

# **IDENTIFYING TRIPS AND ACTIVITIES AND THEIR CHARACTERISTICS FROM GPS RAW DATA WITHOUT FURTHER IN- FORMATION**

8TH INTERNATIONAL CONFERENCE ON SURVEY METHODS IN TRANSPORT:  
ANNECY, FRANCE, MAY 25-31, 2008

*Nadine Schuessler, IVT, ETH Zurich, CH-8093 Zurich*

*Kay W. Axhausen, IVT, ETH Zurich, CH-8093 Zurich*

## **ABSTRACT**

In recent years, studies based on GPS records have become more and more important due to their manifold advantages compared to classic survey methods, such as paper diaries or telephone interviews. However, without additional information, a lot of post-processing work is required to derive data that can be used for analysis and model estimation. These post-processing procedures are still an ongoing research issue. Recently, a couple of new ideas have been published, but the key research questions remain the same.

This paper focuses on the description of a post-processing procedure that is able to determine trips and activities, including several characteristics, such as modes and routes from GPS raw data without any further information. It is applied to GPS records collected in the Swiss cities of Zurich, Winterthur and Geneva with 4882 participants, each of which carried an on-person GPS-receiver for 6.65 days on average. The data outcome is compared to the Swiss Micro-census 2005 to prove that it is ready for further applications, such as discrete choice model estimations.

## 1 INTRODUCTION

In recent years, studies based on GPS records have become more and more important due to their manifold advantages compared to classic survey methods, such as paper diaries or telephone interviews. First, GPS records provide researchers with more detailed information in terms of both spatial and temporal resolution. Second, they prevent under-reporting of trips, a well-known limitation of recollection-based surveys. Third, they reduce participants' burden to a minimum, as long as they are not combined with extensive questioning to derive additional information, such as trip purposes and transport modes. However, without additional information a lot of post-processing work is required to derive data that can be used for analysis and model estimation. The appropriate post-processing procedures are still an ongoing research issue. Recently, a couple of new ideas have been published. However, the key research questions remain: How to detect individual trips and activities? How to derive the modes used by the participants? How to extract the routes chosen on the network?

This paper deals with all of these questions. It is based on GPS records collected in the Swiss cities of Zurich, Winterthur and Geneva. The original study was conducted by a private sector company with the aim to explore whether or not participants pass certain billboards. Hence, not only were all modes included in the study, but also trip chains covering the whole day were obtained. In addition, each of the 4882 participants carried the GPS logger not only for one but for several days, resulting in about 32 000 recorded person-days. However, no additional information was collected. Therefore, an advanced post-processing procedure had to be developed to derive individual trips and activities, means of transport and chosen routes.

Due to the large amount of data, it was decided not to work within a GIS environment but to implement our own post-processing procedures for data cleaning, trip and activity detection and mode identification in JAVA. The approach is not only able to handle the huge amount of data efficiently, it is also independent of the quality of the network and the spatial information available. In particular, there is no necessity to employ a fine-grained multi-modal network, which is rarely available in practise, in the mode detection. However, this comes at the price of not being able to take into account network and other spatial characteristics until the map-matching, which especially affects the mode detection. As described in Section 7, this could be solved in the future by implementing a feedback loop between the map-matching of the different modes to their specific networks and the mode detection.

The input data needed for the post-processing procedure is merely the raw data derived from the GPS receivers consisting of the three-dimensional position and the time stamp. Unlike

most other studies, no information about the quality of the GPS data is employed because it was absent in the data we received. Thus, much attention had to be focussed on filtering the data and smoothing their trajectories, as described in Section 3.

Another important step in the post-processing is the detection of trips and activities within the continuous stream of GPS points. Because person-based GPS points were collected, the data recording often continued during activities, leading to bundles of GPS points. Therefore, time gaps and periods with zero speed are not sufficient to detect all activities. Instead, as demonstrated in Section 4, they have to be combined with an indicator for bundles.

Since each trips can still contain more than one mode, the mode detection, which is presented in Section 5, starts with a segmentation of the trips into single-mode stages. Based on the assumption that walking is required for any mode change, the approach takes advantage of the distinctive features of this mode, such as low speed and low acceleration, and uses it as separation mode. Afterwards, the actual mode detection is executed based on the speed and acceleration characteristics of the stages. Given that the value ranges of these characteristics are overlapping, membership scores and consequent likelihoods are calculated for each stage using a fuzzy logic approach.

In Section 6, the results of the proposed post-processing procedure are compared to the Swiss Microcensus 2005, which was chosen for validation even though it cannot be assured that the sample at hand is representative of the Swiss population due to the absence of information about the survey participants. For each individual step, it is demonstrated that the patterns of the participants' travel behaviour resemble those observed in the Microcensus. This is remarkable since no calibration of the parameters was performed. However, it is also shown that there is room for methodological refinements in all steps of the post-processing procedure as summarised in Section 7.

## 2 RECENT DEVELOPMENTS REGARDING GPS POST-PROCESSING

Since the first GPS studies were conducted in the mid-1990s (e.g. Wagner, 1997; Casas and Arce, 1999; Yalamanchili *et al.*, 1999; Draijer *et al.*, 2000; Wolf, 2000; Pearson, 2001) this new way of surveying individual travel behaviour has gained increasing attention in transportation research. Researchers benefit from more accurate and reliable information about times, geographic locations, and routes while the burden for the participants is reduced significantly. However, so far, most studies asked for a wide range of additional information to ensure the usability of the data. The repertory of additional information ranges from socio-demographic data over travel diaries (Schönfelder *et al.*, 2006) and complete travel logs stating names of travellers, times, trip purposes, origins and destinations (Du and Aultman-Hall, 2007) up to driving events such as streets traveled including the time of entrance, lane changes, possible signal interruptions, etc. (Ogle *et al.*, 2002).

While this information was necessary in the beginning to get to know the new technology, including its advantages and disadvantages, the aim is now to develop appropriate post-processing procedures. These procedures allow the researcher to derive all necessary information, such as start and ending times, mode, and trip purpose in an automated way directly from the GPS records. They lead to a minimum of additional questions to be answered by the participants and thus a real reduction of the burden. Therefore, a lot of the recent research effort in the field has been dedicated to the development of post-processing procedures, some of which are briefly presently in the following.

The choice of the appropriate post-processing approach strongly depends on whether the GPS data was collected vehicle-based or person-based. In vehicle-based studies, such as Schönfelder *et al.* (2006), Ogle *et al.* (2002), Du and Aultman-Hall (2007) or Biding and Lind (2002), the participants' vehicles are often equipped with GPS loggers which record only, when the engine of the vehicle is running. Accordingly, the detection of individual trips is relatively easy based on the time differences between the recorded points. In addition, short stops during which the engine is not turned off can be found fairly reliably by identifying times when the speed of the vehicle is zero. However, there are also some shortcomings related to vehicle-based data. First and foremost, all other modes are omitted, even though they are essential for the analysis of complete trip chains, especially in the context of a European urban environment. Second, the real trip origins and destinations have to be guessed since only vehicle movements are recorded. Third, the analyst is not able to determine which person is driving the vehicle without further information.

Therefore, person-based GPS studies have recently become more popular. However, person-based GPS data raises the requirements for the post-processing procedures considerably. In addition to data filtering, trip detection and map-matching, which are also necessary for vehicle-based GPS data, the analyst has to detect the modes used by the participant, preferably in an automated way. Moreover, the method for trip detection needs refinement and the map-matching has to be done either on a fine-grained multi-modal network or, if this not available, multiple networks.

A couple of authors have started to address these problems (e.g. de Jong and Mensonides, 2003; Chung and Shalaby, 2005; Tsui and Shalaby, 2006; Flamm and Kaufmann, 2007). Basically, all approaches contain individual modules accounting for:

- Data filtering
- Detection of trips and activities
- Mode stage determination
- Mode identification
- Map-matching

Some authors include additional features such as the merging of stages after the mode detection (de Jong and Mensonides, 2003) or a feedback between the map-matching and the mode detection (Tsui and Shalaby, 2006).

Yet, all of these methods have so far only been tested on small samples or test scenarios. de Jong and Mensonides (2003), for example, engaged two researchers travelling for a couple of days through Sidney using several modes and keeping extensive travel logs. Chung and Shalaby (2005) and Tsui and Shalaby (2006) constructed specific test samples with the help of their students containing 60 trips or 59 person-days, respectively, whereas Flamm and Kaufmann (2007) are still building up their sample. However, since they are especially interested in a qualitative analysis of travel behaviour changes during personal life transitions, their sample will remain small, allowing them to detect modes manually.

Due to the size of the dataset available for this project, manual detection is not a feasible option. Thus, the post-processing procedure has to be robust and efficient enough to deal with the large amount of data while still producing reasonable and meaningful results. Therefore, this work will provide an important step towards the large-scale use of GPS data for the analysis of travel behaviour.

### 3 DATA CLEANING AND DATA SMOOTHING

The quality of GPS study results strongly depends on the way the analyst takes into account the accuracy level of the data. The positioning accuracy of GPS receivers under ideal conditions lies between five and ten metres (Wolf, 2006). In reality, however, the accuracy level 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, including the time stamp. Even if there are enough satellites in view, they might be positioned too closely to each other, which is expressed by a high position dilution of precision (PDOP) value (Wolf *et al.*, 1999). Together with the so-called *warm start/cold start problem*, these errors form the group of systematic errors of GPS measurements (Jun *et al.*, 2007). While the first two errors lead to GPS positions that are completely different from the actual position of the receiver, the 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).

The second group of errors are random errors caused by satellite orbit, clock or receiver issues, atmospheric and ionospheric effects, multi-path signal reflection or signal blocking (Jun *et al.*, 2007). Especially burdensome for all means of transport 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 around and often seem to be scattered around the actual position of the receiver. For travel analysis purposes, they are worthless since neither stops in this area nor the route actually taken can be detected (de Jong and Mensonides, 2003). Signal blocking, in contrast, leads to missing GPS points and is of special importance for person-based GPS surveys since it varies between 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). This information can be used in the mode detection.

There are several ways to overcome the problems caused by the GPS errors described above. To get rid of erroneous points, filtering and smoothing techniques can be applied. While filtering methods take care of systematic errors, smoothing techniques remove random errors. All these approaches, however, depend on the information available in the study. Previous studies (e.g. Wolf *et al.*, 1999; Ogle *et al.*, 2002) showed that the number of satellites in view and the PDOP value are fairly efficient measures for determining systematic errors. Unfortunately, they are

not accessible here and therefore, other criteria have to be found that reliably detect erroneous GPS points while omitting true ones. One of these measures is the altitude value. Considering the Swiss topology, all points with an altitude value of less than 200 and more than 4200 metres above sea level are removed. In addition, since the accuracy in altitude is usually lower than in latitude or longitude, the altitude value is left out of any subsequent calculation of distance, speed or acceleration.

Other indications for erroneous GPS points are unrealistic speed and acceleration values. Because speeds from Doppler measurements are not available, speeds and accelerations have to be determined from the positional and temporal difference between consecutive GPS points. For this, the velocity and the three-dimensional acceleration are computed depending the smoothing technique employed. Subsequently, the speed is computed as the two-dimensional length of the velocity and the acceleration as the change in speed in the two-dimensional moving direction. Since speed and acceleration depend on the smoothing method, the filtering for unrealistic speed and acceleration values takes place after the smoothing. This has the advantage that speed jumps due to random errors or actual behaviour do not lead to a removal of the points concerned. Consequently, the boundaries are not too rigorous. Only points with speeds above 50  $m/s$  or accelerations above 10  $m/s^2$  are deleted.

Several approaches were reviewed to select the appropriate smoothing technique. Given that speed and acceleration are calculated from the position and the timestamp of the GPS points and assuming that the timestamps are correct, the position of the GPS points is smoothed rather than the speed values. First, a moving average approach was evaluated as it was applied by Ogle *et al.* (2002) and Chung and Shalaby (2005), but it led to unrealistic speed and acceleration combinations. Second, the application of a modified Kalman filter, as presented by Jun *et al.* (2007), was considered. However, due to the lack of reliable information about the measurement and process noise, it was not employed here.

Instead, a Gauss kernel smoothing approach was implemented. For each coordinate dimension  $c \in x, y, z$  the smoothed value  $\tilde{c}(t)$  at time  $t$  is individually calculated as

$$\tilde{c}(t) = \frac{\sum_j (w(t_j) \cdot c(t_j))}{\sum_j w(t_j)} \quad (1)$$

with  $c(t_j)$  being the raw value of the coordinate  $c$  at time  $t_j$  and  $w(t_j)$  the Gaussian Kernel

function computed for each point of time  $t_j$  by

$$w(t_j) = \exp - \frac{(t - t_j)^2}{2\sigma^2} \quad (2)$$

The Kernel bandwidth is represented by  $\sigma$ , in this case 10 seconds, which results in a 15 second smoothing range, because this is assumed to be a reasonable time frame for real behavioural changes as opposed to signal jumps. Accordingly, the directional speed for each coordinate  $c$  is the first derivative with respect to  $t$  of the smoothed coordinate and the acceleration the second derivative with respect to  $t$ .

So far, no other approaches such as a correction for the warm start/cold start problem (e.g. Stopher *et al.*, 2005) or an interpolation of missing data (e.g. Ogle *et al.*, 2002) have been applied. Further analysis of the resulting data will show if these are necessary.

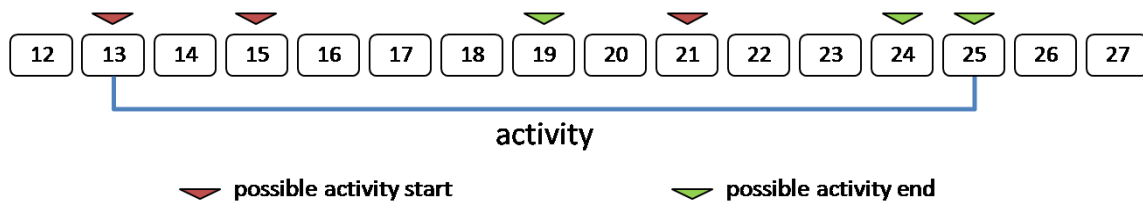
#### **4 TRIP AND ACTIVITY DETECTION**

Subsequently, the filtered and smoothed GPS points have to be subdivided into trips and activities. As the GPS points at hand have been collected person-based, the recording often continued while the person was performing an activity. Therefore, appropriate criteria had to be found which determine whether the GPS point in question belongs to an activity or to a trip. The most common criterion is the so-called dwell time, i.e. the minimum time difference between two GPS points after which it is assumed that an activity took place. The value for this dwell times varies in the literature between 45 (Pearson, 2001) and 300 seconds (e.g. Wolf *et al.*, 2004; Doherty *et al.*, 2001), whereas most studies apply 120 seconds. In this study, however, a 900 second dwell time is used. That is a rather high value, but the examination of the GPS points lead to the conclusion that a shorter dwell would lead to too many wrongly detected activities. These misidentified activities are predominantly caused by bad reception during trips or activities with ongoing GPS recording. The map-matching algorithm will handle data losses during a trip. Accordingly, they do not need to be considered here. Activities with shorter durations or ongoing GPS recording are detected by the following criteria.

First, there are the so-called *bundles of GPS points* (e.g. Doherty *et al.*, 2001; Stopher *et al.*, 2005). As mentioned above, the recording of GPS data often continues during activities. Usually, this results in a bundle of GPS points positioned very close to each other. This bundle typically extends over a diameter of about 30 metres if the person stays at the same place, approximating 3 times the standard deviation of the measurement inaccuracies (Stopher *et al.*, 2005). Thus, a measure for GPS bundles outside a GIS environment was developed. For each GPS point, its point density is determined by counting how many of the 30 preceding and succeeding GPS points are positioned within a 15 metres radius around it. If the sequence of points with a density higher than 15 lasts for at least 10 points or 300 seconds, an activity with ongoing recording is detected. Thereby, also short periods (maximum 15 GPS points) of smaller densities are included in the activity since they are usually caused by measurement errors. However, sequences of GPS points with densities below and above 15 are only considered to be activities, if the ratio of points with a density higher than or equal to 15 against the length of the sequence is at least  $2/3$ .

Additionally, the criterion of zero speed was applied, as in Schönfelder *et al.* (2006) or Tsui and Shalaby (2006). If the speed is smaller than  $0.01\text{ m/s}$  for at least 120 seconds it is assumed that the person stopped to perform an activity. No further criteria have been used so far. A possibly useful extension would be the consideration of heading changes of about  $180^\circ$  (e.g.

Figure 1: Treatment of possible activity start and end points in activity detection



Du and Aultman-Hall, 2007; de Jong and Mensonides, 2003) as they occur if a person only drops off or picks up someone or something. This has, however, not yet been implemented.

Each criterion described above is first used individually to determine possible activity start and end points. The according coordinates are marked as *potential activity start* or *potential activity end*. Thereby, each activity can be detected by more than one criterion. A bundle of GPS points, for example, usually contains also a sequence of points with zero speed. Consequently, the potential activity start and end points detected by the individual criteria are joined as depicted in Figure 1. The outermost potential activity start and end points are considered to be the true activity start and end points regardless of the criterion they belong to. Moreover, if a new activity starts shortly (maximum 15 GPS points) after the last one has ended, the two activities are joined. This rule on the one hand accounts for measurement errors and on the other hand considers that trips of less than 15 GPS points can not be reasonably used for route choice modelling. Furthermore, the first GPS point of a dataset is by default an activity end while the last GPS point is by default an activity start unless they have explicitly been marked otherwise. However, since no information is available about these activities, they are left out of the subsequent analysis.

After finding all activity start and end points, activity and trip objects are generated and stored in separate lists. An activity object contains all GPS points between an activity start and an activity end, including the boundary points, whereas a trip object comprises of all the points between an activity end and an activity start, also including the boundary points. Only, the trip objects are used in the subsequent analysis. The activity objects will be needed later on to determine trip purposes as well as trip and activity chains.

## 5 MODE DETECTION

Determining the modes used by the participants is one of the major research issues for person-based GPS studies. It is the crucial step to make them usable for large-scale applications. However, few approaches for an automated mode detection have been published to date. Stopher *et al.* (2005) worked with a stepwise elimination of modes based on average and maximum speeds, proximity to certain network elements, such as bus stops or train stations, and the deviation from the street network. In contrast to most other authors, they assume that each trip is undertaken in a single mode. de Jong and Mensonides (2003) also segment trips into single-mode stages based on the assumption that a short period of zero speed is necessary for each mode change. The mode of the stages is then determined by employing speed characteristics and the proximity to public transport stops and routes. In addition, they test to see if the derived mode chains are reasonable and prohibit, for example, direct changes from bus to car without an intermediate walking stage.

The approach followed in the study at hand follows the mode detection method presented by Chung and Shalaby (2005) and Tsui and Shalaby (2006). It rests upon the main assumption that walking is required for every mode change. Consequently, each trip is segmented into single-mode stages by finding the points where the mode changes from walk to another mode or vice versa, the so-called *mode changing points (MTP)*. The procedure exploits the uniqueness of the walk mode with consistently low speeds and accelerations. The mode of the remaining stages is determined by applying a fuzzy logic approach based on speed and acceleration characteristics. Based on the circumstances in the study area, five modes are distinguished:

- walk
- cycle
- car
- urban public transport (i.e. bus and tram)
- rail

Three types of MTPs are determined using the same method as Tsui and Shalaby (2006): *end of walk (EOW)*, *start of walk (SOW)* and *end of gap (EOG)* points, the latter indicating the end of a period with GPS signal loss. For each transition from a speed below  $2.78\text{ m/s}$  to above  $2.78\text{ m/s}$ , the algorithm searches backwards until the next point with a speed above  $2.78\text{ m/s}$  or at least three consecutive GPS points with a maximum acceleration of  $0.1\text{ m/s}^2$  are found. In the latter case, the last of these points with small acceleration is marked as a potential EOW point,

otherwise, no EOW point is detected. The procedure for the potential SOW points follows the same logic but in reverse order while each point after a time difference of at least 80 s is marked as a potential EOG point. Consequently, the final MTPs are identified by arranging the potential MTPs into reasonable chains ensuring the following rules:

- an EOW point can be followed by an SOW or an EOG point but not by an EOW point
- an SOW point can be followed by an EOW or an EOG point but not by an SOW point
- an EOG point can either be an EOW or an SOW point
- the speed of the stage between an SOW and an EOW point has always to be smaller than 2.78 m/s and the time difference has to be at least 60 s
- time difference between an EOW and an SOW point has to be at least 120 s

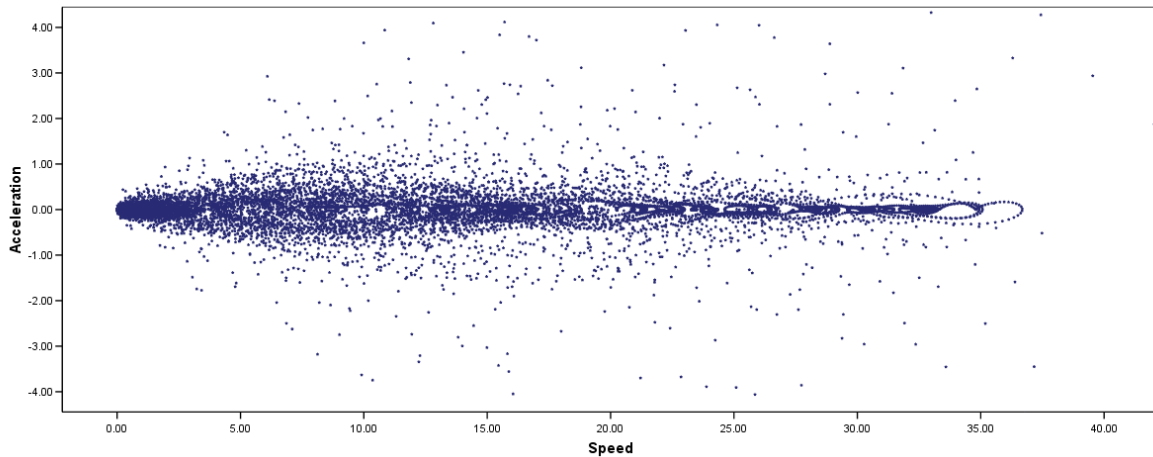
An open source fuzzy engine (Sazonov *et al.*, 2002) is used for the implementation of the fuzzy logic-based mode detection. The crucial elements are the fuzzy variables, the fuzzy rules describing the relationship between the modes and the fuzzy variables, and the membership functions representing the different levels of the fuzzy variables. Three fuzzy variables were chosen, each with three membership functions: the median of speed, and the ninety-fifth percentiles of the speed and acceleration distributions. These statistical location parameters were explicitly chosen over the average speed or the maximum speed and acceleration to make the algorithm more robust against outliers.

The trapezoidal membership functions are described by four key points: Start point, left top corner, right top corner, and end point. They were chosen after an analysis of the available modes and the speed and acceleration characteristics in the GPS data. Figure 2 depicts an example of a distribution of speed and acceleration combinations for two persons with obviously different travel behaviour. Overall, five more or less well-defined clusters can be distinguished:

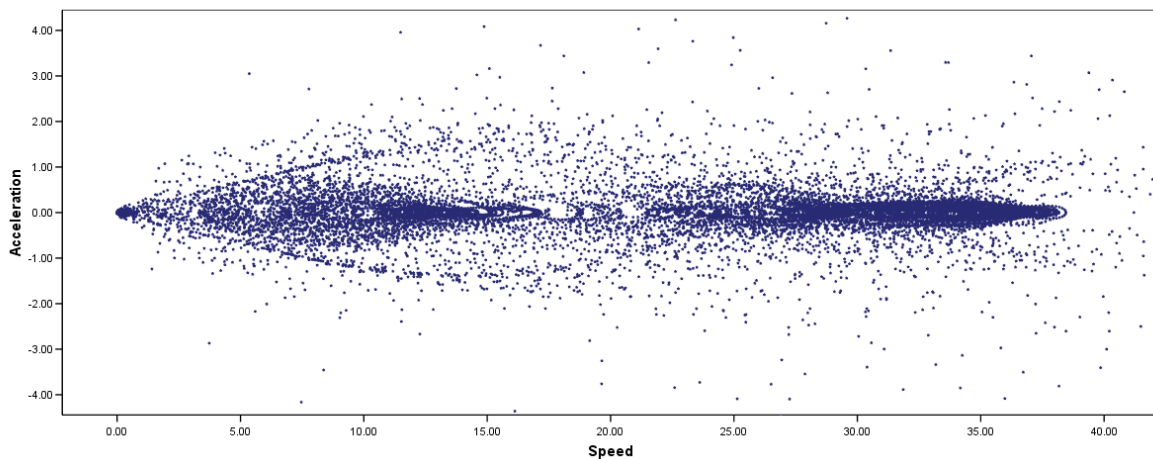
1. speed  $< 2$  m/s and acceleration  $< 0.15$  m/s<sup>2</sup>
2. speed 4-8 m/s and acceleration  $< 0.2$  m/s<sup>2</sup>
3. speed 14-17 m/s and acceleration  $< 0.3$  m/s<sup>2</sup>, with accelerations up to 1 m/s<sup>2</sup> on the way to the cluster
4. speed 20-28 m/s and acceleration  $< 0.3$  m/s<sup>2</sup>, with accelerations up to 1.6 m/s<sup>2</sup> on the way to the cluster
5. speed  $> 30$  m/s and acceleration  $< 0.4$  m/s<sup>2</sup>, with accelerations up to 4 m/s<sup>2</sup> on the way to the cluster

Figure 2: Example of speed and acceleration distribution

(a) Person 1548



(b) Person 16048



Although these clusters cannot be directly assigned to specific modes, they give an idea about the distribution of speed and acceleration of the modes. The first cluster, for example, characterises walking, the fifth cluster contains mainly trips by car on the motorway or by high-speed train, whereas the fourth cluster represents travel on country roads or on the InterRegio or rapid-transit railway system. Clusters 2 and 3 on the other hand cannot be so easily matched to individual modes, since they can be caused by the whole range of urban means of transport, such as cycle, urban public transport and car. They are, however, used to determine the key points of the membership functions, as they are presented in Table 1.

Table 1: Range of fuzzy variables

Level	Start	Left top	Right top	End
Median speed				
Walk	0	0	1.5	2
Low	1.5	2	4	6
Medium	5	7	11	15
High	12	15	100	100
95 percentile speed				
Low	0	0	6	8
Medium	7.5	9.5	15	18
High	15	20	100	100
95 percentile acceleration				
Low	0	0	0.5	0.6
Medium	0.5	0.7	1	1.2
High	1	1.5	25	25

Having established this, the fuzzy rules can be derived. They characterise the different modes with regard to the fuzzy variables. As depicted in Table 2, each mode is described by at least one rule. Since the ranges of the membership functions overlap, more than one rule can apply to an individual stage and the same stage can be linked to different modes. These ambiguities are deliberately included in the outcome of the mode detection and allow for feedback loops with the map-matching. The defuzzify method combines the membership values for each individual mode using the AND operator. Thus, the final score for each mode equals the minimum membership value amongst all its rules. Subsequently, the likelihood for each mode is calculated based on all the mode scores of a stage. For the map-matching to the public transport and car networks, all mode stages with a likelihood higher than a certain threshold for the respective mode will be used.

Table 2: Fuzzy rules for mode detection

Mode	Median speed	95 percentile acceleration	95 percentile speed
Walk	walk	low	—
Cycle	walk	medium	—
Cycle	walk	high	—
Cycle	low	low	low
UrbanPuT	low	low	medium
Car	low	low	high
UrbanPuT	low	medium	—
UrbanPuT	low	high	low
Car	low	high	medium
Car	low	high	high
UrbanPuT	medium	low	—
Car	medium	medium	—
Car	medium	high	—
Rail	high	low	—
Car	high	medium	—
Car	high	high	—

## 6 ANALYSIS OF THE RESULTS

Since no information about the actual trips and activities of the participants is available, the Swiss Microcensus on Travel Behaviour 2005 (MZ 2005) (Swiss Federal Statistical Office, 2006) is used as the basis for the validation of the post-processing procedure. The Swiss Microcensus on Travel Behaviour (MZ) is conducted every five years and delivers a representative and detailed insight into the travel patterns of the Swiss population. 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 the mode-stages and activities they undertook during the course of one specific day. In the following, the results of the post-processing procedure are compared with a sample of the Microcensus of 2005, which comprises respondents living in Zurich, Winterthur or Geneva. After a short analysis of the overall statistics, the effects of the data cleaning and smoothing are shown. Subsequently, the trip and activity detection algorithm is evaluated. The section closes with an assessment of the mode detection procedure.

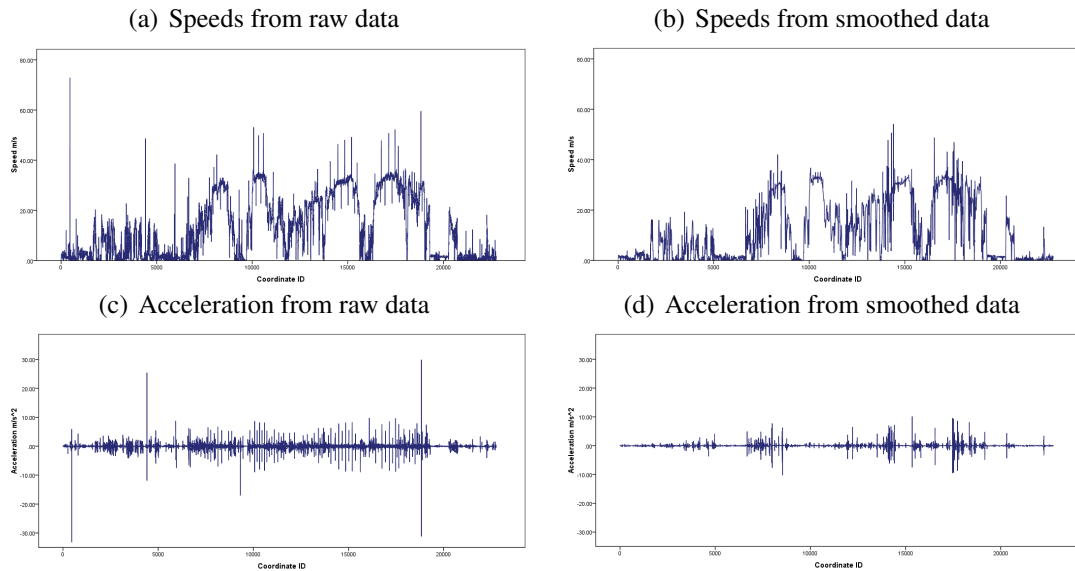
The overall statistics for the three sub-studies for Zurich, Winterthur and Geneva, and the corresponding sample in the MZ 2005 are represented in Table 3. It can be seen that the values vary slightly for each of the cities. The number of days observed per participant, for example, is about seven in Zurich compared to approximately six days in Winterthur. Overall, however, the numbers reveal the same trend. The number of trips per day, for example, is higher for all cities than the one reported in the MZ 2005. Accordingly, the average trip distance and duration are smaller. This is probably due to two reasons. First, the algorithm should guarantee that all trip ends are detected. Thus, sometimes prolonged waiting times at public transport stops or in congestion can be incorrectly identified as trip ends. This problem is inherent to all GPS trip end detection algorithms and difficult to overcome without additional information from the participants or manual post-processing. In comparison, several studies (e.g. Du and Aultman-Hall, 2007; Wolf *et al.*, 2003; Yalamanchili *et al.*, 1999; Bricka and Bhat, 2006) have demonstrated that respondents tend to underreport short trips and stops within trip chains in recollection-based surveys such as CATI. Therefore, a slightly higher trip rate per day than the one given in the MZ 2005 is probably consistent with reality.

The effects of the data cleaning and smoothing are depicted in Figures 3 and 4. Figure 3 shows the development of speed and acceleration over time for a sample individual. In the left column, speed and acceleration are derived from the raw data, whereas in the right column, they are calculated after filtering for unrealistic altitudes and smoothing, but before the final filtering for unrealistic speeds and accelerations. As can be seen, much of the noise in the raw

Table 3: Overall statistics of the GPS study compared to the Microcensus 2005

	Zurich	Winterthur	Geneva	MZ 2005
Number of persons	2 435	1 086	1 361	2940
Number of days per person	7.01	5.99	6.52	1.00
Number of trips per person	35.12	19.95	28.05	–
Number of trips per day	5.01	3.33	4.41	3.28
Average trip distance [km]	7.11	7.36	6.64	8.79
Average daily mileage [km]	37.21	26.77	30.86	32.13
Average trip duration [min]	11.65	12.17	12.42	26.21
Average activity duration [min]	312.13	519.03	333.23	–

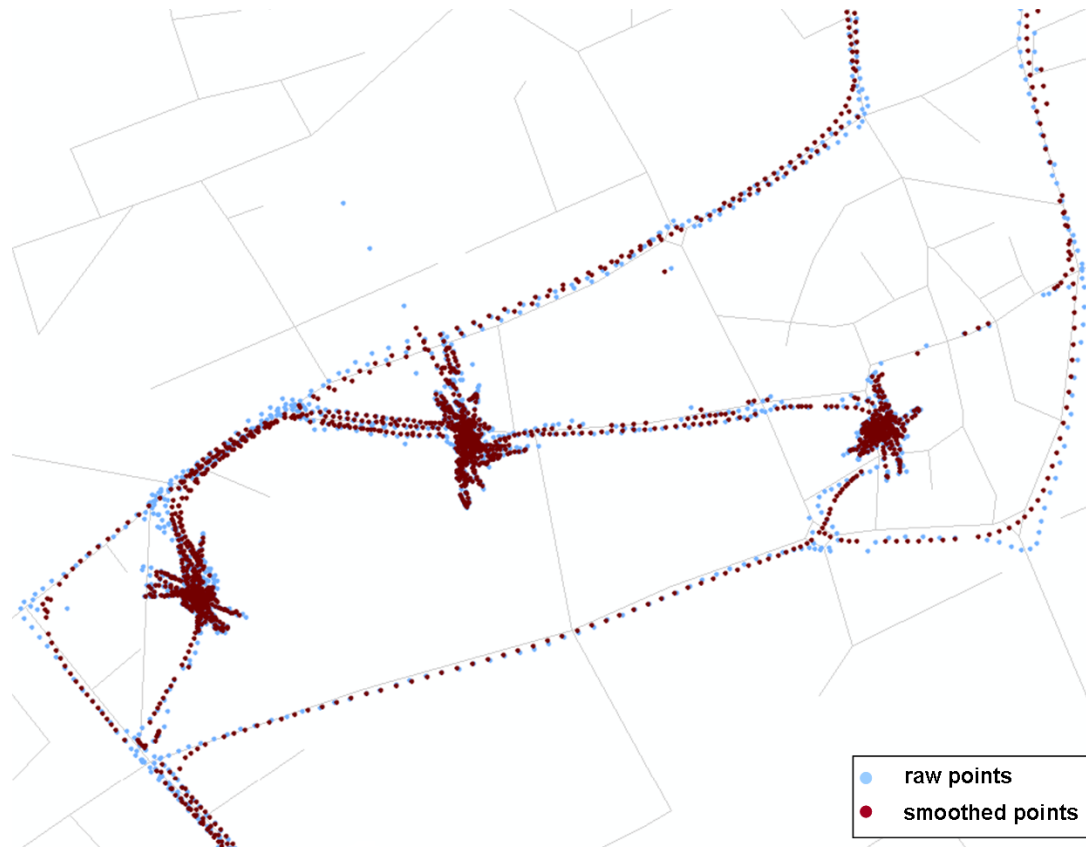
Figure 3: Comparison of speeds and accelerations from raw vs. smoothed data - Person 1548



data could be removed. In particular, the completely unrealistic speed jumps which result in acceleration values higher than  $10 \text{ m/s}^2$  are excluded without being explicitly filtered. The resulting progression of speeds and accelerations provides reasonable patterns, especially for trips in an urban environment.

Several spatial effects that occurred as expected are illustrated in Figure 4. First and foremost, it can be seen that most of the outliers, here especially noticeable at activity locations, are attenuated. Second, the overall movement trajectories can be identified more clearly, since small

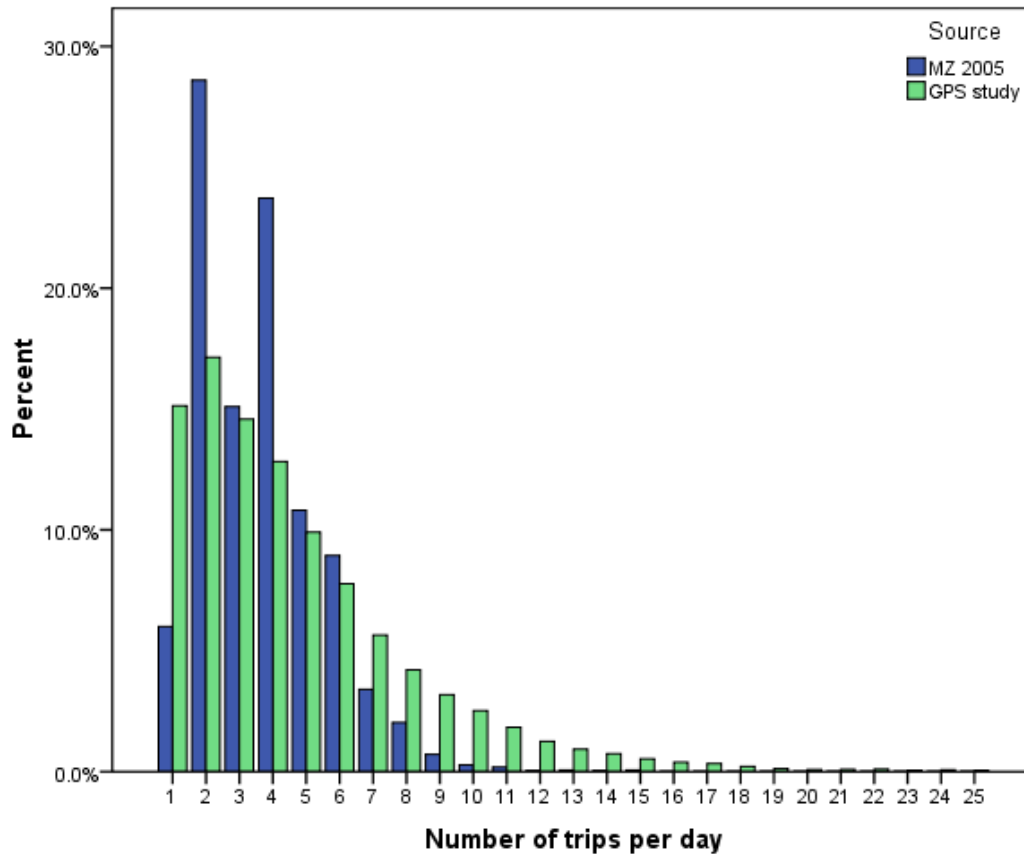
Figure 4: Comparison of point positions raw vs. smoothed - Person 1548



deviations from the general path are diminished. The downside of this is that some corners are cut stronger than in reality due to the 15-second smoothing interval. This is, however, not an important shortcoming because, first, it is not the spatial positions of the GPS points but the speeds and accelerations resulting from the overall movement trajectories that are taken into account in the trip and activity and the mode detection, and, second, the original coordinates are stored along with the smoothed coordinates. Thus, both types of coordinates can be used in the map-matching.

As already discussed in the context of the overall statistics, the average number of trips per day detected here is higher than the one reported in the MZ 2005. To investigate this further and evaluate the results of the trip and activity detection, the distribution of the trips per day is presented in Figure 5. It can be seen that the distributions are similarly skewed right although the tail of the one derived from the GPS data is slightly longer and the two distinct peaks for two and four trips per day that are present in the MZ 2005 are not replicated in the GPS data.

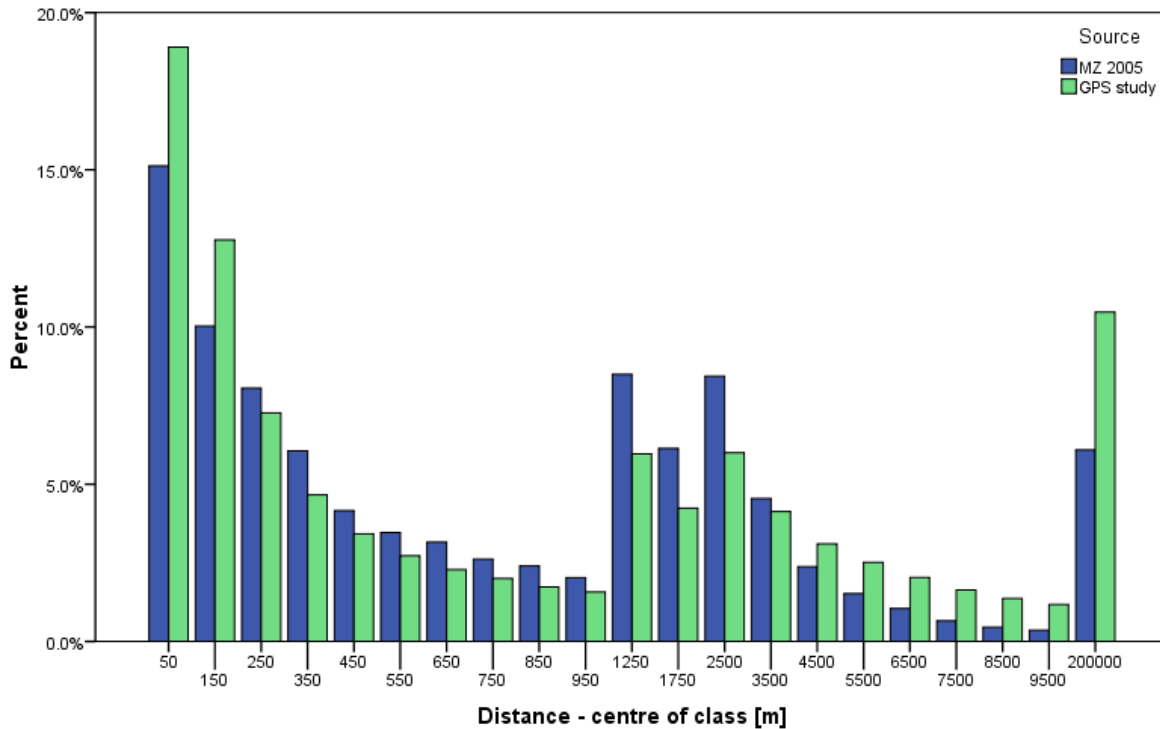
Figure 5: Trips per day compared with the Swiss Microcensus 2005



Both effects were expected. First, it has been frequently observed that short intermediate stops are omitted in recollection-based surveys (e.g. Du and Aultman-Hall, 2007; Wolf *et al.*, 2003; Yalamanchili *et al.*, 1999). Second, the length of the survey period itself influences the shape of the distribution of the trips per days. As recently shown by Madre *et al.* (2007), it approaches the normal distribution in studies covering longer periods of time. Either way, this phenomenon will be scrutinized further in combination with the identification of trip purposes.

The distribution of the stage lengths depicted in Figure 6 reveals very similar patterns in the MZ 2005 and the GPS data. The slight bias towards very short stages, i.e. stages shorter than 100 metres, is probably due to the omission of very short stages in the MZ 2005. Apart from that, the similarity between the distributions confirms that the trip and activity detection as well as the detection of stages within the trips works properly.

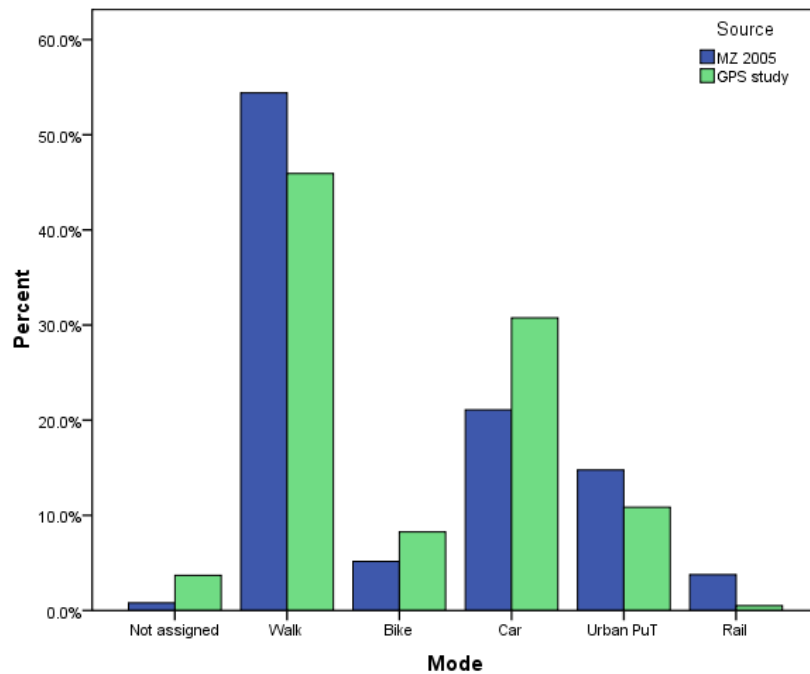
Figure 6: Stage distance distribution compared with the Swiss Microcensus 2005



In Figure 7, the distribution of the mode per stage is compared to the MZ 2005. Since the fuzzy logic algorithm delivers probabilities and not crisp values, an assumption had to be made to derive this figure. A mode is only assigned to a stage if its probability is higher than 50%. This leads to a fraction of stages for which the mode is not yet determined. Figure 7 reveals that rail is fairly under-represented at this phase of the analysis. This might be due to the design of the study which focussed on passing urban bill-boards. Thus, respondents might have been told that is not so important to keep the device active during inter-urban rail trips. However, to ensure that no actual rail trip is missed, special attention will be paid to them during the map-matching and even stages with only a low probability for rail will be matched to the rail network. The distribution of the other modes, however, is very close to the one observed in the MZ 2005.

Even closer to the MZ 2005 is the distribution of modes per stage distances, as depicted in Figure 8. In both surveys, this distribution follows reasonable patterns, e.g. shorter trips are walked or done by bike, trips longer than three *km* are predominantly travelled by car, while

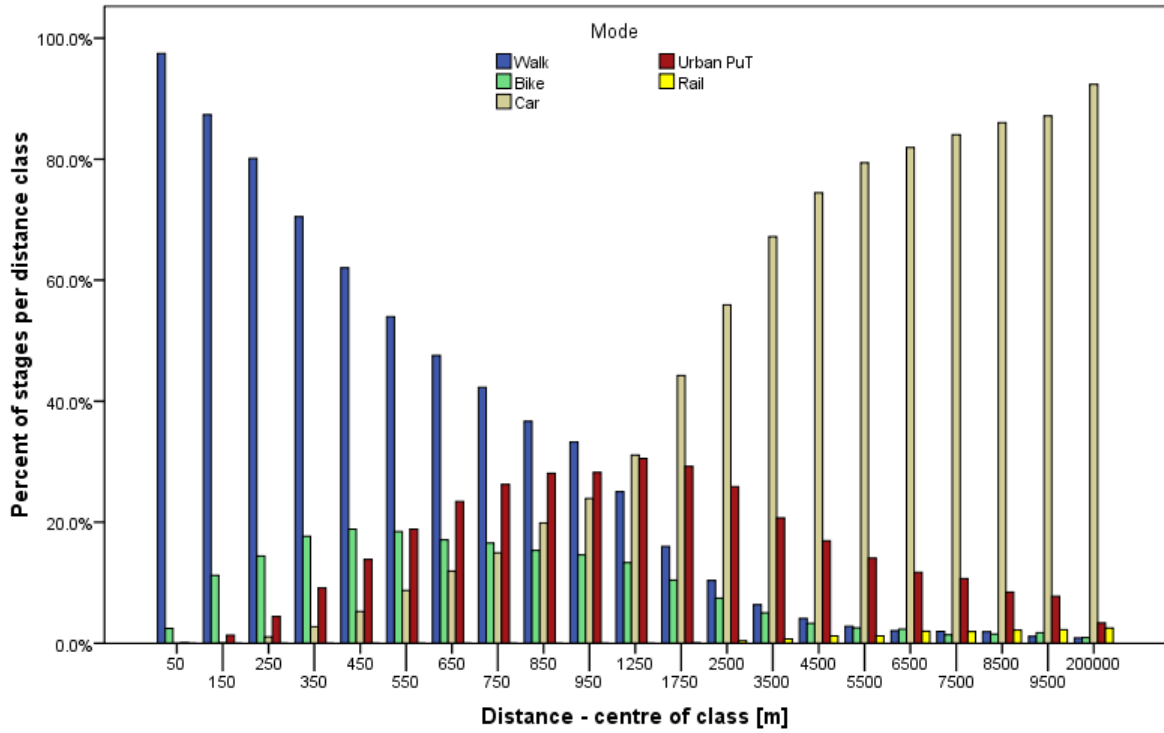
Figure 7: Mode distribution compared with the Swiss Microcensus 2005



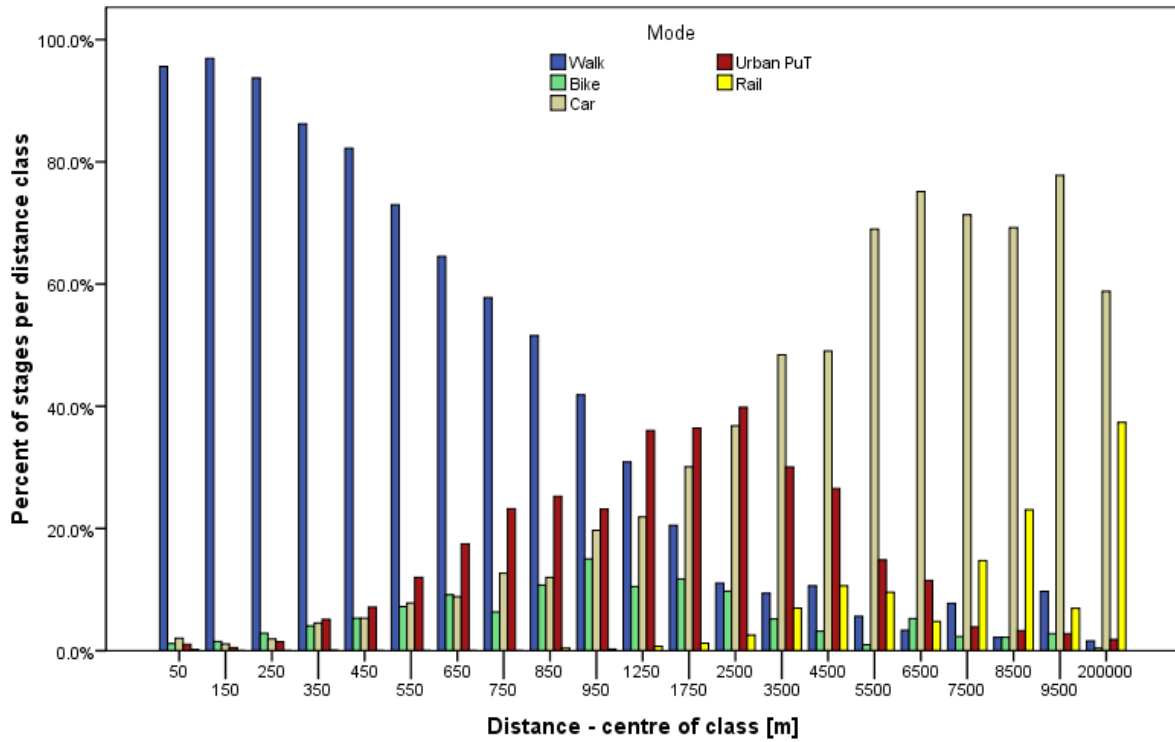
bike and urban public transport stages are most likely to be situated between 0.2 and 3 *km* and 0.5 and 5 *km* respectively. The similarity of these distributions underlines the quality of the mode detection procedure, especially since the stage distance is not used in the procedure. Only rail stands out in the comparison of these, otherwise, fairly similar distributions. Again, it can be seen that the under-representation of rail mentioned above particularly occurs for trips longer than seven *km*.

Figure 8: Mode distribution per distance compared to Microcensus 2005

(a) GPS study



(b) Microcensus 2005



## 7 CONCLUSION AND OUTLOOK

The data obtained from the post-processing procedure will primarily be used to develop models for private and public transport route choice and activity location choice. Thereby, the focus will be put on the different ways to account for similarities in discrete choice modelling. The person-based GPS dataset at hand is ideal for this purpose. It covers complete trip and activity chains over several days, thus describing the participants travel behaviour in the most comprehensive way. In addition, the sample size is large enough to obtain stable choice models. However, until now an appropriate post-processing procedure was missing that could deal with a large amount of data as well as a minimum amount of information available. The procedure proposed here meets both criteria. In addition, it is extendible in a straightforward way to account for new GPS input formats, data cleaning techniques or other post-processing modules.

A key factor for the expedience of a post-processing procedure is the implementation of appropriate filtering and smoothing techniques. Finding the right approach was essential in the study at hand because no information about the numbers of satellites in view or their positioning was available. Filtering based on altitude level, speed and acceleration was not sufficient, therefore, a smoothing method had to be applied as well. However, only a few researchers have looked at smoothing GPS points so far. This is especially surprising for person-based GPS studies because it could be shown here that a suitable smoothing method is vitally important for the adequate identification of modes, particularly if the latter relies on speed and acceleration. This also applies for speeds derived from Doppler measurements as demonstrated, for example, by Jun *et al.* (2007) or Ogle *et al.* (2002). Data cleaning methods that are still missing are the correction for the cold start/warm start problem (e.g. Stopher *et al.*, 2005) and an interpolation of missing data (e.g. Ogle *et al.*, 2002). The necessity of their implementation will be scrutinised in the course of the still outstanding analysis of activity locations and trip purposes.

The activity dwell time of 900 seconds employed in the trip and activity detection is rather high. Shorter dwell times, however, led to too many false trip ends. Instead, the detection of shorter activities is based on the identification of GPS point bundles. This accounts for the fact that in a person-based GPS study, data recording usually continues during short activities even if they are performed indoors. The trip and activity detection, which is topped off by the consideration of zero speeds for more than 120 seconds, delivers results very similar to those derived manually. A useful extension might be the consideration of heading changes of about 180° (e.g. Du and Aultman-Hall, 2007; de Jong and Mensonides, 2003) because trip ends indicated by heading changes are missed occasionally if the associated stopping time is shorter

than 120 seconds.

In Section 6, it could be shown that mode detection yields realistic results although there is no opportunity to validate them against actual behaviour. The fuzzy logic approach follows the general idea of that suggested by Tsui and Shalaby (2006). However, different variables and new rules have been applied. Moreover, the values of the parameter functions have not been automatically calibrated but derived by inspection of the speed and acceleration patterns revealed in the data and reasonable assumptions about the individual modes. In future, a check for plausible mode chains will be implemented to validate the results as well optimise the mode detection procedure. The first ideas for this have been provided by de Jong and Mensorides (2003).

The map-matching will be done with the algorithm described in Marchal *et al.* (2005) and Marchal (2008) because it produces accurate results and efficiently handles large data volumes. As a first step, map-matching will be employed externally. All stages with a certain likelihood for a mode are matched to the relevant network. The final decision on the mode will be made based on these results. In a second step, map-matching should be integrated into the post-processing procedure and feedback to the mode detection should be automated.

The last step of the post-processing procedure will be the determination of activity purposes. A couple of researchers have already worked on that (e.g. Wolf *et al.*, 2001; Schönfelder and Samaga, 2003; Wolf *et al.*, 2004; Stopher *et al.*, 2005). Their input will be reviewed to derive an approach suitable for the data available in this study. A key input will certainly be detailed land-use data. Additional criteria will presumably be time of day and duration of activities as well as their spatial clustering. Finally, the derived activity chains will be put through a plausibility check and used to scrutinise the discrepancy between the number of trips per day in the GPS data and the MZ 2005 which was discussed in Section 6.

Finally, the whole post-processing procedure will be validated with the GPS data collected by Flamm and Kaufmann (2007) as the information about stages, trip purposes and activity locations are thoroughly validated in that dataset. This will also be good opportunity to test the flexibility of the procedure with respect to other input formats and information available.

## **8 ACKNOWLEDGEMENTS**

The authors would like to thank the Swiss National Science Foundation for the funding of this work and Stefan Muff (at that time Endoxon AG) for providing the dataset. A special thanks goes to David Charypar for the invaluable discussions about GPS data and smoothing methods in general and in particular.

## REFERENCES

- Biding, T. and G. Lind (2002) Intelligent Stöd för Anpassning av hastighet (ISA), Resultat av storskalig försöksverksamhet i Borlänge, Lidköping, lund och Umea under perioden 1999–2002, *Research Report*, Vaegverket, Borlaenge.
- Bricka, S. and C. R. Bhat (2006) A comparative analysis of GPS-based and travel survey-based data, *Transportation Research Record*, **1972**, 9–20.
- 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.
- Chung, E.-H. and A. Shalaby (2005) A trip bases reconstruction tool for GPS-based personal travel surveys, *Transportation Planning and Technology*, **28** (5) 381–401.
- de Jong, R. and W. Menonides (2003) Wearable GPS device as a data collection method for travel research, *Working Paper*, **ITS-WP-03-02**, University of Sydney, Institute of Transport Studies, Sydney.
- Doherty, S. T., C. Noel, M. E. H. Lee-Gosselin, C. Sirois, M. Ueno and F. Theberge (2001) Moving beyond observed outcomes: Integrating Global Positioning Systems and interactive computer-based travel behaviour surveys, *Transportation Research E-Circular*, **C026**, 449–466.
- Draijer, G., N. Kalfs and J. Perdok (2000) Global Positioning System as data collection method for travel research, *Transportation Research Record*, **1719**, 147–153.
- Du, J. and L. Aultman-Hall (2007) Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues, *Transportation Research Part A: Policy and Practice*, **41** (3) 220–232.
- Flamm, M. and V. Kaufmann (2007) Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions, paper presented at *the 11th World Conference on Transportation Research*, Berkeley, June 2007.
- Jun, J., R. Guensler and J. Ogle (2007) Smoothing methods to minimize impact of Global Positioning System random error on travel distance, speed, and acceleration profile estimates, *Transportation Research Record*, **1972**, 141–150.

- Madre, J.-L., K. W. Axhausen and W. Brög (2007) Immobility in travel diary surveys, *Transportation*, **3** (1) 107–128.
- Marchal, F. (2008) An open-source toolkit for analysis of GPS data, paper presented at *8th International Conference on Survey Methods in Transport*, Annecy, May 2008.
- Marchal, F., J. K. Hackney and K. W. Axhausen (2005) Efficient map matching of large Global Positioning System data sets: Tests on speed-monitoring experiment in Zürich, *Transportation Research Record*, **1935**, 93–100.
- Ogle, J., R. Guensler, W. Bachman, M. Koutsak and J. Wolf (2002) Accuracy of Global Positioning System for determining driver performance parameters, *Transportation Research Record*, **1818**, 12–24.
- 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.
- Sazonov, E. S., P. Klinkhachorn, H. V. GangaRao and U. B. Halabe (2002) Fuzzy logic expert system for automated damage detection from changes in strain energy mode shapes, *Nondestructive Testing and Evaluation*, **18** (1) 1–17.
- Schönfelder, S., H. Li, R. Guensler, J. Ogle and K. W. Axhausen (2006) Analysis of commute Atlanta instrumented vehicle GPS data: Destination choice behavior and activity spaces, paper presented at *the 85th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2006.
- Schönfelder, S. and U. Samaga (2003) Where do you want to go today? - more observations on daily mobility, paper presented at *the 3th Swiss Transport Research Conference*, Ascona, March 2003.
- 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.

- Tsui, S. Y. A. and A. Shalaby (2006) An enhanced system for link and mode identification for GPS-based personal travel surveys, *Transportation Research Record*, **1972**, 38–45.
- 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, September 1997.
- Wolf, J. (2000) Using GPS data loggers to replace travel diaries in the collection of travel data, Ph.D. Thesis, Georgia Institute of Technology, Atlanta.
- 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., R. Guensler and W. Bachman (2001) Elimination of the travel diary - experiment to derive trip purpose from Global Positioning System travel data, *Transportation Research Record*, **1768**, 125–134.
- 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.
- Wolf, J., M. Oliveira and M. Thompson (2003) Impact of underreporting on mileage and travel time estimates - results from Global Positioning System-enhanced household travel survey, *Transportation Research Record*, **1854**, 189–198.
- Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira and K. W. Axhausen (2004) Eighty weeks of Global Positioning System traces, *Transportation Research Record*, **1870**, 46–54.
- 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.