in sample sizes are associated with increased data preparation effort, and longer model run times. In this context, the main objective of this paper is to investigate the impact of sample size on BSS analysis using data from New York City's BSS (CitiBike). Specifically, the research evaluates the impact of sample size on model parameter estimates, inference measures and prediction capabilities. The findings provide analysts and planners guidelines on the "*minimum*" and "*ideal*" size of data necessary for examining BSS.

There is a gradual increase in the research on bicycle-sharing systems over the past few years (see Fishman (2015) and Ricci (2015) for a review of recent literature on BSS). Under the <u>systems perspective</u>, quantitative studies employed actual bicycle usage data to capture the determinants of BSS usage (Nair et al. 2013; Rixey 2013; Gebhart and Noland 2014; O'Brien et al. 2014; Faghih-Imani et al. 2014; Faghih-Imani and Eluru 2014; Rudloff and Lackner 2014; Zhao et al., 2014; Wang et al. 2015). These studies typically postulate that BSS usage from a system perspective is influenced by various attributes such as BSS infrastructure, transportation network infrastructure, land use and urban form, meteorological data, and temporal characteristics. These studies mostly characterized usage as the number of trips originated and destined to one station or divided the usage of one station in two rates: arrivals (depositing bicycle) and departures (removal of bicycles). The studies focussed on the <u>user perspective</u> contribute to the literature by studying user behavior in response to bicycle-sharing systems.

The BSS operators provide system availability data on their websites. Through relatively simple scripting exercises, it is possible to build a database of bicycle availability across stations for the BSS system. The data thus obtained can provide a glimpse of how BSS usage varies across the day. More recently, in addition to the system availability information, BSS operators release trip information containing details including origin and destination stations, start and end time of the trip for BSS users. The station usage may also be obtained from this information by aggregating trips originated or destined at one station. In addition to usage rate, the information is useful to understand destination choice behavior of BSS users. An important question in the process of developing models to analyze bicycle-sharing systems is to choose the size of the data to be employed for the model estimation. As opposed to the traditional travel demand literature where sample sizes are quite limited, in the context of BSS, the information is available for every minute for multiple days and months. Hence, the selection of appropriate sample for BSS analysis is quite critical.

Employing large samples requires substantial data preparation and model run times. For example, one month of data for a BSS with 300 stations results in 216,000 records of hourly arrivals or departures and about one million trips. The processing of usage or trip data and preparation of station level variables including built environment attributes and other variables such as weather characteristics or temporal attributes are substantially time consuming. In addition to data preparation, a very large sample significantly increase the model run times. On the other hand, employing a smaller sample than appropriate would result in biased or skewed model estimates affecting the planning process. Hence, it would be useful to understand the sample size requirements for examining bicycle-sharing systems. Besides, the data is not always available; knowing the required appropriate sample size prior to collecting data would be beneficial. Due to the relative infancy of BSS, there is little to no guidance on the amount of data necessary for analysis.

Current Study in Context

It is evident from the discussion above that sample size requirements would assist transportation planners in developing appropriate models to study BSS trip generation (arrivals and departures) and distribution (destination). The current study proposes a systematic evaluation of the impact of sample size on model estimates, inference statistics and predictive performance. Towards this end, we evaluate the BSS data from two perspectives: 1) system usage – what contributing factors influence hourly arrival and departure rates at station level, 2) user destination choice – what factors contribute to users' preference of destination station choice.

To examine the system usage, we employ the linear mixed model methodology to determine the factors contributing to BSS usage. The usage is characterized as hourly arrival and departure rates of each station. The traditional linear regression model is not appropriate to study data with multiple repeated observations such as the hourly arrivals and departures for each station in our empirical context. Thus, we employ a linear mixed modeling approach that builds on the linear regression model while incorporating the influence of repeated observations from the same station (see Faghih-Imani et al., 2014). We analyze New York City's BSS (CitiBike) stations' hourly arrivals and departures for various samples. To explore the user destination choice, we employ the Multinomial Logit Model (MNL) to examine the impact of individual bicyclist attributes, trip attributes and destination attributes on destination choice. The most common methodology to study location choice in transportation and related literature is MNL (see Faghih-Imani and Eluru, 2015). We estimate the MNL model for CitiBike system in New York City. For the sake of brevity, we only focus on trips, arrivals and departures made by annual members.

The model estimation exercises for system demand and destination choice are conducted on several samples of data. The performance of these sample models relative to a base sample data is observed. Further, the performance of these sample based models on a hold-out sample relative to the predictive accuracy of the base sample is also compared. In order to account for the randomness of selecting smaller samples, for each smaller sample size, we randomly select two sets of data from that large sample and report the average results. The results would help the analysts to make decision on sample size in order to accurately examine BSS usage.

Sample Formation

A sample formation was necessary in order to obtain the arrivals and departures. Number of trips originated from and destined to one station are equal to the number of departures and arrivals for that station. Thus, we aggregated the number of trips originated from/destined to one station by different type of users at an hourly level to obtain hourly arrivals and departures by members and daily customers at a station level. Further, we normalized stations' arrivals and departures with station capacity to account for the influence of station capacity on demand. In our modeling efforts, we employ logarithm of the hourly normalized arrivals and departures as the dependent variable. We focused on the month of September, 2013; i.e. the peak month of the usage in 2013 for our base analysis. This would give us a base sample consisting of 237,600 records (330 stations × 24 hours × 30 days). We estimate our base model for arrivals and departures and assume the estimate results as the true (base) values and compare the rest of models with these results. Then, from this base sample we randomly select a series of smaller samples. For this purpose, we select random days in the month of September and assign them to each station. We choose 15 days, 7 days, 5 days, 3 days, 2 days and 1 day randomly for each stations to create smaller samples. It must be noted that the random days assigned to each station are different from random days assigned to other stations; thus the sampling approach covers the whole month across the urban region. To account for the impact of randomness, we generate two sets of these random days, estimate the arrival and departure models with both, and then obtain the average results.

For destination choice model, to be consistent with the usage analysis, again we focus on the trips in the month of September. The sample formation exercise also involved a series of steps. First, trips with missing or inconsistent information were removed. Second, trips longer than 90 minutes in duration (only 0.5% of all the trips) were deleted considering that the trips longer than 90 minutes are not typical bicycle-sharing rides and could also be a result of misplacing the bicycle when returning it to the station. At the same time, trips that had the same origin and destination were also eliminated. For trips with the same origin and destination, it is possible that the bicycle was not functioning well and the users returned them to the origin station. Also, to accommodate for intentional same origin and destination trips would require additional trip purpose information and is beyond the scope of this work. Therefore, we focus on trips that were destined outward. CitiBike system has 330 stations. Considering all the stations in the universal choice set of destination station choice model will result in substantial computational burden. Thus, we sample 30 stations from the universal choice set including the chosen alternative (McFadden, 1987; Faghih-Imani and Eluru, 2015). For the evaluation of sampling impact, we consider the sample with 50,000 trips as the base sample. Then from this 50,000 trips, we randomly select two sets of 20000, 10000, 5000, 3000, 2000, 1000 trips generated as our smaller samples. For every sample size, the information for the 30 stations is augmented with the individual trip records.

Linear Mixed Models

We consider a linear mixed model structure for BSS demand (arrivals and departures). Let q = 1, 2, …, Q be an index to represent each station, d = 1, 2, …, D be an index to represent the various days on which data was collected and t = 1, 2, …, 24 be an index for hourly data collection period. The dependent variable (arrival or departure rate over station capacity) is modeled using a linear regression equation which, in its most general form, has the following structure:

$$y_{qdt} = \beta X + \varepsilon \qquad (1)$$

where y_{qdt} is the normalized arrival or departure rate as dependent variable, X is an L×1 column vector of attributes and the model coefficients, β, is an L×1 column vector. The random error term, ε, is assumed to be normally distributed across the dataset.

The error term may consist of three components of unobserved factors: a station component, a day component, and an hour-of-the-day component. We consider the station and the time-of-day to be related common unobserved effects. In this structure, the data can be visualized as 24 records for each Station-Day combination for a total of "Q stations × D days" observations. We parameterize the covariance matrix (Ω). For estimating a parsimonious specification, we assume a first-order autoregressive moving average correlation structure with three parameters σ, ρ, and φ as follows:

$$\Omega = \sigma^2 \begin{pmatrix} 1 & \varphi\rho & \varphi\rho^2 & \cdots & \varphi\rho^{23} \\ \varphi\rho & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi\rho^{23} & \cdots & \cdots & \cdots & 1 \end{pmatrix} \qquad (2)$$

The parameter σ represents the error variance of ε, φ represents the common correlation factor across time periods, and ρ represents the dampening parameter that reduces the correlation with time. The correlation parameters φ and ρ, if significant, highlight the impact of station specific effects on the dependent variables. The models are estimated in SPSS using the Restricted Maximum Likelihood Approach (REML).

Multinomial Logit Model

We consider a multinomial logit structure for destination choice model. Let q = 1, 2, …, Q again be an index to represent each station, j = 1, 2, …, J be an index to represent the BSS users. Then, the random utility formulation takes the following form:

$$u_{jq} = \beta' X_{jq} + \varepsilon_{jq} \qquad (3)$$

Where u_{jq} is the utility obtained by user j by selecting station q from the choice set of 30 stations. X_{jq} is the vector of attributes and β is the model coefficients to be estimated. The random error term, ε, is assumed to be independent and identically Gumbel-distributed across the dataset. The BSS user j will choose the station as destination that offers the highest utility. With this notation, the probability expression takes the typical multinomial logit form given by:

$$P_{jq} = \frac{\exp(\beta' X_{jq})}{\sum_{q=1}^{30} \exp(\beta' X_{jq})} \qquad (4)$$

The log-likelihood function can be defined as:

$$L = \sum_j \ln(P_{jq}) \qquad (5)$$

By maximizing this log-likelihood function, the model parameters β are estimated. The maximum likelihood model estimation is programmed in GAUSS matrix programming language.

Evaluation Measures

Considering that the base models represent the population models, we set forth estimating the same specification using the several samples prepared. As expected, the estimation on the smaller samples provide different values for the estimated parameters and different standard errors of estimates. To account for randomness of sampling process, we estimate the specifications on two sets of samples for each sample size. Based on these two samples an average estimate value is generated. The estimation results for base sample for arrivals, departures and destination choice models are presented in Table 1.

The impact of sample size on the estimation results are examined by the following measures: 1) the capability to produce the same parameters estimate of the base sample, 2) the significance of the parameter represented by the standard error and 3) the prediction capability as a measure of goodness of fit to predict the same results for data hold-out sample. All the measures for evaluating performance of arrivals, departures and destination choice models are presented in Table 2. For usage models, in order to evaluate the parameters estimated, mean absolute percentage error (MAPE) and root mean square error (RMSE) of estimated parameters with respect to the base sample' parameters estimated are calculated. In addition, the MAPE and RMSE for the change in standard error of estimates with respect to base values are also generated (presented in parentheses in front of the values for parameters estimated in Table 2). In order to show the prediction capability of models, we used the data from first week of October (i.e. the next week after our base sample for model estimation) to validate the estimated models by each sample. The same data procedure described in sample preparation for models estimation was repeated in order to compute hourly arrivals and departures. For each sample, the model developed was used to generate predictions of hourly arrival and departure rates and the predictions were compared with the observed rates in validation sample. Again, to compare the prediction performance, we calculate two error metrics of mean absolute error (MAE) and RMSE.

For the destination choice model, the same measures (MAPE and RMSE) are used in order to evaluate the performance of models to produce the estimated parameters of the base sample. However, for prediction capability measure, different procedure was needed. For this purpose, we employ a hold-out sample of 5000 trips as our validation sample. The same data preparation and choice set generation for estimation samples is exercised for the validation sample. The parameters estimated by each sample size models, were used to compute the probability of choosing a station for 30 stations of choice set for each of the 5000 trips. In order to evaluate the performance of models in prediction, two metrics are used: a) the predictive log-likelihood: the sum of log of probability of chosen station across the validation sample, and b) the percentage of correct prediction (correct prediction is defined as assigning the highest probability to the chosen station).

Table 1 Estimation Results

Parameter	Arrivals		Departures		Destination Choice	
	β	**Std. Err.**	**β**	**Std. Err.**	**β**	**Std. Err.**
Constant	-4.0140	0.0309	-4.0305	0.0299	-	-
Built Environment Variables						
Length of Bicycle Facilities in buffer	0.0486	0.0000	0.0713	0.0000	-	-
Presence of Subway Station in Buffer	0.1000	0.0096	0.1529	0.0092	0.0402	0.0102
Presence of Path Train Station in Buffer	0.3563	0.0231	0.3626	0.0222	0.0759	0.0194
Length of Rails in Buffer	-0.0233	0.0000	-0.0385	0.0000	-0.0633	0.0101
Area of Parks in Buffer	-3.3466	0.0000	-3.4852	0.0000	-1.2352	0.4137
Area of Parks in Buffer *Weekend	2.6017	0.0000	2.8177	0.0000	3.0355	0.7924
Number of Restaurants in Buffer	0.7622	0.0001	0.8353	0.0000	0.4743	0.0469
Population Density	13.4928	0.3328	12.4059	0.3210	4.14	0.4729
Population Density*AM	-3.3934	0.3725	4.9233	0.3774	-12.3575	0.8710
Population Density*PM	2.9382	0.3599	-0.4879	0.3635	1.5288	0.7205
Job Density	0.6680	0.0923	0.9033	0.0890	-1.0011	0.1541
Job Density*AM	3.5351	0.1016	-0.3091	0.1029	4.5145	0.2404
Job Density*PM	0.1994	0.0981	2.2847	0.0991	-1.7287	0.2383
Station Capacity	-	-	-	-	0.0161	0.0005
Weather & Temporal Attributes						
Weekend	-0.1301	0.0121	-0.1351	0.0117	-	-
AM	1.0487	0.0129	1.1826	0.0131	-	-
Midday	1.1479	0.0078	1.1807	0.0077	-	-
PM	1.3347	0.0141	1.3205	0.0142	-	-
Even	0.7185	0.0096	0.6991	0.0095	-	-
Temperature	0.0124	0.0009	0.0076	0.0009	-	-
Relative Humidity	-0.6643	0.0002	-0.5904	0.0002	-	-
Rainy Weather	-0.2310	0.0113	-0.2716	0.0115	-	-
Trip Attributes						
Distance	-	-	-	-	-0.5089	0.0198
Distance*Female	-	-	-	-	0.0393	0.0073
Distance*Temperature	-	-	-	-	0.1385	0.0658
Distance*Humidity	-	-	-	-	-0.0964	0.0200
Distance*Rainy	-	-	-	-	-0.1155	0.0415

Evaluation Results

In this section, we discuss the impact of sample size on the performance of models estimated. For the usage models, the effect of sample size in arrivals models and departures models are similar with a slightly better performance for departures (lower RMSE for both parameters estimated and prediction). The performance of models on smaller sample size to produce the estimated parameters of base sample varies by MAPE of 4.78 to 28.76 and 7.10 to 39.26 for arrivals and departures, respectively. The results show that as we choose smaller sample size until two days, we have almost similar results (within 20%) as the base case. We observe a huge jump in MAPE or RMSE when we use only one day to estimate arrivals or departure models. The results for the standard error changes indicate that as sample size decreases, standard error of estimates increases substantially thus altering inference i.e. the variable might be considered insignificant. In terms of prediction capability, we cannot observe any significant difference (loss) due to use of smaller samples.

However, based on the escalation of error relative to standard errors, we suggest an ideal sample wit three days of data for analysis of BSS hourly usage.

Table 2 Evaluation Measures (values for standard errors are in parentheses)

Arrivals	Parameters MAPE	Parameters RMSE	Prediction MAE	Prediction RMSE
Base Sample	-	-	2.89	5.16
15 Days	4.78 (41.58)	0.007 (0.42)	2.89	5.16
7 Days	14.15 (107.07)	0.133 (1.07)	2.89	5.15
5 Days	6.64 (145.15)	0.008 (1.45)	2.9	5.17
3 Days	18.08 (220.69)	0.101 (2.21)	2.89	5.14
2 Days	9.47 (293.47)	0.026 (2.94)	2.89	5.15
1 Days	28.76 (454.63)	0.205 (4.55)	2.9	5.15
Departures	**Parameters MAPE**	**Parameters RMSE**	**Prediction MAE**	**Prediction RMSE**
Base Sample	-	-	2.88	5.24
15 Days	7.1 (41.48)	0.015 (0.41)	2.88	5.24
7 Days	9.72 (108.01)	0.036 (1.08)	2.88	5.23
5 Days	10.7 (146.53)	0.063 (1.47)	2.88	5.25
3 Days	16.13 (219.08)	0.07 (2.19)	2.87	5.21
2 Days	13.23 (291.7)	0.101 (2.92)	2.87	5.21
1 Days	39.26 (449.95)	0.792 (4.5)	2.87	5.21
Destination Choice	**Parameters MAPE**	**Parameters RMSE**	**Predictive LL**	**% of Correct Prediction**
Base Sample			-14798.9	11.9
20000 Trips	13.61 (59.13)	0.174 (0.59)	-14801.9	11.89
10000 Trips	25.83 (122.68)	0.405 (1.23)	-14805.9	11.83
5000 Trips	29.1 (216.06)	0.436 (2.16)	-14807.4	11.82
3000 Trips	27.41 (309.46)	0.359 (3.1)	-14813.1	11.67
2000 Trips	66.08 (399.28)	0.892 (3.99)	-14828.6	11.89
1000 Trips	82.04 (657.65)	1.22 (6.73)	-14829.3	11.85

For destination choice models, as sample size decreases, the error measures increase. The MAPE (RMSE) of model performance to generate the estimated parameters of base model increases from 13.61 (0.174) of sample size of 20,000 trips to 82.04 (1.220) of sample size of 1,000 trips. Again, as expected, the standard error of estimate increases when sample size reduces. Further, we can clearly see that as sample size decreases, the prediction capability of estimated models marginally reduces as highlighted by the predictive LL and percentage of correct prediction measures. In total, based on the results and considering the increase in standard error, we can recommend an ideal sample size of 5000 trips for sample size for examining users' destination choice and more generally users' behaviour towards BSS.

Conclusion

This paper examined the impact of sample size on hourly usage and users' destination choice preferences employing data from New York City CitiBike's. Towards this end, we evaluated the BSS data from two perspectives: 1) system usage – what contributing factors influence hourly arrival and departure rates at station level, 2) user destination choice – what factors contribute to users' preference of destination station choice. For system models, we estimated linear mixed models for hourly arrivals and departures on one month of data as our base model and compared it with the estimation on a set of smaller samples (15 days, 7 days, 5 days, 3 days, 2 days and 1 day – two sets each). For destination choice models, we estimated multinomial logit model on 50,000 trips as our base model and again on two sets of smaller samples (20000, 10000, 5000, 3000, 2000, 1000 trips). We examined the impact of sample size on the estimation results based on three measures: the capability to produce the same parameters estimate of the base sample, the comparison of standard errors and the prediction capability.

As expected, the estimation on the smaller samples provided different values for the estimated parameters and standard error of estimates. For usage, the performance of models on smaller sample size to produce the estimated parameters were within 20% of base case until two days of data. However, the standard error increase was too high. We observed a huge jump in MAPE or RMSE when we used only one day to estimate arrivals or departure models. The results clearly indicated that when sample size decreases, standard error of estimates increases and the confidence in estimated parameters reduces. In terms of prediction capability, we were not able to observe any significant difference due to use of smaller samples. For destination choice models, as sample size decreased, the error measures increased. The MAPE (RMSE) of model performance to generate the estimated parameters of base model increased from 13.61 (0.174) for sample size of 20,000 trips to 82.04 (1.220) for sample size of 1,000 trips. Again, as expected, the standard error of estimate increased when sample size reduced. Further, we observed that as sample size decreases, the prediction capability of estimated models marginally reduced. The results suggested an ideal sample size of three days data to analyse BSS demand and 5000 trips for examining users' destination choice. The results from our paper can be further generalized by conduction such sampling analysis for other BSS.

To be sure, the study is not without limitations. While the two model structures have been extensively tested, these specifications might not be applicable for other regions. Hence, the transferability of sample sizes cannot be generalized to other urban regions. However, considering the guidelines from our research will ensure that the sample sizes employed for analyzing BSS are reasonable even for other urban regions.

References

Faghih-Imani A. and N. Eluru, 2014. Role of Bicycle Sharing System Infrastructure on Usage: Evidence from Montreal, Presented at the 5th Innovations in Travel Modeling Conference, Baltimore, MD.

Faghih-Imani A., and N. Eluru, 2015. Analyzing Bicycle Sharing System User Destination Choice Preferences: An Investigation of Chicago's Divvy System, Journal of Transport Geography, Vol. 44, pp. 53-64.

Faghih-Imani A., N. Eluru, A. El-Geneidy, M. Rabbat, and U. Haq, 2014. How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in Montreal, *Journal of Transport Geography*, Vol. 41, pp. 306-314.

Fishman, E., 2015. Bikeshare: A Review of Recent Literature, forthcoming *Transport Reviews.*

Gebhart, K. and Noland, R., 2014. The Impact of Weather Conditions on Bikeshare Trips in Washington, DC. Transportation, Vol. 41, pp. 1205-1225.

McFadden, D. Modeling the Choice of Residential Location. In *Spatial Interaction Theory and Planning Models* (A. Karlqvist et al., eds.), North Holland Publishers, Amsterdam, Netherlands,1978.

Nair, R., Miller-Hooks, E., Hampshire, R., and A., Busic, 2013. Large-Scale Vehicle Sharing Systems: Analysis of Velib. International Journal of Sustainable Transportation 7, 85-106.

O'Brien, O., Cheshire, J., and M. Batty, 2014. Mining bicycle sharing data for generating insights into sustainable transport systems, Journal of Transport Geography, Volume 34, Pages 262-273.

Ricci, M. (2015). Bike sharing: a review of evidence on impacts and processes of implementation and operation. Research in Transportation Business & Management, 15, 28-38.

Rixey, R., 2013. Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three U.S. Systems. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2387, pp. 46-55.

Rudloff, C. and Lackner, B., 2014. Modeling Demand for BikeSharing Systems: Neighboring Stations as Source for Demand and Reason for Structural Breaks. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2430, pp. 1-11.

Shaheen, S., S. Guzman and H. Zhang, 2010. Bikesharing in Europe, the Americas, and Asia Past, Present, and Future. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2143, pp. 159-167.

Wang, X., G. Lindsey, J. Schoner and A. Harrison, 2015. Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations. Forthcoming *Journal of Urban Planning and Development*.

Zhao, J., W. Deng and Y. Song, 2014. Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in China. *Transport Policy,* Vol. 35, pp. 253-264.