Axis of travel: Modeling non-work destination choice with GPS data

Arthur Huang, Ph.D.
Department of Civil Engineering
University of Minnesota, Duluth
Duluth, MN 55812
Phone: 218-726-6452
E-mail: huang284@umn.edu

David Levinson, Ph.D.
Department of Civil, Environmental, and Geo- Engineering
University of Minnesota, Twin Cities
Minneapolis, MN 55455
Phone: 612-625-6354
E-mail: dlevinson@umn.edu

Abstract

Based on in-vehicle GPS travel data in the Minneapolis - St. Paul Metropolitan Area, this research investigates how land use, road network structure, and route familiarity influence home-based single-destination choice. We propose a new choice set formation approach which combines survival analysis and random selection. Our empirical findings reveal that: (1) Walkable opportunities and diversity of services at the destination influence destination choice. (2) Route-specific network measures such as turn index and speed discontinuity display statistically significant effects on destination choice. (3) The familiarity factors reflected by distance to home, work, and downtown also plays a role. A destination closer to home and work, all else equal, is more likely to be selected. A destination farther away from downtown is more attractive for auto users. This research contributes to methodologies in modeling destination choice using GPS data. The results enhance our understanding of non-work travel behavior and have implications for transportation and land use planning.

Keywords: GPS data, non-work trips, land use, axis of travel, destination choice

1 Introduction

Non-work destinations, including a spectrum of trip purposes: social, recreational, shop, family, personal, school, and church activities, comprise approximately 90% of trips (NHTS, 2009).

This research seeks deeper understanding about the factors shaping non-work travel behavior.

Advances in GPS and GIS technologies provide new opportunities and challenges for investigating non-work travel behavior. GPS devices have advantages over traditional paper-and-pencil diary methods:

1. Real-time spatial and temporal information of a trip is available, such as distance, travel times, travel speed, and route information;
2. Fewer misreporting or underreporting of trips;
3. Data are stored in digital formats;
4. The subjects' burden of reporting travel information is reduced Draijer et al. (2000).

However, there are also several challenges:

1. How to appropriately define destinations for spatial analysis at the microscopic level?

2. How to connect GPS data with other types of data sets for modeling destination choice?

3. How to select models and form choice sets for modeling destination choice in a large metropolitan area with a balance of computational precision and calculation time?

Consistent with the a long line of destination choice research extending the gravity model, it is hypothesized that (1) all else equal, a destination with shorter travel time is more likely to happen, and that (2) all else equal, a destination with more walkable opportunities at the destination is more favorable.

More significantly, we test a number of novel hypotheses regarding the routes that people use and the places people choose. We posit that travelers will select destinations on routes that are less complicated, as complicated routes are perceived as farther away and require more thought on the part of the travelers. We also posit that travelers will select non-work destinations that are on or near to routes that are often used on axis between home and work because travelers are more familiar with those destinations, and the paths connecting to them, from seeing them on a more regular basis.

This research aims to model home-based non-work destination choice at the microscopic level using GPS data. Two methodological contributions include a new method of forming choice sets for the non-work destination choice problem and a new procedure to justify the choice set size for destination choice.

The rest of this paper is organized as follows. Section 2 introduces the in-vehicle GPS data. The home-based, non-work trips are extracted in 3 for our GPS data. The non-work destinations are further defined in Section 4. Section 5 describes the independent variables, following which the choice set formation approach is introduced. Section 8 formulates the mixed-effects model for analysis and the results are shown in Section 9. The key findings are summarized in Section 10.

2 GPS travel data

The in-vehicle GPS data collection process lasted from September to December of 2008, during which 141 surveyed subjects made over 20,000 trips (Zhu, 2010). The data collection consisted of three stages (Figure 1). The first stage was to recruit the subjects. The announcements on recruiting subjects were posted on various media such as Craigslist.com and Citypages.com and were sent out via other forms such as postcards handed out in downtown parking ramps and emails sent to about 7000 University of Minnesota staff (excluding students and faculty). The second stage was to collect the data by installing GPS devices in participants' vehicles.

The third stage was to create GPS trip trajectories. The trip trajectories were drawn based on the GPS points in the underlying the Metropolitan Planning Network. The techni-

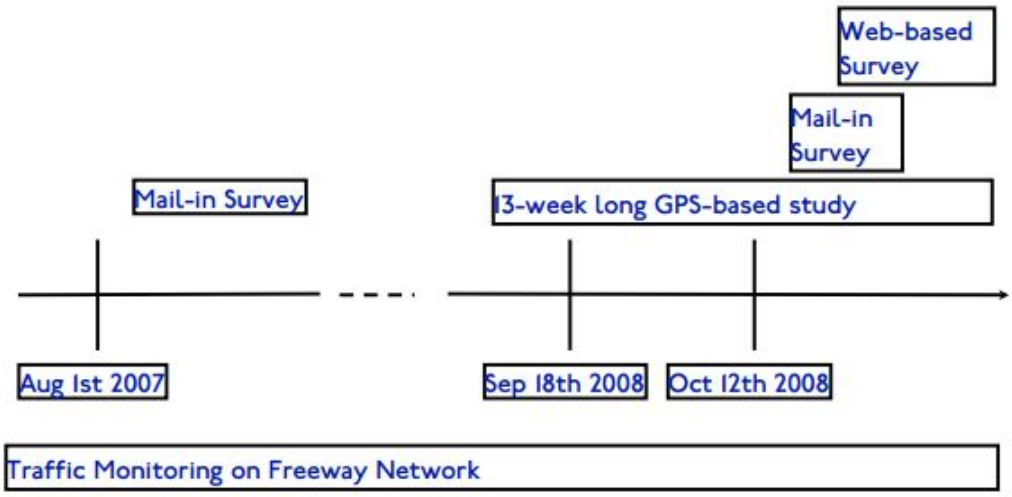

Figure 1: The timeline of the GPS Travel data collection process. (Source: Zhu (2010))

cal details on creating such trajectories can be found in Zhu (2010). Figure 2 shows an exemplary non-work trip trajectory.

3 Define non-work trips

In the GPS travel data, only a small proportion of the trip purposes are available, as individuals were only required to report such information on selected dates, from a sample of the data. Therefore, the first tasks are to identify non-work trips and to identify home-based non-work trip chains. To achieve this goal, the in-vehicle GPS trips with available travel diaries are analyzed. It is important to measure how far the subjects parked from home/work for the trips that were indicated as home trips or work trips. The procedure is as follows:

1. Match the time stamp of the in-vehicle GPS data and one individual's travel diary, and create a "trip purpose" attribute for the in-vehicle GPS data.
2. Select the trips whose purpose is work or home from the in-vehicle GPS data.
3. For a home trip, calculate the Euclidean distance between the trip destination and home. For a work trip, calculate the Euclidean distance between the trip destination and one's work address.

The percentiles of parking distances from home for home trips in the GPS data are calculated to identify an appropriate threshold (Figure 3). The maximum distance from home

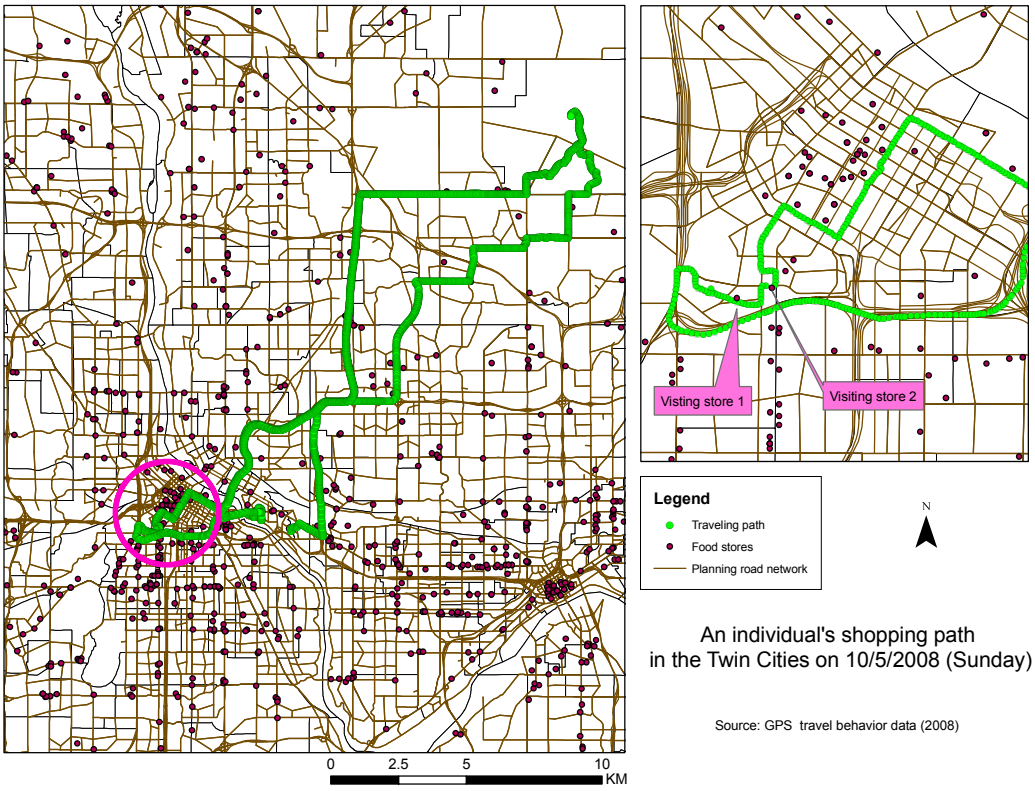

Figure 2: An individual's shopping trip trajectory captured by an in-vehicle GPS device on October 5th, 2008 in Minneapolis, MN.

Table 1: The subjects' socio-demographics in the in-vehicle GPS data

Variable	Category	GPS data (%)
Gender	Male	41.25
	Female	58.75
Education	11th grade or less	0
	High School	13.09
	Associate	24.99
	Bachelor	45.22
	Graduate	16.69
Household Income	$< \$49,999$	20.20
	$\$50,000 - \$74,999$	30.73
	$\$75,000 - \$124,999$	29.44
	$> \$124,999$	20.16
Race	White	83.06
	Black	7.36
	Others	9.58

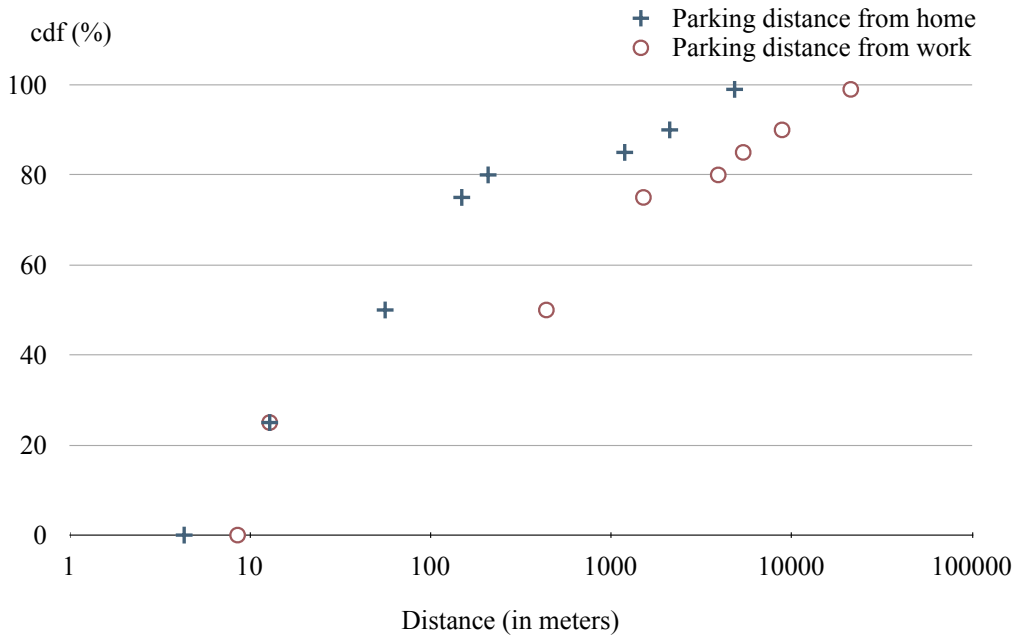

Figure 3: The percentiles of parking distances (in meters) from home and work in the GPS data.

and the distances at the 90th threshold and above clearly subject misreporting of trip purposes. We focus between the 85th percentile, where the distance is about 1190m, and the 80th percentile is about 208m

Based on this information, we use 800m as the maximum parking distance from home for home trips.

Figure 3 further shows the percentiles of parking distances from work. The range of distances is much wider than that of distances from home. Besides the possibility of misreporting trip purposes, it may be because workers did not actually go to the reported work address for work, as traveling to another site for work-related purposes is possible. Yet such information about other possible work sites is unknown to us. We use 1000 m as the maximum parking distance from the reported workplace for work trips.

4 Definition of destinations

In the literature, the definition of a destination ranges from counties, cities, traffic analysis zones, parcel-based locations, to store-based destinations. Since this study focuses on non-work trips (e.g., shopping, recreational, visiting friends), we prefer finer granularity of locations to larger granularity because finer granularity can provide more insights about microscopic land use.

In this research, the centroids of Census blocks are used to define destinations for the following two reasons:

1. The Census block data provide better precision of locations than other larger scale definitions. The block-level data are the finest geographical definition of locations in the US Census. In the 2010 US Census, the Twin Cities have 16851 Census blocks with at least one establishment, far more than 1165 traffic analysis zones, 182 cities, and 7 counties in the metropolitan area. In addition, even though we do know which store one visited, we can measure the land use around a destination.

2. The Census block-based definition of destinations creates more precise travel paths once mapped to the road network data. The shortest travel paths are created with the ArcGIS Network Analyst tool which locates the centroid of a Census block to its nearest road. The more granular a destination is, the more precisely travel time can be calculated.

It is important to check whether there are many repeated destinations visited by the same person. In the modeling destinations, if there exist repeated destinations visited by the same person, the modeling results may be biased. This is because there may be unknown reasons (such as an individual's preference for a particular store or service) that explain the choice of a repeated destination, and such information is unavailable to us. Therefore repeated destinations should be examined before applying the model.

We calculate the Euclidean distance (in meters) between destinations visited by the same person and identify the percentiles for the whole data set (Figure 4). If we use 100 m as the threshold for defining repeated destinations, repeated destinations account for only about 10% of all destinations. Thus, its effect on the modeling results is marginal.

5 Independent variables

The independent variables used in this study include land use, transportation network measures, traveler familiarity measures, and interaction terms between socio-demographics and land use/transportation network measures.

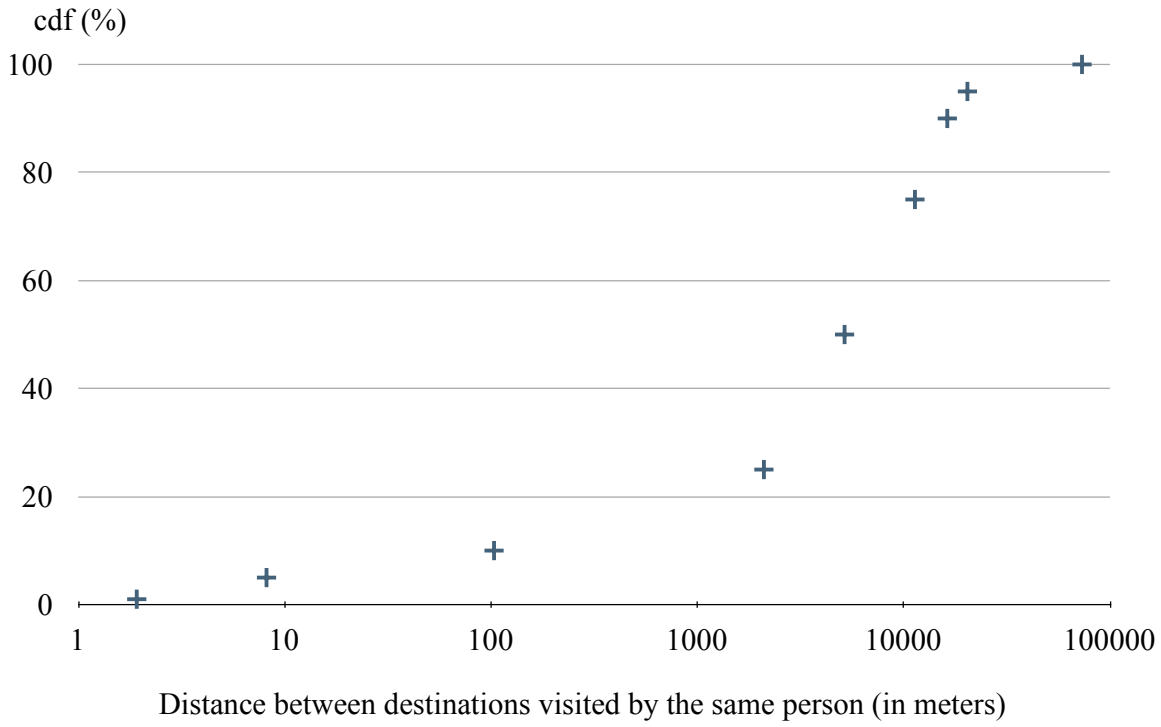

Figure 4: The cumulative probability distribution of distances between destinations visited by the same individual in the GPS data.

5.1 Land use measures

The key land use measures include accessibility and diversity of services (land use mix). In the literature, there are several accessibility measures: cumulative opportunities measure, gravity-based accessibility measure, and random utility-based measure. They offer different trade-offs between simplicity and the sophistication with which the activities and transportation system are characterized (Handy and Clifton, 2001). In this research, we are interested in walkable opportunities around a destination. After leaving one's car, one might walk to multiple stores and might not even visit the store closest to the parking spot due to parking constraints.

Considering the characteristics of shopping and for the simplicity of measurement, it is decided to use cumulative opportunities (A_k) to measure accessibility at destination k. The empirical tests reveal that its natural log form produces greater goodness of fit of the model. Therefore $ln(A_k)$ is adopted as an independent variable.

It is hypothesized that more walkable opportunities around a destination enhances its attractiveness.

The next question is to define the size of the walking area. Burke and Brown (2007) found that the 85th percentile of the walking distance to a shop is 1.24 km. If we assume the average walking speed is 5.4 km/hr (3.4 mi/hr) (Krizek et al., 2009), the walking time is around 15 minutes. The walking time considers the fact that auto travelers must walk farther when there is not parking near the destination. We calculate the total number of establishments within 15-min walking area around a destination.

The diversity of services or land use mix at destination k is typically measured by the entropy index (Shannon, 1948) which can be written as:

$$H_k = -\sum_{v=1}^{V} \rho_{kv} ln(\rho_{kv}) \tag{1}$$

Where ρ_{kv} is the proportion of service type v in destination k's walking area. The service type of a store is defined by the 6-digit North American Industry Classification System (NAICS) code. V is the total number of services in the destination's walking area. The greater H_k is, the more diverse services a destination has. All else equal, a destination with greater entropy indicates greater diversity of services, which supports multipurpose shopping and reduces the average travel time it takes to finish per task compared with making several single-destination trips. It is therefore hypothesized that greater diversity of services, all else equal, is associated with greater attractiveness of a destination.

Our further analysis shows that the entropy index and walkable opportunities at a destination are highly correlated (Pearson $r = 0.94$. Therefore, we desire to modify the traditional entropy index in order to obtain less biased estimates when incorporating the two measures in the model.

Mathematically the diversity of services measure can be written as:

$$H_k = \begin{cases} 0 & \text{if } A_k = 0 \text{ or } A_k = 1 \\ \frac{-\sum_{v=1}^{V} \rho_{kv} ln(\rho_{kv})}{ln(A_k)} & \text{if } A_k > 0 \end{cases} \tag{2}$$

5.2 Transportation network measures

The road network measures used in this study include speed discontinuity (Levinson and El-Geneidy, 2009) and turn index.

5.2.1 Speed discontinuity

Speed discontinuity, first proposed and applied in Xie and Levinson (2007) and Parthasarathi et al. (2012), was described as the changes of speed along the fastest path between an origin and a destination divided by the length of this route. In this study travel time is used as an independent variable. In order to reduce the correlation with travel time, speed discontinuity in this study is defined as the changes of speed along the fastest path between an origin and a destination divided by trip travel time. Mathematically it can be written as:

$$\psi_k = \begin{cases} ln\frac{0.5}{T_k} & \text{if } \sum(|v_q - v_{q+1}| = 0 \\ ln\frac{\sum(|v_q - v_{q+1}|)}{T_k} & \text{if } \sum(|v_q - v_{q+1}| > 0 \end{cases} \tag{3}$$

Where v_q is the travel speed on road link q and T_k refers to travel time. $|v_q - v_{q+1}|$ indicates the absolute value of the speed difference on two consecutive links q and $q+1$. $\sum |v_q - v_{q+1}|$ measures the sum of the absolute value of the changes of speed along a route. When $\sum(|v_q - v_{q+1}|) = 0$, we use the midpoint of 0 and 1 to replace 0 to make the definition meaningful.

Speed discontinuity serves as an index for measuring a destination's reachability. A trip with greater speed discontinuity is considered less comfortable and requires more mental effort on the part of travelers (Parthasarathi et al., 2013), and thus is perceived to be longer than actual and thus reduces the perceived reachability. It is therefore hypothesized that greater speed discontinuity on the route dampens the attractiveness of the trip's destination.

5.2.2 Turn index

We propose another measure: turn index (ϑ_k). It measures the number of turns a drivers needs to make from home to a destination. If the acute angle between every two con-

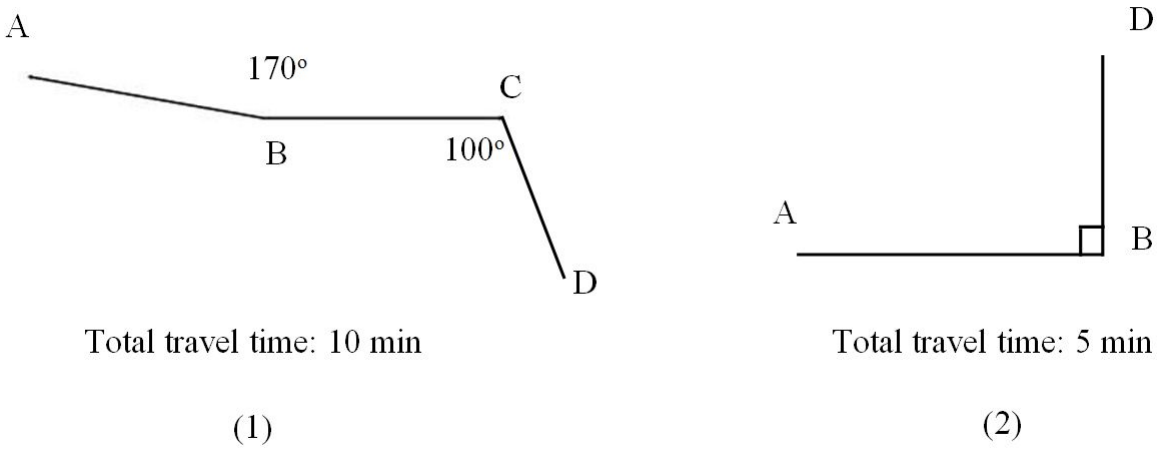

Figure 5: Two examples of calculating turn index.

nected road links is between 170 degrees (inclusive) and 180 degrees (inclusive), a driver is considered as not having to make any turning maneuver to transition from one link to the other; otherwise, a driver is considered as having to make a turn. Turn index (ϑ_k) is calculated as the cumulative number of turns one drivers needs to make on a route divided by the total travel time (in order to reduce the correlation with travel time). Our further test reveals that its natural log form also produces a higher log-likelihood value for the model. Therefore, turn index (ϑ_k) used in this research are defined as:

$$\vartheta_k = \begin{cases} ln\frac{0.5}{T_k} & \text{if } \Gamma_k = 0 \\ ln\frac{\Gamma_k}{T_k} & \text{if } \Gamma_k > 0 \end{cases} \tag{4}$$

Where Γ_k is the total number of turns on the route to visit destination k. Two simple examples of calculating turn index are shown in Figure 5. The route in Figure 5-(1) consists of three links. The angle between link AB and BC equals 170 degrees; therefore there is no turning maneuver. The angle between link BC and CD equals 100 degrees; thus, a driver needs to make a turn to go from BC to CD. The total number of turns equals 1. Given that the total travel time equals 10 minutes, the turn index of the route equals $ln(1/10) = -2.3$. In Figure 5-(2), the route consists of two links. The angle between link AB and link BD equals 90 degrees; thus, there is one turn between the two links. Given that the total travel time equals 5 minutes, the turn index of the route equals $ln(1/5) = -1.6$. The greater this value is, the more turns per unit travel time a route requires.

A greater turn index suggests more turns one needs to make per unit time, which makes a trip less desirable. It is hypothesized that a greater turn index reduces the convenience of driving on the route, and thus lowers the attractiveness of a destination.

5.3 Familiarity

The measures on about traveler familiarity include travel time between destination and work and travel time between destination and the nearest downtown. For each individual, we measure the fastest-path travel time between destination k and workplace, which is represented by $T_{w,k}$ (the symbol w represents work). This measure indicates one's familiarity with the destination. One may be more familiar with destinations adjacent to work, and therefore may be inclined to select these destinations. We hypothesize that all else equal, a non-work destination closer to work is more likely to be selected due to a person's greater familiarity with the destination.

The travel time between home and the nearest downtown ($T_{d,k}$, and the symbol d represents downtown) implies one's consideration of greater parking constraints, narrower streets, and more traffic lights which are common in the downtown area. It is hypothesized that all else equal, for auto users a non-work destination closer to the nearest downtown is less likely to be selected because of these nuisances.

6 Choice set formation

Choice set formation concerns how to form choice sets based on all destinations in the Metropolitan Area. We propose a new method of choice set formation which combines survival analysis and random sampling.

In recent years, several studies used the hazard-based analysis to calculate average work distance for housing location choice set (Rashidi et al., 2012), the length of stay in golf tourism (Barros et al., 2010), and the deterministic distance constraint for residential destination choice (Zolfaghari et al., 2012).

We are interested in how likely a trip will happen given some travel time and the destination's land use characteristics.

The survival model used in this research aims to produce the selection probability for each possible destination based on distance and walkable opportunities. Although we lack information about individuals' preferences of destinations, our intuition tells us that travel time is an important factor in destination choice. All else equal, a person is more likely to consider a closer destination. If the probability distribution function to visit various destinations for a traveler can be formulated, we can estimate the "importance" of a destination to the traveler by measuring the probability of visiting a destination. This way of understanding matches the purpose of survival analysis which is used to model the probability of the occurrence of events for longitudinal data.

The generic form of survival function given the duration of time (t) can be written as:

$$\Omega(t) = P(t > T) \tag{5}$$

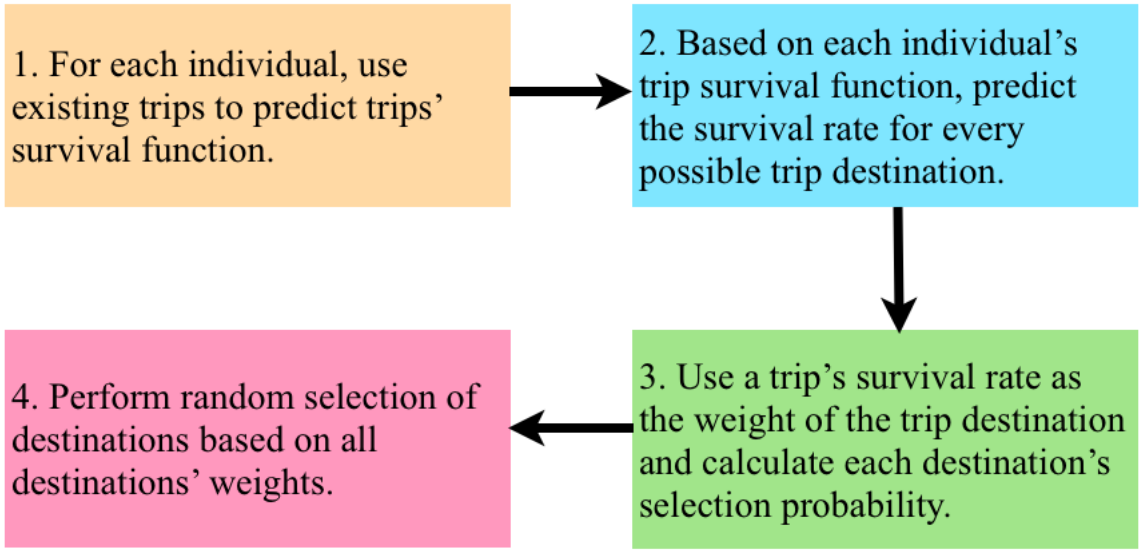

Figure 6: The procedure to form choice sets which combines survival analysis and random sampling.

For one individual, let T_k be the travel time of the fastest route from home to destination k. It can be written as:

$$ln(T_k) = \beta_0 + \beta_1 ln(A_k) + \sigma\epsilon_k \tag{6}$$

Where ϵ_k is a random error term. β_0 and σ are parameters to be estimated. Note that if $\sigma = 1$ and $ln(T_k)$ follows a normal distribution, the model is the same as the ordinary linear model.

The survival function is a useful tool for describing the probability distribution for the time of event occurrence (Allison, 2010). The simplest function is that the hazard is constant over time ($h(t) = \lambda$), meaning that during any period of time with a fixed length, the expected number of event occurrences is the same. Then its survival function follows exponential distribution. If the natural logarithm of hazard presumably equals $h(t) = \mu + \alpha log(t)$, the time of event occurrence is said to follow the Weibull distribution. Other distributions for survival analysis include log-normal distribution, log-logistic distribution, and the Gamma distribution. The exponential, Weibull, and log-normal distributions are special cases of the generalized Gamma model. In addition, the generalized Gamma model can also take a U shape or a bathtub shape.

The next step is to select an appropriate distribution function for $ln(T_k)$. The tested distribution functions include: Weibull, log-normal, log-logistic, exponential, and the Gamma model. We test which distribution function is the best fit for each individual's trips by

Table 2: A comparison of log-likelihoods of different distributions of $ln(T_k)$ for a single subject with GPSID 1019.

Distribution of T_k	Distribution of ϵ_k	log-likelihood	Nagelkerke R^2
Gamma	Log-gamma	-45.20	0.58
Log-logistic	Logistic	-48.69	0.55
Log-normal	normal	-48.13	0.54
Weibull	Extreme value (2 parameters)	-55.90	0.48
Exponential	Extreme value (1 parameter)	-74.63	0.28

Table 3: A comparison of different distributions of $ln(T_k)$ for all subjects

Distribution of T_k	Distribution of ϵ_k	log-likelihood
Gamma	Log-gamma	-123.97
Log-normal	normal	-124.19
Log-logistic	Logistic	-124.69
Weibull	Extreme value (2 parameters)	-128.77
Exponential	Extreme value (1 parameter)	-142.10

comparing the models' log-likelihood values. We separately estimate the probability density function of $ln(T_k)$ for each individual. To illustrate how to select the distribution function, an individual with GPSID 1019 is used as an example.

Table 2 compares the log-likelihoods of different distributions of $ln(T_k)$ for an individual with GPSID 1019. The Gamma distribution produces the largest log-likelihood, which suggests the best fit among the candidates. The next step is to test whether the differences of the log-likelihoods are statistically significant by performing the log-likelihood ratio test. The null hypothesis is that the log-likelihood of another model equals the log-likelihood of the Gamma model; the alternative hypothesis is that the log-likelihood of another model is smaller than the log-likelihood of the Gamma model.

To compare the goodness of fit of different distribution functions, the test statistic used is defined as twice the difference of two log-likelihood values (i.e., -2 ln(likelihood for another model model) + 2 ln(likelihood for the Gamma model)). The value of the test statistic is later compared with Chi-squared distribution with df = 1 at a level of significance of 0.01. In all these tests, we reject the null hypothesis that the log-likelihood of another model equals the log-likelihood of the Gamma model. Therefore the Gamma distribution is chosen to fit the distribution of travel time for the individual with GPSID 1019. The generalized Gamma distribution also produces a Pseudo-R^2 of 0.58, which shows satisfactory goodness of fit compared with other distribution functions. Based on the Gamma distribution function and AFT model, the coefficient of $ln(A_k)$ is estimated. Given the estimated probability density function, we can predict the survival probabilities for trips to all destinations (Allison, 2010).

The procedure for forming choice sets is shown in Figure 6. We reject several existing methods. First, the deterministic boundary setting of the selection area is not adopted because we do not have specific data to help define individuals' selection boundary. Second, the stratified sampling approach is rejected because our predicted probabilities have implied the selection weight for each destination and we do not want to add a new parameter (the number of strata) to the selection process. The random sampling of all destinations in the Metropolitan Area is selected because it is simple to use and it gives every destination an opportunity to be considered, as we lack more information about an individual's selection boundary. We integrate it with the survival analysis so that the random selection is based on the estimated weights of destinations. If a destination has a higher survival probability, it carries a heavier weight (i.e., a higher chance) to be selected into the choice set.

7 Choice set size

After the method of constructing choice sets is determined, a key question is to decide the choice set size M. A large number of destinations (Census blocks) in the Metropolitan Area make it computationally difficult to include all destinations in a choice set. But too small choice sets can result in inconsistent estimates (Auld and Mohammadian, 2011). It is therefore necessary to test different sizes of choice set to decide an appropriate choice set size needed for this research. In this study we propose a systematic method to test choice set size based on the weighted RMSE value of each model.

The traditional RMSE value is defined as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (\hat{\kappa}_n - \kappa_n)^2}{N}} \tag{7}$$

Where $\hat{\kappa}_n$ is the predicted probability for the nth observation in a data set. κ_n is a binary dependent variable which equals 1 if the destination in observation is visited and 0 otherwise. N is the total number of observations in the data set. The smaller the RMSE is, a better fit the model is claimed to be.

The traditional RMSE has one defect. If we increase the choice set size by adding less attractive destinations (such as very far destinations), the RMSE value may decline because such destinations have low selection probability anyway. Nevertheless, it does not necessarily mean that the model's actual predictability is enhanced. To control for this situation, we first separately measure the RMSE for chosen destinations and RMSE for non-chosen destinations in choice sets.

If there are N_1 actually chosen destinations in the data set, the RMSE of actual destinations can be written as:

$$RMSE_{chosen} = \sqrt{\frac{\sum_{n=1}^{N_1} (\hat{\kappa}_{chosen,n} - \kappa_{chosen,n})^2}{N_1}} \tag{8}$$

If there are N_2 unchosen destinations in the data set, the RMSE of all non-chosen destinations can be written as:

$$RMSE_{unchosen} = \sqrt{\frac{\sum_{n=1}^{N_2} (\hat{\kappa}_{unchosen,n} - \kappa_{unchosen,n})^2}{N_2}} \tag{9}$$

This function better balances the accuracy of predicting chosen destinations and the accuracy of predicting non-chosen destinations. The RMSE of the model is defined as the average of $RMSE_{chosen}$ and $RMSE_{unchosen}$. In other words, we assign 50% weight to $RMSE_{chosen}$ and the other 50% weight to $RMSE_{unchosen}$. This definition reduces the effects of having more undesirable destinations in a choice set on RMSE.

$$RMSE_{model} = p \cdot RMSE_{chosen} + (1 - p) \cdot RMSE_{unchosen} \tag{10}$$

Where $p = 0.5$. The choice set sizes tested range from 10 to 200, with an increment of 10. Even larger sizes such as 500, 1000, 2000, 5000, and 10000 are investigated. The RMSE values of models with different choice set sizes are further compared to decide an appropriate choice set size.

8 Model formulation

In the literature, traditional utility-based models have been developed to model non-work destination choice. The basic structure is the multinomial logit model (MNL). McFadden (1978) showed that the MNL model can consistently estimate parameters from a sample of alternatives through maximizing the conditional likelihood function, a feature that makes MNL widely used in modeling discrete choices.

Other revisions of this model include the generalized extreme value (GEV) model and mixed multinomial logit (MMNL) model (Bhat and Guo, 2004).

Table 5 summarizes some exemplary studies on discrete choice models applied in studying shopping destination choice. Such studies tend to use traffic analysis zones or specific stores (such as big supermarkets or malls) as destinations. In modeling destination choice, the utility-based MNL model and its variations are widely used.

Since the GPS data are panel data with repeated choices for individuals, there exists unobserved heterogeneity. To tackle this issue, we apply the mixed-effects logit model to investigate individuals' home-based non-work, single-destination destination choice.

The utility for one individual to visit destination k is defined as:

$$U_k = f(ln(T_k), \Lambda_k, \Theta_k, \Upsilon_k, b) \tag{11}$$

Where T_k is travel time of the fastest route from home to destination k. Λ_k represents a vector of land use variables. Θ_k represents a vector of transportation network measures. Υ_k represents the interaction of the individual's socio-demographics and transportation network measures and land use at destination k. b is an extra random effect term generated from a standard normal distribution with mean 0 for this individual.

9 Results and analysis

9.1 Choice set size

Figure 7 shows the RMSE of the mixed-effects models given different choice set sizes. As the choice set size increases, the RMSE of the model decreases in the beginning but then floats around a certain value. The computational cost rises with the increase of choice set size. Figure 8 further exhibits the RMSE values for different choice set sizes. The RMSE value floats around 0.48 as the choice set size ascends from 60 to 200. As the choice set size increases to 2000, the RMSE only lowers by 0.01 but the computational cost increases exponentially. The RMSE value for the chosen destinations shows similar values as the choice set size increases. The RMSE value for the chosen destinations also show similar values as the choice set size becomes greater than 40. After balancing the level of accuracy and computational cost, it is decided to use choice set size 60 for modeling non-work, single-destination choice, as it produces an appropriate level of accuracy with reasonable computational cost.

9.2 Modeling results

The results of the mixed-effects multinomial logit model are shown in Table 4. Model 1 includes all variables of interest. Model 2 excludes all interaction terms and turn index and Model 2 excludes all interaction terms and speed discontinuity, thanks to the correlation between turn index and speed discontinuity.

As shown in Model 1, travel time has a significant effect on destination choice. Longer travel time, all else equal, lessens the attractiveness of a destination. In Model 1, the walkable opportunities measure has a positive and statistically significant coefficient, indicating that an increase of stores at a destination makes it more attractive. The interaction term between male and accessibility has a negative coefficient, implying that a destination's increase of accessibility, all else equal, is more attractive to women than men. In

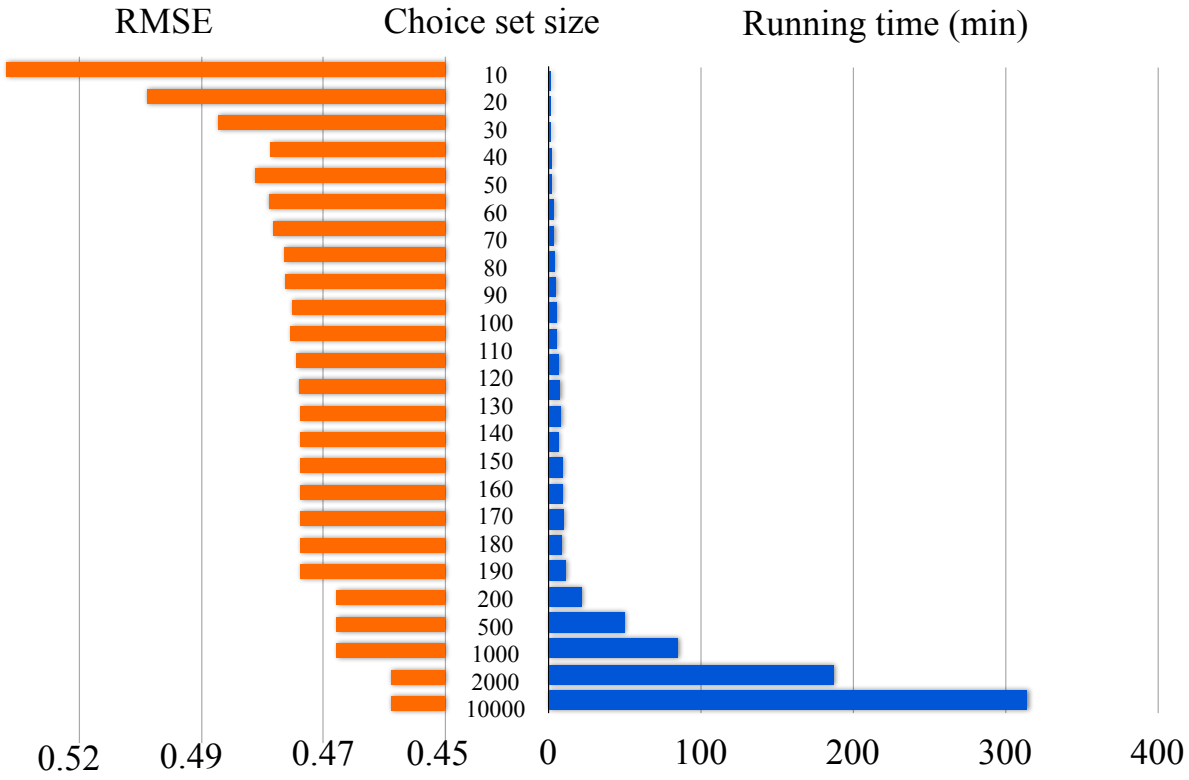

Figure 7: The root mean squared error (RMSE) value and running time for models of different choice set sizes.

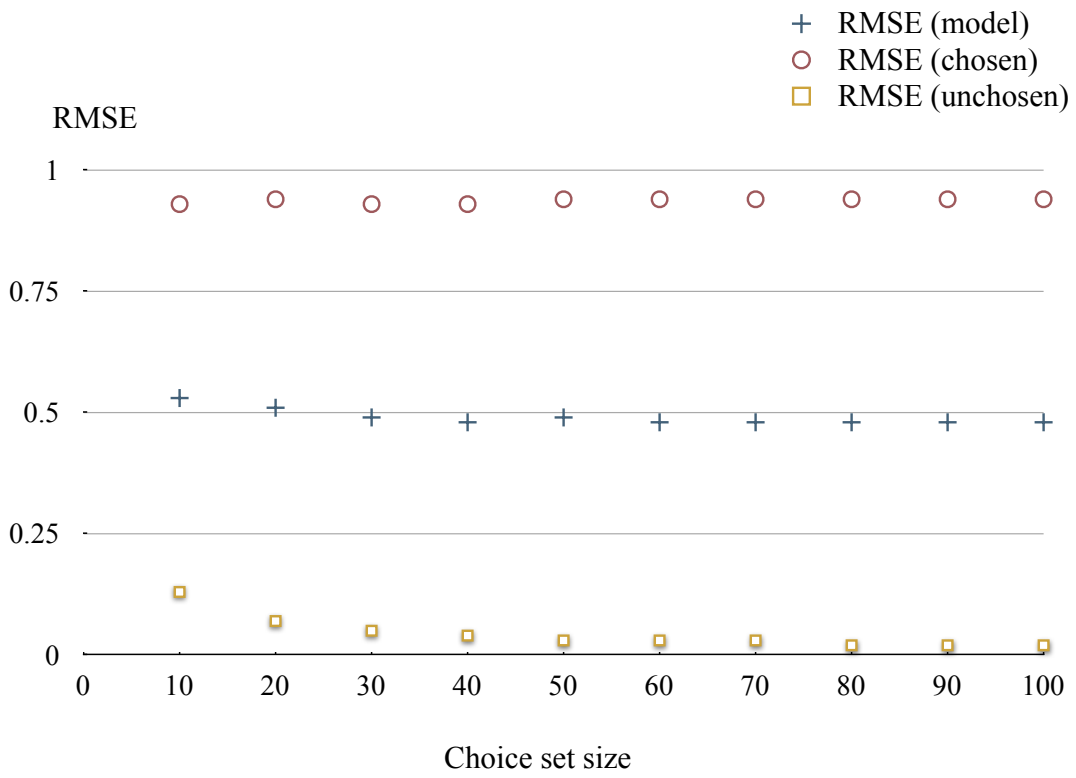

Figure 8: RMSE values for single-destination choice models of different choice sizes.

Table 4: Modeling single-destination choice for non-work vehicle trips

Model type		Mixed-effects logit model				
Number of observations used		73381				
Independent variables		Model 1	Model 2	Model 3	Model 4	Model 5
Land use	Walkable opportu. ($ln(A_k)$)	0.27 ***	0.17 ***	0.18 ***	0.01	0.10
	Male $\times$ $ln(A_k)$	-0.29 ***				
	Inclevel2 $\times$ $ln(A_k)$	0.10 ***				
	Inclevel3 $\times$ $ln(A_k)$	-0.12 ***				
	Diversity of services ($ln(H_k)$)	0.19	0.14	0.14	0.26 **	0.13
	Male $\times$ $ln(H_k)$	0.11***				
	Inclevel2 $\times$ $ln(H_k)$	-0.36***				
	Inclevel3 $\times$ $ln(H_k)$	3.69 ***				
Network features	Travel time ($ln(T_k)$)	-0.51***	-0.10 ***	-0.60 ***	-0.56***	-0.56***
	Speed discontinuity (ψ_k)	-0.03 *	-0.67 ***		-0.03 **	
	Turn index (ϑ_k)	-1.45**		-1.48 ***	-1.38 ***	
	Male $\times$ ψ_k	0.01				
	Inclevel2 $\times$ ψ_k	0.11**				
	Inclevel3 $\times$ ψ_k	-0.13***				
	Male $\times$ ϑ_k	0.19 **				
	Inclevel2 $\times$ ϑ_k	0.45 *				
	Inclevel3 $\times$ ϑ_k	0.28 **				
Familiarity	Time to work ($ln(T_{w,k})$)	-0.80 **	-0.68 ***	-0.77 ***		
	Time to downtown ($ln(T_{d,k})$)	0.55 ***	0.56 ***	0.77 ***		
	Male $\times$ $ln(T_{w,k})$	-0.24				
	Inclevel2 $\times$ $ln(T_{w,k})$	-0.20				
	Inclevel3 $\times$ $ln(T_{w,k})$	0.25				
	Male $\times$ $ln(T_{d,k})$	-0.60 **				
	Inclevel2 $\times$ $ln(T_{d,k})$	0.93 ***				
	Inclevel3 $\times$ $ln(T_{d,k})$	0.24				
	Time from home $\times$ time to work (ln)				-0.43 ***	
Evaluation	AIC	11225	12134	11473	11547	12396
	log-likelihood	-5586	-6059	-5728	-5767	-6139
	McFadden's R^2	0.11	0.04	0.09	0.09	0.02
	Nagelkerke R^2	0.12	0.04	0.10	0.09	0.02

Inclevel 1: $< \$100,000$; Inclevel 2: $\$100,000 - \$149,999$; Inclevel 3: $> \$149,999$
*** significant at 0.01; ** significant at 0.05; * significant at 0.1.

addition, a destination's increase of accessibility, all else equal, is attractive to an individual with household income $\$100,000 - \$149,999$ than an individual whose household income is below $\$100,000$. We also have examined a model using continuous household income variable in the interaction terms. This model reports a smaller log-likelihood value and smaller McFadden's R^2 than the using income levels as groups, and therefore is not adopted.

Though not statistically significant (but close to 0.10 level of significance), the coefficient of the diversity of services is also positive. In fact when the walkable opportunities measure is excluded, the coefficient of the diversity of services becomes statistically significant. The interaction term between male and diversity of services has a positive sign, meaning that a destination's increase of diversity of services, all else equal, is more attractive to men than women.

In network structure measures, as hypothesized, the turn index has a negative coefficient which indicates that a destination reached via a route with more turns per unit time dampens its attractiveness. The interaction term between male and speed discontinuity has a positive coefficient, and so does the interaction term between male and turn index. The findings reveal the attractiveness of a destination drops more for a woman than for a man, as the route requires more turn per unit time. Speed discontinuity here has a negative coefficient but is not statistically significant. Further investigation reveals that it may be due to the correlation between speed discontinuity and turn index. When turn index is excluded from the model, the coefficient of speed discontinuity becomes statistically significant (see Model 2). The interaction terms also suggest that the changes of speed discontinuity have different effects on gender and income groups in single-destination choice.

Regarding the axis of travel, travel time between destination and work has a negative coefficient, meaning that a destination closer to work, all else equal, is more likely to be selected. In addition, as hypothesized, travel time between destination and the nearest downtown has a positive coefficient. It suggests that all else equal, a destination closer to the nearest downtown is less attractive, which may be due to greater parking cost or limited parking space. All else equal, men are more likely to choose a destination far away from downtown than women. We further test a new variable which equals the multiplication of travel time from home and travel time to work. It aims to quantify the distance from a destination to the axis between work and home. It is hypothesized that the greater this term is, the less attractive the destination is. As this multiplicative term is correlated with trip chain's travel time and travel time to work, these two variables are excluded in the model when the multiplicative term is included. The results are shown in Model 4 in Table 4. The coefficient of the multiplicative term is negative which supports our hypothesis. The results further reveals that models with network measures or axis of travel measures have greater goodness of fit than a model with only land use variables (Model 5).

The elasticity of key independent variables are further calculated (Figure 9). Considering

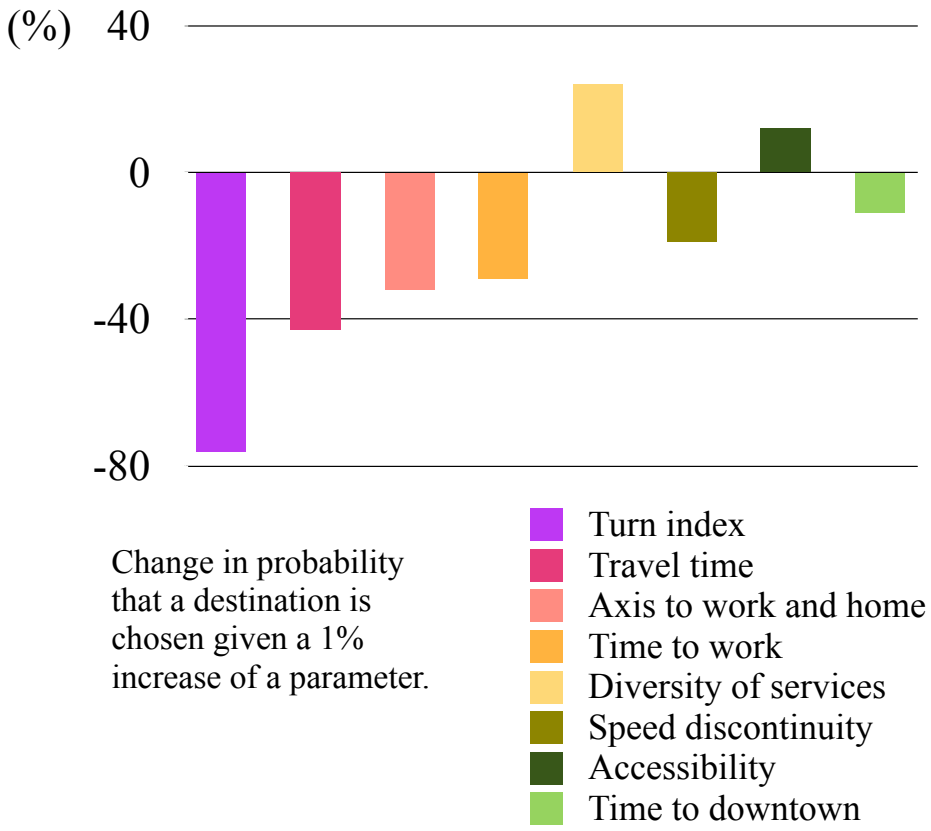

Figure 9: The elasticities of the odds of destination choice for key independent variables.

the correlations among variables, we first run the mixed-effect model on one variable at a time, and then calculate the elasticity for each estimated coefficient. The variable that has the highest absolute value of elasticity is turn index, following which is travel time. A one percent increase of the number of turns per unit distance for the travel route, all else equal, is associated with an about 76% decrease of the likelihood of selecting this destination. A one percent rise of travel time, all else equal, is associated with an about 43% decline of the likelihood of selecting this destination. The same interpretation applies to other variables. We further test a new independent variable which is the multiplication of travel time to home and travel time to work. The idea is to investigate the impact of a destination's relative distance to home and work on destination choice. Its elasticity equals -32%, suggesting that the farther away a destination is from the axis between home and work, the less attractive the destination is.

10 Discussion

This research proposes a new approach that combines survival analysis and random sampling to form choice set for non-work destination choice using the in-vehicle GPS data. A systematic investigation of appropriate choice set sizes is also performed. The mixed-effects multinomial logistic models are used to model single-destination choice. In these models, we examine the effects of land use and transportation networks on destination choice. The key findings are:

1. The two most influential factors on single-destination choice are turn index and travel time from home to destination.

2. More walkable opportunities and greater diversity of services, all else equal, make a destination more attractive.

3. A destination reached by a route with greater changes of speed per unit time or more turns per unit time is less attractive.

4. Individuals' socio-demographics such as gender and household income, interacting with land use and route network measures, also affect destination choice.

5. A destination's travel time to work and home influences its attractiveness. All else equal, a non-work destination closer to work is more attractive to travelers. A destination closer to the nearest downtown is less attractive to travelers, which may be due to a greater parking cost and other nuisances near downtown.

6. The above variables have different effects on gender and income groups in destination choice.

In summary, based on the in-vehicle GPS travel data, this paper proposes a new approach to select choices and a systematic method to decide the choice set size. Further, we test some hypotheses which were not tested before. The results suggest that land use, travel time, path familiarity, a destination's reachability, axis of travel, and individuals' income and socio-demographics all together influence non-work destination choice. A future extension of this research is to investigate home-based trip chains with multiple destinations. In a multi-destination scenario, the spatial interactions of different destinations may influence both the choice set formation and the destination choice process in a trip chain.

References

Allison, P. (2010), *Survival analysis using SAS: A practical guide*, SAS Institute Inc.

Arentze, T. A., Oppewal, H. and Timmermans, H. J. (2005), "A multipurpose shopping trip model to assess retail agglomeration effects", *Journal of Marketing Research* , Vol. 42, pp. 109–115.

Auld, J. and Mohammadian, A. (2011), "Planning-constrained destination choice in activity-based model", *Transportation Research Record: Journal of the Transportation Research Board* , Vol. 2254, pp. 170–179.

Barros, C., Butler, R. and Correia, A. (2010), "The length of stay of golf tourism: A survival analysis", *Tourism Management* , Vol. 31, pp. 13–21.

Bernardin, V., Koppelman, F. and Boyce, D. (2009), "Enhanced destination choice models incorporating agglomeration related to trip chaining while controlling for spatial competition", *Transportation Research Record: Journal of the Transportation Research Board* , Vol. 2132, pp. 143–151.

Bhat, C. (1998), "Analysis of travel mode and departure time choice for urban shopping trips", *Transportation Research Part B: Methodological* , Vol. 32, pp. 361–371.

Bhat, C. and Guo, J. (2004), "A mixed spatially correlated logit model: Formulation and application to residential choice modeling", *Transportation Research Part B: Methodological* , Vol. 38, pp. 147–168.

Burke, M. and Brown, A. (2007), "Distances people walk for transport", *Road & Transport Research: A Journal of Australian and New Zealand Research and Practice* , Vol. 16, p. 16.

De Palma, A., Dunkerley, F. and Proost, S. (2010), "Trip chaining: Who wins who loses?", *Journal of Economics & Management Strategy* , Vol. 19, pp. 223–258.

Draijer, G., Kalfs, N. and Perdok, J. (2000), "Global positioning system as data collection method for travel research", *Transportation Research Record: Journal of the Transportation Research Board* , Vol. 1719, pp. 147–153.

Fotheringham, A. S. (1986), "Modelling hierarchical destination choice", *Environment and Planning A* , Vol. 18, pp. 401–418.

Handy, S. and Clifton, K. (2001), "Evaluating neighborhood accessibility: Possibilities and practicalities", *Journal of Transportation and Statistics* , Vol. 4, pp. 67–78.

Krizek, K., Iacono, M., El-Geneidy, A., Liao, C. and Johns, R. (2009), Access to destinations: Application of accessibility measures for non-auto travel modes. Access to Destinations Study Series Report: Mn/DOT 2009-24, University of Minnesota.

Leszczyc, P., Sinha, A. and Timmermans, H. (2000), "Consumer store choice dynamics: An analysis of the competitive market structure for grocery stores", *Journal of Retailing* , Vol. 76, pp. 323–345.

Levinson, D. and El-Geneidy, A. (2009), "The minimum circuity frontier and the journey to work", *Regional Science and Urban Economics* , Vol. 39, pp. 732–738.

McFadden, D. (1978), Modeling the choice of residential choice, *in* A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull, eds, 'Spatial Interaction Theory and Planning Models', Amsterdam: North Holland, pp. 75–96.

Newman, J. and Bernardin, V. (2010), "Hierarchical ordering of nests in a joint mode and destination choice model", *Transportation* , Vol. 37, pp. 677–688.

Parthasarathi, P., Hochmair, H. and Levinson, D. (2012), "Network structure and spatial separation", *Environment and Planning Part B* , Vol. 39, pp. 137–154.

Parthasarathi, P., Levinson, D. and Hochmair, H. (2013), "Network structure and travel time perception", *PloS one* , Vol. 8, Public Library of Science, p. e77718.

Pellegrini, P., Fotheringham, A. and Lin, G. (1997), "An empirical evaluation of parameter sensitivity to choice set definition in shopping destination choice models", *Papers in Regional Science* , Vol. 76, pp. 257–284.

Pozsgay, M. and Bhat, C. R. (2001), "Destination choice modeling for home-based recreational trips: Analysis and implications for land use, transportation, and air quality planning", *Transportation Research Record: Journal of the Transportation Research Board* , Vol. 1777, pp. 47–54.

Rashidi, T., Auld, J. and Mohammadian, A. (2012), "A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection", *Transportation Research Part A: Policy and Practice* , Vol. 46, pp. 1097–1107.

Recker, W. W. and Kostyniuk, L. P. (1978), "Factors influencing destination choice for the urban grocery shopping trip", *Transportation* , Vol. 7, pp. 19–33.

Shannon, C. E. (1948), "A mathematical theory of communication", *Bell System Technical Journal* , Vol. 27, pp. 379–423, 623–656.

Timmermans, H. (1996), "A stated choice model of sequential mode and destination choice behaviour for shopping trips", *Environment and Planning A* , Vol. 28, pp. 173–184.

Wang, L. and Lo, L. (2007), "Immigrant grocery-shopping behavior: Ethnic identity versus accessibility", *Environment and Planning A* , Vol. 39, p. 684.

Xie, F. and Levinson, D. (2007), "Measuring the structure of road networks", *Geographical Analysis* , Vol. 39, pp. 336–356.

Zhu, S. (2010), The Roads Taken: Theory and Evidence on Route Choice in the Wake of the I-35 W Mississippi River Bridge. Ph.D. Dissertation, Department of Civil Engineering, University of Minnesota.

Zolfaghari, A., Sivakumar, A. and Polak, J. (2012), "Choice set pruning in residential location choice modelling: A comparison of sampling and choice set generation approaches in greater london", *Transportation Planning and Technology* , Vol. 35, pp. 87–106.

Table 5: Summary of selected studies on destination choice from literature

Study	Data	Topic	Model	Key findings
Timmermans (1996)	Travel survey data in Eindhoven, Netherlands	Sequential mode and destination choice for shopping trips	MNL	Mode choice does not influence the choice of shopping centers. Bigger shopping centers are more attractive than smaller ones.
Pellegrini et al. (1997)	Phone survey data on shopping trips in Gainesville FL	Parameter sensitive to choice set specification for shopping destination choice	MNL	The stability of parameter estimates can be sensitive to choice set size and composition.
Bhat (1998)	1990 San Francisco Bay Area Household Travel Survey	Travel mode and departure time choice of shopping trips	MNL for mode choice and MNL-OGEV for departure time choice	In estimating travel time choice, nested logit model outperforms the MNL model and MNL-OGEV model outperforms the nested logit model in terms of data fit.
Leszczyc et al. (2000)	Grocery shopping data in Springfield, MO	Consumers' store choice and trip time choice	Hazard model and MNL	Store choice and shopping time choice are interdependent. Spatial competition between stores affects consumers' store choice and switching behavior.
Pozsgay and Bhat (2001)	1996 Dallas-Fort Worth household activity survey	Destination choice for home-based recreational trips	nonlinear-parameter MNL	Agglomeration effects are prominent in affecting recreational attraction-end choice.

Bernardin et al. (2009)	Household survey data in Knoxville, Tennessee	Destination choice of home based maintenance trips and home-based other trips	MNL and ACDC (agglomerating and competing destination choice models)	The ACDC model reflects the effects of trip chaining and spatial agglomeration whereas MNL cannot.
Newman and Bernardin (2010)	2000 Knoxville Urban Area Household Travel Behavior data	Mode choice and destination choice for work tours	Hierarchical ordering nested logit	Hierarchical ordering of decision nesting trees is important for modeling location and mode choice; employing a reverse ordering can be a good choice.
Recker and Kostyniuk (1978)	Survey data of 1500 households in Buffalo, NY	Shopping destination choice	MNL	Individuals' attitudes toward the store and its operation, perception of destinations' reachability, and the number of destinations at the destinations influence destination choice.
Fotheringham (1986)	Immigration data in 62 US Cities (1965-1970)	Hierarchical destination choice	Competing destination model	Competing destinations model perform better than gravity-based model for hierarchical destination choice. The model is limited by the lack of consideration of trip chaining behavior.

Arentze et al. (2005)	Household survey data in Northern Brabant in the Netherlands	Combined choices of trip purpose and destination	Nested logit	(1) The presence and size of different types of stores influence location choices. (2) Consumers prefer shopping centers they are more familiar with and prefer to visit a single large agglomeration of stores to conduct multi-purpose shopping.
Wang and Lo (2007)	Stated shopping preference data of Chinese immigrants in Toronto	Supermarket destination choice	MNL	Immigrant shoppers' socio-demographic attributes more affect their shop destination choice than distance and accessibility.
De Palma et al. (2010)	Numerical examples	Modeling both retail location choice and consumers' destination choice with the consideration of trip chaining behavior	Decision tree and logit model	Trip chaining option reduces the profit margins in the short run but increases welfare for firms.
Auld and Mohammadian (2011)	Activity travel survey of households in Chicago	Non-work destination choice	MNL	Choice set formation that considers planning constraints improves prediction accuracy.