Choice Set Generation and Model Identification for Route Choice Using GPS-Data from Smart Phones

Authors:

Eileen Mandir[1], Juliane Pillat[1], Markus Friedrich[1], Christian Schiller[2]

Affiliations:

1) Universitaet Stuttgart
2) German Aerospace Center

Keywords:

GPS, trajectories, route choice, choice set

INTRODUCTION

To advance the state-of-the-art in traffic control systems and reliable driver information, not only methods for monitoring the current traffic state but also route choice models to forecast driver behavior are crucial. The development and calibration of good route choice models is heavily depending on profound data. Floating car data has been considered a promising data source for monitoring time-space trajectories of single travelers for the past years. The deficits of low market penetration of equipped vehicles and high communication costs could soon be overcome by the success story of smart phones with internet flat rates and the boom of downloadable applications. This paper shows the capabilities of floating car data, if the same drivers can be monitored over a long period of time [1].

The research project *wiki*, funded by the German Federal Ministry of Economics and Technology, aims at identifying the impact of driver information on route choice in metropolitan areas on the example of the Munich region. The research project which is scheduled to run until 2011 intends to prove that driver information can contribute to the reduction of total time spent as well as fuel consumption. Therefore a large survey was conducted in the Munich area, monitoring the behavior of 300 commuters over a period of 8 weeks. The survey data is used as an empirical basis for identifying, estimating and calibrating a profound route choice model. The data source not only contains information on the impact of driver information on route choice but also on the size of the choice set which is critical for model estimation [2].

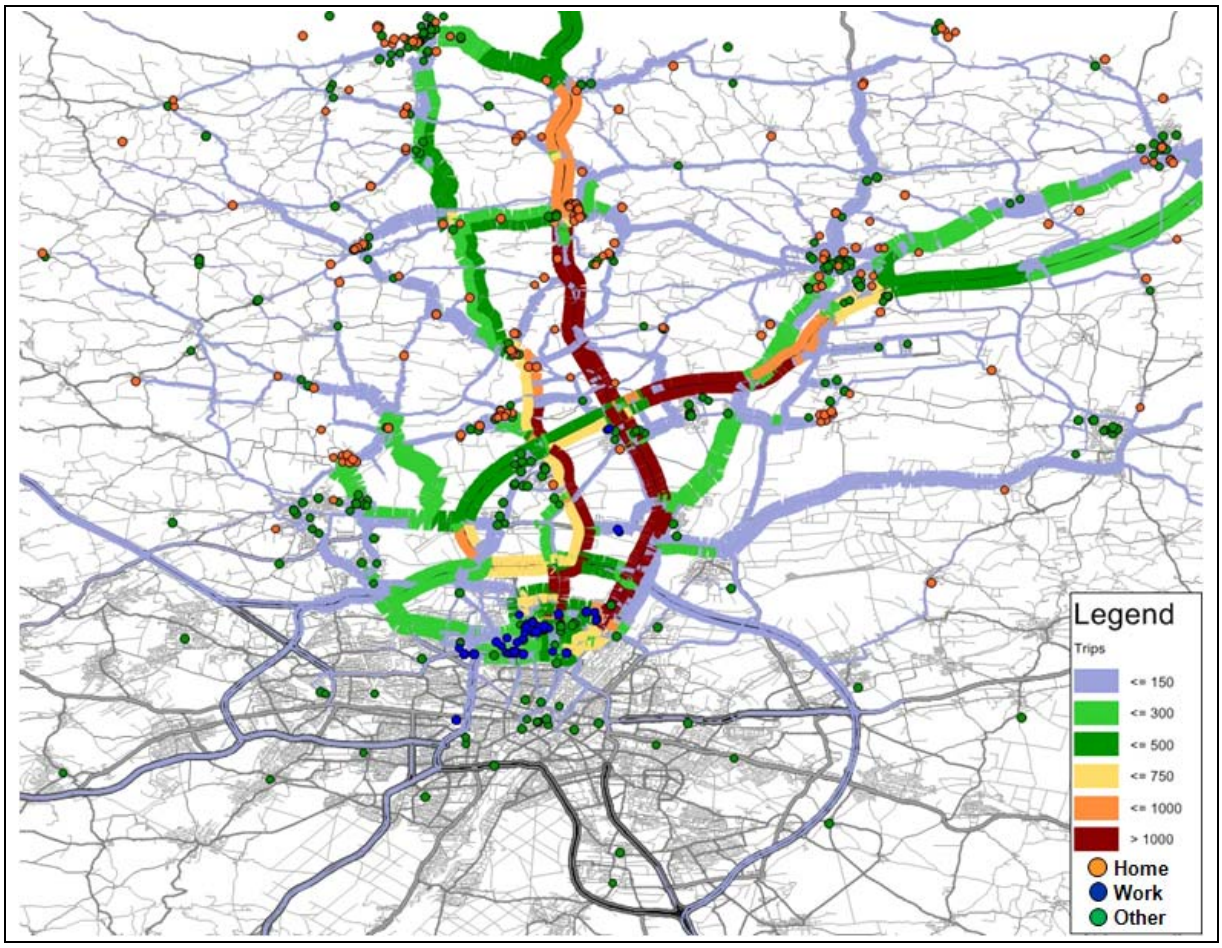


Figure 1: Overview of survey area with GPS trajectories and activity locations

SMART PHONE BASED DATA COLLECTION

Data collection platform

During the survey 300 participants were equipped with a smart phone and GPS sensor. A software application was developed for receiving and storing data from a GPS sensor and transmitting it to a server during a trip. The application started automatically when the smart phone was switched on. The GPS sensor calculated a position every second which's latitudinal and longitudinal coordinates, date/time and speed were stored on the smart phone via Bluetooth connection. Every five minutes a GPS data package was transmitted to a server via the GSM mobile phone network. Because each smart phone's SIM card was registered on the server, GPS data can be analyzed person specific. Sending data through the mobile phone network rather than storing it locally has the benefit that future participants theoretically could use their own smart phone, register online and download the application themselves. On the other hand sending data to a server can result in data loss. This happens if the smart phone is switched off at the destination before the current data package is sent or if the GSM network connection fails in between a trip.

The GPS files stored on the server include data recorded between switching on and switching off the smart phone. This may include several trips. In a second step the trips are map matched. The problem in map matching GPS data on a digital map is that mostly at the trip ends the location cannot be found because minor roads are likely to be missing in the network model [3]. Due to this some further processing is needed to advance from link trajectories to actual trips. Table 1 gives an overview of the collected data volume.

Data Volume		
	Total over 300 participants	Per person
Total time of detection	8,850 hours	29.5 hours
Number of detected trajectories	20,000	66

Table 1: Data volume from GPS trajectories

Identification of trip ends from GPS trajectories

For identifying trip ends within a map-matched GPS track [4] two major problems need to be addressed:

1. Gaps in track due to data loss
2. Sections in track with speed equal to zero

Gaps are identified by a jump in the time stamp or position of subsequent data points (time and space gap). They result from data loss due to unavailable GSM network during a trip or from a trip done while the smart phone was switched off. For solving these problems, rules for time and space gaps are defined and validated for the survey data.

Figure 2 shows rules for identifying a trip end in the time-space-plane. If a trip includes a time gap larger than 5 min ($t_{critical}$) a trip end is detected. The critical time threshold is determined by matching the GPS data with traffic states from road-side detectors. Time gaps of less than 5 minutes are mostly due to stopping at traffic signals or other sorts of congestion.

The critical space gap ($x_{critical}$) is not only dependent on the distance travelled but also the time spent. If a trip contains a space gap of 0.5 kilometers for duration of 10 minutes, no new trip end is detected. Large space gaps in relatively short time can for example result from travelling through a tunnel. If the time spent during the space gap is small, there is a high probability that no activity took place but rather the entire time was spent to travel to the new location. The larger the space gap, the larger the allowed time gap. Only if the time gap increases beyond t_{max} a new trip end is detected. t_{max} is thereby

defined as the maximum possible direct speed between the two locations of the space gap depending on the speed limit of the predominant road class of the shortest path connection between the two locations.

Sections in tracks with speed equal to zero occur from error in the positioning of the GPS sensor or from a stop of the vehicle while the smart phone remains switched on. The stop can either be caused by congestion or by an activity, typically pick up and drop off or food shopping. This problem is solved by analyzing current speed, detour factor and location of points of interest such as airport, train stations etc. Table 2 displays the number of trips detected from the map-matched GPS trajectories.

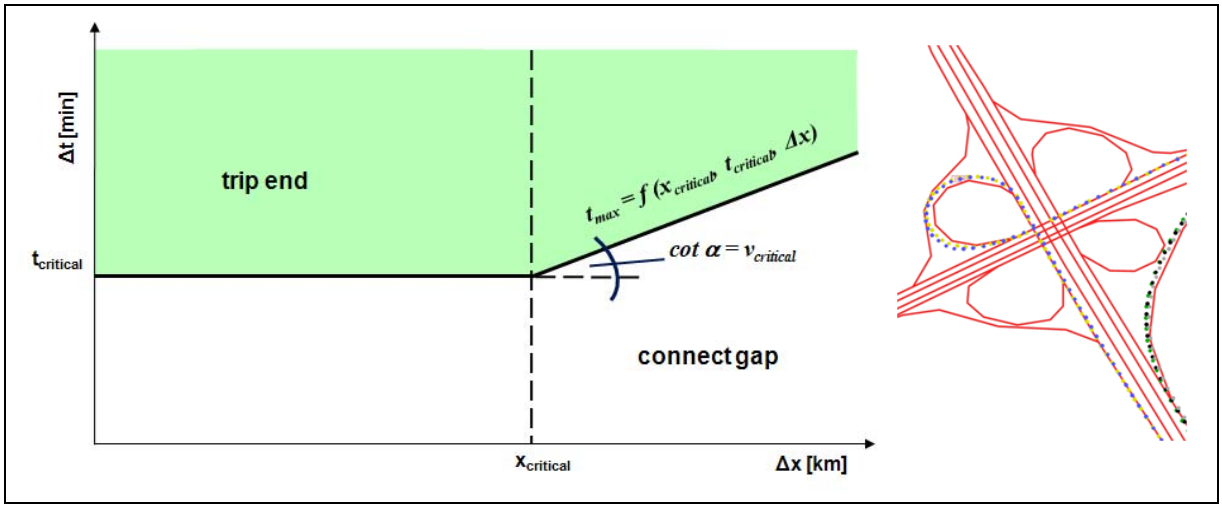


Figure 2: Identifying trip ends with time-space-gaps

Data Volume		
	Total over 300 participants	Per person
Total number of GPS trajectories	20,000	66
Number of identified trips	24,000	80

Table 2: Number of trips identified

Identification of activity locations at trips ends

The identified trips from the previous step are still missing one crucial thing for being used in further applications, such as generation of choice sets for discrete choice model estimations. For most trips data is missing at the beginning and end. This results from the smart phone being switched on a little late at the origin or being switched off at the destination before the current data package is sent to the server or from already mentioned problems in map-matching. Therefore, the exact activity locations such as Home and Work need to be determined in an extra step. Three criteria are analyzed for identifying activity locations:

- position matching
- time matching
- point of interest matching

First, all start and end points of trips are clustered in groups within a 2.5 kilometer perimeter. The centroids of these groups are set as the activity locations, see Figure 3. In a second step all start and end points assigned to an activity location are checked for their time stamp. According to the predominant start and end time the activity purpose is defined. For example, an activity with predominant starting time at 8 am and predominant end time at 5 pm is set as Work. For all other activities, apart from Home and

Work, the location is furthermore checked for having a point of interest close by, for example airport etc. Again according to the predominant time stamp different activity purposes become more or less likely, e.g. food shopping is more likely after work than before work. This processing step allocates distinct start and end locations to trips. These locations are the relevant origin-destination pairs that need to be considered in choice set generation. Table 3 displays the number of trips for which an OD-pair could be determined.

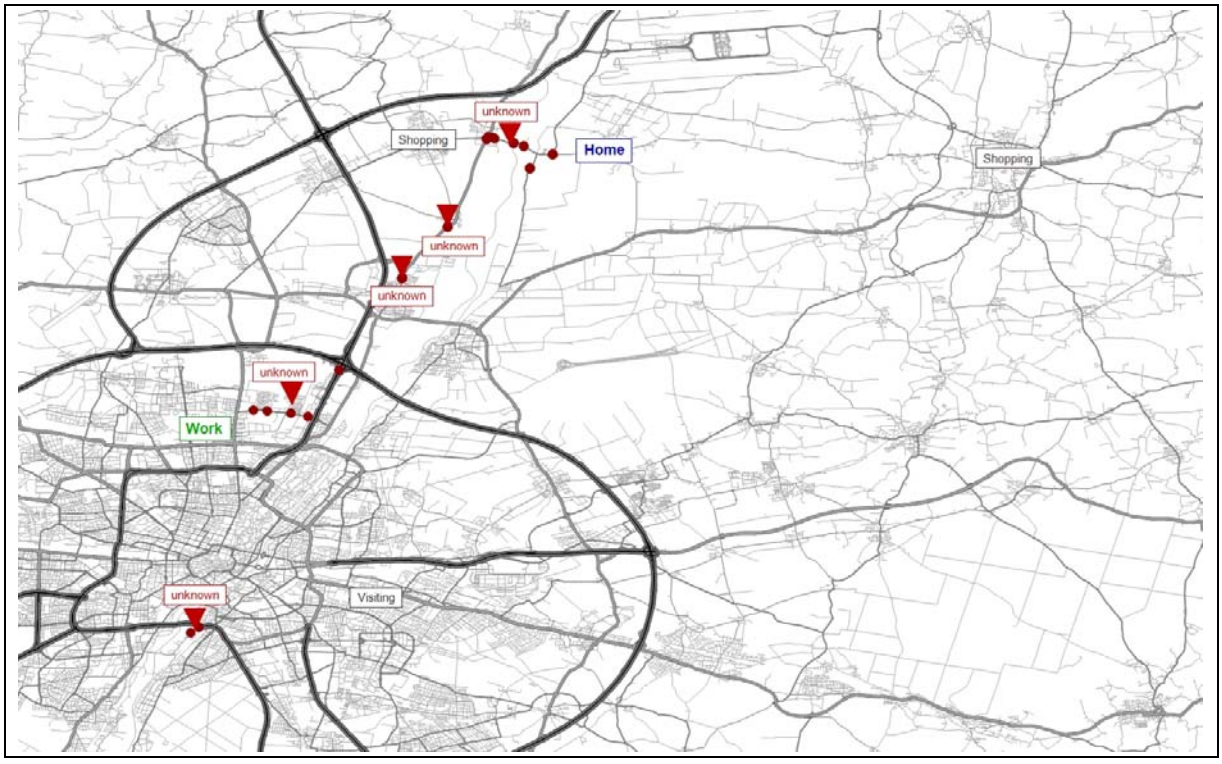


Figure 3: Clustered activity locations

Data Volume		
	Total over 300 participants	Per person
Total number of trips	24,000	80
Number of trips between identified activity locations	18,300	61

Table 3: Number of activity locations identified

CHOICE SET GENERATION

Due to the duration of the survey numerous trips and routes were logged for many OD-pairs and especially for home to work and vice versa. Figure 4 shows the chosen routes from home to work for one participant. Although the participant experienced very different travel times on his way to work the number of chosen routes are comparatively low. Figure 5 shows the level of service experienced by the participant on trips during the survey period. The trips from home to work and vice versa are encircled.

Additionally, all participants were asked to draw the routes they knew from their home to work on a map in an interview. Table 4 shows a comparison of the number of chosen routes (revealed

preference) to the number of known routes (stated preference). Although both values are similarly low, a general tendency to more known than chosen routes can be observed.

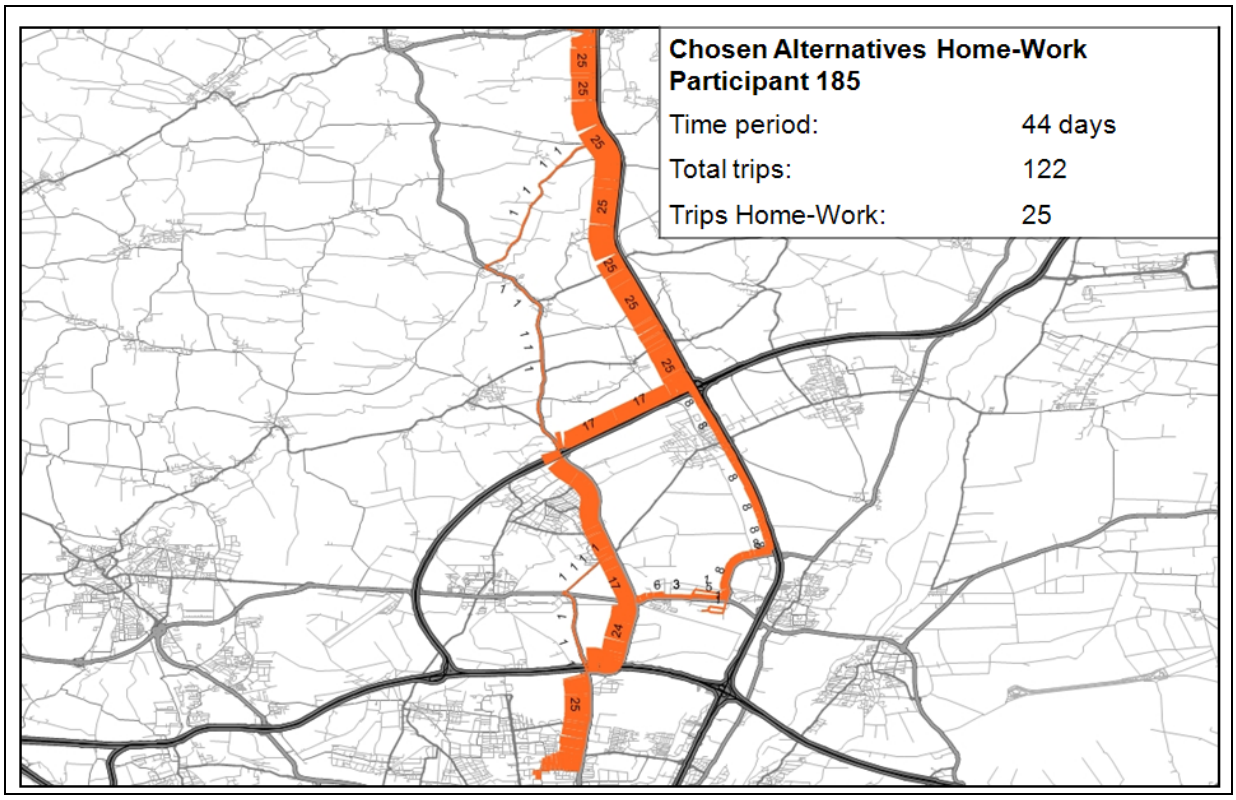
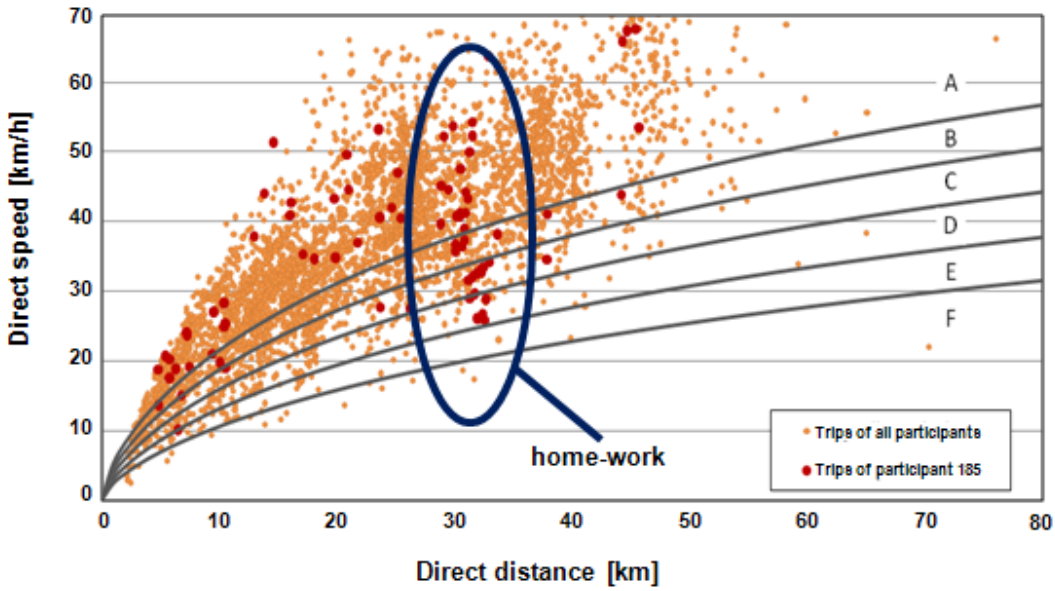


Figure 4: Chosen alternative routes from home to work for one participant

This empirically surveyed and verified choice set as well as the thereby used routes allow the conclusion that, especially for compulsory activities like Work, only a marginal number of alternative routes are perceived and used. This may conflict with today's assignment procedures which generally use a choice set generator to produce a large number of potential alternative routes. Dugge [6] shows that for a similar metropolitan network approximately 30 to 35 routes per OD-pair are found. It is unlikely that a traveler is actually aware of such a high number of alternatives for compulsory activities. Further analysis over all participants on common OD-pairs will show if the total number of routes per OD-pair is realistic for aggregated models.

One of the main reasons for the large number of generated alternative routes is that choice set generators do not distinguish between routes on the main and minor road network. Thus, merely generalized costs are examined without accounting for the traveler's subjective perception of the network hierarchy. In order to obtain realistic traffic flows, optimizing choice sets for traffic assignment is of high importance.



Level of Service	Description
A	Excellent Quality
B	Good Quality
C	Satisfactory Quality
D	Sufficient Quality
E	Insufficient Quality
F	Inadequate Quality

Figure 5: Level of Service, German guideline directive for integrated network design (RIN [5])

Routes from home to work		
	Total over 300 participants	Per person
Number of chosen routes (GPS)	670	2.2
Number of known routes (interview)	880	2.9

Table 4: Comparison of number of chosen and known routes

MODEL IDENTIFICATION

To calculate the route split, normally more or less modified discrete choice models, e.g. C-Logit-Model from Cascetta [7], are used. These models calculate the expected value of traffic flow on the routes. The basic approach to calculate the probability of a route is:

$$P_r = \frac{f^C \cdot \exp\left(-\beta \cdot G_r\right)}{\sum_{r'} f^C \cdot \exp\left(-\beta \cdot G_{r'}\right)}$$

β	Parameter
f^C	Cascetta coefficient
G	Generalized cost
P	Probability
r	Route

Figure 6 shows the choice probabilities of two alternative routes in a classical Logit model. On the abscissa the generalized costs of route 1 are displayed from 0 to 100, contrarily the generalized costs of route 2 are displayed from 100 to 0. The disadvantage of this route split rule is that even marginal deviations of the routes generalized costs result in major shifts in route split. Yet, in reality marginal deviations between alternatives have only minor impact on traffic flow on routes. This is an important fact because the surveyed route choice behavior shows that travelers change their route choice only if the deviation of generalized costs goes beyond a certain threshold.

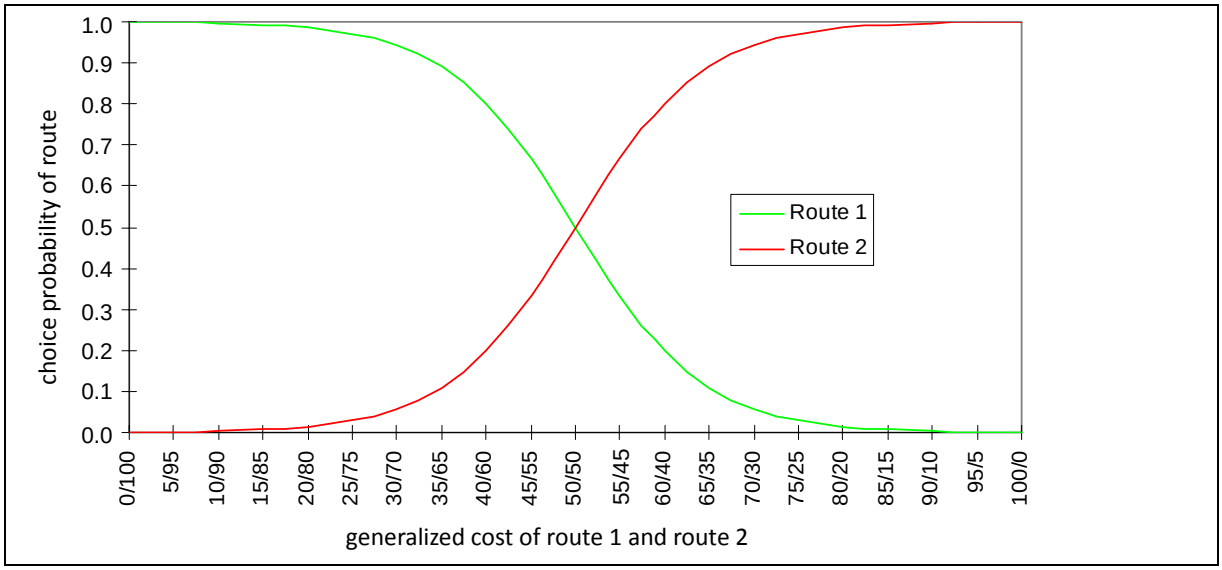


Figure 6: Logit model with β=0.07

A first step to address this problem is to use a modified route split rule which has a relatively low elasticity for very small deviations of generalized costs. In respect to Figure 6 this results in a nearly horizontal curve progression around route costs of 50/50. Thus, the choice probability stays almost constant for small cost deviations. Figure 7 shows such a modified Logit model which uses a transformation of the original generalized costs.

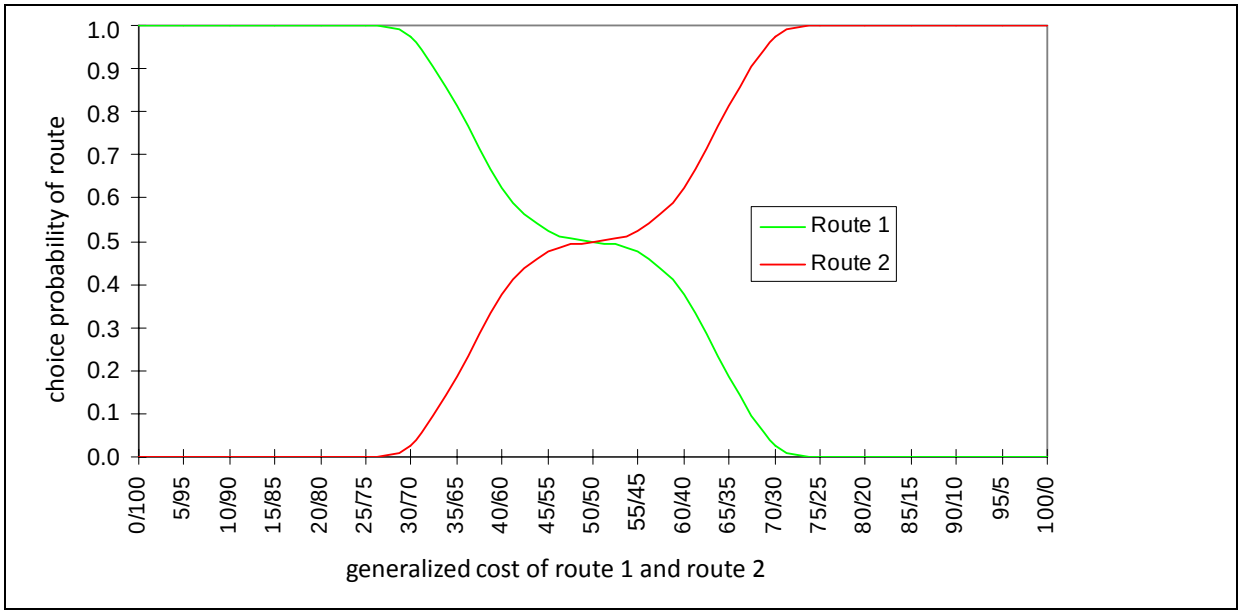


Figure 7: Modified Logit model with β=2 and γ=2

A first approach for a modified Logit was shown by Gobiet [8]:

$$P_r = \frac{\exp\left(-\beta \cdot \left(\frac{G_r}{G_r^1} - 1\right)^{\gamma}\right)}{\sum_{r'} \exp\left(-\beta \cdot \left(\frac{G_{r'}}{G_{r'}^1} - 1\right)^{\gamma}\right)}$$

β, γ	Parameter
G	Generalized cost
G^l	Generalized cost of route with lowest generalized cost
P	Probability
r	Route

The mentioned threshold of cost deviation is included by scaling each route's generalized cost to the lowest generalized cost of the respective route in the choice set. This approach is a starting point for ongoing research within and beyond the project at hand. However, first analyses point out that this approach significantly improves the results of choice probabilities compared to classical approaches.

The project *wiki* has a key advantage in having a profound revealed preference data base of GPS routes for model identification. Travel times, traffic information and other route specific attributes are available for a large number of different route types so that choice rules can be identified on manifold variable values. Typically, stated preference data from driver interviews is used for model identification. Yet, besides the often lacking ability of the questioned participants to picture the actual situation, only few variable values are included in the interview in order to keep the number of choice situations and therefore the stress for the participant as low as possible. Furthermore, GPS data is very beneficial for calibrating the estimated model parameters by comparing route choice in traffic assignment to the actual paths of the travelers.

REFERENCES

1. Axhausen, K.W., S. Schönfelder, J. Wolf, M. Oliveira and U. Samaga, 80 weeks of GPS traces: approaches to enriching the trip information, *Transportation Research Record 1870*, 46–54, 2003.
2. Prato, C. G. and S. Bekhor, Modeling route choice behavior: How relevant is the composition of choice set?, *Transportation Research Record*, 64–73, 2007.
3. Bierlaire, M. and E. Frejinger, Route choice modeling with network-free data, *Transportation Research Part C 16*, 187–198, 2008.
4. Schüssler, N. and K.W. Axhausen, Processing Raw Data from Global Positioning Systems Without Additional Information, *Transportation Research Record*, 28-36, 2009
5. FGSV – Forschungsgesellschaft für Straßen- und Verkehrswesen: *Richtlinie für die integrierte Netzgestaltung (RIN)*, Köln, 2009.
6. Dugge, B., Ein simultanes Erzeugungs-, Verteilungs-, Aufteilungs- und Routenwahlmodell, *Dissertation, Lehrstuhl für Theorie der Verkehrsplanung der TU Dresden*, Dresden, 2006.
7. Cascetta, E., A. Nuzzolo, F. Russo and A. Vitetta, A modified Logit route choice model overcoming path overlapping problems specification and some calibration results for interurban networks, *1STTT conference*, Lyon, 1996.
8. Gobiet, W., Die Verkehrsprognose für Straßennetze, *Dissertation, Technische Hochschule Graz, Fakultät Bauingenieurwesen und Architektur*, Graz, 1969.