# A LATENT SEGMENTATION MULTINOMIAL LOGIT APPROACH TO EXAMINE BICYCLE SHARING SYSTEM USERS' DESTINATION PREFERENCES

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

## Introduction

A bicycle-sharing system (BSS) is intended to provide increased convenience to individuals because they can use the service without the costs and responsibilities associated with owning a bicycle for short trips within the service area of the system. These systems are recognized to have traffic and health benefits such as flexible mobility, physical activity, and support for multimodal transport connections (Shaheen et al., 2010). Given the recent rapid growth of bicycle-sharing systems as a viable and sustainable mode of transportation for short trips, there is substantial interest in identifying contributing factors that encourage individuals to use these systems. This paper looks at BSS behavior at a trip level to analyze bicyclists' destination preferences using a random utility maximization approach. Understanding the individuals' decision processes in adoption and usage of bicycle-sharing systems will enable bicycle-sharing system operators/analysts to enhance their service offerings. Specifically, we study the decision process involved in identifying destination locations after picking up the bicycle at a BSS station. There have been several location choice studies in traditional travel demand literature that adopt a random utility maximization approach for understanding destination/location preferences (Chakour and Eluru 2014 Waddell et al. 2007; Sivakumar and Bhat, 2007). In this paper, we adapt this approach to the bicycle-sharing system data.

The decision process is studied using a random utility maximization approach where individuals choose the destination that offers them the highest utility from the universal choice set of stations in the study region. In the random utility maximization approach, the destination station utility is affected by temporal and weather characteristics (such as time period of the day or temperature), trip attributes (such as trip distance), users attributes (such as age and gender), and origin and destination characteristics (such as capacity, and built environment attributes around the stations). The assumption that the influence of exogenous variables remain the same for the entire spectrum of trips might lead to biased estimation for several reasons. First, individuals might use BSS for different trip purposes such that one exogenous attribute might be appealing for only a segment of users and not for the rest of users. For example, stations around park areas are more likely to be the destination of recreational trips while they are not favorable stations for work commute trips. Second, different users may have different preferences within their destination choice decision process. For example, female users might consider specific barriers towards the use of BSS for commute to work. Also, dedicated cyclists might have proclivity to a regular set of routes and subsequently stations while leisure cyclists might consider bicycle specific infrastructures. Third, there might be heterogeneity in trips made by BSS users which may be unobserved to the analyst.

In this study, rather than homogenize the entire spectrum of trips we examine the presence of potential segments within the trip population. Specifically, by developing a Latent Segmentation Multinomial Logit Model (LSMNL), we allow for various segments and segment specific destinations choice models to enhance our understanding of the destination choice behavior. The proposed approach also addresses a specific limitation of the traditional MNL model. In a destination choice based MNL model only attributes that vary across alternatives within choice set can be considered for model estimation i.e. the model is limited to destination attributes. The consideration of socio-demographics and other attributes fixed across all alternatives (such as origin variables or temporal and meteorological characteristics) is possible only through their interaction with destination attributes. However, within the LSMNL framework, we can

account for such fixed attributes through the segmentation component of the model. In this paper, we employ data from the New York City bicycle-sharing system (CitiBike) for 2014.

**Methodology**

Multinomial Logit Models (MNL) are commonly employed to analyze location choice in the transportation literature. The Latent Segmentation Multinomial Logit framework allows us to probabilistically classify trips into latent segments based on a host of characteristics including trips, origin and destination attributes. A brief description of the LSMNL model employed in our study is provided below.

Let us consider S homogenous segments of trips (the optimal number of S is to be determined). The utility for assigning a trip $q$ *(*1,2,..*Q)* to segment $s$ is defined as:

$$U_{qs}^{*} = \beta_s' z_q + \xi_{qs} \quad (1)$$

$z_q$ is a (M x 1) column vector of attributes that influences the propensity of belonging to segment $s$, $\beta_s'$ is a corresponding (M x 1) column vector of coefficients and $\xi_{qs}$ is an idiosyncratic random error term assumed to be identically and independently Gumbel-distributed across trips $q$ and segment $s$. Then the probability that trip $q$ belongs to segment $s$ is given as:

$$P_{qs} = \frac{\exp(\beta_s' z_q)}{\sum_s \exp(\beta_s' z_q)} \quad (2)$$

Now let us assume $k$ (1,2, … K, in our study K=30) represents destination station alternatives, then the random utility formulation takes the following form when a trip probabilistically assigned to a segment s and station k is chosen as destination:

$$U_{qk} | \, s = \alpha_s' \, x_q + \varepsilon_{qk} \quad (3)$$

$x_q$ is a (L x 1) column vector of attributes that influences the utility of destination choice model. $\alpha$ is a corresponding (L x 1)-column vector of coefficients and $\varepsilon_{qk}$ is an idiosyncratic random error term assumed to be identically and independently Gumbel-distributed distributed across the dataset. Then the probability that trip $q$ chooses station $k$ as destination within the segment $s$ is given as:

$$P_q(k) \mid s = \frac{\exp(\alpha_s' x_q)}{\sum_k \exp(\alpha_s' x_q)} \quad (4)$$

Within the latent segmentation framework, the probability of trip $q$ to be destined at station $k$ is given as:

$$P_q(k) = \sum_{s=1}^{S} (P_q(k) \mid s)(P_{qs}) \quad (5)$$

Therefore, the log-likelihood function for the entire dataset is:

$$L = \sum_{q=1}^{Q} \log(P_q(k_q^*)) \quad (6)$$

where $k_q$* represents the chosen destination station for trip $q$. By maximizing this log-likelihood function, the model parameters $\beta$ and α are estimated. The maximum likelihood model estimation is programmed in GAUSS matrix programming language.

**Data**

New York's CitiBike system is the latest major public BSS around the world and the largest in the United States. The service was launched in May 2013 with 330 stations and 6,000 bicycles in the lower half of Manhattan and some part of northwest of Brooklyn. The data used in our research was obtained primarily from the CitiBike website (https://www.citibikenyc.com/system-data). The CitiBike website provides trip dataset for every month of operation since July 2013. The trip dataset include information about origin and

destination stations, start time and end time of trips, user types, and the age and gender for members' trips. Additionally, the stations' capacity and coordinates as well as trip duration are also provided in the dataset. The built environment attributes such as bicycle routes and subway stations were derived from New York City open data (https://nycopendata.socrata.com) while the socio-demographic characteristics of resident population were gathered from US 2010 census and the weather information corresponding to the Central Park station was retrieved from the National Climatic Data Center (http://www.ncdc.noaa.gov/data-access).

We employed data for trips for the year 2014. We separated trips made by members and daily customers; about 90% of all the trips were made by members. For the sake of brevity, we focus on members for this paper. The sample formation exercise also involved a series of steps. First, trips with missing or inconsistent information were removed. Second, trips longer than 2 hours in duration (only 0.5% of all the trips) were deleted considering that the trips longer than 2 hours 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, modeling intentional same origin and destination trips would require additional trip purpose information which is not available in the dataset currently. Therefore, we focus on trips that were destined outward. Further, to obtain a reasonable sample size for model estimation, 10,000 trips were randomly selected from entire year of 2014. This sample size was adopted to maintain a reasonable data processing and model estimation related computational effort.

CitiBike system has 332 stations across the city in 2014. From each origin station, individuals have 331 other stations to choose to return the bicycle to. However, considering all the stations in the universal choice set will result in substantial computational burden. Hence, for the purpose of the modeling effort, for every destination choice record, a sample of 30 alternatives from the universal choice set including the chosen alternative was randomly selected. The process of random sampling does not affect the parameter estimates in multinomial logit models (see McFadden, 1987; Faghih-Imani and Eluru, 2015). A descriptive summary of sample characteristics is presented in Table 1.

*Table 1 Descriptive Summary of CitiBike Sample Characteristics*

| **Continuous Variables** | **Min** | **Max** | **Mean** | **Std. Deviation** |
|---|---|---|---|---|
| **Temperature (°C)** | -15 | 33.3 | 16.35 | 9.09 |
| **Relative Humidity (%)** | 13 | 100 | 56.34 | 18.06 |
| **Length of Bicycle Facility in 250m Buffer (m)** | 0.0 | 3355.3 | 1027.08 | 591.25 |
| **Area of Parks in 250m Buffer ($m^2$)** | 0 | 95209.9 | 10186.22 | 15159.82 |
| **Number of Restaurants in 250m Buffer** | 0 | 545 | 54.20 | 92.11 |
| **Number of CitiBike stations in 250m Buffer** | 0 | 4.00 | 1.24 | 1.01 |
| **Capacity of CitiBike stations in 250m Buffer** | 0 | 169.00 | 44.15 | 38.85 |
| **Station Capacity** | 3.00 | 67.00 | 34.40 | 10.79 |
| **Pop Density (people per $m^2$ ×1000)** | 0.01 | 67.20 | 24.90 | 14.69 |
| **Job Density (jobs per $m^2$ ×1000)** | 0 | 432.52 | 55.98 | 53.79 |
| **Trip Distance (km)** | .05 | 10.32 | 2.00 | 1.33 |
| **Trip Duration (min)** | 1.02 | 117.83 | 11.92 | 8.44 |
| **Members Age** | 17.00 | 90.00 | 37.99 | 11.25 |
| **Categorical Variables** | **Percentage** | | | |
| **Weekends** | 21.5 | | | |
| **Subway Station in 250m Buffer** | 49.7 | | | |
| **Path Train Station in 250m Buffer** | 4.2 | | | |
| **Female Members** | 22.7 | | | |

**Results**

This section discusses the estimation results of latent segmentation multinomial logit model (LSMNL) with two segments to understand the different factors influencing users' choice of destination in the New York City's CitiBike bicycle-sharing system. The final Log-likelihood values of the LSMNL model is -28266.2 while the corresponding value for the simple MNL model is -28450.98. We use the Bayesian Information Criterion (BIC) to compare two models. BIC penalizes the modelling framework for additional parameters. For a given empirical model, $BIC = Kln(Q) - 2ln(L)$ where *K* is the number of parameters, $Q$ is the number of observations and *ln(L)* is the log–likelihood value at convergence. The model with the lowest value of BIC is preferred. The BIC value for LSMNL model is 56882.35 and the BIC value for MNL model is 57012.48. The improvement in the data fit clearly illustrates the superiority of the LSMNL based destination choice models; providing strong evidence in favour of our hypothesis that BSS users' decision process can be better investigated through segmentation of trips. Moreover, the LSMNL allows us to capture a behaviourally richer contributing factors. The model specification process was guided by intuition and parsimony considerations. The estimation results for two segments LSMNL model is presented in Table 2.

It is important to discuss the overall segmentation attributes. We can estimate the trips share across the two segments as well as the distribution of independent variables within each segment. These estimates are presented in the top panel of Table 2. The probability of trips belonging to segment 1 is substantially higher than the probability of being allocated to segment 2. Further, trips in segment 2 are pursued more during weekends and when temperature is higher. Female users are more likely to be assigned to segment 2. Stations near park areas are more likely to be origins of trips in segment 2. Further, trips in segment 2 are more likely to depart from and arrive ar stations located in lower job density areas. Based on these differences in the mean values of the exogenous variables across the two segments, we can characterize our segments. Regular and daily commute trips are assigned to segment 1 while casual and recreational trips are assigned to segment 2.

The latent segmentation component determines the probability that a trip is assigned to one of the two segments identified. In our modelling effort we select segment 2 as the base and the estimated coefficients correspond to the utility for being a part of the segment 1. The positive constant term indictaes a larger likelihood for trips to be assigned to segment 1. Origin attributes including presence of transit station in 250m buffer, station elevation, population density, area of parks in 250m buffer and distance to CBD influence the segmentation estimation. Users' gender, as well as temporal and weather characteristics are also found to have significant impact on the segment share.

The destination choice component provides contributing factors in users' destination station preferences. The positive coefficients for station capacity variable in both segments demonstrate that stations with higher capacity are more likely to be chosen as they are likely to have more available docking stations. Moreover, people tend to easily remember larger stations. The number of stations and the capacity of stations within the buffer variables take into account the impact of neighbouring stations on destination choice. It must be noted that the coefficients of number and capacity of stations in the buffer should be examined as a combination recognizing that as the number of stations in the buffer increases we simultaneously increase the capacity in the buffer. The length of rails variable has negative impact on the propensity of choosing a station in segment 1; it is expected as railway tracks typically act as barriers to pedestrian and bicyclist movements. The area of park in 250 meter buffer variable has negative influence on destination choice propensity in segment 1 while has positive influence in segment 2. The opposite influence in two segments is in agreement with our interpretation of the two segments. The segment with more casual and recreational trips have positive coefficient for the area of park in buffer variable. A simple MNL model would not have been able to disentangle this difference across segments.

*Table 2 LSMNL estimation results*

| **Segment Characteristics and Mean Values of Segmentation Variables** | **Segment 1** | | **Segment 2** | |
|---|---|---|---|---|
| Trips Share | 0.80 | | 0.20 | |
| Female | 0.21 | | 0.30 | |
| Temperature | 15.8 | | 18.6 | |
| Weekend | 0.20 | | 0.27 | |
| Area of Parks in 250m Buffer of Origin Station | 0.0081 | | 0.0104 | |
| Job Density of Origin Station | 0.0672 | | 0.0592 | |
| Job Density of Destination Station | 0.0661 | | 0.0625 | |
| **Latent Segmentation Component** | **Segment 1** | | **Segment 2 (Base)** | |
| | Coefficient | t-stat | Coefficient | t-stat |
| **Constant** | 2.7806 | 8.293 | - | |
| **Temporal and Weather Variables** | | | | |
| AM | -0.6002 | -2.733 | - | |
| Midday | -0.5064 | -2.359 | - | |
| PM | -0.7273 | -3.606 | - | |
| Weekend | -0.3630 | -2.656 | - | |
| Temperature | | | | |
| **Origin Built Environment Attributes** | | | | |
| Presence of Transit Station in 250m Buffer | 0.3310 | 2.631 | - | |
| Elevation | 0.1000 | 1.787 | - | |
| Population Density | -0.1373 | -2.311 | - | |
| Area of Parks in 250m Buffer | -0.1980 | -3.833 | - | |
| Distance to CBD | 0.3802 | 3.976 | - | |
| **Trip Attributes** | | | | |
| Female | -0.3676 | -2.770 | - | |
| **Destination Choice Component** | **Segment 1** | | **Segment 2** | |
| | Coefficient | t-stat | Coefficient | t-stat |
| **Destination Built Environment Attributes** | | | | |
| Destination Station Capacity | 0.1656 | 10.184 | 0.3084 | 6.506 |
| Number of Other Citibike Station in 250m Buffer | - | - | -0.5694 | -4.004 |
| Capacity of Other Citibike Station in 250m Buffer | -0.0010 | -2.438 | 0.0078 | 2.138 |
| Length of Rails in 250m Buffer | -0.0371 | -2.455 | - | - |
| Area of Parks in 250m Buffer | -0.0463 | -2.706 | 0.0738 | 1.676 |
| Number of Restaurants in 250m Buffer | 0.0206 | 1.739 | - | - |
| Population Density | 0.0387 | 2.136 | 0.3172 | 6.143 |
| Population Density * AM | -0.2195 | -6.198 | -0.2140 | -2.083 |
| Job Density | -0.1062 | -5.270 | - | - |
| Job Density * AM | 0.3170 | 9.208 | 0.4602 | 6.610 |
| Distance to CBD | - | - | -0.6406 | -8.416 |
| Elevation | - | - | -0.2373 | -4.400 |
| **Trip Attributes** | | | | |
| Dummy Network Distance to Destination < 750m | -1.0341 | -14.664 | -1.3059 | -4.551 |
| Network Distance (750m< &<4000m) | -0.9285 | -29.245 | -1.1787 | -7.884 |
| Network Distance (750m< &<4000m)*Temperature | 0.9064 | 5.983 | - | - |
| Network Distance (750m< &<4000m)*Female | 0.1017 | 2.978 | - | - |
| Dummy Network Distance to Destination > 4000m | -5.0699 | -25.705 | -1.9481 | -6.978 |

In segment 1, stations with higher number of restaurants in the vicinity are more likely to be selected as destinations. In both segments, population density has positive impact on propensity of choosing a CitiBike station. However, the impact is negative in AM period given that trips in AM period are most likely originated from residential areas and are destined to work locations. In segment 1, job density variable has negative impact except in AM period. In fact, for both segments, job density variable has significant positive impact on likelihood of choosing a station as destination. For segment 2, CitiBike users select stations that bring them closer to CBD as highlighted by negative coefficient of destination station distance to CBD. Moreover, they prefer stations with lower elevations indicating preference of downhill trips within segment 2.

The most important variable in destination station decision making process in a BSS is expected to be the distance of trip between origin and destination. In general, it is expected that the likelihood of choosing a station very close to origin station or very far from origin station is lower than stations in between. In order to better model the distance impact on the utility of choosing a station, we distinguish the very short distance and very far distance by indicator variables and a continuous variable for distance in between. We tried several different distance combinations determining thresholds for distance to define close and far stations from the origin. Indicators variables identifying stations within 750m or farther than 4000m of origin and a continuous form of distance for stations within 750m to 4000m from origin provided better results. As expected, the network distance variables have negative impact on likelihood of choosing a station as destination for both segments. In segment 1, the interaction of distance variable with gender and temperature variables are significant. The results show that female members are more likely to have longer trips. This might be due the fact that in New York CitiBike system, only about 22.7% of members are female. It is possible that women who join the system are actually regular bicyclists and are more likely to be fit and pursue longer trips. Further, when temperature increases, it is expected that users pursue longer trips as indicated by positive variable of temperature and distance interaction variable. The huge difference in coefficients of dummy variable for long trips between two segments is also another evidence in support of our behaviour interpretation of segments. The casual and recreational trips are less sensitive to very long distance trips.

**Conclusion**

This paper presented a Latent Segmentation Multinomial Logit Model to examine BSS users' destination choice preferences. We allowed for various segments and segment specific destinations choice models to enhance our understanding of the BSS users destination decision behavior. The models were estimated using data from the New York City bicycle-sharing system (CitiBike) for 2014. The LSMNL model outperformed the simple MNL model in terms of goodness of fit indicating the presence of unobserved heterogeneity within trips spectrum and superiority of examining BSS users' destination choice preferences through endogenous segmentation framework. Based on distribution of exogenous attributes we characterized our segments as: segment 1 is more likely to have regular and daily commute trips while segment 2 is more likely to be assigned to casual and recreational trips. Within destination choice component, coefficients of area of park in the buffer variable and long distance trip dummy variable also illustrated our interpretation of the two segments.

**References**

Chakour V. and N. Eluru, 2014. Analyzing Commuter Train User Behavior: A Decision Framework for Access Mode and Station Choice, *Transportation*, Vol. 41 (1), pp. 211-228.

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.

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.

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.

Sivakumar, A., and C.R. Bhat, 2007. A Comprehensive, Unified, Framework for Analyzing Spatial Location Choice, *Transportation Research Record: Journal of the Transportation Research Board,* No. 2003, pp. 103-111.

Waddell, P., C.R. Bhat, N. Eluru, L. Wang and R.M. Pendyala, 2007. Modeling the Interdependence in Household Residence and Workplace Choices, *Transportation Research Record: Journal of the Transportation Research Board,* No. 2003, pp. 84-92.