DEMOGRAPHIC FEATURES AND MODE TRANSFER PATTERNS OF TNC USERS: EVIDENCE FROM 2017 NHTS DATA

Bowen Xiao
Statistics Department
University of Washington
Tel: 206-291-0453
Email: xiaobw95@uw.edu

Xiatian Wu
Department of Civil & Environmental Engineering
Department of Urban Design & Planning
University of Washington
Tel: 313-888-4037
Email: wxt47@uw.edu

Yiran Zhang
Department of Civil & Environmental Engineering
University of Washington
Tel: 206-240-7662
Email: yiranz94@uw.edu

Submission Date: September 30th, 2018

Paper submitted for NHTS Data Challenge

INTRODUCTION

New mobility services provided by transportation network companies (TNCs) have significantly transformed people's mode choice and travel behavior. Most of TNCs have initiated some programs to enhance their roles in mode connection, activity engagement and social equity promotion. For instance, Uber has launched a pilot to subsidize rides within Orlando to and from commuter rail stations, Lyft has partnered with Phoenix for dealing with 'first mile, last mile' dilemma, and non-emergency medical transportation delivering service has also been reshaped in the era of TNCs.

Despite the explosive growth of TNC services, few empirical researches have adequately specify the characteritics of TNC users and TNC trips. The challenges of analyzing TNC services may include followings: (1) temporally, they only exist for a relatively short time, thus their impacts are still inconclusive, (2) spatially, the services are so far constrained on urban areas, thus the observations maybe less representative for certain regions and groups of population, (3) TNCs are strongly caucious about data publicity for business strategies and privacy concerns. To our best knowledge, 2017 National Household Transportation Survey (NHTS) is the first comprehensive survey since the explosion of TNC mobility services nationally, thus it is quite timely to analyze their impacts now. In this study, both taxi/limo (Uber/Lyft) and rental car (Car2go/Zipcar) are taken into analysis, which results in 1,474 TNC users with 3,116 TNC trips among 7,458 daily trips in total in the dataset.

The aim of this study is to explore the role of taxi/limo and rental car on residents' daily trip chain patterns. In particular, the study includes the following tasks: (1) to summarize the characteristics of TNC trip chain from temporal, spatial and demographical perspectives, (2) to investigate the relationship between demographic variables and TNC usage by constructing Bayes Belief Network (BBN), (3) to estimate the variables that contribute to TNC trips with Poisson model; and (4) to examine the mode transfer patterns of TNC users by hierarchical semantic model, (5) to put forward some suggestions on how TNC can play a better role on current mobility network.

CHARACTERISTICS & TYPOLOGIES OF TNC TRIP CHAIN

Trip chain refers to a sequence of trips to single or multiple anchor destinations, begin and end at home *(1,2)*. The rules of connecting TNC user trips segments to form trip chains were derived and adjusted based on the classification of activities: (1) Mandatory activities (m) which has fixed frequency, location and timing, (2) flexible activity (f) which is perfomed on a regular basis but have some flexible characteritics, (3) optional activity (o) which can vary for all characteristics. Depending the priority of the activity, mandatory activity is also referred to primary activity (p), while flexible and optional activity are referred to secondary activities (s). TABLE 1 summarizes the typologies of trip chains.

Table1 Typologies of trip chains

Configration		Note
Work Based		
1	h-p-h	Simple trip chain only with mandatory activities (eg. work, school)
2	h-(s)-p-(s)-h	Flexible/optional activities on the way to or from mandatory activities
3	h-(s)-p-s-p-(s)-h	Flexible/optional activities taking place during mandatory activities
Non-Work Based		
1	h-f-h	Simple trip chain only with flexible activities (eg. banking, shopping)
2	h-o-h	Simple trip chain only with optional activities (eg. social, recreation)

3	h-f/o-h	Trip chain with both flexible and optional activities

FIGURE 1 demostrates the corresponding distribution among all trip chains of TNC users, by applying weights. It indicates that the work based and non-work based trip chains have a share of 57.92% and 44.85%, respectively. The simple 'h-p-h' trip chains gain the largest share of 36.25%.

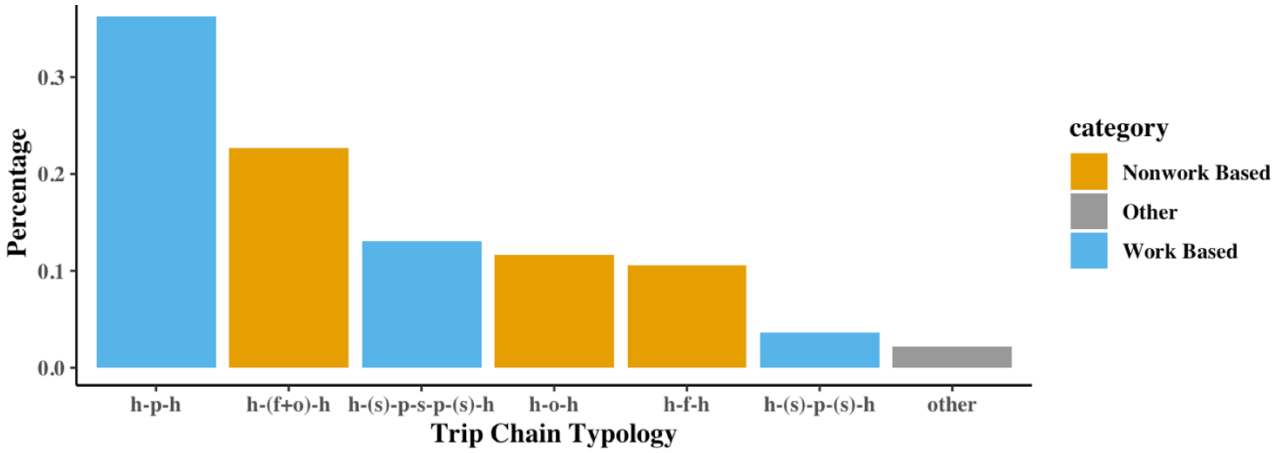

Figure1 Weighted distribution of TNC trip chains

RESAMPLING FITTING

This part illustrates the reason to apply bootstrap instead of just applying weight for processing survey data before modeling.

Jackknifing is a converntional way to reduce bias, which, in our case, can be specified as the following equation. However, it is not ideal for models like Poisson Regression, since the weighted counts are not integers anymore. Inspired by the idea of function fitting, a framework called *Resampling Fitting* was developed: Firstly, a ratio as hyperparameter was chosen (0.1 for this project), then the weighted m-out-of-n bootstrapping was used to mimic *Jackknifing*, resulting with very close results in estimation and corresponding standard error of marginal and conditional probabilities (some results are shown in `src/bootstrape.html`). The results were also compared in more complicated calculation, like Cramer's V. With the validation of its accuracy, this bootstrapping strategy was applied in BBN and Poisson Regression. One obvious advantage is that any statistics can be calculated based on bootstrapping since it returns as a distribution. Moreover, bootstrapping is known to be consistent in more cases and also implemented in some modern models/algorithms. Thus, this method helps to broaden and deepen the exploration of the survey data.

$$SE = \sqrt{\sum_{i=1}^{98} \left(\frac{6}{7}\right) [REP_i - x]^2} \quad [1]$$

MODELING & FINDINGS

Missing Value Network of TNC Usage

First of all, the correlations between TNC usage and personal demographic features (education level, gender, race, age level and health condition) were explored. Most of the variables were derived from the original survey dataset, while 'R_AGE' was discretized to 3 levels which was denoted by 'AGE_LEVEL'. Besides, the use of TNC was denoted by 'USES_TNC' accroding to the value of 'RIDESHARE' variable. (see `data/csv/2017/ derived_variable_config.csv`).

To measure the associations between USES_TNC and demographic features, Cramer's V was calculated based on Pearson's chi-squared statistic (see Appendix A-1). The results range from 0 (no association between the variables) to 1 (complete association) [4]. It turned out that education level, age level and health condition have stronger associations than others, see FIGURE 2(left).

However, since demographic features are not necessarily independent from each other, a probability network (see Appendix A-2) was constructed beside of pair-wise associations . To make the best use of dataset, Random Forrest was implied in nonparametric imputation to handle the missing values. It is found that out-of-bag proportion of falsely classified of *USES_TNC*, *AGE_LEVEL* and 'R_SEX' (which contains ~19% missing values after weighted) is smaller than 50%. Therefore, we followed this imputation strategy. Entries with unimputed missing values were dropped and new boolean variables were generated to encode their missing patterns (see `data/csv/2017/ derived_variable_config.csv`). Finally, the remaining data was trained in BBN learning and the network was demonstrated by FIGURE 2 (right).

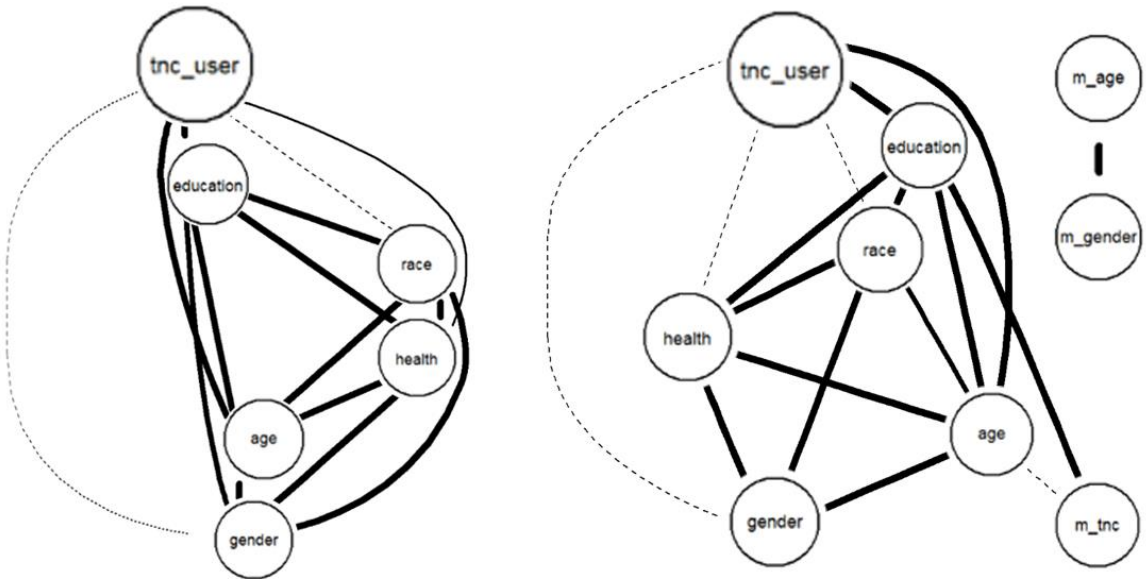

Figure2 BBN without (left) & with (right) missing pattern

(the thicker the line is, the stronger the association is beween the two variables in two ends)

Findings: With Comparison, BBN with missing pattern encoding indicates a weaker association between 'tnc_user' and 'health', while the associations with 'education' and 'age' remains the same. Overall, the correlations among the selected variables suggest that the education has a stronger correlation for people to decide whether to use TNC or not. It is also interesting to find that race is less directly associated with the usage of TNC, which possibly indicates that TNC services are not strongly preferred by people from certain racial group.

Bootstrapped Poisson Regression of TNC Trip Count

Secondly, the correlation between TNC trip counts and other variables were estimated by implementing a Poisson model for futher exploration. The number of TNC trips for each person was acquired based upon the trip table from the survey data. One issue as mentioned previously is that when applying the weights into the trip counts, the number of TNC trip per person would no longer be integer. To solve this problem, the bootstrap has been introduced so that we can resample the data: the resampling time has been set as 1,000 and the sample rate is settled as 0.1.

Variables were selected based on our literature review, including age, education, income, health and occupation, URBAN (urban classification), MSASIZE (Population size category of the Metropolitan Statistical Area), OUTOFTWN (whether the trip is away from home for entire day). The model is built as:

$$\begin{aligned}\text{model} <- \text{glm}(&\text{number of TNC} \sim \text{Education} + \text{Age} + \text{Urban} + \text{OUTOFTWN} + \text{Income} \\ &+ \text{Occupation} + \text{Health} + \text{MSASIZE}, \text{data} = \text{temp}, \text{family} = 'poisson')\end{aligned}$$

Findings: The results indicated that people from different education level may choose TNC differently (see Appendix A-3). People with older age may not use TNC quite frequently, this is sensible since Uber and Lyft are more popular for people with younger age. In addition, people with higher income may be more likely to choose TNC possibly because they are less price-sensitive. Another interesting finding is that people with sale or service-related job are inclined to choose TNC more, which is also explainable since they may generate more trips for their business during the day. Besides, people with very poor health condition (HEALTH 05) may hardly choose TNC and people who live very far from the metropolitan area may generate more TNC trips.

Hierarchical Semantic Model of Travel Mode Transferring

Finally, the mode transferring patterns were examed. As people may have several trips in a day and it is reasonable to assume that there are some connections between each transportation mode transferring. Furthermore, inspired by semantic analysis in NLP, we see a discrete time series of transportation modes of one's daily trips as a *sentence*, and each mode is a *word*. But unlike typical NLP problems, people in our cases have different demographic features, that is, people cluster. So, a multi-lateral language model was built.

The model can be seen as a 1-gram model or Markov Chain, that is, each transportation mode is only dependent on the previous mode. Denote $P_{i\to j|k}$ as probability of transformation from mode i to model j under condition of education level k, then three models can be written as following:

$$P_{i\to j|k}^{fully} = \frac{\#\{mode\ j\ immediately\ after\ mode\ i\}}{\#\{mode\ i\}} \quad [2]$$

$$P_{i\to j|k}^{no} = \frac{\#\{mode\ j\ immediately\ after\ mode\ i|education\ level\ k\}}{\#\{mode\ i|education\ level\ k\}} \quad [3]$$

$$P_{i\to j|k}^{partial} = P_{i\to j|k}^{fully} + C_k \quad [4]$$

where all the counts are weighted by trip weights and C_k is a constant defined by the following optimization problem.

$$C_k = argmin_{C_k} \sum_{i\to j|k} \left(P_{i\to j|k}^{partial} - P_{i\to j|k}^{no}\right)^2 \quad [5]$$

Solve the above optimization problem, we can get the analytic form of C_k.

$$C_k = \sum_{i \to j|k} P^{no}_{i \to j|k} - P^{fully}_{i \to j|k} \quad [6]$$

To compare the performance of three models, we applied 5-fold cross validation by dividing TNC users into five folds and using each fold to test the model obtained by training on the other four. As a result, the partial-pooling model outperforms the other two with regard to both mean square error and mean absolute error (see Appendix A-4). The reason behind it should be the trade-off between bias and variance. Thus, we applied partial-pooling model to get the probabilities of transportation transfer for each education level people (see in `result/data/ transportation_transformation.csv`).

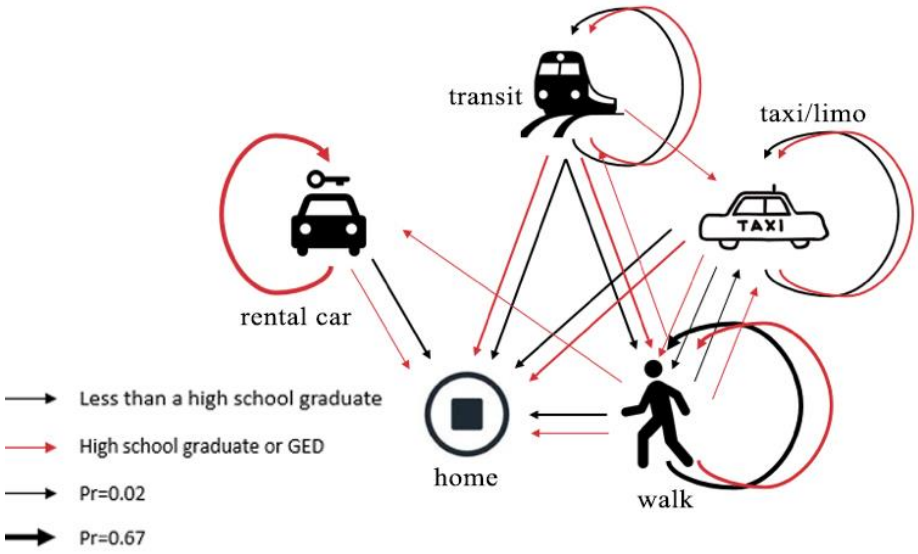

Figure 3 Partial-pooling Model (local visualization)

Findings: It is worth noting that probability property $\sum_{j|k} P^{partial}_{i \to j|k} = 1$ may not holds. And now we always give $P = 0$ to unobserved cases. One interesting finding is people with higher education seem to have a more diverse mode transfer patterns than less educated ones. More specifically, the formers have a higher probability of chaining trips with rental cars. Besides, if transit is intermediate mode of going home, then the formers are more likely to transfer to taxis. In general, if taking a taxi, people, whatever their education background, prefer to arrive home directly, other than taking an extra walk trip.

POLICY IMPLIMENTION

The results from the analysis reveal that certain group of population, especially senior and people in a poor health condition, are less likely to choose TNC as their travel modes. For the former group of population, one possible reason is that they don't have access to smart phones, or lack of knowledge of new mobility services. For the latter group of population, possibly because the current TNC services don’t provide ADA standards so that disabled people cannot receive enough assistances from both the vehicle and driver. Hence it is suggested for TNCs to apply some paratransit programs for ADA users to upgrade the level of services, promote social equity, as well as open a wider business market.

APPENDIX

A-1 The results of Cramer's V calculation

Var	CV.RW	CV.Boot	sd.RW	sd.Boot
Education	0.2028	0.2030	0.0048	0.0058
Race	0.0468	0.0491	0.0060	0.0071
Health	0.1328	0.1334	0.0054	0.0059
Age Level	0.1551	0.1553	0.0050	0.0049
Sex	0.0229	0.0225	0.0031	0.0062

Table 2 Cramer's V (replicate weights vs bootstrap)

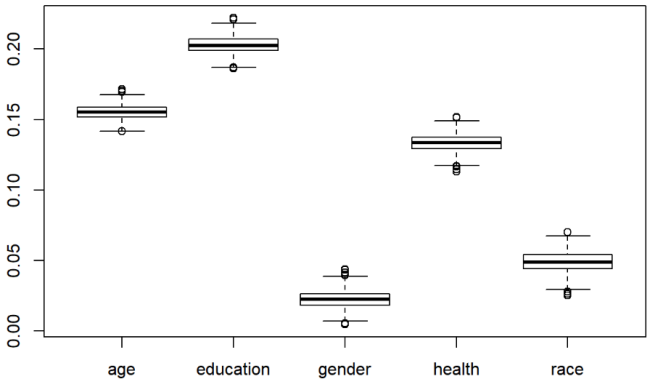

Figure4 Cramer's V (bootstrap)

A-2 The process of building probability network

At first, we simply dropped entries with missing value (which contains ~16% data after weighted) and obtained the following network. The widths of linked lines show the strength of association (the wider the line is, the stronger the association is), which is computed as the probability of observing the association in the bootstrap replicates. Combined with our Resampling Fitting, we were implementing a double-bootstrapping here. A 10% stratified sampling with replacement followed by a n-out-of-n bootstrapping in BBN algorithm.

A-3 Coefficients of Bootstrapped Poisson Regression

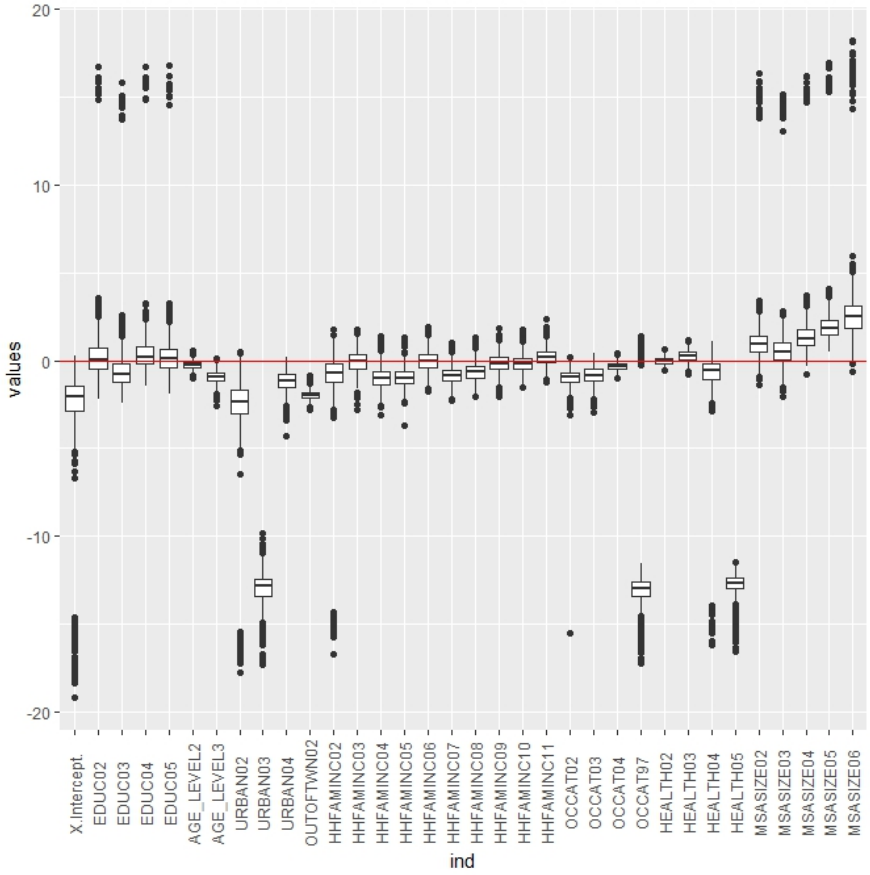

Figure6 Correlation among trip counts and selected variables

A-4 Comparison of three models of travel mode transferring

To compare the performance of three models, we applied 5-fold cross validation by dividing TNC users into five folds and using each fold to test the model obtained by training on the other four. And the results are shown as following.

Table3 Model Comparison

Method	MSE	MAE
Fully-pooling	0.026	0.0776
No-pooling	0.021	0.0723
Partial-pooling	0.019	0.0717

REFERENCES

1. Hensher, D.A., Reyes, A.J. Trip chaining as a barrier to the propensity to use public transport. Transportation 27, 2000, pp. 341–36.
2. McGuckin, N., Nakamoto, Y. Trips, Chains, and Tours—Using an Operational Definition. NHTS Conference, Nov 1-2, 2004.
3. Federal Highway Administration. 2017 NHTS Data User Guide. 8 Mar. 2018.
4. “Cramér's V.” Wikipedia, Wikimedia Foundation, 6 July 2018, en.wikipedia.org/wiki/Cram%C3%A9r%27s_V#Calculation.