


Choice set generation from GPS data set for grocery shopping location choice modelling in canton Zurich

Comparison with the Swiss Microcensus 2005

Working Paper

Author(s):

Kawasaki, Tomoya; [Axhausen, Kay W.](#) 

Publication date:

2009-11

Permanent link:

<https://doi.org/10.3929/ethz-a-005939093>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

Originally published in:

Arbeitsberichte Verkehrs- und Raumplanung 595

Choice set generation from GPS data set for grocery shopping location choice modelling in canton Zurich: Comparison with the Swiss Microcensus 2005

Tomoya Kawasaki
Kay W. Axhausen

Working Paper 595

November 2009

Arbeitsberichte Verkehrs- und Raumplanung

Working Paper 595

Choice set generation from GPS data set for grocery shopping location choice modeling in canton Zurich: Comparison with the Swiss Microcensus 2005

Tomoya Kawasaki	Kay W. Axhausen
Department of International	IVT
Development Engineering	ETH Zurich
Tokyo Institute of Technology	CH-8093 Zurich
2-12-1, O-okayama, Meguro-	Phone: +41-44-633 39 43
ku, Tokyo	Fax: +41-44-633 10 57
Phone: +81-3-5734-3468	axhausen@ivt.baug.ethz.ch
Fax: +81-3-5734-3468	
kawasaki@tp.ide.titech.ac.jp	

November 2009

Abstract

This paper describes a set of new procedures used to generate choice sets for grocery shopping destination choices observed in a large scale GPS data set. The procedures are a rule-based identification of the shopping trips, defining places from the observed locations using a hierarchical-clustering algorithm and finally a rule-based identification of the place of home and work/education and implicitly with this the imputation of the trip purposes.

The GPS traces were collected by volunteers: 2,435 persons in Zurich and 1,086 persons in the smaller city of Winterthur. No socio-demographic information such as age, sex, income, etc. was available for this research, as the data was collected for a different purpose. The choice sets are generated by transport mode used for the shopping trip; car, walking and urban public transport whereas trips by intercity rail and bicycle were excluded since their numbers are too low for an estimation of a destination choice model, the ultimate purpose of the overall project.

The results were compared with the Swiss Microcensus (MZ) 2005, the national travel behaviour survey, for time-of-week, time-of-day, duration, duration of the shopping activity. While the quality of the results was high, the comparisons highlight instructive differences resulting from the different data generation process.

Keywords

Choice set generation, GPS, destination choice, grocery shopping, cluster analysis, purpose imputation

Preferred citation style

Kawasaki, T. and K.W. Axhausen (2009) Choice set generation from GPS data set for grocery shopping location choice modelling in canton Zurich: Comparison to Swiss Microcensus 2005, *Arbeitsberichte Verkehrs- und Raumplanung*, **595**, IVT, ETH Zürich, Zürich.

1 Introduction

Since the mid 1990's Global Positioning System (GPS) based studies have dramatically expanded the research opportunities of travel behaviour research, as they reduce response burden substantially, if they are not combined with elaborate questioning for additional information, such as for trip purposes or expenditures, while providing very precise time and location information. Current research focuses upon the development of GPS post-processing procedures that allow the researcher to derive all necessary information, such as start and ending time, mode, and trip purpose directly from the GPS records. Moreover, as continuously growing sample sizes naturally lead to an increasing demand for automated procedures with low computational cost. Several studies (e.g. Axhausen *et al.*, 2003; Wolf, 2006; Stopher *et al.*, 2005) have been proposing the state-of-art ideas for automated GPS data processing, but no standards have yet emerged.

For the purpose of modelling individual behaviour from processed GPS data, one needs to identify trip purposes and to detect the place of home and work/education of each individual and finally generate the relevant choice sets. This study aims to generate choice sets for grocery shopping destination choice modelling from large GPS data sets and compare the results of the automated processes with the Swiss Microcensus 2005, the best available data source for Swiss travel behaviour.

The paper will first describe the data set employed, followed by a discussion of the proposed new automated new procedures. The comparisons of the results is the final section of the paper, which also includes an outlook for further work.

2 Data

2.1 GPS data sets used

Since the first GPS studies in the mid 1990s (e.g. Wagner, 1997; Casas and Arce, 1999; Yalamanchili *et al.*, 1999; Draijer *et al.*, 2000; Pearson, 2001), this new way of surveying individual travel behaviour has gained increasing attention in transport research.

The GPS data sets in canton Zurich are processed from the most basic person-based GPS raw data with three dimensional positions and the corresponding timestamps, as described in Schuessler and Axhausen (2009). In their study, each individual's trip and activity were identified and trips were segmented by transport mode used (car, walking, urban public transport, rail, bike). The detection of transport mode was carried out by a fuzzy logic approach. The resulting data set is summarized as Table 1.

Table 1 Overall statistics of the GPS data and MZ 2005 data for canton Zurich

	GPS Trace		MZ 2005
	Zurich	Winterthur	
Number of persons in the sample	2,435	1,086	3,005
Number of days per person	6.99	5.96	1.00
Number of trips per day	4.50	3.40	3.65
Average trip distance [km]	7.72	7.37	8.79
Average daily mileage [km]	34.74	23.20	32.13
Average trip duration [min]	15.17	13.71	26.21

Source: Schuessler and Axhausen (2008)

GPS data of some 2,435 individuals resident in the Swiss city of Zurich and 1,086 individuals in nearby smaller Winterthur (total 3,521 individuals) were collected by logging their locations second-by-second for one week each using a personal GPS unit between April, 2004 and April, 2005. Data sets were divided into subsets with the trip and activity descriptions: trip starting and ending point, corresponding date and time, trip distance and duration, and

transport mode used are available for the trip, whereas activity starting point and ending point, corresponding date and time and activity duration are available for the activities. Starting and ending point for each activity were obtained based upon two ideas; activities with ongoing GPS recording and activities with signal loss defined by Schuessler and Axhausen (2008). No participants' socio-demographic data are available, as this information was not available anymore.

2.2 Swiss Microcensus (MZ) 2005

The survey for the MZ is conducted about every five years and delivers a representative and detailed insight into the travel patterns of the Swiss population by Swiss Federal Statistical Office (FSO) and Federal Office for Spatial Development (ARE) since 1974. In 2005, 33,390 individuals reported in the course of a computer-assisted telephone interview (CATI) on their socio-economic background, their mobility tools, and stage-based recorded, trips and activities on the reporting day. As a novelty, MZ 2005 data newly captures and provides geo-coordinates for every trip's origin and destination as well as for the home location of the participants and their work places. This allows us to model in a more disaggregate manner than ever before. MZ 2005 includes information at a national level. A subset of the trips starting and ending in canton Zurich for general shopping was extracted. This subset contains 3,005 trips. Since this paper is just focussing on grocery shopping, only these relevant trips were retained. Fortunately, this is possible as the MZ 2005 included the question "What kind of shopping did you carry out in this place?" which explicitly ask for the type of shopping performed. As Carrasco (2008) summarized, from the initial shopping trip subset of 3,005 observations, 1,170 (38.9% of the valid observations) were for grocery shopping. This new subset was used for the later comparisons with the grocery shopping trips identified in the GPS traces.

3 Detecting grocery shopping trips

The GPS traces do not include activity purpose information as mentioned earlier. They need to be identified as best possible using the information in the traces. The available signals of the trip for grocery shopping are (1) space, (2) time and (3) activity duration. The geo-coordinates of grocery store as well as hectare data¹ are available (Carrasco, 2008). Activity duration and time-of-day of the activity start is recorded in the GPS data. Ideally, to identify the shopping trips, one should exploit all the available information in one step as pointed out by Axhausen *et al.* (2004), but at this moment, sequential approach is sufficient because of the different levels of precision and lack of a suitable simultaneous approach. The heuristics used to identify the grocery trips are shown in box 1;

Box 1 Heuristics to detect grocery shopping trips

1. Extract trip if trip ending is within 30 meter from a grocery store. In case more than two stores, the nearest store is chosen.
2. Extract trip if activity is done during the opening hours of the grocery store
3. Extract trip if activity is less than 180 minutes
4. Extract trip if activity matches with relevant employment in the hectare data

As a first heuristic, trips ending within 30 meter radius from grocery stores were extracted. Several studies point out the problem of positioning accuracy of GPS data. Wolf (2006) concludes that the GPS positioning error lies between five and ten meter under the ideal condition. In reality, however, it is usually much worse due to several error sources. For instance, there might be less than the four satellites in view that are required to precisely calculate a three-dimensional position. Even if there are enough satellites in view, they might not be ide-

¹ Hectare data refers to an official data set providing the number of employees by General Classification of Economic Activities (NOGA) code group and the number of firms by NOGA code for each hectare in Switzerland. It also provides the number of residents.

ally positioned, which is expressed by a high position dilution of precision (PDOP) value (Wolf *et al.*, 1999). While this leads to GPS positions that are rather different from the actual position of the receiver, the so-called *warm start/cold start problem* results in missing GPS points at the beginning of the trip due to the time the GPS receiver needs to acquire the position of at least four satellites in view (Stopher *et al.*, 2005). In addition, there are random errors caused, for example, by satellite or receiver issues, atmospheric and ionospheric disturbances, multi-path signal reflection or signal blocking (Jun *et al.*, 2007). Especially burdensome are multi-path errors, also called urban canyoning errors because they typically appear in urban canyons. The GPS signal is reflected by buildings, walls or surfaces and the corresponding GPS positions jump and are often widely scattered around the actual position of the receiver. Signal blocking, on the contrary, leads to missing GPS points and is of special importance for person-based GPS surveys since it varies systematically with the different means of transport. While GPS reception is generally good when the participant is walking, cycling and or travelling by car, it varies considerably for public transport journeys, depending on the proximity of the person to the nearest window (Draijer *et al.*, 2000; de Jong and Mensonides, 2003). Ashbrook and Starner (2003) summarize that GPS error is 15 meter and Liao *et al.* (2007) conclude 10 meter.

As Carrasco (2008) mentioned, geo-coordinates of grocery store might also contain an error in terms of positioning accuracy, which are collected from Google Earth. Considering the papers mentioned above, this study adopts a conservative 30 meter radius around the grocery store in order to overcome problems caused by GPS positioning errors.

However, of course, to rely only on geo-coded data is too simple minded since some of other places, for example home, might be located within 30 meters from the store. Thus we use the “time” information for next heuristics posterior to the usage of “space” information. People can visit shops only during their opening hours. Opening times of the stores in the region are available from Carrasco (2008, and see his sources there). Some of them were however missed. That missing information was manually collected. The differences between opening

times across the day of the week and the presence of lunch breaks are taken into account. In the MZ 2005 data the activity duration of grocery shopping is less than 180 min. In other words, we can infer that activities of more than 180 min are very likely to be other activities than grocery shopping (e.g. working, leisure, etc.). However, this step excludes few trips since most of activity durations after the first heuristic are of less than 180 min duration.

As a final heuristic, trips were checked against the hectare data, which contains information about the firms residing in each hectare. Fortunately, the grocery store's information is classified according to store size. Unless the hectare of the observed trip end includes a grocery store, it will be excluded from the list. This step also excluded few trips similarly to the previous step. As a result of applying the heuristics above, 1,243 grocery shopping trips by 790 individuals were obtained.

4 Home and Working/Educational (w/e) Locations

4.1 Clustering places into locations

For modelling purposes, one needs to detect an individual's home and working/educational (w/e) locations. First of all, all trip ending *places* of each person should be clustered into *locations* to detect these two locations. The clustering is necessary due to GPS errors discussed above. The definition of *places* and *locations* are as follows;

- “Place”: each point of GPS data *prior to* clustering
- “Location”: each point of GPS data *posterior to* clustering

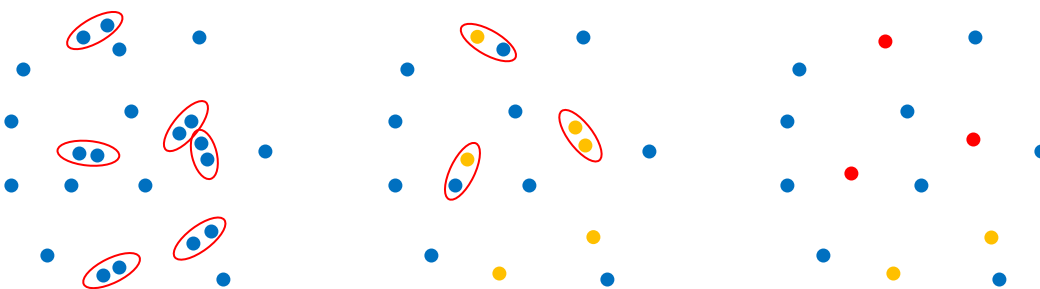
For the technique of clustering *places*, several algorithms have been proposed. For example, Ashbrook and Starner (2003) employ k-means clustering across the persons. These locations are then incorporated into a Markov model to finally find POIs by trip purpose. The k-means

clustering method is one of the best known methods among non-hierarchical clustering approaches. It clusters *places* into of number of *locations* which is determined prior to analysis. It is preferable when handling very large number of data. However in this study, our interest is just to know individual's home and w/e location. In addition, this study does not analyze across the persons but for each individual separately. The average number of trips for each individual is relatively low:

- Winterthur: 18.2 activities/person-survey (3.14 activities/person-day)
- Zurich: 29.6 activities/person-survey (4.23 activities/person-day)

Thus a hierarchical approach is, in this case, preferable for POI detection. Since trip ending points of each individual are statistically independent and equally weighted, and in addition, clustering is just based on distance between places, an agglomerative clustering method, which is one of the most well-known methods among the hierarchical approaches, is preferable to cluster *places* into the *locations*. The basic idea of agglomerative clustering method used in this study is shown in following figure 1 and the clustering algorithm is presented in Box 2.

Figure 1 Basic idea of the clustering algorithm



Box 2 Clustering Algorithms

A set of clusters: $C = C_1, \dots, C_n$

1. Cluster C_i, C_j if Euclidian distance of two places (locations) is the nearest *and* less than 30m
2. Create C_{new} and delete C_i, C_j from the cluster set
3. Estimate the mean point of C_{new} and add to the cluster set

At this stage, distance error between each place (location) is again set as 30 m Euclidian distance as for the GPS positioning accuracy. Thus, geo-coded locations within 30 meter are regarded as belonging to the same cluster. Distances between clusters can be calculated by Un-weighted Pair Group Method with Arithmetic mean (UPGMA), which is a suitable approach as each point of this GPS data is equally weighted.

$$d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \quad (1)$$

Where x_i, y_i denotes location and A, B is for location (cluster). Posterior to applying the clustering algorithm, the trip ending points of the city of Winterthur, Zurich and canton Zurich were grouped as follows.

- Winterthur: 19,761 *places* to 16,159 *locations*
- Zurich: 72,011 *places* to 33,125 *locations*
- Canton Zurich (total): 91,772 *places* to 49,284 *locations*

4.2 Detecting home and w/e locations

Next, we attempt to infer the individual's home location and w/e locations. For setting up the heuristics to identify home and w/e locations of every respondent two further pieces of information were used: (1) frequency of visit to the location and (2) activity duration at the location. As for duration at locations, average duration of corresponding activity was simultane-

ously calculated in the previous step. By using this information including its time profile across the day, the most probable POIs are identified. Four items of information: trip frequency, activity duration, day of week, time of day are given in the datasets and used for this heuristic. Weighting values (1.5, 1.0, 0.7 and 0.4) were derived from the study of Wolf *et al.* (2004). The heuristics established in this study is shown in Box 3.

Box 3 Heuristic to identify location of home and w/e

1. Clusters that are frequented at most were given the weight 1.5, 2nd the weight 1.0, 3rd the 0.7, 4th the weight 0.4.
2. Clusters that have the longest duration were given the weight 1.5, 2nd the weight 1.0, 3rd the 0.7, 4th the weight 0.4.
3. Clusters, which have the highest weight at this moment and matched with hectare data were categorized as home. In case clusters were not matched with hectare data, the 2nd most frequented clusters were assigned as home.
4. Clusters, which were assigned as home were deleted from the list.
5. Clusters, which included weekend activity were deleted from the list.
6. Clusters, which have the highest weight at this moment and matched with hectare data were categorized as w/e. In case clusters were not matched with hectare data, the 2nd most frequented clusters were assigned as w/e.

During the survey period (6.99 days/person in Zurich and 5.96 days/person in Winterthur as shown in table 1), a person living might visit home most frequently among all potential locations and w/e place (university, high school, elementary school, etc.) might be the 2nd most frequented visited location. For w/e place, it is impossible to distinguish between the two purposes due to the lack of individual socio-demographic information. Thus, we give them the same weight. This is the first heuristic.

Secondly, activity duration at the location is used for the next heuristic. Obviously, the activity with the longest duration is very likely to be “home” whereas “w/e” is likely to have the 2nd longest duration. Thus, clusters that have the longest duration were given the weight 1.5,

2nd the weight 1.0, 3rd the 0.7, 4th the weight 0.4. This is the second heuristic. As the next step, clusters, which have the highest weight at this moment and with matching hectare data, were categorized as home. In case clusters were in hectares without residential addresses, the 2nd most frequented clusters were assigned as home. After deleting “home” locations, locations visited only on the weekend were also deleted from the location list. Normally, activities done in the weekend are very unlikely to be “work”. After that, clusters, which have the highest weight at this moment and matching hectare information were categorized as w/e. In case clusters were not confirmed by hectare data, the 2nd most frequented clusters were assigned as w/e.

Informal tests were performed to check the validity of obtained data. We assume each individual is very likely to be at home at 2am whereas individuals are very likely to be at w/e place between 10am and 2pm. Result of these tests are shown in Table 2.

Table 2 Validity of purpose imputation: Results of informal tests

	GPS Trace	
	Zurich	Winterthur
Home location identified	1,435	1,066
At home location at 2am	1,406 (97.98% of all)	1,026 (96.25% of all)
w/e location identified	1,402	995
At w/e location	1,346	934
between 10am and 2pm	(96.01% of all)	(93.87% of all)

From the informal test produces convincing fits as shown in table 2. In 97.98% of locations detected as home the persons were at home at 2am in the city of Zurich, similarly, activities done at 2am in Winterthur also matched at a very high rate. The same is true for the w/e locations. The following Figures 2 and 3 show the distributions of the activity durations at the identified locations. These distributions look reasonable with regards to their form, as well as with regards to their means.

Figure 2 Activity duration distributions at home in canton Zurich

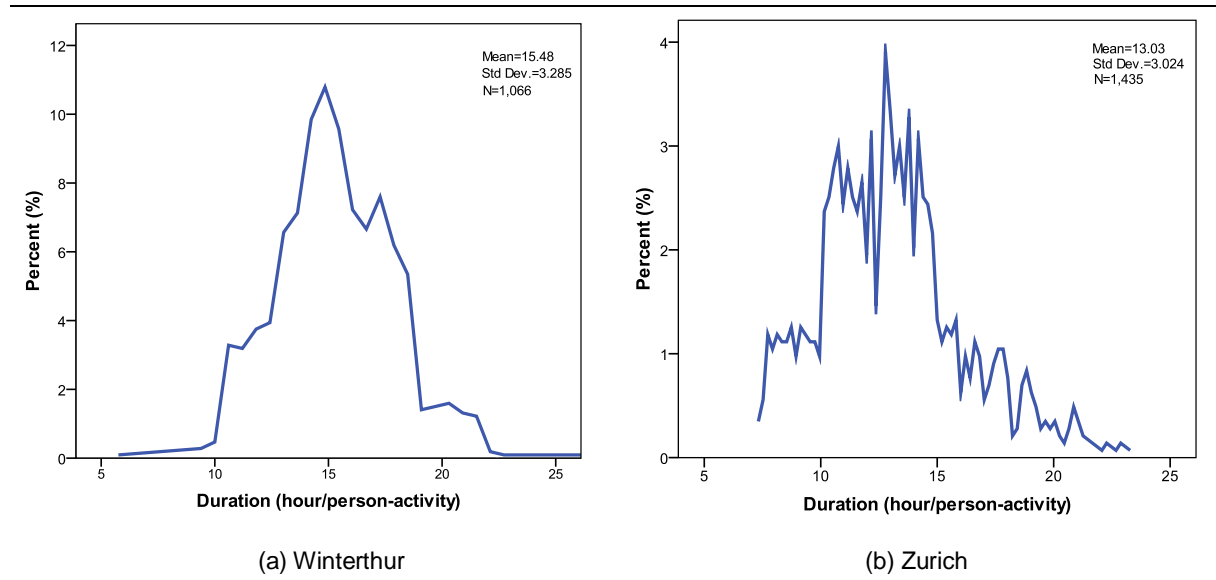
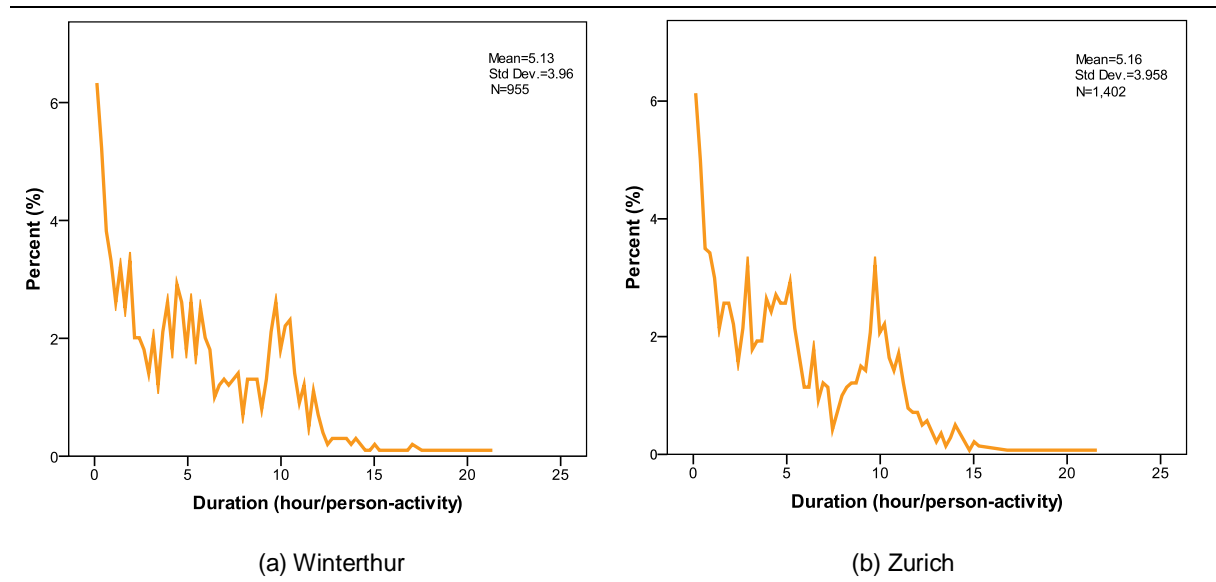


Figure 3 Activity distribution at w/e locations in canton Zurich



5 Choice Set Generation

In preparation for a later choice modelling exercise three choice sets were generated: one for the grocery shopping trips undertaken by car, one for trips where people walked to the store and one for trips undertaken by public transport. In the following section, the choice set generation process is described.

5.1 Initial data

Various datasets are required to produce the choice sets. In the first place, three data files (one for each type of choice set – car, walk and public transport) were generated, containing information about the grocery shopping trips (individual trip ID) with the geo-coded location of departure, the location of the activity posterior to the shopping trip, location of home-based trip and the available time to travel for each person, or travel time budget (TTB) which will be mentioned later in this section. The trips undertaken by other transport modes (14 trips for intercity rail and 127 trips for bicycle) are excluded from the choice sets due their low absolute number, which does not allow further choice modelling. An additional dataset included the shopping opportunities, containing the geo-coded location of the universe of alternatives (1,250 stores). Using the latitude and longitude information, event files were created for all the origins, all the destinations and all the shopping opportunities locations. At this point, locations of activities posterior to the grocery shopping which were outside of the study area (canton Zurich) were excluded from the data.

5.2 Generation of travel time matrices and feasible alternative selection

5.2.1 Time-geographic approach

In Hagerstrand's (1970) time-geographic approach, as described by Pagliara *et al.* (2006), an individual's activity profile in space and through time is confined to a "prism" shaped by three constraints defined as "capacity", "authority" and "coupling". Given its focus on constraining the choice environments, the time-geographic approach has grown into a suitable framework for the choice set generation processes in shopping destination choice problems. The potential path area (PPA), which represents the projection of potential path space (PPS) on planar space, reflects how far an individual can reach given a time budget. It can be assumed that opportunities within the PPA are the destination choice set.

In the context of location choice models, it is common practice to distinguish between *primary* and *secondary* activities. Horni *et al.* (2008) follow similar classification but due to the lack of a consistent definition in the literature, they classify activities as *fixed* and *flexible*, where *flexible* activities are the activities for which location choice is applied. In this study, PPA based on TTB is estimated by using the following rule: We know the locations and the planned start and end times of the *fixed* activities and the duration of the *flexible* activity from GPS trace. In turn, this means that the TTB can be calculated as the remainder. However, some people might be able to reach stores with longer travel time (TT) than observed. For this reason, we use an adjustment factor. Due to the lack of reasonable evidence, we arbitrarily set it to 10% to allow a wider search radius. As long as the total travel time is smaller than the TTB, stores will be included in the list of feasible alternatives.

5.2.2 Generation of travel time matrices

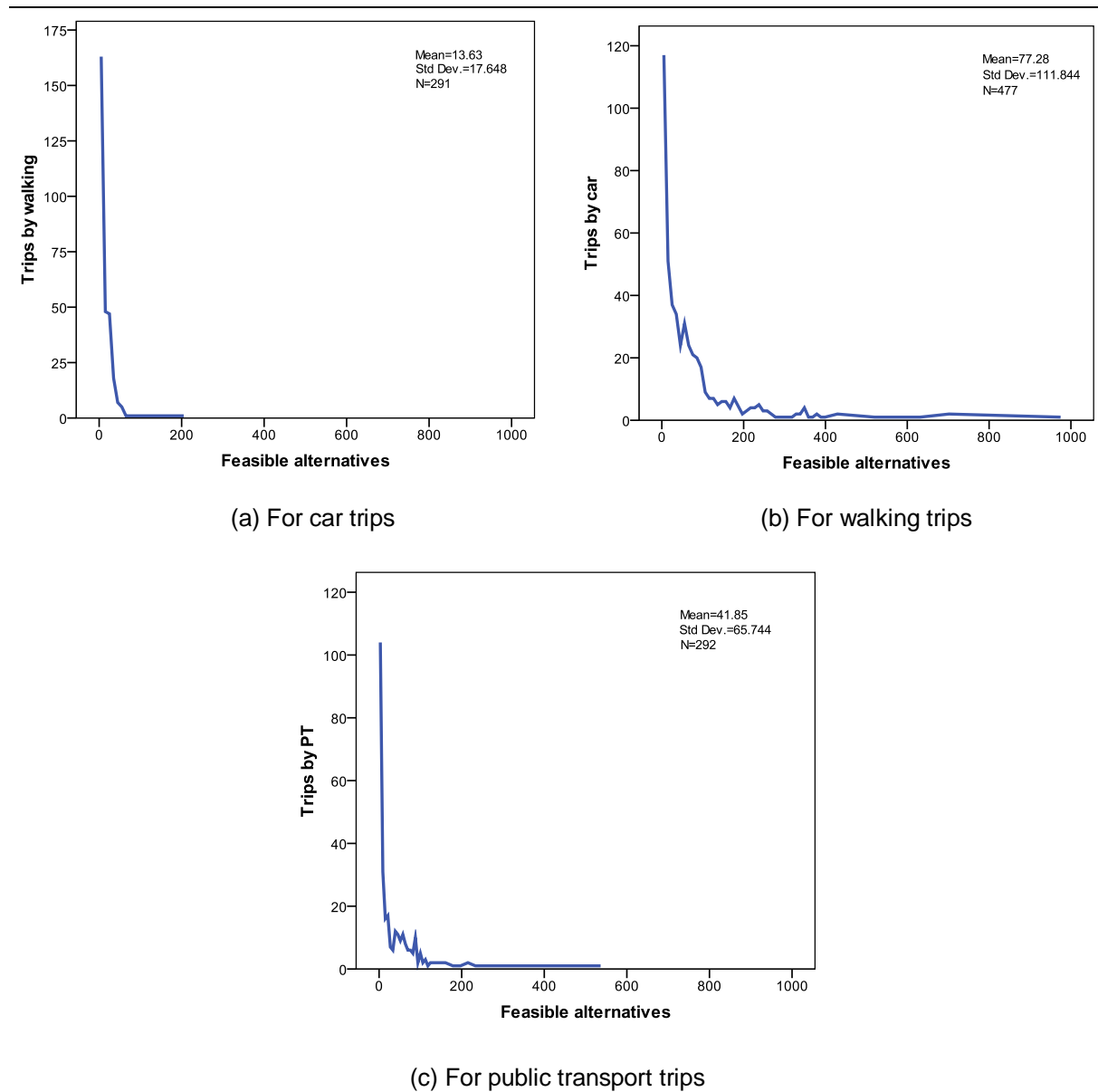
Once the three event files (origins, destinations and shopping opportunities) were generated, two origin-destination (OD) matrixes were set-up for each mode (walk, car and public trans-

port), the first containing the travel time from all origins to all shopping alternatives, and the second one with the travel time from all shopping alternatives to all destinations.

Once the two OD travel time matrixes for each mode were calculated, the travel time from the origin to all the shopping alternatives was then added to the travel time from all the shops to the destinations of the trips. Once the travel time to all shopping destinations was calculated, it was compared to the actual TTB of the individual, that is, the available time that a person has for travel. Only those stores for which the value of travel time plus the adjustment factor was lower than the TTB were included as a feasible alternative and therefore made part of the choice set for a particular trip.

In this way, using the start and end locations, speed and the TTB as spatial and temporal constraints, the concept of space-time prism could be applied to the problem of determining which locations are feasible to each and every individual pursuing a grocery shopping trip in canton Zurich. The number of feasible alternatives for each individual's grocery shopping trips is shown in Figure 4.

Figure 4 Distribution of the number of feasible alternatives for shopping trips



5.2.3 Variables used for grocery destination choice modelling

Table 3 shows the variables used in the modelling process. The variables added to the choice sets were determined based on the reviews of the previous literature on shopping behaviour and are available for our choice sets from the GPS data or other sources. Retailer name is used as a proxy for product quality and selection of goods. This variable is mentioned by

number of authors (Schuler 1979; Kahn and Schmittlein 1989; Innes *et al.* 1990; Baltas and Papastathopoulou 2003; Erath 2005). Store size was confined to categories, and is mentioned to have an impact by Erath (2005). Store hours of operation per week is mentioned by Innes *et al.* (1990), and this information was available in the existing datasets.

Two distance variables are provided. One distance variable is the detour distance variable which was defined for shopping trips inside larger tours and the another was defined for individual shopping trips, where a person went from home to the shop and back home. This variable is mentioned by Uncles (1996). Carrasco (2008) also finds the effect of this variable to be different from the detour distance, i.e. the difference between the distances to and from the stores and the shortest distance between the two locations before and after the shopping location. The distances from bus/tram stops were also calculated.

Simma *et al.* (2004) find a positive effect of the total sales floor area of a municipality on the mode selected for shopping for groceries. It was added to the files. The information regarding closeness to urban public transport (bus and tram) station stops as well as distance from home to the closest grocery shop was also obtained.

The accessibility is mentioned by Robinson and Vickerman (1976), as well as by Recker and Kostyniuk (1978). Moreover, accessibility measures were calculated for every store using the following function, which is widely used for grocery shopping accessibility indicator.

$$f(x) = \sum \exp(-d, \beta) * X \quad (2)$$

Where the equation is computed for distance d using values for β of 0.2, 1.0 and 2.0 km. The value of X represents the number of opportunities, which was set to one in all cases. The sum of all values corresponds to the accessibility measure of each store, one for each of the β values used.

Table 3 Description of variables in choice set

Variable	Description	Vallue
Choice	Chosen alternative	1: Chosen 0: Otherwise
RetID	Retailer ID – categorial *Dummy Variables	1: Coop 2: Migros 3: Denner 4: Spar 5: Pick Pay 6: Primo 7: Volg 8: Warehouses 9: Supermarkets
RetArea	Total floor area - categorial *Dummy Variables	1: below 100 m ² 2: 100 - 300 m ² 3: 300 - 1000 m ² 4: 1000 - 3000 m ² 5: over 3000 m ²
H_week	Store opening hours per week	Scalar variable
dist_H_S_car	Distance from home to shop (car)	Scalar variable
add_dist_O_S_car	Distance from origin to shop in trip chains (car)	Scalar variable
dist_H_S_walk	Distance from home to shop (walk)	Scalar variable
add_dist_O_S_walk	Distance from origin to shop in trip chains (walk)	Scalar variable
dist_H_S_PT	Distance from home to shop (Public transport)	Scalar variable
add_dist_O_S_PT	Distance from origin to shop in trip chains (Public transport)	Scalar variable
dist_BusTram_PT	Distance from bus/tram stops to stores	Scalar variable
aAlt	Store accessibility variable	Scalar variable

5.3 Random sampling of alternatives

Once the feasible shopping opportunities for each individual were determined, it was observed that the distribution of the number of feasible alternatives, due to the differences in TTB for every individual, was quite heterogeneous. For a few individuals, a very large number of the shopping opportunities were available (e.g. 95% of all possible alternatives), whereas for others only a few alternatives were viable. Figure 4(a) displays the number of feasible alternatives for the trips by car, from which it can be appreciated that for a few individuals up to 975 alternatives are accessible, but the number decreases rapidly for more people.

In the case of walking trips, Figure 4 shows the distribution of the number of feasible alternatives. For a very low number of persons a large number of alternatives are available. Only 1.3% of the individuals can reach more than 100 alternatives within the time constraints.

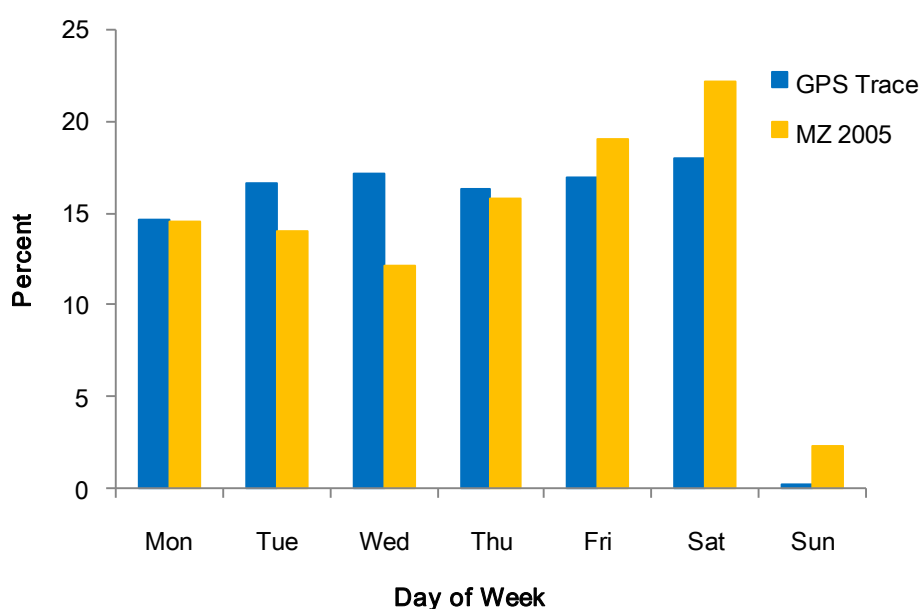
Given the high variation in the number of alternatives and the time constraints at the moment, for modelling purposes, a random sample was taken from the choice sets and the number of alternatives was fixed to 20 for car trip (the chosen destination plus 19 alternatives), 10 for walking trip (the chosen destination plus 9 alternatives) and 15 for PT (the chosen destination plus 14 alternatives). For those alternatives that had less than 20 shopping opportunities after the choice set determination process, the availability variable within the model itself was given the value of zero in the BIOGEME model file. This sampling procedure, proposed by McFadden (1977) is accepted in the literature, as the analysis of decisions can be carried out with a limited number of choices from the full set of alternatives without biasing the model parameters.

6 Comparison with MZ 2005

Since no information about the actual trips and activities of the participants is available, the MZ 2005 is used as the validation of the choice set generation process. In the following, the results of the procedure are compared to a sub-sample of the MZ 2005, which comprises the respondents living in canton Zurich. The summarized result for the canton Zurich (city of Zurich and Winterthur) and the corresponding sample in the MZ 2005 are depicted in Figure 5-8 and table 4.

The comparison of distribution of grocery shopping trips by day of the week is depicted figure 5. The highest percent share of the visits is on Saturday for both the GPS traces and MZ 2005 as expected. However, the distribution of GPS trace data is obviously flatter than the one of MZ 2005, which has an obvious low on Wednesday. On the other hand, GPS traces show us a relatively “flat” distribution across the days. In addition, approximately 2.5% of customers visited on Sunday in MZ 2005 case, whereas GPS trace records quite a low percentage of less than 1%.

Figure 5 Distribution of grocery shopping trips by day of week



Distribution of grocery shopping trips by time of day is illustrated in Figure 6. This is very similar in skew although differences can be observed. Only at one point, which is a between 10 and 10:30, MZ respondents went shopping more than twice as frequently as the participants in the GPS study. One has to keep the very strong rounding tendency of the diary respondents in mind, which is visible by the massive drop for 10:30 to 11:00.

Figure 6 Distribution of grocery shopping trips by time of day

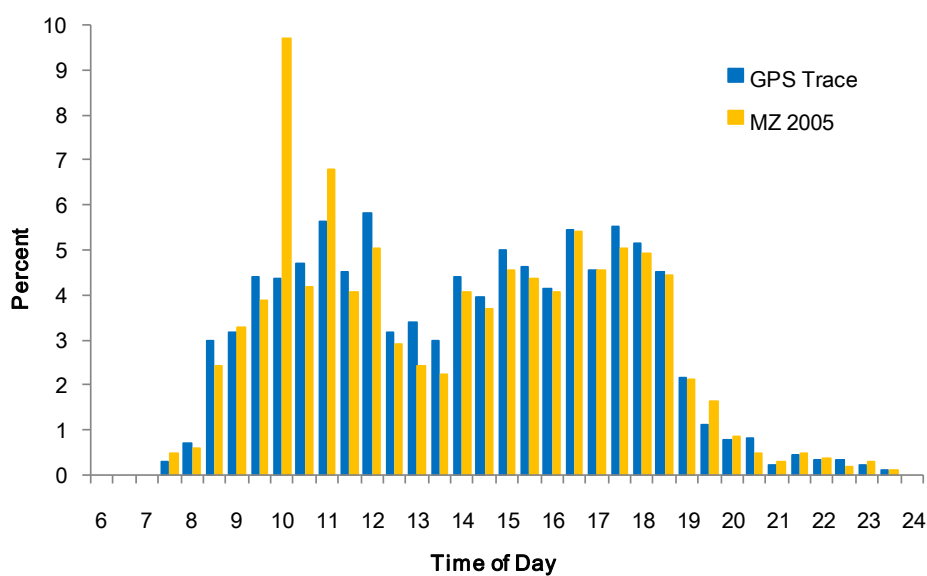
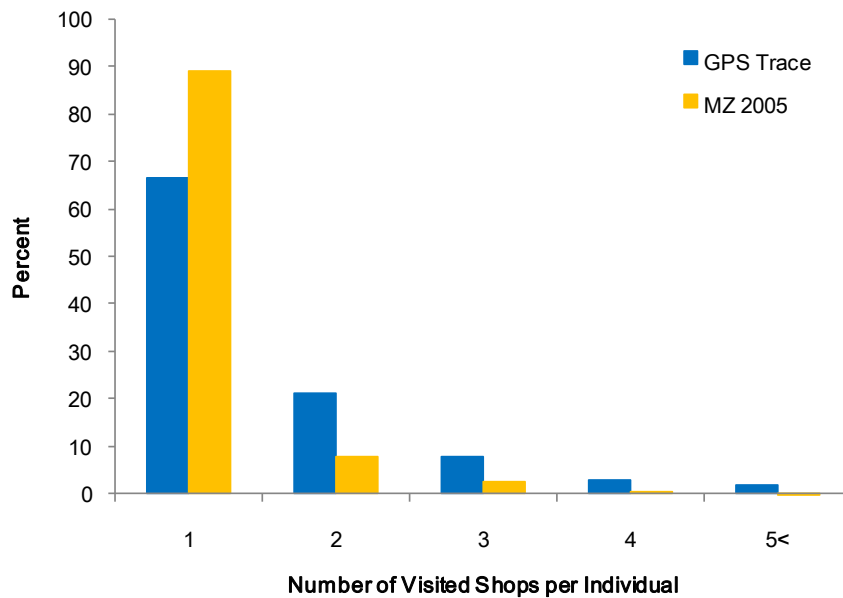


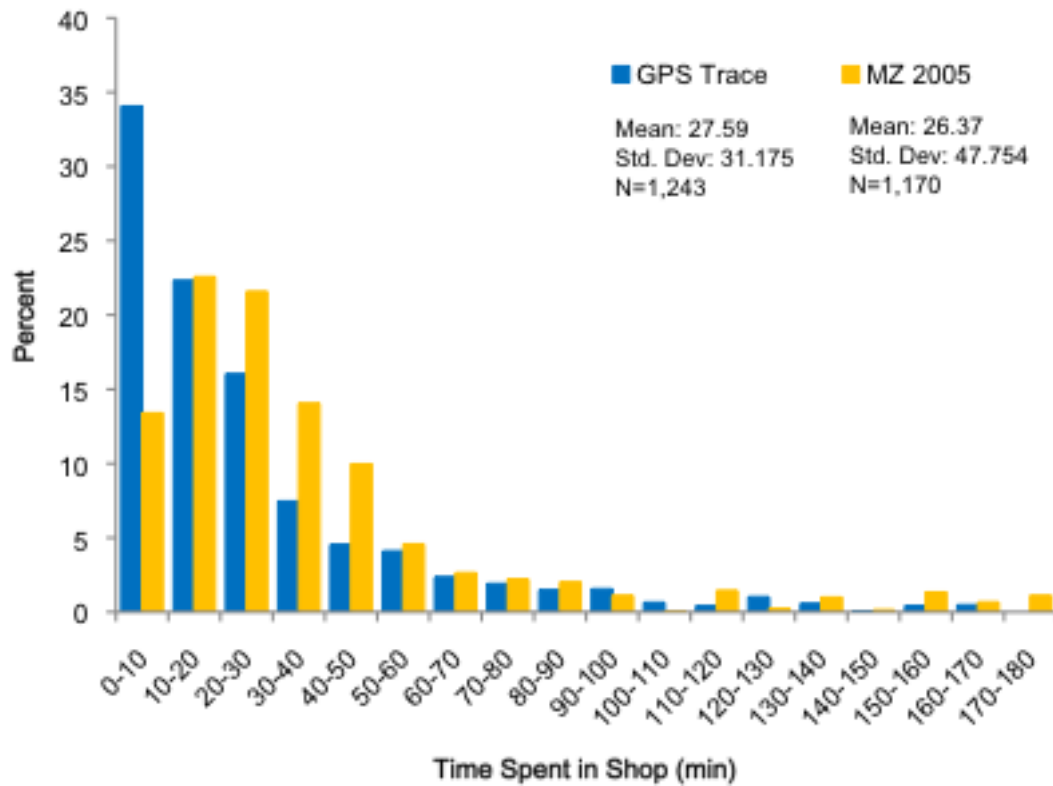
Figure 7 depicts compares the distribution of number of visited shops per individual. Since we distinguished each trip when clustering, it is possible to observe how many times one individual visits a grocery store during the survey period. For both the MZ 2005 and GPS traces, the highest share is the one time visit. Still, the higher average number of trips is noticeable. This is consistent with the general concern with travel diary data suppressing the report of short duration trips, which is borne out in the next figure.

Figure 7 Distribution of number of visited shops per individual



The distribution of time spent in shop per individual is illustrated in Figure 8. They are differently skewed; however, the mean of each distribution is almost same (GPS trace: 27.59 and MZ 2005: 25.37). The comparison shows the underreporting of short shopping trips in the MZ, which we had suggested as the reason for the lower frequency discussed above.

Figure 8 Distribution of time spent in shop per individual



The modal shares are compared in Table 4. While the car share is well reproduced, the automated process (Schüssler and Axhausen, 2009) identifies too many public transport and cycling trips, while underestimating the share of walk trip. This problem will be addressed with the integration of map matching into the overall automated processing.

Table 4 Grocery shopping trips by transport mode used

	GPS trace		MZ 2005	
	Frequency	Percent	Frequency	Percent
Car	510	41.0%	487	41.6%
Walking	314	25.3%	364	31.1%
PT (Bus and Tram)	278	22.4%	151	12.9%
Bicycle	127	10.2%	81	6.9%
Rail	14	1.1%	79	6.8%
Motorcycle	n/a	-	7	0.6%
Others	n/a	-	2	0.2%
Total	1,243	100%	1,170	100%

7 Conclusions and outlook

In this study, we generate choice sets for grocery shopping choice models from a large GPS data set by using additional data available in principle to the analyst, and readily available at the IVT. They will be primarily used to develop grocery shopping choice models. The detection of the home and w/e location seems to work well. The paper also presented a well working heuristic to identify detecting trips for grocery shopping. This approach can be applied to most of GPS data set since most GPS trace data contains the starting time, duration and frequency of activity.

Due to the lack of a direct validation data set, the data were compared with the MZ 2005, which includes day of week visited store, time of day, visited shops per individual, time spent in shop. The overall fit is surprisingly good, when acknowledging well known weaknesses in the travel diary data (rounding of clock times; suppression of minor activities). The differences in the modes are due to the algorithms used to identify them. The differences in the

day-of-week pattern is difficult to judge given the known trend of diary surveys to suppress short duration trips.

While this is a very promising start using very little additional information and overall simple heuristics, it is clear, that more effort needs to be spent on a detailed sensitivity studies and validation of the results. There might be any number of interactions, which produce false positives leading to an overly positive impression of the results. Suitable validation data will be collected at the IVT in 2010/11. It would be worthwhile to develop a heuristic, which does not use the expensive to collect geo-codes of the universe of grocery stores. The trade-off between the costs of better point-of-interest data bases and less accurate imputation results is not clear yet.

8 Acknowledgements

This work was done at IVT, ETH Zurich as part of the exchange program between ETH Zurich and Tokyo Institute of Technology (Tokyo Tech) funded by Tokyo Tech Centennial Academic Research Fund. The first author would like to express his extraordinary gratitude to respected Prof. Dr. Kay W. Axhausen (ETH) and respected Assoc. Prof. Dr. Shinya Hanaoka (Tokyo Tech) for giving me such a wonderful opportunity and extremely valuable discussion on this research. His deep gratitude goes to every staff in IVT for making his daily life extremely joyful and fruitful. Particularly, his office mates, Mr. Christof Zöllig and Mr. Michael Löchl for their extraordinary kind contribution for facilitating Swiss life for a stranger from Japan.

9 References

- Ashbrook, D. and Starner, T. (2003) Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing*, **7** (5) 275-286, 2003.
- Axhausen, K.W., S. Schönfelder, J. Wolf, M. Oliveira and U. Samaga (2003) 80 weeks of GPS-traces: Approaches to enriching the trip information, *Transportation Research Record*, **1870**, 46 -54.
- Baltas, G. and P. Papastathopoulou (2003) Shopper characteristics, product and store choice criteria: A survey in the Greek grocery sector, *International Journal of Retail & Distribution Management*, **31** (10) 498-507.
- Carrasco, N. (2008) Deciding where to shop: Disaggregate random utility destination choice modeling of grocery shopping in canton Zurich, Master Thesis, IVT, ETH Zürich, Zürich.
- Casas, J. and C. Arce (1999) Trip reporting in household travel diaries: A comparison to GPS-collected data, paper presented at the *78th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 1999.
- de Jong, R. and W. Mensonides (2003) Wearable GPS device as a data collection method for travel research, *Working Paper*, **ITS-WP-03-02**, Institute of Transport Studies, University of Sydney, Sydney.
- Draijer, G., N. Kalfs and J. Perdok (2000) Global Positioning System as data collection method for travel research, *Transportation Research Record*, **1719**, 147–153.
- Erath, A. (2005) Zeitkosten im Einkaufsverkehr, Diplomarbeit, IVT, ETH Zurich, Zurich.
- Hagerstrand, T. (1970) What about people in regional science?, *Papers of the Regional Science Association*, **24** (1) 7-21.
- He, S.Y. (2006) Modeling destination choice for shopping using a GIS-based spatio-temporal framework: An investigation of choice set generation and scale effects, Master Thesis, School of Geography and Earth Sciences, McMaster University, Hamilton, Ontario.

- Horni, A., Scott, D.M., Balmer, M. and Axhausen, K.W. (2008) Location choice modeling for shopping and leisure activities with MATSim: Combining micro-simulation and time geography, *Arbeitsberichte Verkehrs- und Raumplanung*, **527**, IVT, ETH Zurich, Zurich.
- Innes, D., M. Ircha and D. Badoe (1990) Factors affecting automobile shopping trip destinations, *Journal of Urban Planning and Development*, **116** (3) 126-136.
- Kahn, B.E. and D.C. Schmittlein (1989) Shopping trip behavior: An empirical investigation, *Marketing Letters*, **1** (1) 55-69.
- Liao, L., D. Fox and H. Kautz (2007) Extracting places and activities from gps traces using hierarchical conditional random fields, *The International Journal of Robotics Research*, **26** (1) 119-134.
- McFadden, D. (1977) Modelling the choice of residential location, *Cowles Foundation Discussion Paper*, **477**, Cowles Foundation for Research in Economics, Yale University, New Haven.
- Pagliara, F., E. Cascetta and K.W. Axhausen (2006) Dominance attributes for alternatives' perception in choice set formation: an application in spatial choices, *Arbeitsberichte Verkehrs- und Raumplanung*, **371**, IVT, ETH Zürich, Zürich.
- Pearson, D. (2001) Global Positioning System (GPS) and travel surveys: Results from the 1997 Austin household survey, paper presented at *8th Conference on the Application of Transportation Planning Methods*, Corpus Christi, April 2001.
- Recker, W.W. and L. Kostyniuk (1978) Factors influencing destination choice for the urban grocery shopping trip, *Transportation*, **7** (1) 19-33.
- Robinson R. and R. Vickerman (1976) The demand for shopping travel: A theoretical and empirical study, *Applied Economics*, **8** (4) 267-281.
- Schuler, H. (1979) A disaggregate store-choice model of spatial decision-making, *The Professional Geographer*, **31** (2) 146-156.
- Schüssler, N. and K.W. Axhausen (2009) Processing GPS raw data without additional information, *Transportation Research Record*.
- Simma, A., P Cattaneo, M. Baumeler and K.W. Axhausen (2004) Factors influencing the individual shopping behaviour: The case of Switzerland, *Arbeitsberichte Verkehrs und Raumplanung*, **247**, IVT, ETH Zurich, Zurich.

- Stopher, P. R., Q. Jiang and C. FitzGerald (2005) Processing GPS data from travel surveys, paper presented at *2nd International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications*, Toronto, June 2005.
- Swiss Federal Statistical Office (2006) Ergebnisse des Mikrozensus 2005 zum Verkehrsverhalten, Swiss Federal Statistical Office, Neuchatel.
- Uncles, M. (1996) Classifying shoppers by their shopping-trip behaviour: A polythetic divisive method, *Marketing Intelligence & Planning*, **14** (1) 35-44.
- Wagner, D. P. (1997) Lexington area travel data collection test: GPS for personal travel surveys, Final Report, Office of Highway Policy Information and Office of Technology Applications, Federal Highway Administration, Battelle Transport Division, Columbus.
- Wolf, J. (2006) Applications of new technologies in travel surveys, in P.R. Stopher and C. C. Stecher (eds.) *Travel Survey Methods - Quality and Future Directions*, 531–544, Elsevier, Oxford.
- Wolf, J., S. Hallmark, M. Oliveira, R. Guensler and W. Sarasua (1999) Accuracy issues with route choice data collection by using Global Positioning System, *Transportation Research Record*, **1660**, 66–74.
- Yalamanchili, L., R. M. Pendyala, N. Prabakaran and P. Chakravarty (1999) Analysis of Global Positioning System-based data collection methods for capturing multistop trip chaining behavior, *Transportation Research Record*, **1660**, 58–65.