Preliminary results: Route choice analysis from multi-day GPS data

Lara Montini

Kay W. Axhausen

Transport and Spatial Planning

May 2015

STRC | **15th Swiss Transport Research Conference**

Monte Verità / Ascona, April 15 – 17, 2015

Transport and Spatial Planning

Preliminary results: Route choice analysis from multi-day GPS data

Lara Montini
IVT
ETH Zürich
CH-8093 Zürich
phone: +41-44-633 30 88
fax: +41-44-633 10 57
lara.montini@ivt.baug.ethz.ch

Kay W. Axhausen
IVT
ETH Zürich
CH-8093 Zürich
phone: +41-44-633 39 43
fax: +41-44-633 10 57
axhausen@ivt.baug.ethz.ch

May 2015

Keywords

route choice model, bicycle, car, pedestrian, GPS travel diaries, dedicated device

1 Introduction

In recent years recording GPS data has become more common. Location data is collected as part of travel surveys but also as part of different kinds of smartphone applications, such as journey planners, health or running apps and self-tracking applications. Highly accurate, second-by-second GPS data has the main advantage that detailed routes can be observed. The goal of this paper is to describe and model route choice observed in a multi-day multi-modal GPS study of the greater Zurich area.

The paper is structured as follows. First, the data requirements and the route data set are described. In Section 3 the used methods for map-matching, choice set generation and modelling are introduced. Followed by the results for car, bicycle and pedestrian route choice models. The paper concludes with summary remarks and an outlook on future work.

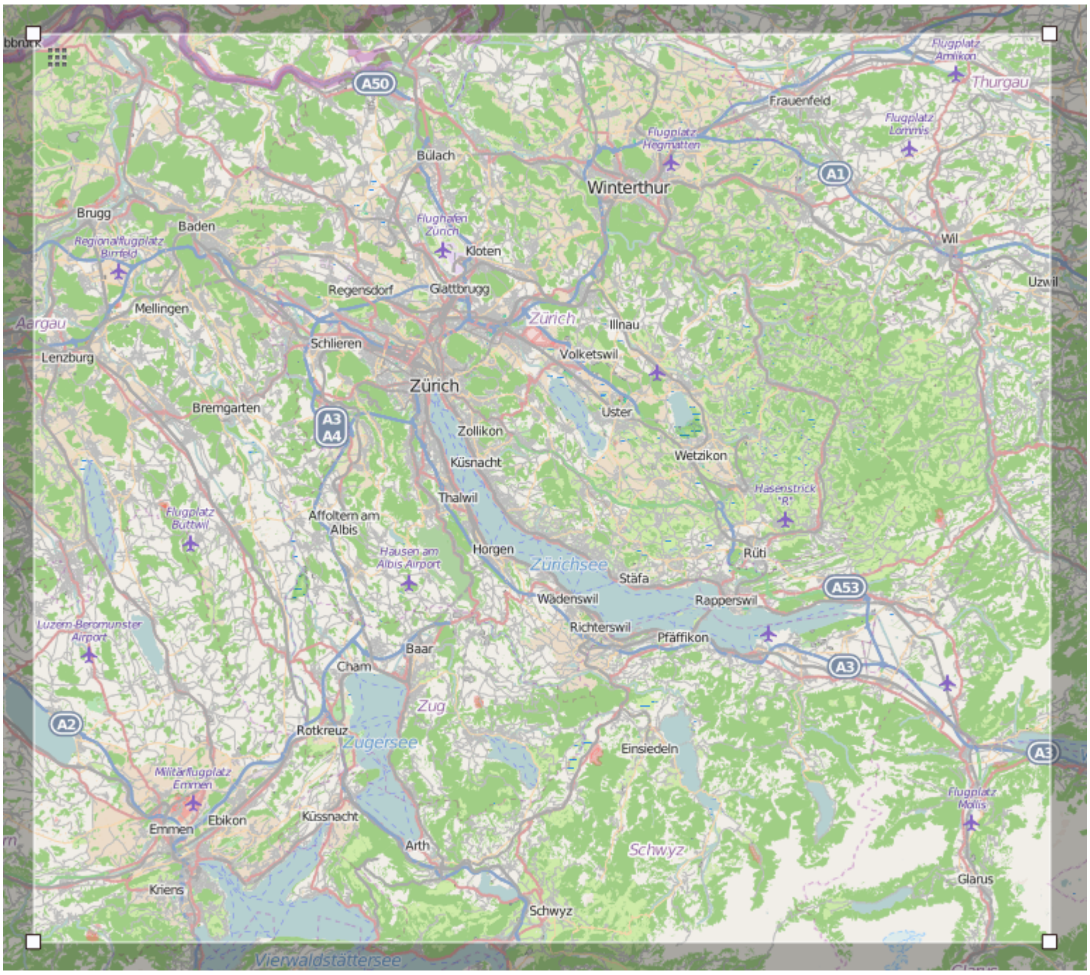
2 Data

Following the network data, the elevation model used and the route data extracted from GPS observations are introduced.

2.1 Network data

Route choice in this preliminary analysis is restricted to the area around Zurich depicted in Figure 1, it was chosen such that most everyday travel of the survey participants was included. All map data was extracted from OpenStreetMap (2015). The network for cyclists and pedestrian includes all links except motorway and trunk links, this results in approximately 3 million links. Figure 2(b) shows an extract around the city of Zurich, where cycling routes are highlighted in green. For the car network on the other hand, pedestrian and cycling only links were excluded resulting in 2.4 million links as depicted with road types in Figure 2(a). Using the map data a network in Matsim format MATSim (2015) is created, mainly based on the highway attribute and taking into account the oneway tag, as the network is directed. The conversion code can be found as part of the POSition DAta Processing project on sourceforge (POSDAP, 2012).

Figure 1: Area around Zurich included in analysis



Source: www.openstreetmap.org

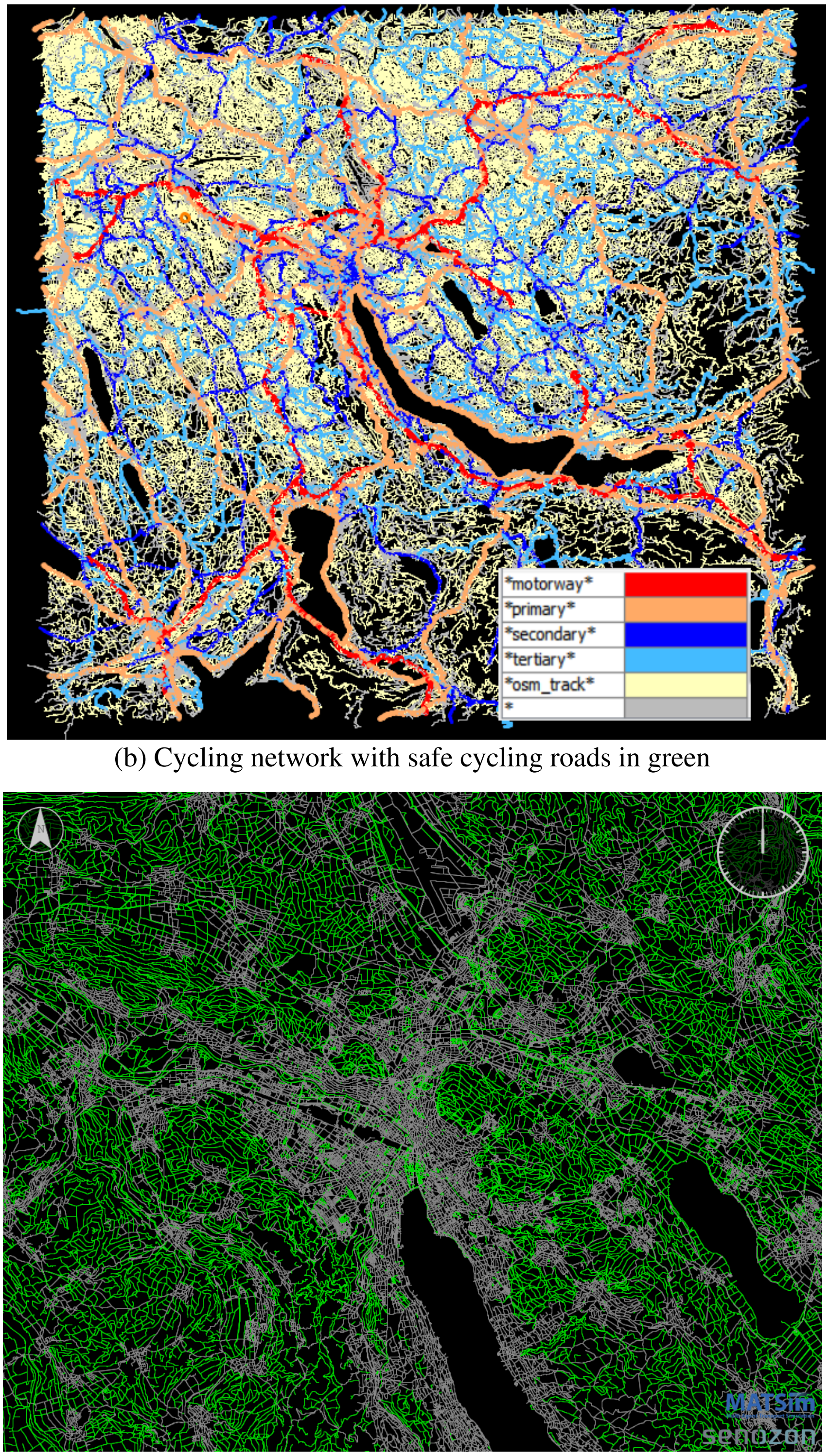
2.2 Elevation model and measures

Elevations for the canton of Zurich were released very recently Open Source under the GIS-ZH licence, the digital terrain model is available with a resolution of 0.5 meters (Office for Spatial Development of the Canton of Zurich, 2015) and is therefore used whenever possible. Outside the canton the digital elevation model with a resolution of 25 meters by swisstopo is used (Federal Office of Topography swisstopo, 2012). Each node of the network is assigned the elevation of the nearest measurement point.

The elevation measures, that is maximum and average rise as well as maximum and average fall are then calculated per route. For every link longer than 20 meters of a route the slope is calculated directly, if a link is too short it is joined with the next links until the sum of link lengths is greater than 20 meters, the slope is then calculated for the joined segment. The

Figure 2: Networks generated from OSM data for car and bicycle

(a) Car network with road types of OSM



(b) Cycling network with safe cycling roads in green

Source: www.openstreetmap.org, Visualisation: Senozon Via

Table 1: Kilometers driven by car per road type

Road Type	Driven [km]
Motorway	5998
Trunk	484
Primary	4230
Secondary	4325
Tertiary	3252
Residential	1429
Track	214
Other	488
Total	20420

average rise is then calculated as the average of all positive segment slopes, the average fall is the absolute of the average of all negative segment slopes. Accordingly, the maximum fall is the absolute value of the most negative slope.

2.3 Routes

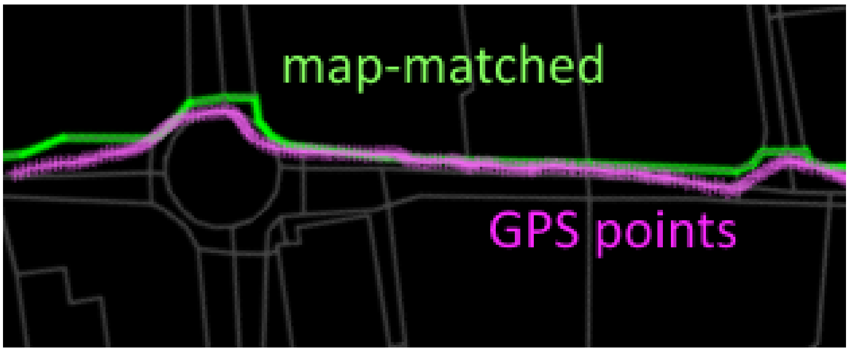
Routes are extracted from a data set collected in and around Zurich in 2012/13 comprising of approximately 1 week of data for 150 participants (Montini et al., 2013). Data was collected with dedicated GPS trackers and participants corrected the processed travel diaries using a web-based prompted recall tool. They could change times, travel mode and trip purpose. All tracks were double checked by student assistants after the survey period. Based on these corrected diaries 7233 stages that are part of 5128 trips are observed.

Within the defined area, 2250 car stages remain for modelling. Table 1 shows that most kilometres are driven on the motorway, followed by primary, secondary and tertiary roads.

In total, 410 bicycle stages are used for modeling. From the original data 82 stages are filtered as being too short (less than 500 meters), being round-trips with same start and end location or being round-trip suspects (chosen distance > 2.5 * shortest distance). Filtering round-trips is especially important as these caused positive distance parameters.

To model pedestrian route choice, 985 stages are available. From the original data round-trips and trips with an average speed higher than 5 m/s are excluded.

Figure 3: Map-matching walk, GPS points on roundabout weighting ensures that pedestrian route is chosen



3 Method

3.1 Map-matching

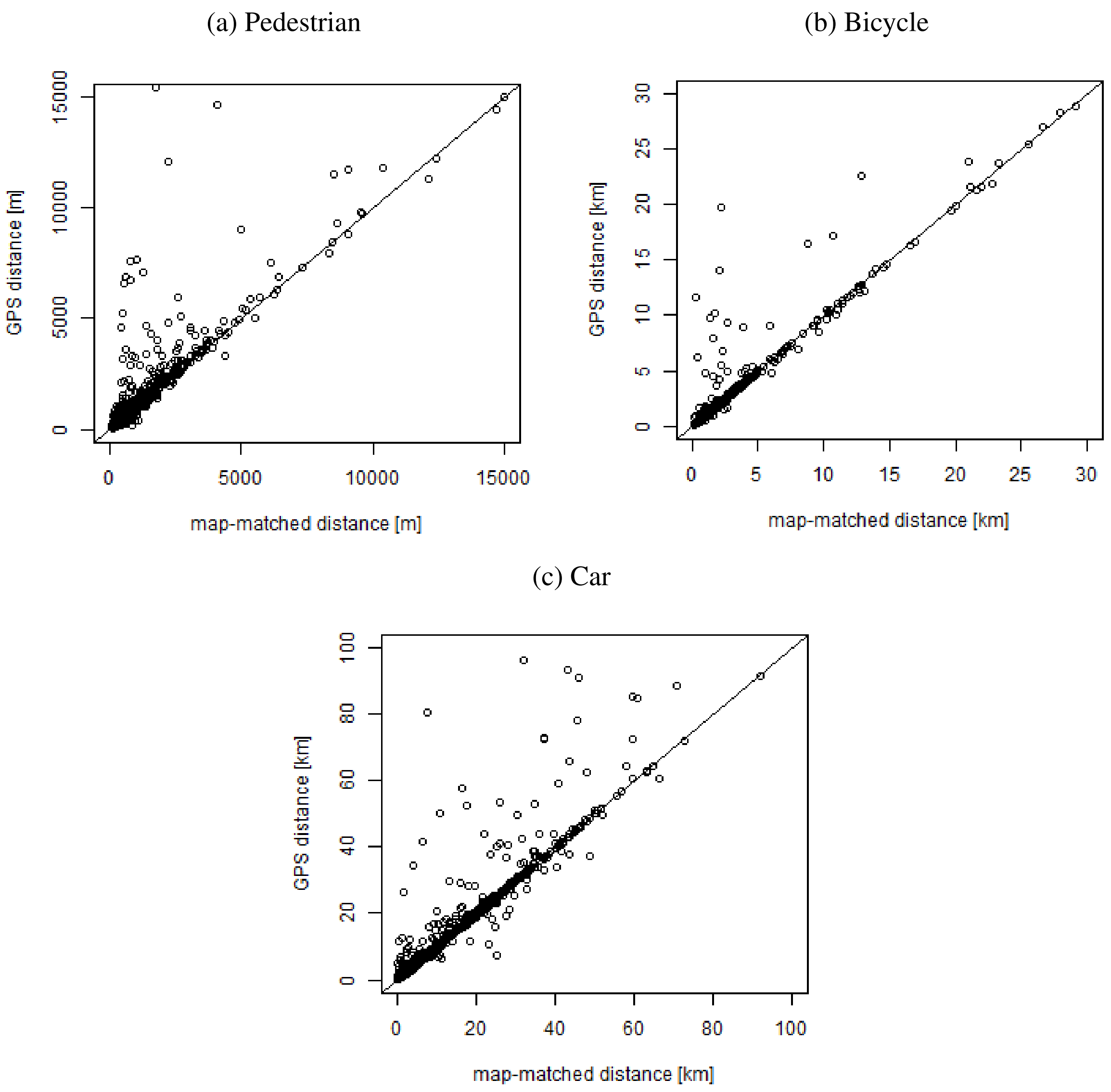
Matching the GPS points to a given network was done based on the work by Schüssler and Axhausen (2009) which is implemented in POSDAP (2012). The original map-matching routines were developed for navigation networks. Networks used for this paper on the other hand are extracted from OSM and are spatially correct but therefore have many small links to e.g. represent a curved street. Therefore, some constraints, like number of points per link to get a valid match had to be loosened. Further, for map-matching of walk stages links that are pedestrian only were weighted slightly more, to ensure that people walked on side-walks and not in the middle of the street or on the roundabout as shown in Figure 3.

To ensure valid map-matching, the map-matched distance is compared to the distance computed from the GPS points. Figure 4 shows the comparison for all travel modes considered here. Most distances are the same, and if they do not correspond, the map-matched distance is lower than the GPS distance, which is preferable, as invalid coordinates can cause high deviations and it shows that no detours are generated due to missing links.

3.2 Choice set generation

For choice set generation the Breadth First Search on Link Elimination algorithm (BFS-LE) as described in Schüssler et al. (2010) was used. The general idea is that given a cost function the shortest paths of the network is computed, to generate the next alternative one link of this shortest

Figure 4: Map-matched distance compared to GPS distance



(a) Pedestrian

(b) Bicycle

(c) Car

path is eliminated from the network, then the shortest path on this subnetwork is computed, this is repeated until the desired number of paths or all possible paths are computed. This algorithm was shown to be computationally very efficient while as well producing relevant routes, e.g. for bicycle routes in Halldórsdóttir et al. (2014).

Most important for our work with high resolution OSM networks is the second performance optimisation described inSchüssler et al. (2010), where a topologically equivalent network is created before choice set generation, that is vertices that are not a junction, intersection or a dead-end are removed and links are joined per direction. For the car network this means instead of 2’363’307 links 499’928 segments are processed.

For this paper, 50 alternatives were generated per route and the chosen path was added if not generated by the algorithm. For car 52 % of the chosen alternatives were generated in the procedure, for bicycle route only 28 % and for pedestrians the rate of reproduced shortest paths was highest with 58 %. For some models, the choice sets was reduced to 20 or 10 routes, starting with the chosen routes, alternatives were added such that the sum of path sizes was maximised. The idea of this reduction procedure is, to create a choice set with least overlap.

3.3 Route choice model

For the route choice models for car, walk and bike route choice the path size logit formulated by Ben-Akiva and Bierlaire (1999) is used, a multinomial logit model (MNL) with the path size (PS) as adjustment term to correct for overlapping routes. The path size is given in equation (1), the path size is one if a path does not overlap with any other in the choice set and it is very small if there is a lot of overlap. The general form of the deterministic part of the utility function is given in equation (2), the path size is transformed logarithmically such that it is very negative for very overlapping routes and 0 for routes without overlap.

$$\mathrm{PS}_{\mathrm{route,set}} = \sum_{\mathrm{link} \in \mathrm{route}} \left(\frac{\mathrm{length}_{\mathrm{link}}}{\mathrm{length}_{\mathrm{route}}} \right) \frac{1}{\text{\# routes in set via link}} \tag{1}$$

$$\begin{aligned} V =& \beta_{\mathbf{TT}} * \text{travel time} + \beta_{\mathbf{RT}} * \text{distance road type} + \beta_{\mathbf{FRAC}} * \text{fraction road type} + \\ & \beta_{\mathbf{ELEV}} * \text{elevation measure} + \beta_{\mathbf{PS}} * \ln(\text{path size}) \end{aligned} \tag{2}$$

To estimate the models the python version of BIOGEME (Bierlaire, 2003) was used.

4 Results

Three travel modes were analysed: car, cycling and walking. Following for each travel mode preliminary modelling results are presented.

4.1 Car route choice model

The car route choice model presented in Table 2 considers two different travel time coefficient one for stages starting during rush hour (07:00 - 09:00 and 16:00 - 18:00) and one for all other stages. Further, distance is split per road type.

Table 2 shows that several parameters have an unexpected sign. First, travel time is positive for all choice sets and for both parameters. This might be partly due to the high correlations with the distances per road type, which have negative signs. The travel time parameter for stages starting during rush hour are higher, which might suggested that routes with best free flow travel times are less attractive during rush hour as they might be more congested. Second, the path size is positive for the choice sets with 10 and 20 alternatives respectively, which is unexpected as in theory the path size corrects routes with overlaps such that they are less likely to be chosen. But it has already been shown in Frejinger and Bierlaire (2007), that positive path sizes are not uncommon, they argue that the path size has an ambiguous function also capturing behavioural aspects. That is, overlapping paths might be more attractive as they e.g. provide more possibilities for route switching.

At first glance the model for 10 alternatives is much better as the ρ^2 is much higher, but it has to be noted that due to the different choice sets the models cannot be compared. In general, it is intuitive that a choice can be better explained if less alternatives are provided, that are as different from the chosen alternative as possible.

The ratios of the parameters on the other hand stay very similar over all three choice sets, the ratio of travel time parameters is 1.46 for the 10 alternative choice set and goes up to 1.73 for the 51 alternatives, comparing the different road types to motorway, clearly residential and track roads are least attractive in all models (for CS10 1 km of residential road correspond to 12 km on motorways), for secondary roads it is not so clear as for CS10 1 km on those correspond to 1.5 km on motorways, assuming speeds of 50 km/h and 120 km/h respectively, according to the model time is spent preferably on secondary roads (1.6 minutes secondary road corresponding to 1 minute motorway).

Table 2: Car routes

	CS 10		**CS 20**		**CS 51**	
Free flow time (rush) [min]	0.646		0.753		0.693	
Robust Std err | Robust t-test	0.103	6.3	0.11	6.85	0.0964	7.19
Free flow time (non rush) [min]	0.443		0.509		0.401	
Robust Std err | Robust t-test	0.1	4.41	0.112	4.55	0.096	4.18
Distance motorway	-0.161		-0.148		-0.207	
Robust Std err | Robust t-test	0.0525	-3.07	0.0592	-2.5	0.0482	-4.29
Distance trunk	-0.376		-0.421		-0.482	
Robust Std err | Robust t-test	0.105	-3.57	0.112	-3.76	0.0992	-4.86
Distance primary road	-0.28		-0.246		-0.346	
Robust Std err | Robust t-test	0.0899	-3.11	0.103	-2.39	0.083	-4.17
Distance secondary road	-0.242		-0.196 (*)		-0.304	
Robust Std err | Robust t-test	0.0953	-2.54	0.111	-1.77	0.0892	-3.41
Distance tertiary road	-0.327		-0.29		-0.376	
Robust Std err | Robust t-test	0.0978	-3.35	0.11	-2.62	0.0892	-4.21
Distance residential/track/other	-1.89		-2.01		-1.93	
Robust Std err | Robust t-test	0.154	-12.25	0.176	-11.41	0.147	-13.18
ln(path size)	-2.3		-1.15		0.195	
Robust Std err | Robust t-test	0.12	-19.26	0.097	-11.89	0.051	3.81
Sample size	2250		2250		2250	
Init log-likelihood $\mathcal{L}(\beta_0)$	-5178.9		-6736.8		-8760	
Final log-likelihood $\mathcal{L}(\hat{\beta})$	-3843		-5759.5		-8071.9	
Likelihood ratio test	2671.87		1954.57		1376.34	
ρ^2	0.258		0.145		0.079	
$\bar{\rho}^2$	0.256		0.144		0.078	

Values with (*) are not significant on a 95-% level

4.2 Bicycle route choice model

In the bicycle route choice model elevation measures (average and maximum fall and rise) are considered, as well as the fraction driven on bike paths and the fraction on residential roads for safety considerations. In Model 1 one parameter for the distance is estimated in Model 2 on the other hand parameters are estimated for the distance depending on trip purpose. Results in Table 3 are estimated on a choice set with 20 alternatives, whereas the alternatives were reduced from the generated 50 alternatives by maximising the sum of the path sizes

For Model 1, as expected, the travel distance parameter is negative and path size is positive, contrary to the car route choice model, but in line with choice theory. Travel distance parameters in the second model are interacted with trip purpose, which on the one hand gives promising results, as for leisure distance has a less negative effect than for access or egress stages, but for work trips on the other hand the parameter is positive and not significant. The elevation measures are mostly not significant, except for the average rise which decreases utility as expected, as you need to be fitter for steep paths. The magnitude of the travel distance and rise average parameters is similar, but the value range of the two variables is different therefore the influence of travel distance is much higher (travel distance: 0.5 - 30 km, average 3.2 km; rise average: 0 - 0.1 with an average of 0.01). Parameter estimates for the bicycle path and residential road fractions are low but significant. The bicycle path fraction is unexpectedly negative, this might be due to a too sparse cycling network.

The adjusted ρ^2 of the two models are almost the same, whereas the log-likelihood test confirms that the second model performs slightly better.

4.3 Pedestrian route choice model

For the pedestrian route choice model, round-trips and walks with speeds higher than 5 m/s are excluded. The choice set consists of 10 alternatives (reduced from the base set of 50). Table 4 shows the results for the model with all data, and two models where the data is split into access- and egress and all other stages. It is assumed that access and egress stages are more likely to be the shortest possible path.

The main variable again is distance, which is split into pedestrian only areas, pedestrian/cycling shared areas and rest, which mostly corresponds to side-walks. Unexpectedly the pedestrian only areas have a higher negative effect than the normal side-walks, the shared areas on the other hand have a positive effect, but are not significant in the split models. Surprisingly, the model for access and egress stages has a more negative value for side-walks and a less negative value for pedestrian areas, even though it was expected that minimising distance is independent of safe paths. It might be that access and egress stages are in the city, where there are more pedestrian areas.

As in the bicycle model the elevation measures are also considered in the pedestrian models. In that case almost all measures are significant. The maximum fall and rise have a negative sign, which means that very steep paths are avoided, but the effect is very small. The positive effect of a higher average fall, is probably because this corresponds to shorter paths. The average rise is positive as well, in the split model it can be seen, that this is mostly due to the not access /

Table 3: Bicycle routes (choice set with 20 alternatives)

	Model 1		**Model 2 (purpose)**	
Travel distance [km]	-0.665		-	
Robust Std err | Robust t-test	0.274	-2.43		
Distance * isLeisure [km]	-		-0.872	
Robust Std err | Robust t-test			0.38	-2.29
Distance * isWork [km]	-		0.604 (*)	
Robust Std err | Robust t-test			0.457	1.32
Distance * isAcc-/Egress[km]	-		-1.24 (*)	
Robust Std err | Robust t-test			0.837	-1.48
Distance * isOther [km]	-		-1.07	
Robust Std err | Robust t-test			0.494	-2.16
Fraction bike path	-0.0825		-0.0834	
Robust Std err | Robust t-test	0.00873	-9.45	0.00915	-9.12
Fraction residential road	0.0352		0.0354	
Robust Std err | Robust t-test	0.0051	6.89	0.00507	6.98
Fall max	-0.0583(*)		-0.059 (*)	
Robust Std err | Robust t-test	0.0298	-1.96	0.0304	-1.94
Fall average	-0.365 (*)		-0.359(*)	
Robust Std err | Robust t-test	0.355	-1.03	0.349	-1.03
Rise average	-0.872		-0.912	
Robust Std err | Robust t-test	0.308	-2.83	0.294	-3.11
Rise max	-0.0196 (*)		-0.0214 (*)	
Robust Std err | Robust t-test	0.0235	-0.83	0.0211	-1.02
ln(path size)	1.16		1.19	
Robust Std err | Robust t-test	0.196	5.92	0.196	6.07
Number of estimated parameters	8		11	
Sample size	410		410	
Init log-likelihood $\mathcal{L}(\beta_0)$	-1228.2		-1228.2	
Final log-likelihood $\mathcal{L}(\hat{\beta})$	-544.138		-539.959	
Likelihood ratio test for the init. model	1368.121		1376.479	
ρ^2	0.557		0.56	
$\bar{\rho}^2$	0.55		0.551	

Values with (*) are not significant on a 95-% level egress model, which is a little bit surprising and does not support the assumption of efficient access and egress stages.

The log-likelihood ratio test of the all model versus the combination of the two split models confirms that the latter is slightly better.

Table 4: Pedestrian routes: Not round trip nor access/egress

	all	**Not access / egress**	**Access / egress**
Distance pedestrian areas [km]	-2.13	-2.43	-1.86
Robust Std err | Robust t-test	0.574 -3.71	0.967 -2.51	0.724 -2.57
Distance pedestrian/bicycle [km]	0.422	0.468 (*)	0.436 (*)
Robust Std err | Robust t-test	0.208 2.03	0.385 1.22	0.241 1.81
Distance rest (side-walks) [km]	-1.16	-0.474 (*)	-1.69
Robust Std err | Robust t-test	0.382 -3.04	0.592 -0.8	0.518 -3.25
Rise max	-0.0717	-0.0877	-0.0601
Robust Std err | Robust t-test	0.0176 -4.08	0.0251 -3.49	0.0253 -2.37
Rise average	0.215	0.431	0.0626 (*)
Robust Std err | Robust t-test	0.08 2.69	0.125 3.45	0.107 0.58
Fall average	0.268	0.204 (*)	0.286
Robust Std err | Robust t-test	0.088 3.05	0.145 1.41	0.114 2.52
Fall max	-0.0755	-0.0757	-0.074
Robust Std err | Robust t-test	0.0154 -4.91	0.023 -3.29	0.0208 -3.56
ln(path size)	-2.02	-2.12	-1.97
Robust Std err | Robust t-test	0.172 -11.8	0.306 -6.92	0.208 -9.45
Sample size	985	370	615
Init log-likelihood $\mathcal{L}(\beta_0)$	-2265.5	-851.04	-1414.5
Final log-likelihood $\mathcal{L}(\hat{\beta})$	-1893	-716.46	-1169.5
Likelihood ratio test	744.97	269.158	489.903
ρ^2	0.164	0.158	0.173
$\bar{\rho}^2$	0.161	0.149	0.168

Values with (*) are not significant on a 95-% level

5 Conclusion and Outlook

The bicycle model seems to be the most promising model of the three travel modes under investigation. The explanatory power of distance and average rise is quite good, and the model could be slightly improved by considering trip purpose ($\bar{\rho} = 0.551$). The pedestrian model's explanatory power is lower ($\bar{\rho} = 0.161$). The signs are mostly reasonable but the distinction between access/egress stages and others was not fully as expected, the assumption that access and egress stages are more optimised could not be strengthened. But this could be due to the choice set and location of the stages, which should be investigated in more detail. For the car route choice model the positive travel time parameters are problematic, therefore, the choice set and the generation procedure has to be analysed in more detail.

From the modelling perspective, several approaches should be tested for the final results. On the one hand, several choices of the same person are collected with GPS travel surveys, therefore, panel effects should be accounted for. Further, the very successful model using subnetworks formulated by Frejinger and Bierlaire (2007) should be tested. A possible subnetwork is the motorway network, and in that context it is important that routes between cities are not lost due to network size constraints, one potential approach could be intelligent loading of network parts and also considering different levels of resolution, such as the motorway and primary road network in between cities and the complete network within the city.

Content wise, socio-demographic and attitudinal data as well as weather information is available that should be included in the analysis, especially if mode choice models are estimated. Further, the authors plan to estimate public transport connection models. To conclude, route analysis could be extended to two more data sets that are available. First, a data set collected with dedicated devices for a poster company in 2006 with routes in Zurich, Winterthur and Geneva and second data collected in Vienna and Dublin with smartphones for the PEACOX project (Montini et al. (2014)).

6 References

Ben-Akiva, M. E. and M. Bierlaire (1999) Discrete choice methods and their applications to short-term travel decisions, in R. Hall (ed.) *Handbook of Transportation Science*, chap. 2, 5–34, Kluwer, Dordrecht.

Bierlaire, M. (2003) BIOGEME: A free package for the estimation of discrete choice models, paper presented at the *3rd Swiss Transport Research Conference*, Ascona, March 2003.

Federal Office of Topography swisstopo (2012) Dhm25, webpage, February 2012, http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/height/dhm25.html.

Frejinger, E. and M. Bierlaire (2007) Capturing correlation with subnetworks in route choice models, *Transportation Research Part B: Methodological*, **41** (3) 363–378.

Halldórsdóttir, K., N. Rieser-Schüssler, K. W. Axhausen, O. A. Nielsen and C. G. Prato (2014) Efficiency of choice set generation methods for bicycle routes, *European Journal of Transport and Infrastructure Research*, **14** (4) 332–348.

MATSim (2015) Multi-Agent Transportation Simulation, webpage, http://www.matsim.org.

Montini, L., S. Prost, J. Schrammel, N. Rieser-Schüssler and K. W. Axhausen (2014) Comparison of travel diaries generated from smartphone data and dedicated GPS devices, paper presented at the *10th International Conference on Transport Survey Methods*, Leura, November 2014.

Montini, L., N. Rieser-Schüssler and K. W. Axhausen (2013) Field Report: One-Week GPS-based Travel Survey in the Greater Zurich Area, paper presented at the *13th Swiss Transport Research Conference*, Ascona, April 2013.

Office for Spatial Development of the Canton of Zurich (2015) Dtm gis zh, webpage, February 2015, http://maps.zh.ch/?topic=LidarZH&offlayers=dom2014hillshade&over=UpBackgroundZH.

OpenStreetMap (2015) The Free Wiki World Map, webpage, http://www.openstreetmap.org.

POSDAP (2012) Position Data Processing, webpage, http://sourceforge.net/projects/posdap/.

Schüssler, N. and K. W. Axhausen (2009) Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT), *Working Paper*, **568**, IVT, ETH Zurich, Zurich.

Schüssler, N., M. Balmer and K. W. Axhausen (2010) Route choice sets for very high-resolution data, paper presented at the *89th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2010.