FUSING MULTIPLE SOURCES OF DATA TO UNDERSTAND RIDE-HAILING USE

Felipe F. Dias
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 512-471-4535; Email: fdias@utexas.edu

Patrícia S. Lavieri
The University of Melbourne
Department of Infrastructure Engineering
Grattan Street, Parkville, Victoria, 3010, Australia
Tel: +61-3-9035-3274; Email: patricia.lavieri@unimelb.edu.au
and
The University of Texas at Austin, Austin, TX 78712, USA

Taehooie Kim
Arizona State University
School of Sustainable Engineering and the Built Environment
660 S. College Avenue, Tempe, AZ 85287-3005, USA
Tel: 480-727-3613; Email: taehooie.kim@asu.edu

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 512-471-4535; Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Ram M. Pendyala
Arizona State University
School of Sustainable Engineering and the Built Environment
660 S. College Avenue, Tempe, AZ 85287-3005, USA
Tel: 480-727-4587; Email: ram.pendyala@asu.edu

ABSTRACT
The rise of ride-hailing services has presented a number of challenges and opportunities in the urban mobility sphere. On the one hand, they allow travelers to summon and pay for a ride through their smartphones while tracking the vehicle's location. This helps provide mobility for many who are traditionally transportation disadvantaged and not well served by public transit. Given the convenience and pricing of these mobility-on-demand services, their tremendous growth in the past few years is not at all surprising. However, this growth comes with the risk of increased vehicular travel and reduced public transit use, increased congestion, and shifts in mobility patterns that are difficult to predict. Unfortunately, data about ride-hailing service usage are hard to find; service providers typically do not share data and traditional survey data sets include too few trips for these new modes to develop significant behavioral models. As a result, transport planners have been unable to adequately account for these services in their models and forecasting processes. In an effort to better understand the use of these services, this study employs a data fusion process to gain deeper insights about the characteristics of ride-hailing trips and their users. Trip data made publicly available by RideAustin is fused with census and parcel data to infer trip purpose, origin/destination information, and user demographics. The fused data is then used to estimate a model of frequency of ride-hailing trips by multiple purposes.

Keywords: ride-hailing services, trip characteristics, user characteristics, mobility patterns, data fusion, secondary data.

1. INTRODUCTION

Ride-hailing services have experienced very rapid growth in the past few years. Among all new mobility services, ride-hailing has the highest penetration in the United States and a number of other countries. In 2017, Uber alone had more than 10 times the number of active subscribers of all North American car-sharing programs taken together (*1*, *2*). The level of market penetration and adoption of these services reflects the convenience that they offer to users as a reliable, lower cost (compared to traditional taxi services), on-demand, and door-to-door transportation service. Furthermore, these services do not require subscription fees and involve minimal cognitive and physical efforts on the part of the traveler when compared to other alternatives, such as driving or taking the bus.

Despite the rapid growth of these services in the urban transportation ecosystem, very little is known about the trips undertaken by these services and the travelers who use these services on a regular or irregular basis. As a result, many researchers in the travel behavior field have attempted to study ride-hailing trips through a variety of primary or secondary data sets that contain any information about these trips. Although ride-hailing companies undoubtedly collect user information and detailed trip characteristics, such data is usually not shared due to privacy concerns and proprietary value. There is, therefore, a significant gap in knowledge regarding the characteristics of ride-hailing trips and users, and this study employs a data fusion exercise in an effort to shed deep insights about these aspects of ride-hailing services. In particular, the study involves the use of data on more than one million trips undertaken using RideAustin, a ride-hailing company that began servicing the Austin, Texas metropolitan area during a period when the bigger service providers, Uber and Lyft, had temporarily ceased operations in the region. The trip data furnished by RideAustin does not contain any information about trip purpose or the traveler. In order to infer trip purpose, the fuzzy location coordinates are superimposed on a parcel map to check land use characteristics of trip origins and destinations. Trip coordinates of frequent users are mapped in order to infer home locations; and the inferred home location information is fused with census block group data to impute users' characteristics. Through such data fusion exercises, this study aims to provide detailed profiles of ride-hailing service users, at least in the context of RideAustin.

The results of the data fusion exercise are used to draw some conclusions on the nature of ride-hailing trips and users. In order to characterize and understand factors that contribute to various types of ride-hailing trips, a multivariate ordered probit model is estimated and presented; this model essentially estimates the number of ride-hailing trips for each purpose (location) while accounting for error correlations that may exist across trip purposes.

The remainder of this paper is organized as follows. The next section offers a more detailed review of recent literature. Section 3 offers a description of the RideAustin data, and the data fusion methodology adopted in this study. Section 4 presents a descriptive characterization of ride-hailing trips. Section 5 describes the modeling effort undertaken in this study. Conclusions are presented in Section 6.

2. REVIEW OF RECENT WORK ON RIDE-HAILING SERVICE USAGE

It should be recognized that a growing body of literature has been devoted to the study of ride-hailing services. Given the scarcity of publicly available data from service providers, many studies rely on specialized user surveys (*3*) or special purpose online surveys (*4-8*). Other studies have utilized the limited information available in large scale household travel surveys (*9*, *10*) or aggregate data released by service providers (*11*, *12*). More recently, some studies have begun to

utilize data released by ride-hailing companies (*13*-*17*). A few of these prior studies are discussed in greater detail in this section.

Kooti et al. (*13*) analyze large scale data covering 59 million rides over a span of seven months. The data were extracted from e-mail receipts sent by Uber and collected on Yahoo servers, allowing the researchers to examine the role of demographics (e.g., age and gender) on participation in the ride-sharing economy. They also examine the influence of surge pricing and income on ride-hailing service usage and present comparisons of usage patterns by age group, gender, and income. Gerte et al. (*14*) used one year Uber data for New York City, aggregated by taxi zone and by week (weekly pick up counts per taxi zone in the borough of Manhattan) to model ride-hailing trip generation.

Lavieri et al. (*16*) used six months' worth of trip data from an Austin-based ride-hailing company, RideAustin, together with secondary data sets such as traffic analysis zone (TAZ) level demographic data, census data, and General Transit Feed Specification (GTFS) data, to estimate a spatially lagged multivariate count model, which was used to estimate the number of trips generated in a specific zone, and a fractional split model to spatially allocate the trips based on the characteristics of candidate destination zones that attract ride-hailing trips. Komanduri et al. (*15*) conduct an exploratory analysis of the same RideAustin data and examine several characteristics of ride-hailing use, such as trip-making patterns by month and day of the week; and average trip distances and durations. Zheng et al. (*17*) use ride-hailing data provided by DiDi (based in China) and questionnaire data collected from passengers to explore the short, medium, and long-term impacts of ride-hailing on personal vehicle usage and vehicle ownership.

In general, it appears that ride-hailing services not only serve as a substitute for traditional taxi, auto, and public transit trips (*3, 6, 8, 17*), but also induce new trips that would not have taken place in the absence of such services. Rayle et al. (*3*) reports that around eight percent of users indicated that their last ride-hailing trip would not have been performed if the service were not available. Lavieri and Bhat (*8*) report that six percent of the ride-hailing trips in their sample were induced, and most of them had shopping/errand trip purposes. There is also some evidence to suggest that ride-hailing services are used to compensate for poor transit access and level of service (*5, 16*).

A few studies have attempted to characterize users of ride-hailing services and the types of trips that they are undertaking when using the services. Although there are some disparate findings reported in the literature, the general pattern reported in various studies is that ride-hailing services are used to a greater degree by younger adults (18-30 years), individuals residing in urban areas, those with a college education, and people residing in medium to high income households (*4*, *5*, *9*, *13*). The evidence suggests that the majority of ride-hailing users own personal vehicles (*4*, *9*), but vehicle ownership is lower among frequent users of ride-hailing services (*4*, *6*, *8*). Additionally, when asked if they had made the choice to dispose one or more vehicles, frequent users were more likely to answer in the affirmative (*5*).

Ride-hailing services appear to be most used for trips that are social-recreational in nature (*3, 7, 8*). For the cases where ride-hailing is being used as a substitute for driving a private car, the most frequent motivations include the desire to avoid the hassle of parking and/or to avoid driving while intoxicated (*5*). Kooti et al. (*13*) found that, although older riders use ride-hailing services less frequently than younger individuals, they tend to make longer trips and choose more expensive services.

Overall, it can be seen that there is a growing body of literature devoted to understanding the characteristics of and demand for trips by ride-hailing services. This study aims to add further

insights in this domain by fusing publicly available ride-hailing trip data with secondary data sources.

3. DATA DESCRIPTION AND DATA FUSION METHODOLOGY

This section provides a description of the data sets used in this study and the data fusion methodology adopted to obtain a more complete profile and characterization of ride-hailing usage. A map of Austin's central area can be seen in Figure 1.

3.1. Brief Description of Datasets

Several public data sets were compiled to undertake the analysis in this paper. They are as follows:

- **RideAustin Data**: The primary data source originated from RideAustin, a ride-hailing company operating in Austin, Texas. RideAustin entered the market in late May 2016, shortly after Uber and Lyft shut down their operations in the city due to disputes over local regulations. The RideAustin data (*18*) contains trip-level information, including the spatial coordinates and timestamps of pickups and drop-offs as well as anonymized user IDs that allow for the identification of the multiple rides performed by each user. To protect users' privacy, RideAustin truncated the location coordinates at the third decimal place. The original dataset contains trips that occurred between June 4, 2016 and April 13, 2017. Since ridership during the first few months was limited, the analysis in this paper only includes data from August 1, 2016 onwards. RideAustin held approximately one third of the market share of all ride-hailing trips performed in Austin during this period.

- **Austin Zoning and Parcel Data:** The City of Austin provided parcel-level data with zoning information for Austin's central area, in which most of the RideAustin trips took place. Although very detailed, the zoning and parcel data required extensive processing in order to be usable for the current study. After processing, the zoning and parcel data was used to infer trip purpose or location type. This data was also used to infer the residence location for frequent users of ride-hailing services.

- **2016 American Community Survey:** The 2016 American Community Survey (ACS) data provided by the Census Bureau was used to derive socio-economic and demographic profiles of locations where frequent ride-hailing users were believed to reside. After inferring the household locations of frequent users, their demographic characteristics were imputed using the ACS data.

3.2. Data Fusion Methodology

The data fusion and imputation methodology consisted of a series of logical steps. First, trip purpose/location type was inferred and appended to the ride-hailing trip records by mapping coordinates to the parcel level land use data, similar to Bohte and Maat (*19*). This was done to better characterize how, when, and why ride-hailing trips were made. After purpose/location information was inferred, the household locations were derived for frequent riders. By mapping locations that are visited very frequently by the same users and examining the land use/parcel profile of those frequently visited locations, it was possible to identify travelers' potential residential locations. Finally, these riders (whose residence locations were inferred and identified) had their demographic characteristics imputed from the 2016 ACS data based on their household

location mapped to census block groups. The fully fused and integrated data set was eventually analyzed to quantify the number of ride-hailing trips undertaken for various trip purposes (locations). A multivariate ordered probit model was then estimated to identify factors that influence number of trips by purpose (location).

3.2.1. Inferring Trip Purpose and Location Type

Trip purposes were inferred and assigned based on mapping origin and destination coordinates in the RideAustin data to the City of Austin's zoning and parcel database. Using land use information in the parcel map, trip purposes and location types were classified into one of the following six groups:

- *Residential*: These are mostly residential areas, where a rider is probably going to or leaving a residence (either their own or somebody else's). It should be noted that this area may also contain educational establishments such as schools.
- *Commercial*: These are mostly commercial areas with shops, restaurants, or bars. Trips to/from this location type are likely to be for shopping as well as social-recreation activities. This category might also include trips by employees of commercial establishments, but these seem much less likely than customers.
- *Work*: Trips marked with this location type or purpose are associated with industrial areas and office buildings, rendering the identification of this trip purpose quite straightforward.
- *Education*: These parcels mainly encompass trips from/to the University of Texas central campus area and the Texas School for the Deaf. As mentioned above, grade schools and other more common educational establishments are likely to be located in the Residential land use category.
- *CBD (Central Business District)/DMU (Downtown Mixed Use)*: Given the mixed use of Austin's central business district (CBD), inferring actual trip purposes is not trivial. These areas include shops, bars, restaurants, houses (mostly apartment complexes) and office buildings. Therefore, this location type was treated as a separate category to avoid potential confounding effects due to the presence of mixed land uses.
- *Airport*: Trips to and from the airport were marked as a separate category given the ease of identifying their purpose.
- *Recreation*: These trips have origins or destinations in green areas and public parks; therefore the trip purpose is quite easily identifiable as recreation.
- *Other*: Trips that had no associated zoning information and therefore could not have their purpose inferred were marked with the "Other" trip purpose.

These categories were defined as a direct consequence of the land use classifications in the parcel/zoning data. The categories listed above are those that represented a reasonable aggregation of the original zoning designations in the parcel data.

The spatial coordinates of trip origins and destinations were truncated to the third decimal place in the RideAustin data. Given the spatial fuzziness associated with such truncated values, square buffers were created by taking the coordinates and adding/subtracting 0.0005 degrees (which translates to a square with approximate side dimensions of 100 meters). This buffer size was adopted to encompass the entire region in which a trip may have originated or terminated.

The square buffer of a trip location was then spatially cross-referenced with the zoning/parcel file to obtain a list of the location's potential land uses. In cases where multiple uses were found, locations were randomly assigned to one single use. This was done at the individual-location level.

3.2.2. Inferring Residential Location

A user's residential location was inferred by analyzing each individual's most visited location. Given that this process of residential location identification was based on users making multiple trips to the same place, it was possible to infer residential locations only for the most frequent ride-hailing users. For each rider, the most frequented location was identified; if the most frequented location was visited more than 10 times within the time span of the data set and belonged to a "residential" area, then this location was designated as the rider's household location. The frequency of 10 visits was set based on the distribution of frequency of visits (by the same person) to different locations. In general, it was found that only a small fraction of locations (users) met the criteria of being visited 10 or more times and fell within a residential area. It was therefore considered safe to designate these locations as residential locations for this subset of users.

The square buffers corresponding to these locations were cross-referenced with the Census block groups. For buffers that intersected multiple block groups, the final association was made by randomly assigning the location to one of the intersected block groups. The random association was done in proportion to the intersected area as illustrated in Figure 2.

In the example shown in Figure 2, the buffer of Location A (LOC A) intersects three block groups (BG1, BG2, and BG3). Suppose 60 percent of the buffer falls within BG1, 10 percent falls within BG2, and 30% falls within BG3. Location A is randomly assigned to one of the block groups using the proportions as probabilities; thus the probability that Location A would be assigned to BG1 is 0.6, BG2 is 0.1, and BG3 is 0.3. This process is repeated for every square buffer (location).

The association between locations and block groups allowed the fusion of block group demographic data to the RideAustin location data. Demographic data, aggregated from the 2016 American Community Survey database, was appended to each identified residential location based on the block group to which it was assigned. Then, proportions of various demographic segments (e.g., age, gender, race) were calculated, and a random proportion-based assignment procedure was implemented to assign demographic information to each frequent rider for whom a residential location was identified. The segments for age and sex were considered jointly, as were the segments for race and Hispanic ethnicity. All other demographic segments were considered independently. After imputing rider demographics, the total number of trips for each trip purpose/location category was computed and grouped into frequency categories. It is these grouped frequencies that served as the dependent variables for the multivariate ordered probit model estimated in this study.

4. DESCRIPTIVE SUMMARY OF RIDE-HAILING TRIPS

The final trip data set with inferred trip purposes had 1,475,596 trips and 258,022 unique riders or users. The final "frequent rider" data set (which includes only those individuals for whom census demographic information could be imputed) contains 13,895 riders. Table 1 (Section A) presents the overall profile of the entire set of ride-hailing trips for which trip purpose/location information was appended.

It can be seen that a majority of the trips (over 55 percent) are made at night between the hours of 10 PM and 7 AM. The remaining trips are rather evenly distributed among other periods of the day. In addition, a large proportion of trips (approximately 46 percent) is made on weekend days (Saturday and Sunday) with another 15.3 percent of trips made on Friday. Only 38 percent of trips are made on Monday through Thursday. An examination of the trip purpose/location distribution shows that trip origins and destinations are rather symmetric (signifying two-way or round-trip travel by the ride-hailing service) in their distributions, with high concentrations of trips in the CBD/DMU, residential, and commercial categories. Together, these three categories account for more than 75 percent of the trip origins/destinations of ride-hailing trips. It appears that individuals use the ride-hailing services to a greater degree on weekend nights to travel between residential areas (home locations) and commercial areas or CBD/DMU establishments. This signifies that individuals are using these services primarily for social-recreational trips on weekends, possibly to avoid driving under the influence of alcohol. Generally, work, airport, education, and recreation (visits to green spaces and parks) purposes account for only small fractions of the ride-hailing trips analyzed in this study. Another interesting finding is that about one-third of the users are single-time users; nearly one half of the users are infrequent users (2-10 times per year); and the remaining 17 percent of users are weekly or monthly users. In general, it appears that many users are single time or infrequent users of ride-hailing services (although that could change as market penetration and adoption continues to grow).

Table 1 (Section B) also presents a profile of frequent ride-hailing users. It should be recalled that only records of frequent users of the ride-hailing service could be geo-tagged with a home location and appended with census demographic data. Thus, the sample size for this table is just 13,895, which corresponds to the number of frequent users with an identified residential location obtained using the procedures outlined in the previous section.

The imputation process predicts that frequent users are more likely to be younger cohorts with just eight percent of users being 65 years or above in age. It also shows that the split between males and females is virtually equal. A vast majority seem not to have children in the household, suggesting that ride-hailing services are used to a greater degree by individuals in earlier lifecycle stages. About 85 percent of the frequent users make use of the service on a weekly basis. The imputation also suggests that the majority of users are White, and less than five percent are Black/African American, suggesting that there may be some equity concerns in the use of ride-hailing services. The income distribution suggests that nearly one-half of the frequent users have a household income less than $50,000 per year. This finding is rather counter to previous literature that suggests ride-hailing users are of higher income strata. It is possible that individuals in lower income households (which may be vehicle deficient) find this service useful and reasonably cost-effective in meeting mobility needs, particularly in areas where transit service is not reliable and does not provide high levels of service.

5. ANALYSIS OF RIDE-HAILING SERVICE USAGE PATTERNS

To further understand the influence of various socio-economic and demographic factors on the number of ride-hailing trips by destination trip purpose/location type, a multivariate ordered probit model was specified and estimated in this study. The number of ride-hailing trips undertaken by riders for each destination purpose/location type was computed and then modeled as a function of demographic variables. The modeling framework accounts for correlations among unobserved effects (error terms), essentially recognizing that there may be common unobserved attributes affecting the number of trips undertaken for different purposes/locations. For this analysis, trips to

"Residential" locations were excluded from the model to focus the analysis on travel undertaken to places other than home. In addition, the analysis excludes trips with a classification of "Other" because of the lack of information about this trip category.

Given that the individuals in the data set presented a wide range of trip counts and that the data set is considerably large (13,000+ riders), the higher trip counts were aggregated into categories. These categories are not identical across all destination purposes, but were set such that approximately 200 riders would fall into each category, thus enabling a more computationally feasible model estimation process. The final table containing the number of individuals per frequency category was omitted for the sake of brevity.

5.1. Modeling Framework

The modeling framework used in this study is very similar to that presented in Ferdous et al. (*20*). Let q be an index for individuals (q = 1, 2, …, Q), and let i be the index for destination purpose (i = 1, 2, …, I, where I denotes the total number of destination purposes for each individual; in the current study, I = 6). Let the number of frequency categories for destination purpose i be $K_i + 1$ (*i.e.*, the frequency categories of destination purpose i are indexed by k and belong in {0, 1, 2, …, K_i}). Following the usual ordered response framework notation, it is possible to write the latent propensity (y^*_{qi}) for each destination purpose as a function of relevant covariates and relate this latent propensity to the observed frequency outcome (y_{qi}) through threshold bounds (see McKelvey and Zavoina (*21*)):

$$y^*_{qi} = \boldsymbol{\beta}_i' \boldsymbol{x}_{qi} + \varepsilon_{qi}, y_{qi} = k \text{ if } \theta_i^k < y^*_{qi} < \theta_i^{k+1} \quad (1)$$

where $\boldsymbol{x}_{qi}$ is a (L×1) vector of exogenous variables (not including a constant), $\boldsymbol{\beta}_i$ is a corresponding (L×1) vector of coefficients to be estimated, ε_{qi} is a standard normal error term, and θ_i^k is the lower bound threshold for frequency category k of destination purpose i ($\theta_i^0 < \theta_i^1 < \theta_i^2 ... < \theta_i^{K_i+1}$; $\theta_i^0 = -\infty$, $\theta_i^{K_i+1} = +\infty$ for each destination purpose i). The ε_{qi} terms are assumed independent and identical across individuals (for each and all i). Due to identification restrictions, the variance of each ε_{qi} term is normalized to 1. However, correlations are allowed in the ε_{qi} terms across destination purposes i for each individual q. Specifically, define $\boldsymbol{\varepsilon}_q = (\varepsilon_{q1}, \varepsilon_{q2}, \varepsilon_{q3}, \ldots, \varepsilon_{qI})'$. Then, $\boldsymbol{\varepsilon}_q$ is multivariate normal distributed with a mean vector of zeros and a correlation matrix as follows:

$$\boldsymbol{\varepsilon}_q \sim N\left[\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1I}\\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2I}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_{I1} & \rho_{I2} & \rho_{I3} & \cdots & 1\end{pmatrix}\right], \text{ or } \boldsymbol{\varepsilon}_q \sim N[\mathbf{0}, \boldsymbol{\Sigma}] \quad (2)$$

The off-diagonal terms of $\boldsymbol{\Sigma}$ capture the error covariances across the underlying latent continuous variables of the different destination purposes; that is, they capture the effect of common unobserved factors influencing the propensity of usage of ride-hailing for each destination purpose. Thus, if ρ_{12} is positive, it implies that individuals with a higher than average propensity in their peer group to engage in ride-hailing trips for the first destination purpose are

also likely to have a higher than average propensity to engage in ride-hailing trips for the second destination purpose. If all the correlation parameters (*i.e.*, off-diagonal elements of $\boldsymbol{\Sigma}$), which can be stacked into a vertical vector $\boldsymbol{\Omega}$, are identically zero, the model system in Equation (1) collapses to independent ordered response probit models for each destination purpose.

The parameter vector of the multivariate probit model is $\boldsymbol{\delta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_I'; \boldsymbol{\theta}_1', \boldsymbol{\theta}_2', ..., \boldsymbol{\theta}_I'; \boldsymbol{\Omega}')'$, where $\boldsymbol{\theta}_i = (\theta_i^1, \theta_i^2, ..., \theta_i^{K_i})'$ for $i = 1, 2, ..., I$. Let the actual observed frequency category for individual q and destination purpose i be m_{qi}, and let $\phi_I(v_1, v_2, ..., v_I \mid \boldsymbol{\Sigma})$ represent the standard I-variate normal probability density function computed at the abscissae values of $v_1, v_2, ..., v_I$, and with a correlation matrix $\boldsymbol{\Sigma}$. Then, the likelihood function for individual q may be written as follows:

$$L_q(\boldsymbol{\delta}) = \Pr(y_{q1} = m_{q1}, y_{q2} = m_{q2}, ..., y_{qI} = m_{qI})$$

$$= \int_{v_1=\theta_1^{m_{q1}}-\beta_1' x_{q1}}^{\theta_1^{m_{q1}+1}-\beta_1' x_{q1}} \int_{v_2=\theta_2^{m_{q2}}-\beta_2' x_{q2}}^{\theta_2^{m_{q2}+1}-\beta_2' x_{q2}} \cdots \int_{v_I=\theta_I^{m_{qI}}-\beta_I' x_{qI}}^{\theta_I^{m_{qI}+1}-\beta_I' x_{qI}} \phi_I(v_1, v_2, ..., v_I \mid \boldsymbol{\Sigma}) dv_1 dv_2 ... dv_I \quad (3)$$

Calculating the high-order I-dimensional rectangular integral above can prove to be computationally challenging. In order to sidestep these problems, this study employs a pairwise marginal likelihood estimation approach, which involves using a composite marginal approach based on bivariate margins (see Ferdous et al. (*20*) for details).

5.2. Model Estimation Results

Model estimation results are presented in Table 2. All statements made in this section are predicated on the assumption that the imputation process was indeed appropriate. Also, we suppress the threshold values (in the ordered-response models) in Table 2 to conserve on space and reduce clutter. All of these threshold values were, however, statistically significant. The results show patterns that are largely consistent with expectations, given the nature of the data set and the geographic context from which it is derived. Alternative model specifications were tested and the final model specification that was adopted is presented in the table. The model exhibits a goodness of fit that is typical of models of this nature. Log likelihood values and the value of the likelihood ratio test (LRT) statistic are presented at the bottom of the table. All model coefficients were found to be statistically significant, generally indicating that socio-demographic variables are key determinants of number of ride-hailing trips undertaken by individuals for various purposes/locations. A number of error correlation terms (just above the goodness-of-fit measures) are significant, suggesting that a multivariate ordered response model that accommodates error correlations is appropriate in this particular context. The correlations suggest the presence of significant unobserved attributes that simultaneously affect multiple ordinal response variables. For example, it is found that the error correlation for Recreation and CBD/DMU is positive and significant (0.304 with a t-stat of 36.342). This signifies that unobserved attributes that contribute to recreational trips are positively correlated with unobserved attributes that contribute to CBD/DMU trips. This positive correlation may arise due to unobserved personality traits that affect both of these trip purposes/locations. For instance, an individual who is outgoing and seeks adventure and out-of-home recreational activities may undertake more recreational trips and more CBD/DMU trips (to frequent establishments that are in Downtown).

Compared with younger individuals aged 18-24 years, older age groups are less inclined to undertake ride-hailing trips for work, education, and commercial purposes. However, they are more inclined to undertake ride-hailing trips for airport and CBD/DMU purposes. This is consistent with expectations, especially in light of the pattern seen for income. Essentially, younger people have lower incomes and are hence inclined to undertake ride-hailing trips for more utilitarian purposes when compared with older and higher income individuals. Higher income individuals are also likely to have greater access to an automobile, thus reducing the need to utilize these services for utilitarian trips. Conversely, older age groups are likely to have higher incomes than younger individuals; they are likely to engage in more airport trips and CBD/DMU trips – both of which are likely to increase with income levels. Higher income individuals probably use their own personal vehicles for utilitarian trip purposes, but utilize ride-hailing services to a greater degree (than lower income individuals) for the sake of convenience and to avoid driving under the influence.

Somewhat similar to income trends, minorities exhibit a propensity to make fewer ride-hailing trips for airport and CBD/DMU than White individuals. Hispanics are less inclined to make ride-hailing trips for all purposes except the commercial trip purpose; it is possible that Hispanics rely on the service to a greater degree than their non-Hispanic counterparts for shopping and running errands. The presence of children in the household contributes to lower propensity to make ride-hailing trips; households with children are likely to own personal vehicles and use their own vehicles for trip-making. A few findings merit further investigation. For example, it is found that minorities are inclined to make more work and commercial trips than White individuals. It is possible that minorities have fewer personal cars and hence use RideAustin to a greater degree for these utilitarian purposes.

Vehicles owned seem to have heterogeneous impacts: higher vehicle ownerships are associated with increased ride-hailing trips to the airport, CBD/DMU and to recreational sites, while also being associated with fewer ride-hailing trips to work, educational and commercial establishments.

Here we also have evidence of the question of whether or ride-hailing services substitute or induce transit use. For work trips, ride-hailing seems to be more common with individuals who live in locations with less access to transit (i.e. larger distances to transit stops), suggesting that the pattern in this case is substitutional. However, for all other purposes/locations, the association is reversed, suggesting that there might be a synergetic effect between ride-hailing and transit.

Finally, individuals who live in denser neighborhoods are associated with higher levels of ride-hailing to work, CBD/DMU, education and commerce, while also being associated with fewer trips to the airport and to recreational locations.

6. DISCUSSION AND CONCLUSIONS

The motivation for this study stems from the very limited knowledge about the characteristics of trips undertaken by ride-hailing services and the characteristics of their users. Despite the enormous growth in the use of ride-hailing services, there is a limited understanding of the nature of these trips due to the scarcity of detailed data about such trips. In addition, metropolitan planning organizations are unable to adequately account for the demand for ride-hailing services in their travel demand forecasting models. In an effort to shed light on the nature and characteristics of ride-hailing trips, this study involves the fusion of trip level data made available by a ride-hailing company with land use and parcel data to infer trip purpose and location information. The study also involves the identification of frequent users of the ride-hailing service and the determination

of their home locations based on a set of criteria that help isolate the residential location for such frequent users. The data fusion methodology presented in the paper then involves fusing socio-demographic data from the American Community Survey with frequent service users in the ride-hailing trip database to obtain their demographic profiles.

The data used in this study originated from RideAustin, a service provider in the Austin metropolitan area in the United States. The trip origins and destinations were mapped to a land use parcel and zoning map to determine the purpose and location type for all trip origins and destinations. Analysis of the augmented data showed that large percentages of trips either come from or go to residential locations, commercial locations, and CBD/DMU (Downtown Mixed Use) locations. Trips in this data set appear to be largely taking place at night and on weekends, suggesting that they are undertaken for social purposes and a desire to avoid driving under the influence. Frequent users of the service (for whom residential locations could be identified and census demographic data could be appended) appeared more likely to be young or middle-aged, White, non-Hispanic, and in lower income households with no children. It appears that lower income individuals are more inclined to use the service because they may have vehicle availability constraints, and the ride-hailing service essentially lifts that constraint by providing on-demand mobility.

A multivariate ordered probit model was estimated to examine the influence of different variables on the number of ride-hailing trips undertaken by frequent users of the system for various trip purposes/locations. It is found that socio-demographic variables are important determinants of trip frequencies. Findings largely corroborated what was found in the descriptive analysis, but provided deeper insights into the differential effects of various socio-demographic variables on the numbers of ride-hailing trips for different trip purposes. For example, lower income individuals appear to make more ride-hailing trips than their richer counterparts for utilitarian purposes, while higher income individuals are prone to making more ride-hailing trips for airport, CBD/DMU, and recreational purposes.

The paper presents a data fusion methodology that agencies can potentially implement in their own jurisdictions (as long as they can obtain some trip data from a ride-hailing service company). Through the use of the data fusion methodology, it is possible to answer questions on why (for what purpose) ride-hailing services are used, where ride-hailing users reside, and how many ride-hailing trips will be undertaken in different time periods. Answers to such questions can help shape transport policy and enhance the ability of travel forecasting models to reflect the demand for ride-hailing services. For example, the analysis shows that ride-hailing services are being used heavily at nights and on weekends. This may call for the development of special night or weekend travel forecasting models so that bottlenecks and congestion points can be better identified, predicted, and managed before issues arise. Similarly, minorities and low income individuals are not making the same number of ride-hailing trips for discretionary purposes as their higher income counterparts. Policies that aim to make these services more accessible and affordable to minorities and lower income individuals may go a long way in enabling their participation in social and recreational activities that enhance quality of life. The quantification of the number of ride-hailing trips by purpose/location is of direct application in travel forecasting models that purport to estimate travel demand in a region.

The methodology presented in this paper constitutes a promising start to fusing multiple data sources to paint a more holistic picture of demand for emerging mobility services. However, a number of limitations and caveats apply. The methodology presented in this paper, although quite robust and logical, has not been empirically tested and verified against a validation sample.

As such, the accuracy of the fused data remains uncertain until comparisons against real-world data can be made, which point to validation efforts being a promising area for future research. The algorithms embedded in the methodology (in terms of matching trip locations to census block groups and parcels, or attributing socio-demographic data to specific users) can be further enhanced to make them more robust and sophisticated through the use of repeated simulations and more advanced matching algorithms. The data generated by the imputation method is stochastic in nature, potentially introducing errors into the multivariate ordered probit model. The influence of these errors might be reduced by estimating models for each of the repeated simulations and reporting the models' the average effects. Furthermore, the data is drawn from a very specific vendor in a specific location; hence the findings and results presented in this study may not be generalizable across multiple geographic contexts. These limitations should be addressed in future research efforts.

ACKNOWLEDGEMENTS

The authors would like to thank RideAustin for publicly sharing their data. This research was partially supported by the Center for Teaching Old Models New Tricks (TOMNET) (Grant No. 69A3551747116) as well as the Data-Supported Transportation Operations and Planning (D-STOP) Center (Grant No. DTRT13GUTC58), both of which are Tier 1 University Transportation Centers sponsored by the US Department of Transportation. The second author acknowledges funding support from CAPES and the Brazilian Government. The authors are grateful to Lisa Macias for her help in formatting this document, and to four anonymous referees who provided useful comments on an earlier version of the paper.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: F.F. Dias, P.S. Lavieri, T. Kim, C.R. Bhat, R.M. Pendyala; data collection: RideAustin; analysis and interpretation of results: F.F. Dias, P.S. Lavieri, T. Kim, C.R. Bhat, R.M. Pendyala; draft manuscript preparation: F.F. Dias, P.S. Lavieri, T. Kim, C.R. Bhat, R.M. Pendyala. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Statista. Number of Uber Users in the United States as of April 2017, by Device (in Millions) [online]. 2018a. Available at: https://www.statista.com/statistics/715236/us-uber-users-by-device/. Accessed: February 7, 2018.
2. Statista. North American Carsharing Member Growth from 2000 to 2016 (in 1,000s) [online]. 2018b. Available at: https://www.statista.com/statistics/263847/carsharing-growth-members-in-north-america/. Accessed: February 7, 2018.
3. Rayle, L., D. Dai, N. Chan, R. Cervero, and S. Shaheen. Just a Better Taxi? A Survey-based Comparison of Taxis, Transit, and Ridesourcing Services in San Francisco. *Transport Policy*, 2016. 45: 168-178.
4. Smith, A. Shared, Collaborative and On Demand: The New Digital Economy. Pew Research Center, Washington, D.C., 2016. Available at: http://www.pewinternet.org/2016/05/19/the-new-digital-economy/. Accessed: February 6, 2018.
5. Clewlow, R.R. and G.S. Mishra. Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States. Institute of Transportation Studies, University of California, Davis, Research Report UCD-ITS-RR-17-07, 2017.

6. Alemi, F., G. Circella, P. Mokhtarian and S. Handy. On-Demand Ride Services in California: Investigating the Factors Affecting the Frequency of Use of Uber/Lyft. Transportation Research Board 97th Annual Meeting Compendium of Papers, No 18-05563, 2018.
7. Hampshire, R.C., C. Simek, T. Fabusuyi, X. Di, and X. Chen. Measuring the Impact of an Unanticipated Suspension of Ride-sourcing in Austin, Texas. Transportation Research Board 97th Annual Meeting Compendium of Papers, No. 18-03105, 2018.
8. Lavieri, P.S., and C.R. Bhat. Investigating Objective and Subjective Factors Influencing the Adoption, Frequency, and Characteristics of Ride-hailing Trips. Technical paper, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, May 2018.
9. Dias, F.F., P.S. Lavieri, V.M. Garikapati, S. Astroza, R.M. Pendyala, and C.R. Bhat. A Behavioral Choice Model of the Use of Car-sharing and Ride-sourcing Services. *Transportation*, 2017. 44(6): 1307-1323.
10. Lavieri, P.S., V.M. Garikapati, C.R. Bhat, R.M. Pendyala, S. Astroza, and F.F. Dias. Modeling Individual Preferences for Ownership and Sharing of Autonomous Vehicle Technologies. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2665: 1-10.
11. Li, Z., Y. Hong, and Z. Zhang. An Empirical Analysis of On-demand Ride Sharing and Traffic Congestion. Presented at the 2016 International Conference on Information Systems, Dublin, September 2017.
12. Ward, J.W., J.J. Michalek, I.L. Azevedo, C. Samaras, and P. Ferreira. On-Demand Ridesourcing Has Reduced Per-Capita Vehicle Registrations and Gasoline Use in U.S. States. Transportation Research Board 97th Annual Meeting Compendium of Papers, No. 18-05185, 2018.
13. Kooti, F., M. Grbovic, L.M. Aiello, N. Djuric, V. Radosavljevic, and K. Lerman. Analyzing Uber's Ride-sharing Economy. *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 574-582. International World Wide Web Conferences Steering Committee.
14. Gerte, R., K.C. Konduri, and N. Eluru. Is There a Limit to Adoption of Dynamic Ridesharing Systems? Evidence from Analysis of Uber Demand Data from New York City. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672(42): 127-136.
15. Komanduri, A., Z. Wafa, K. Proussaloglou, and S. Jacobs. Assessing the Impact of App-Based Ride Share Systems in an Urban Context: Findings from Austin. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672(7): 34-46.
16. Lavieri, P.S., F.F. Dias, N.R. Juri, J. Kuhr, and C.R. Bhat. A Model of Ridesourcing Demand Generation and Distribution. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672(46): 31-40.
17. Zheng, H., X. Chen, and X.M. Chen. How Does On-Demand Ridesplitting Influence Vehicle Use and Ownership? A Case Study in Hangzhou, China. Transportation Research Board 97th Annual Meeting Compendium of Papers, No. 18-04327, 2018.
18. RideAustin. Data file and code book [online]. 2017. Available at: https://data.world/ride-austin/ride-austin-june-6-april-13. Accessed July 26, 2017.
19. Bohte, W., and K. Maat. Deriving and Validating Trip Purposes and Travel Modes for Multi-day GPS-based Travel Surveys: A Large-scale Application in the Netherlands. *Transportation Research Part C*, 2009. 17(3): 285-297.

20. Ferdous, N., N. Eluru, C.R. Bhat, and I. Meloni. A Multivariate Ordered Response Model System for Adults' Weekday Activity Episode Generation by Activity Purpose and Social Context. *Transportation Research Part B*, 2010. 44(8-9): 922-943.
21. McKelvey, R.D., and W. Zavoina. A Statistical Model for the Analysis of Ordinal Level Dependent Variables. *Journal of Mathematical Sociology*, 1975. 4(1): 103-120.

LIST OF FIGURES

LIST OF TABLES

FIGURE 1 Austin's central area.

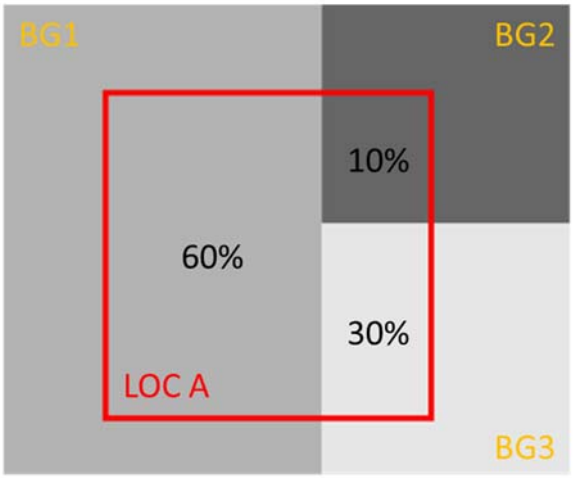

FIGURE 2 Proportional randomization for block group demographic data.

TABLE 1 Characteristics of All Ride-Hailing Trips (Section A) and of Frequent Ride-Hailing Users (Section B)

Section A: All ride-hailing trips (n=1,475,596 trips)

Time of Day	%	Month	%	Trip Location/Purpose	Origin %	Destination %
Morning peak (7am-10am)	8.76	Aug '16	2.51			
Morning off peak (10am-4:30pm)	13.82	Sep '16	6.46	Airport	4.10	5.94
Evening peak (4:30pm-7:30pm)	11.50	Oct '16	13.05	CBD/DMU	21.51	22.48
Evening off peak (7:30pm-10pm)	10.72	Nov '16	10.69	Education	1.33	1.27
Night (10pm-7am)	55.20	Dec '16	10.71	Residential	35.95	32.62
Day of the Week	%	Jan '17	13.21	Recreation	4.92	5.29
Monday to Thursday	38.36	Feb '17	14.58	Commercial	22.11	21.60
Friday	15.32	Mar '17	21.35	Work	7.12	7.00
Saturday to Sunday	46.32	Apr '17	7.44	Other	2.96	3.80

Residential Densities and Distances to Transit	Mean	Rider Frequency (n=258,022)	%
Residential density at origin (pop/km²)	2,508	Single time	33.77
Residential density at destination (pop/km²)	2,376	Infrequent (2 to 10 per year)	49.36
Distance to closest transit stop at orig. (km)	0.33	Monthly (11 to 24 per year)	10.45
Distance to closest transit stop at dest. (km)	0.38	Weekly (25+ per year)	6.42

Section B: Frequent ride-hailing users (n=13,895 users)

Age	%	Race	%
Age 18 to 24	22.99	White	78.58
Age 25 to 44	46.77	Black / African American	5.18
Age 45 to 64	21.76	Asian	6.71
Age 65 and up	8.48	Other (including mixed race)	9.53
Gender	%	Household Income	%
Female	48.31	Below $50,000	46.95
Male	51.69	Between $50,000 $99,999	26.41
Has a Child in the Household	%	Between $100,000 $149,999	12.49
No	83.92	$150,000 and above	14.15
Yes	16.08	Vehicles in Household	%
Frequency of Use	%	0 vehicles	9.54
Monthly (11 to 24 per year)	14.29	1 vehicle	47.82
Weekly (25+ per year)	85.71	2 or more vehicles	42.64
Is Hispanic	%	Continuous Variables	Mean
No	74.72	Residential density (pop/km²)	3142.57
Yes	25.28	Distance to closest transit stop (km)	0.28

TABLE 2 Results of Multivariate Ordered Response Model for Number of Trips by Location/Purpose

Variables	Work		Airport		CBD/DMU		Education		Commercial		Recreation	
	Value	t-stat	Value	t-stat	Value	t-stat	Value	t-stat	Value	t-stat	Value	t-stat
Age *(Base: Between 18 and 24)*												
Between 25 and 44	-0.132	-33.802	0.176	43.008	0.111	21.392	-0.470	-83.465	-0.090	-17.214		
Between 45 and 64	-0.132	-33.802	0.176	43.008	0.130	21.988	-0.470	-83.465	-0.041	-6.843		
65 and older	-0.132	-33.802	0.176	43.008	0.160	21.483	-0.470	-83.465	-0.077	-10.140		
Gender *(Base: Female)*												
Male	0.013	3.467			0.029	8.084	0.031	5.836			0.046	12.320
Race *(Base: White)*												
Black/African American	0.016	1.914	-0.234	-25.375	-0.193	-24.053			0.046	5.737	-0.077	-8.940
Asian	0.113	14.923	-0.034	-6.315			0.032	3.199	0.017	2.349		
Other	0.045	7.135	-0.034	-6.315	-0.049	-8.027					-0.051	-7.713
Hispanic Ethnicity *(Base: No)*												
Yes	-0.044	-9.770	-0.073	-15.844	-0.126	-29.992	-0.113	-17.837	0.036	8.780	-0.041	-9.335
Presence of Children in Household *(Base: No)*												
Yes	-0.077	-14.689	-0.041	-7.616	-0.158	-31.852	-0.104	-12.885	-0.021	-4.234	-0.065	-12.372
Income *(Base: Up to $49,999)*												
Between $50,000 and $99,999	0.034	7.171	0.066	13.282			-0.108	-15.611	-0.045	-10.195	0.032	8.151
Between $100,000 and $149,999	-0.043	-6.971	0.153	24.130	0.094	17.597	-0.145	-20.065	-0.073	-12.753	0.032	8.151
$150,000 or more	-0.065	-10.836	0.229	38.671	0.255	50.010	-0.145	-20.065	-0.115	-20.684	0.032	8.151
Vehicles in Household *(Base: 0 vehicles)*												
1 vehicle	-0.026	-5.465	0.029	4.219	0.071	11.310	-0.095	-14.794	-0.017	-2.715	0.040	6.060
2 or more vehicles	-0.026	-5.465	0.029	4.219	0.020	3.101	-0.095	-14.794	-0.047	-7.325	0.020	2.978
Continuous Variables												
Residential density (pop x 1.000/km²)	0.015	22.968	-0.005	-6.667	0.013	21.084	0.031	39.219	0.016	25.404	-0.005	-7.993
Distance to closest transit stop (km)	0.073	17.955	-0.156	-37.673	-0.383	-101.848	-0.261	-40.162	-0.165	-40.876	-0.191	-39.508
Correlation Terms												
Airport	-0.048	-4.181										
CBD/DMU	0.044	4.510	0.183	18.163								
Education	0.084	5.558										
Commercial	0.257	30.271	-0.035	-3.241	0.050	5.474	0.161	12.055				
Recreation	0.095	8.628	0.098	8.685	0.324	40.483	0.063	4.361	0.133	13.619		

Goodness-of-Fit Statistics: Predicted Log likelihood of Full Model: -140,427
Predicted Log likelihood of Null Model (only thresholds and no correlation terms): -152,241
Likelihood Ratio Test Statistic (χ^2) = 23,628; p-value = 0.000
All coefficients are statistically significant at the 0.05 level
Variables with identical values represent categories that were combined during estimation due to not being statistically different from each other