


# 80 weeks of GPS-traces

## Approaches to enriching the trip information

**Working Paper****Author(s):**

[Axhausen, Kay W.](#)  Schönfelder, Stefan; Wolf, Jean; Oliveira, Marcelo; Samaga, Ute

**Publication date:**

2003-08

**Permanent link:**

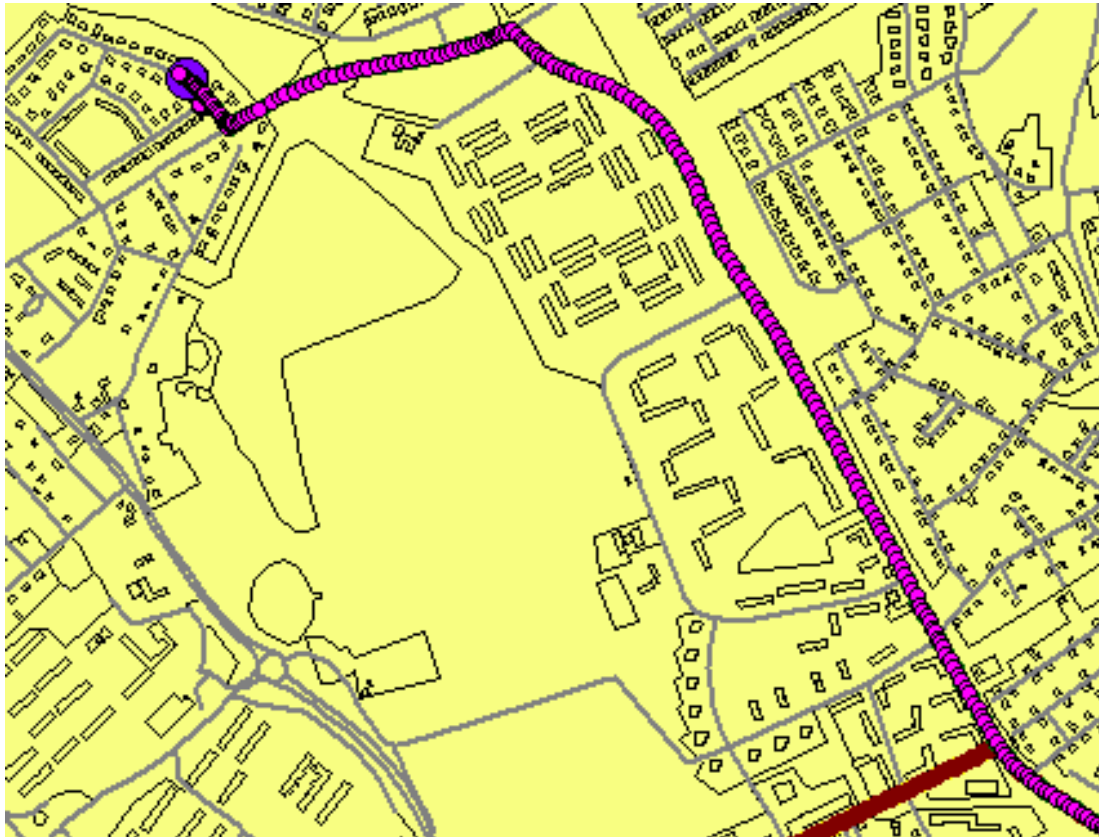
<https://doi.org/https://doi.org/10.3929/ethz-a-004570614>

**Rights / license:**

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

**Originally published in:**

Arbeitsberichte Verkehrs- und Raumplanung 178



---

Submitted to the 83<sup>rd</sup> Transportation Research Board meeting

## 80 weeks of GPS-traces: Approaches to enriching the trip information

KW Axhausen, S Schönfelder

J Wolf, M. Oliveira

U Samaga

Arbeitsbericht Verkehrs- und Raumplanung 178

August 2003

Arbeitsbericht Verkehrs- und Raumplanung

## 80 Wochen GPS-Aufzeichnungen: Ansätze zur Vervollständigung der Wegebeschreibungen

KW Axhausen, S Schönfelder  
und U Samaga  
IVT  
ETH  
CH – 8093 Zürich

J Wolf und M Oliveira  
GeoStats  
USA – Atlanta, GA 30318

Telefon: +41-1-633 3943      Telefon: +1-404-588 1004  
Telefax: +41-1-633 1057      Telefax: +1-404-588 1227  
axhausen@ivt.baug.ethz.ch      jwolf@geostats.com

August 2003

### Kurzfassung

Seit Jahren experimentiert die Mobilitätsforschung erfolgreich mit dem Einsatz von Global Positioning Systemen (GPS) bei der Erhebung des individuellen Verkehrsverhaltens. Die Nutzung von GPS in Verbindung mit geographischen Informationssystemen (GIS) ermöglicht für ausgewählte Charakteristika von Wegen und Aktivitäten (z.B. Routenwahl, Distanzen und Dauern) ein Mass an Exaktheit, das herkömmliche Verkehrsbefragungen auf Basis von Interviews oder Wegetagebüchern nicht erreichen können. Die Länge der Erhebungszeiträume und die Stichprobengrösse der Befragten bzw. ausgerüsteten Fahrzeuge sind dabei stetig gewachsen.

Dieser Arbeitsbericht stellt einen aktuellen GPS-Datensatz vor, der im Rahmen der schwedischen ISA-Verkehrssicherheitskampagne (Intelligent Speed Adaptation) erhoben wurde und von dem eine substantielle Teilstichprobe derzeit für die Analyse des Verkehrsverhaltens aufbereitet wird. Die zu analysierenden Daten basieren auf Aufzeichnungen der Bewegungen von ca. 190 privaten Fahrzeugen aus der mittelschwedischen Stadt Borlänge, die über einen Zeitraum von bis zu 2 Jahren mit einem GPS-Empfänger ausgestattet wurden. Für die Untersuchung des langfristigen Verkehrsverhalten der Fahrer stehen sekundengenaue Weginformationen wie Abfahrts- und Ankunftszeiten, Weg- und Aktivitätendauern sowie Positionen zur Verfügung.

Der Bericht stellt die angewandten Verfahren zur Ergänzung strukturell fehlender Angaben wie eindeutiger Wegdestination und Wegzweck sowie der Korrektur von Übertragungsfehlern vor. Schwerpunkt ist die Imputation des Wegezwecks, sowie die Identifikation ursprünglich nicht erkannter Wege.

### Schlagworte

GPS Aufzeichnungen, Wege, Imputation, Wegezweck, Ziele, Borlänge

### Zitierungsvorschlag

Axhausen, K.W., S. Schönfelder, J. Wolf, M. Oliveira und U. Samaga (2003) 80 weeks of GPS-traces: Approaches to enriching the trip information, *Arbeitsbericht Verkehrs- und Raumplanung*, **178**, Institut für Verkehrsplanung und Transportsysteme, ETH Zürich, Zürich.

Working paper

## 80 weeks of GPS traces: Approaches to enriching the trip information

KW Axhausen  
S Schönfelder  
U Samaga  
IVT  
ETH  
CH – 8093 Zürich

J Wolf  
M Oliveira  
GeoStats  
USA – Atlanta, GA 30318

Telephone: +41-1-633 3943  
Telefax: +41-1-633 1057  
axhausen@ivt.baug.ethz.ch

Telephone: +1-404-588 1004  
Telefax: +1-404-588 1227  
jwolf@geostats.com

August 2003

### Abstract

The recent Swedish Intelligent Speed Adaptation (ISA) study included a component that involved the installation of GPS-based units in hundreds of cars in three Swedish cities (Borlänge, Lund and Lidköping) and observed these vehicles for up to two years. In Borlänge, the speed and location data of each vehicle were transmitted at regular intervals to a central server and stored for later analysis. This dataset contains a wealth of travel behavior information that, to date, has never been available. However, a dataset of this magnitude introduces a major need for automated processes that can glean the travel behavior details out of the trip summary and GPS point files collected for such a large scale study. This paper presents a summary of the characteristics and issues with the Borlänge GPS dataset, which includes 186 personal vehicles that have at least 30 days of travel data and corresponding household socio-demographic data. (These 186 vehicles recorded 49,667 vehicle days of travel and 240,435 trips inside the study area.) The paper then presents automated methodologies for imputing trip purpose for these trips once the trip destinations are identified, as well as correcting the GPS traces and identifying missing trips (or trip ends) within these trips. Results of these automated processes for a subset of the ISA study vehicles are included.

### Keywords

GPS traces, trip and destination identification, purpose imputation, Borlänge

### Preferred citation style

Axhausen, K.W., S. Schönfelder, J. Wolf, M. Oliveira and U. Samaga (2003) 80 weeks of GPS-traces: Approaches to enriching the trip information, *Arbeitsbericht Verkehrs- und Raumplanung*, **178**, Institut für Verkehrsplanung und Transportsysteme, ETH Zürich, Zürich.

# 1 Too much of a good thing?

The travel behavior researcher faces GPS data like a child faces a candy store – there is so much there, that it is difficult to get started. While previous GPS studies tended to be small in sample size and short in duration (Wagner, 1997; Wolf, 2000), recent studies have increased both substantially – from dozens to hundreds of study participants and from a couple of days to months and years in length. One of these recent studies is the Rätt Fart project (part of the Swedish Intelligent Speed Adaptation (ISA) study, see <http://www.isa.vv.se/index.en.htm><sup>1</sup> or see Vägverket (2000), which installed GPS-based units in hundreds of cars in three Swedish cities (Borlänge, Lund and Lidköping) and observed the vehicles for up to two years. Using the Borlänge dataset as a basis, this paper presents highly automated methods suitable for processing large-scale datasets for subsequent travel behavior analyses.

The ISA study is concerned with the traffic safety effects of in-car speed information systems. The three systems, a different one installed in each of cities, informed the driver in real-time about violating the posted speed limits by a blinking light, by a sound, by increasing the resistance of the gas pedal, or by combinations thereof. In each case, an in-vehicle unit measured the location and speed of the vehicle by GPS, looked up the posted speed limit from a suitably enriched network database, and informed the driver of speed violations when they occurred. In Borlänge, the speed and location data of each vehicle were transmitted at regular intervals to a central server and stored for later analysis. While speed, location and time-of-day are the only variables needed for the analysis of safety impacts, these are not sufficient for travel behavior analyses. Specifically, this dataset is missing the clear identification of trip destinations and their corresponding trip purposes. In addition, since trip ends were only recorded for engine on / off events, one has to check the GPS traces to detect abandoned trips and trips associated with short duration activities for which the engine was not switched off.

Earlier studies addressed these issues through a combination of manual and computer-assisted data processing steps (Wolf, Guensler and Bachman, 2001; Wolf, Loechl, Thompson and Arce, 2003). This is impossible for a total of about 50,000 vehicle days (private cars only), as available in the Borlänge data set. This paper reports on initial steps to fully automate these data cleaning and data imputation tasks, which include complementing and correcting the traces, trip end detection, identification of destinations, and trip purpose imputation.

---

<sup>1</sup> See Biding and Lind (2002) for the main study results.

## 2 The Borlänge dataset

As mentioned previously, the Borlänge data were collected as part of the ISA project. The project aimed to recruit 300 private and commercial vehicles for its test of the in-vehicle speed advisory system, which combined a light and sound warning that was activated whenever the system detected a posted speed limit violation. The study recruited drivers and their vehicles from households in which each licensed driver had his or her own car in an attempt to minimize vehicle sharing or swapping.

The system matched every valid GPS point to a road network link and generated a linear measurement from the start of the link. (Obviously, the accuracy of the underlying network had a major impact on the outcome of this ‘onboard’ link matching process.) The link, offset, and speed combinations were stored at one-second or ten-second frequencies, first in the in-vehicle unit and later, permanently, on the central server. The start and end of a trip were detected and recorded based on ignition on / off events. While necessary from the point of view of the ISA study objectives, this data storage process is sub-optimal for all subsequent analysis, as all errors performed by the logic of the in-vehicle unit are imbedded in the only saved version of the data, while none of the original GPS data (such as x, y coordinates, heading, UTC date and time stamp, number of GPS satellites visible to the GPS antenna) were kept. The most critical errors contained within the data are:

- Gaps in the trace (no identification of link and linear measurement)
- Erroneous identification of the link (wrong approach at an intersection, wrong direction along a road with a median, wrong road segment on a roundabout)
- Trip truncation due to out of area travel
- Off network travel matched to nearest network link

These last two errors occurred because the in-vehicle system was designed: 1) to stop recording the GPS trace whenever the vehicle left the study area (the city plus a radius of about 25 km surrounding it); and 2) to snap or match all non-public network travel to the nearest network link (since links were the standard storage unit and off-network links and areas were not included in the onboard road network). Both conditions generate problems, albeit different ones. A substantial number of the trips recorded left the study area, not only for commutes to neighboring towns, but also for other types of trips (visits to the holiday home, for business, for shopping, etc.). This means that only part of the total travel behavior of the participants, or, more precisely, their vehicles, has been captured.

The lack of recording on non-public roads means that the final destination in these cases cannot be confidently identified. Examples of this are visits to large housing estates or apartment complexes, to factories, or to shopping facilities with private parking lots. In addition, due to the 10-second recording intervals, there is incomplete network travel details in some traces prior to the point at which the driver leaves the public network. Finally, in the event of GPS cold start delays, it is difficult, if not impossible, to recover these points from the trace of the next trip.

The system identified the start and end of a trip by the engine on/off signal it received. This methodology for trip end identification could lead to erroneous or false trip ends (such as engines switched off while waiting for a queue to dissolve or for a light to turn green, or engine stalls) and to missing trip ends when the engine is not switched off (e.g., dropping somebody off, picking somebody up, brief shopping activities or personal business, or posting a letter). The lack of an engine off event in these cases could be due to the very short duration of the activity or to the outside temperature (Borlänge is situated at 61°N, which is slightly south of Anchorage, Alaska). In addition, this trip identification methodology does not identify abandoned trips, i.e. detours without any obvious purpose, which are present in typical travel behaviors.

Trip purpose and final destination identification are missing by definition from a vehicle-based GPS trace. Where an individual parking lot or space can be associated with a particular building/complex this is less of a problem, but when this is not possible, as in multiple-use environments in traditional urban cores, or when a person might park at different locations due to the lack of a private parking space at home or because of overcrowding of the dedicated parking lot, then it is necessary to group different final recorded locations into an unique destination. Only afterwards is it possible to impute a purpose to the specific destination.

### **3 Brief statistical summary of the complete dataset**

The Borlänge dataset contained 186 private vehicles for which GPS traces as well as minimum socio-economic information of the driver were available and which were observed more than 30 days. The statistics presented in this section are based on the data for these 186 vehicles, with account for nearly 50,000 vehicle days and a quarter of a million trips. Here are a few statistics of this dataset for vehicles that had more than 30 days of observations recorded between 22 June 2000 and 4 March 2002.

- 186 private vehicles traced (with minimum socio-economic data of the driver(s) available)
- 49,667 vehicle days of vehicles
- 240,435 trips inside the study area (starting and ending within the monitoring area)
- 9873 trips ending or beginning outside the study area

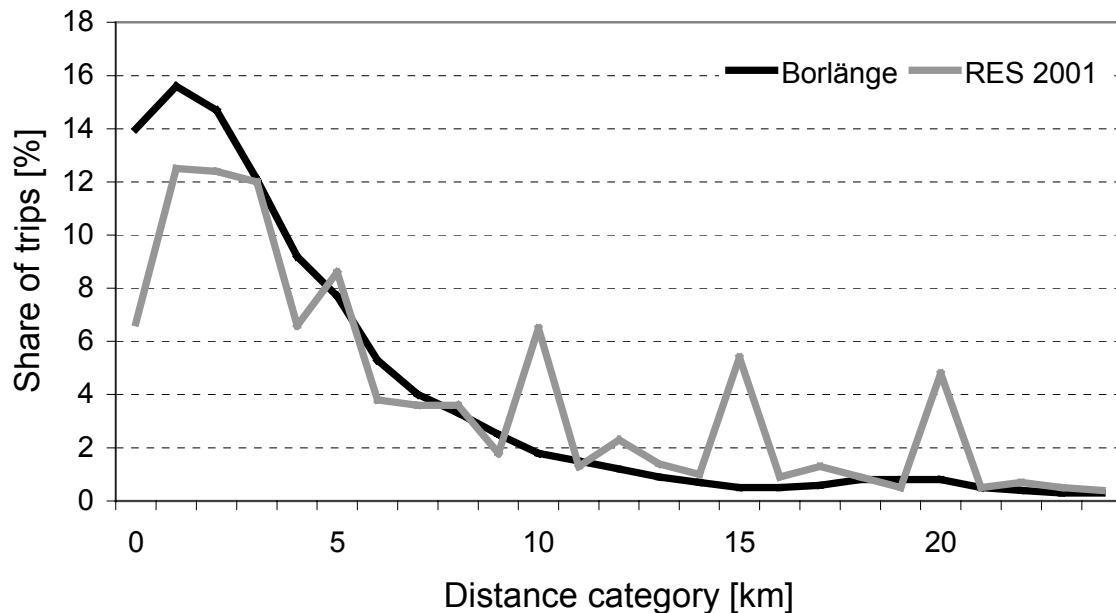
The average number of 5.0 daily trips per vehicle (mobile days only) is not consistent with the estimate of 3.3 from the 2001 sample of the Swedish National Travel survey RES<sup>2</sup>. The estimate for the Borlänge data is based on the raw data. The difference might be due to a number of reasons, including self-selection of highly mobile drivers into this study, underreporting in the RES, local variation, and small sample effects.

Due to the cut-off at the 25-kilometer radius from the center of Borlänge, it is difficult to compare the complete trip length distribution, but up to these terminating lengths, the distribution is comparable to national Swedish numbers (see Figure 1). The graph clearly shows the rounding effects for travel distances in usual travel diary data such as the Swedish RES. Reported trip distances are not distributed continuously but with significant peaks around the 5, 10, 15 and 20 kilometer (km) marks. This is not the case for the exact distance measurements provided by GPS data collection methods.

---

<sup>2</sup> The RES was provided by SIKa Institut. See also SIKa Institut, 2001 for details on sampling and other results. Base: Car trips < 25km, respondents with permanent car availability

Figure 1 Comparison of trip length distributions up to 25km (Borlänge GPS and 2001 RES)



Borlänge: 232,842 trips without trips of zero distance

RES: car trips only of respondents with permanent car availability (5,567 trips)

## 4 Activity Purpose Imputation

For the activity purpose imputation task, a dataset was created containing all data for 39 vehicles/persons. These 28 fulltime workers and 11 retirees were selected because each had a full description of their socio-demographic background and because their exact home addresses were provided. A side benefit of using two disparate groups such as these is that their obviously different travel demand structure enables better testing of the data imputation procedures presented here. These differences are especially true with respect to the different levels of regularity in location choice and the amount of disposable time for leisure and shopping.

The trip purpose imputation process was conducted at ETH by IVT and consisted of the identification of unique origins and destinations based on the final positions of the observed vehicle movements, and a multi-stage approach to impute the missing trip purposes (see Schönfelder and Samaga, 2003 for full details).

## 4.1 Identification of destinations

The coordinates of the final points captured for trips to the same destination can vary significantly for several reasons, including the use of different parking spaces at different occasions. In addition, in the Borlänge data set, the final point recorded when leaving the public road network might vary because of the ten second interval between recordings or because the person approached the private estate from different entrances. One would therefore expect that spatial clusters of such final observation points are associated with unique destinations of the travelers, destinations that need to be identified before one can proceed to impute a purpose for the trip.

The Borlänge land use maps identified public parking lots as well as some associated with private shopping and leisure facilities. Given this information, it was useful to treat those trips for which the last reported position is in a parking lot separately from the others. The remaining final recorded trip positions were grouped into locations. This grouping was neither too vigorous nor too light to avoid having either too few locations or too many, implying also too few or too many different activity types.

Clustering approaches of various complexities were conceivable. The one chosen here is straightforward (Schönfelder and Samaga, 2003). The distances from each last location to all others within a radius of 200 meters from the respective location were calculated using ARC/INFO, a geographic information system (GIS). Those that have the most neighbors (plus the smallest average distance to all other locations considered) were classified as cluster centers. All other less central locations were assigned to the predefined cluster medians. In cases where a location was associated with more than one cluster, it was assigned to the nearest cluster center.

## 4.2 The imputation process

GPS traces obviously do not include activity purpose information. Requesting this information in parallel with GPS instrumentation would defeat the purpose of observing the driver/traveler as unobtrusively as possible over a period as long as possible. For shorter duration studies this logic might change (see Stopher, Bullock and Horst, 2003 and Wolf, 2003 for state-of-the-art prompted recall techniques). In these feedback methods, the GPS information could be complemented with the provision of habitual destinations by the participants, which could facilitate the purpose imputation process.

The imputation of activity purposes based on the information about the trips associated with the activities has rarely been addressed in the literature. Waßmuth (2001) looked at the issue in the context of having surveys of different coding details for the activity purpose. Wolf's (Wolf, 2000; Wolf *et al.*, 2001) approach is structurally similar to the situation here, as she imputed purposes based on GPS traces, but she could exploit the relatively unambiguous land-use patterns of US suburbia (which have a close match between parking and activity opportunity). In addition, she based her work on a very detailed, parcel fine, land use database.

In the attempt to impute the activity purposes, the analyst has various sources of information available, which he can exploit (see Table 1). Ideally, one should exploit all the available information in one step, but due to the different levels of precision and the lack of a suitable approach, sequential approaches are sufficient at this time. One of the largest stumbling blocks to the development of a suitable parametric estimation approach is the lack of spatial information in traditional travel diaries, especially the lack of information about the distance between parking and activity locations.

Table 1 Information sources for activity purpose imputation

Variable	Comment
From the GPS traces (internal)	
Location (of parked car)	Within the limit of the identification of unique locations
Duration	Assuming that there is no change of activity type and location between the two observed (car) trips
Time of day	See above
Day of week	
Frequency of visit by type of day	Precision depends on the duration of the observation period and the level of detail of the classification (type of day, duration, time of day)
From associated surveys	
Age, sex, education	
Profession and working hours	
Hobbies and social commitments	These items are unusual, but would be very helpful for the identification of leisure activities
Home locations of friends and relatives	See Schlich, Kluge, Lehmann and Axhausen, 2002 for an example
From external sources	
Relevant cross sectional travel diary survey	Provide purpose probabilities for activity classes by trip duration, activity duration, time-of-day, day-of-week and person type  Provide information about the transition probabilities between different activity purposes by time-of-day and day-of-week.  Normally no information about the distance between activity location and parked car; often no detailed location information at all
Relevant panel travel diary surveys	Above and by frequency and variability of the other dimensions
Parcel information	Type of use; maybe even type of store or facility (chain)
Land use map	Type of use by generic category

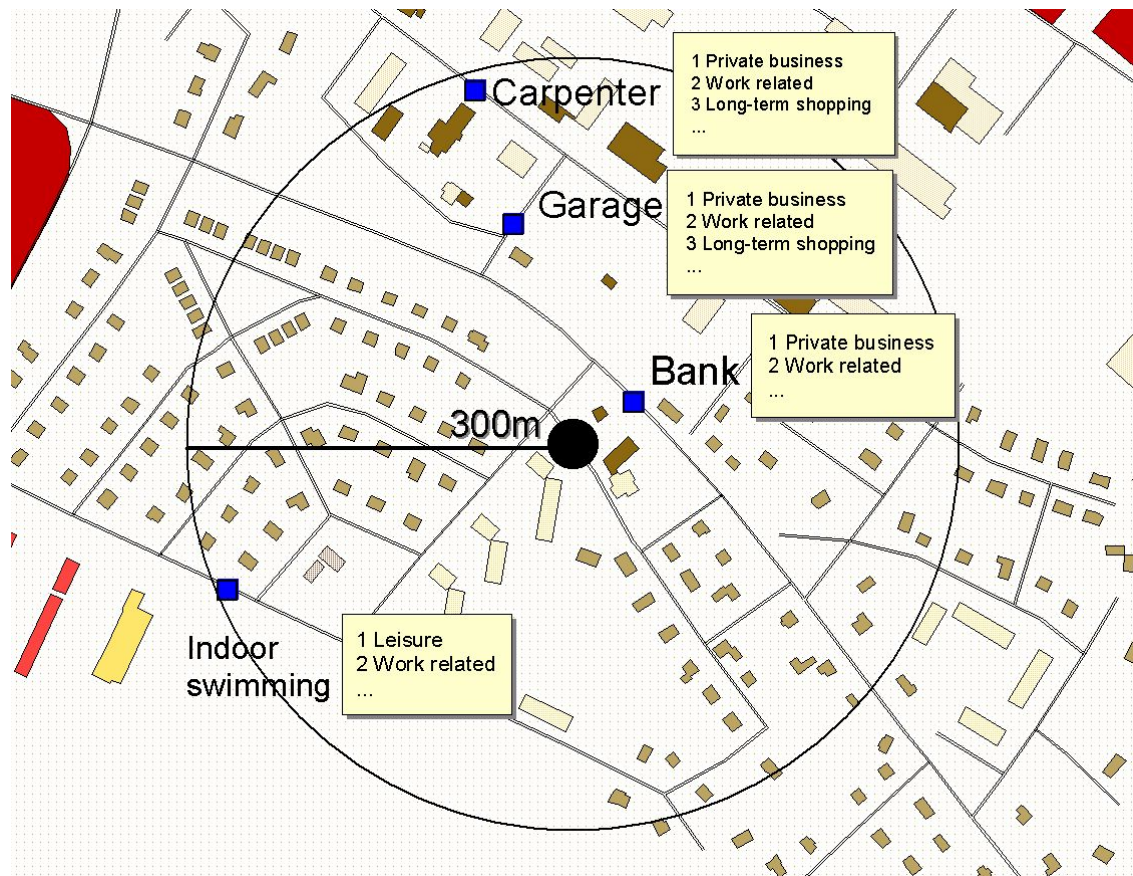
A natural framework for the integration, or fusion, of the different data sources is a Bayesian approach (see for example Jensen, 1996 in combination with a Monte-Carlo selection of the purposes. This would also lead naturally to multiple imputations (Little and Rubin, 1987; Rubin, 1987). In the initial approach developed so far, probabilistic elements have not been included. The most likely outcome is chosen deterministically.

A straightforward identification of the activity purpose is possible when the cluster center is spatially identical or very close to the driver's household location (<200m). In these cases, the cluster centers as well as the underlying trips are assigned with the activity purpose *Home* or *Return Home*. Although some other activities are performed in the direct vicinity of home, there is a high probability that the car is driven home if parked nearby. Probable misassignments that result from this rigid assumption were accepted.

The assignment of the remaining activity purposes followed a multi-stage approach and focused on a deeper analysis of the cluster centers identified in the preceding step. The approach considered the socio-demographic background information available for the drivers, the existing land use data for the town of Borlänge, and the national travel survey (RES).

As a second step, the cluster centers were associated with the available point of interest (POI) data such as restaurants, and petrol stations, and the Borlänge land use pattern (provided in polygons); this association can be seen in Figure 2. Initially, each POI and land use class was given a certain probability for a potential activity purpose – by a user-defined / consistent tabulation.

Figure 2 Identification of potential trip purposes by land use



Source: Schönfelder and Samaga, 2003

For the POI analysis, all cluster centers were buffered within the GIS with a distance of 300m (a distance likely to be accepted for a walk from a vehicle's parking place to a destination) in order to define an evaluation area for potential activities. If a POI was found within the predefined radius, the related purpose probabilities were assigned to the cluster center and stored. If more POI's were found within the radius, the probabilities were summed by the distinct purposes with a higher weight given to points of interest closer to the cluster center. As a first heuristic, points within 50m were given the weight 1.5; between 50m and 100m the weight 1, between 100m and 200m the weight 0.7 and 0.4 between 200m and 300m.

The cluster center to land use polygon comparison was implemented in a similar way. Again, the cluster centers were buffered using a 200m radius to consider the adjoining land use type

and the predefined activity purposes in the evaluation area. The purpose probabilities (or values) of all distinct land uses found were ordered by highest purpose probability and stored.

Finally, the structural and temporal characteristics of the trips were compared with the 2000 and 2001 Swedish national travel survey data (Sika Institute, 2001). The logic behind this approach was: Given a car movement with certain temporal attributes such as a travel time of A and an activity duration of B made by a person with particular socio-demographic attributes X, Y and Z, what is the most probable activity purpose for this combination.

To find a reasonable linkage for the Borlänge GPS data, a multi-dimensional table was created from the Swedish national travel diary data. The tabulation included the variables sex and car availability of the traveler, his/her occupation status, day of week, trip starting time and the activity duration, and yielded for each of the table cells the mode for the activity purpose. This value was assigned to each of the clusters as a potential purpose.

The preceding analysis and comparison steps yielded a range of purpose assignments that had to be consolidated in the final step for each cluster center. This was done using five sequential rules:

1. Clusters that could be associated with the traveller's home address were automatically categorised as *Home*.
2. For fulltime workers, the purpose *Work* was assigned if a) cluster center was the second most frequented of all, b) the comparison with the national travel data was positive for the purpose *Work* and c) the activity took place on a weekday (Monday to Friday)
3. Clusters that yielded the same purpose probability by land use, POI and temporal characteristics comparisons were assigned this purpose.
4. If there existed a difference between the POI, land use based purpose assignment and the temporal matching, the POI, land use categorization was preferred. This rule was ignored if the temporal characteristics yielded the activity purpose *pick up / drop off* which is independent of land use at the location of the stopped vehicle.
5. If there was no clear POI/land use assignment possible, the purpose assignment followed the categorization offered by the temporal characteristics.

The steps were entirely implemented in the ARCINFO GIS environment using the ARC Macro language (AML) (see Samaga, 2003). The whole procedure was designed to allow easy modification of the imputation framework by adding more sophisticated identification

steps such as *discriminant analysis* for the temporal matching or a Bayesian updating of the probabilities.

### 4.3 Some initial results

The procedures were applied to the previously described data collected from 39 vehicles. For this first post-processing of the trips, rigid selection rules were applied (in retrospect, overly rigid rules). The technical shortcomings of the data collection led to a large amount of observed trips and related activities (i.e. dwell times between the trips) with unrealistic attributes such as too short trip durations or improbable speeds. The solutions to these problems were already comprehensively discussed elsewhere (Wolf, 2000; Wolf *et al.*, 2001; Pearson, 2001), especially by defining thresholds of (minimum/maximum) durations which initiates elimination rules. The selection rules were:

- total travel time greater than 30 seconds
- activity duration greater than 120 seconds
- trip length shorter than 25km
- average speed lower than 50 km/h
- trip starts AND ends in the monitoring area

The last requirement potentially leads to ambiguous results. On one hand, it removes trips for the further analysis with no definable / clear destination and activity duration. On the other hand, it “destroys” complete trip chains that can lead to a misinterpretation of the overall daily activity patterns. This will be shown later and has to be improved in further, more sophisticated, analysis steps.

The imputed GPS trip data does not entirely correspond with available cross-sectional travel data that is not illogical if considering the unique longitudinal structure of the Rätt Fart data, the limitations of the small test sample, and the ad-hoc extraction of the reference database. Whereas the number of car trips in this sub-sample is consistent with the information provided by the Swedish national travel survey (RES) (after applying the rigid cleaning procedures), the Borlänge drivers made considerably shorter drives with respect to both trip distance and trip duration (see Table 2). At this stage of the analysis, it is difficult to assess if the GPS data collection method yields systematically different results compared to ordinary paper based travel diaries or if the recruited test drivers in fact show a dissimilar travel behavior including a different structure of daily car usage. Further investigations are required.

Table 2 Post-processed Borlänge GPS data base compared with 2000/2001 RES data (Swedish national travel survey) \*: Selected mobility characteristics

Variable	Retirees		Fulltime workers	
	RES	Borlänge GPS	RES	Borlänge GPS
Mean number of daily trips (Std.)	3.8 (3.8)	4.3 (2.8)	4.8 (5.7)	3.9 (2.4)
Mean daily trip distance [km] (Std.)	24.6 (33.7)	16.3 (12.5)	32.0 (45.3)	13.2 (11.3)
Mean daily trip duration [min] (Std.)	44.1 (55.2)	31.8 (22.4)	58.7 (82.1)	24.4 (18.2)

\* Local car trips made by respective groups; Sample sizes: RES weighted by sex and age, N(Fulltime workers) = 1516, N(Retirees) = 440; Borlänge: N(Fulltime workers) = 28, N(Retirees) = 11

The shares of the single purposes (10 categories) principally show the same pattern as the (weighted) Swedish reference data (see Table 3). The main problems are the differences in the shares of the PRIVATE BUSINESS, WORK RELATED BUSINESS and DAILY SHOPPING purpose. This may be due to many things. For example, the aerial land use information and the available point of interest repertoire that are important bases for the trip purpose assignment procedure allow an ambiguous and partly fuzzy categorization of the detected trip ends. Buildings with a certain land use categorization may be visited by the drivers because of several reasons – for example, a bank can act as work place, a place to do private business, or a location where family members may be picked up or dropped off. Even if integrating the temporal characteristics of the respective trip (e.g. start time or activity duration), the trip purpose assignment may be misleading in some cases. Furthermore, the land use database as well as the points of interests available appears to be insufficient at this stage and have to be completed in the next imputation steps. Finally, it should be noted that the test sample is extremely small and therein may contain considerable biases.

Table 3 Initial Borlänge GPS trip purpose assignment compared with 2000/2001 RES data (Swedish national travel survey data): Trip purpose shares [%]

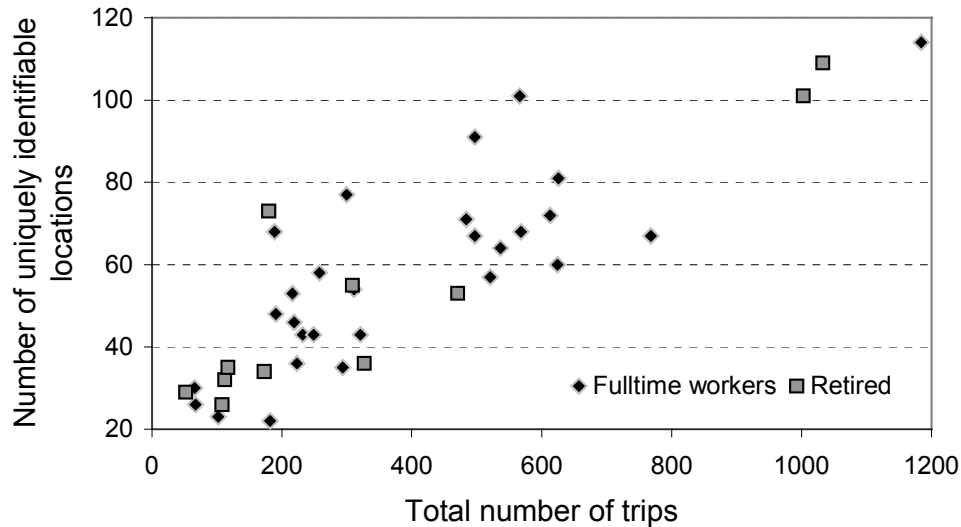
Trip purpose	Retirees		Fulltime workers	
	RES	Borlänge GPS	RES	Borlänge GPS
Pick up / Drop off	6.8	7.0	8.9	8.6
Private business	4.6	10.0	3.7	8.2
Work related	0.1	9.8	8.3	5.7
School	-	0.1	0.0	0.1
Work	0.3	-	16.0	10.2
Daily shopping	12.4	4.4	6.1	1.5
Long-term shopping	8.6	7.3	5.7	7.9
Leisure	20.7	23.6	10.6	20.4
Other	5.2	-	4.1	-
Home	41.3	37.8	36.6	37.4

\* Local car trips made by respective groups; all days; RES weighted by sex and age, N(retirees) = 1516, N(fulltime workers) = 440; Borlänge: N(Fulltime workers) = 28, N(Retirees) = 11

Nevertheless, the imputed Borlänge trip data confirms earlier findings of the *Mobidrive* analysis work on long-term spatial mobility. This is true for both, the number of unique locations the test drivers go to and the amount of variety seeking in location choice. The share of trips to unique locations was previously unknowable, as cross-sectional surveys cannot provide a credible estimate of this parameter. The available long-term travel data now permit an impression of this aspect of spatial choice behavior (see Figure 3). If the number of unique locations grows consistently with the number of trips, then variety seeking, for its own sake, becomes a credible explanation of these choices<sup>3</sup>.

<sup>3</sup> The Borlänge data indicates that there seems to be an almost unlimited number of places people know, because even after half a year of reporting there are still places the drivers “discover” as new (i.e. never visited before) destinations

Figure 3 Total number of trips versus the total number of uniquely identified locations



Source: Adopted from Schönfelder and Samaga, 2003

## 5 Complementing and correcting the traces

GeoStats, a company specializing in GIS-based processing of GPS travel data, performed an automated trip review and identification analysis on a subset of the Borlänge GPS data set. This section presents an initial assessment of the Borlänge GPS dataset with respect to this task, the data processing procedures developed for this dataset, and some preliminary results.

Trip Identification and Analysis System (TIAS) is GeoStats' in-house GPS travel data processing software system. It was built using Visual Basic .NET, ESRI's MapObjects 2.1 GIS library and is supported by a relational database server. The initial versions of TIAS were designed for working with somewhat small data sets (travel over one day to one week), so modifications were needed to properly deal with the massive size and time coverage of the Borlänge data sets. The updated version computes trip data one travel day at a time, saves the results incrementally, and performs summary computations and trip numbering at the end of the process.

TIAS was developed based on the assumption that the input data would consist of GPS points. The linear referenced nature of the Borlänge data goes against this assumption, so a data conversion process was developed to translate the vehicle-log files into points. The link ‘points’ were processed sequentially and converted to absolute coordinates by combining the relative position of each point to its matched network link with the absolute position of that link.

Once this conversion was complete, an initial review of the data points indicated that link match problems during data collection were frequent. These problems occurred more often around intersections, where the GPS point was often matched to the wrong side of the intersection. Another frequent matching problem occurred along parallel road segments. In these situations the point matches alternate between parallel roads, indicating lateral motion that did not occur.

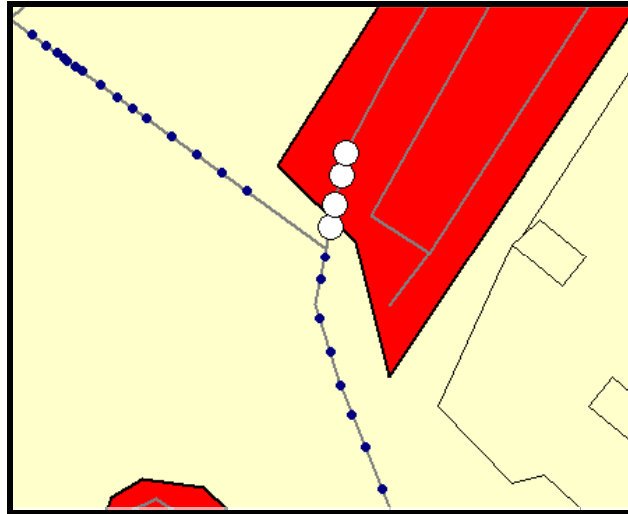
One other problem found in the dataset was inconsistent logging intervals, which occurred at either 1-second or 10-second frequencies. The 10-second data were much more difficult to analyze. Based on the information provided, the equipment logging frequency was determined by an attribute present on the road network links and the GPS computed speed of the vehicle, such that vehicle positions were logged at one-second intervals whenever the vehicle was moving above the posted speed limit. However, this rule does not appear consistent with the frequencies evident in the dataset.

## 5.1 Point Filtering Procedures

A set of filtering procedures was developed to deal with the link matching issues; these filters were rule-based and employed consistency checks on the sequence of individual points and links. Figure 4 shows an example of one of the cases covered by the filters (note that the filtered points are displayed as white circles). Points that were recognized as erroneous were marked as ‘filtered’ and were not included in the subsequent trip end detection and analysis process.

Figure 4 False deviation from path

---



---

## 5.2 Data Set Summary Statistics

A total of 24 vehicle-log files were made available to GeoStats and imported into the TIAS database. These vehicles were not selected from the set of 39 used for the trip purpose imputation processes; instead, they were created from the original database containing 169 vehicles – this occurred as a result of analysis task timing constraints and technical difficulties in extracting data from the original ISA database. Table 4 summarizes the data from the 24 datasets that were successfully imported into TIAS.

Table 4 Summary of Vehicle File Import Results for 24 Vehicles

Vehicle	Trips recorded	Days observed	Number of points	Percent logged at 1 second	Filtered points	Share of filtered points
1	395	121	173572	100.0%	1428	0.8%
2	897	186	325408	100.0%	3978	1.2%
3	617	146	202173	100.0%	2828	1.4%
4	645	152	275116	100.0%	2824	1.0%
5	271	50	17093	51.0%	58	0.3%
6	247	55	48270	80.4%	142	0.3%
7	633	135	239159	100.0%	2708	1.1%
8	115	15	50821	98.2%	660	1.3%
11	478	93	35754	56.2%	32	0.1%
15	403	88	125817	100.0%	1311	1.0%
17	259	58	31911	63.6%	245	0.8%
22	1936	208	507337	100.0%	5652	1.1%
24	1025	117	232140	100.0%	4305	1.9%
28	734	142	226775	100.0%	2781	1.2%
68	279	89	137867	100.0%	1691	1.2%
88	629	76	75270	81.7%	546	0.7%
102	333	73	45936	100.0%	824	1.8%
131	640	156	208225	96.3%	2524	1.2%
154	523	95	57748	70.3%	102	0.2%
155	450	74	131619	100.0%	1138	0.9%
164	493	72	75755	72.6%	408	0.5%
175	497	67	62790	79.6%	108	0.2%
194	474	78	57659	73.8%	152	0.3%
210	401	75	49355	65.7%	47	0.1%

Table 4 also indicates that the majority of the points were logged at 1-second frequencies and that only a small percentage of the points were filtered (~1%) as part of the import process. These points contained very unrealistic vehicle movements and would have introduced unwanted noise into the trip end detection process.

## 6 Trip end identification

The trip end identification steps performed within TIAS eliminate false trip ends, identify missing trip ends, and tag abandoned trip ends. The trip end identification algorithms in TIAS analyze sequences of points and their time-space relationships while factoring in trip attributes such as duration, distance, and start and end coordinates, and leveraging a composite ‘habitual destination’ database built from the initial engine-off trip ends.

### ***Step 1: Load database and perform initial trip end classification***

As part of the import process of the vehicle-log files, engine-off trip ends were placed at the last point of every original trip number within each vehicle file. These trip ends were only overridden if the stop time associated with them was considered to be too short for it to be a real trip ( $\leq 4$  seconds). These short engine-off stops are converted to engine-stalls and do not increment the trip count. In addition, trips terminate at the external edges of the road network have their trip ends classified as ‘out-of-area’. Finally, a habitual destination database is created from the original ‘engine-off’ trip ends for use in later trip end identification processes.

### ***Step 2: Place trip ends based on dwell times***

The basic trip end detection algorithm evaluates dwell times between consecutive, non-zero speed points, and then places trip ends based on these dwell times. Typically, TIAS places ‘confident’ trip ends for points with dwell times greater than 300 seconds (or five minutes) and ‘probable’ trip ends for dwell times greater than 120 and less than or equal to 300 seconds. Points with dwell times greater than 20 and less than or equal to 120 seconds are tagged as ‘suspicious delays’, requiring further examination, but do not dictate the end of a trip. For the Borlänge study, the minimum suspicious delay threshold was lowered to five seconds to increase the candidate set of potential trip ends.

### ***Step 3: Check for travel path circuitry / overlap***

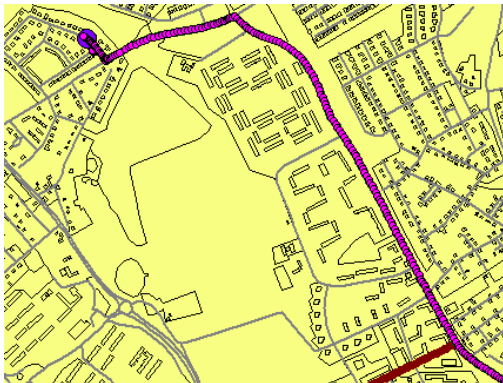
In Step 3, ‘confident’ and ‘probable’ dwell time trip ends are supplemented with path circuitry / overlap trip ends. Circuitry is defined as the distance covered by the vehicle over a path divided by the Euclidian distance between the start and end points of the same path. This process analyses the paths between each trip end for quick changes in heading and for path overlaps. Heading changes are representative of directional shifts that occur at and around intersections, while path overlaps indicate the presence of a loop in the path. TIAS identifies path

overlaps by rounding the point positions into a grid that covers the points' coordinate space. This approach accounts for typical GPS path noise and positional shifts.

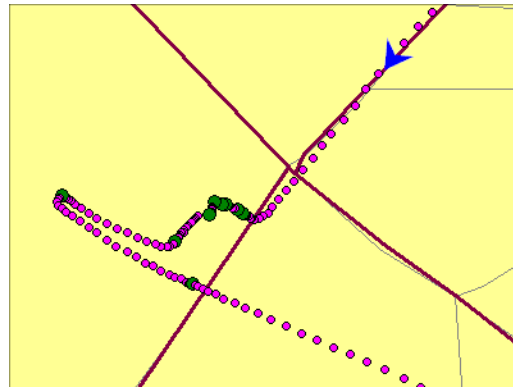
Next, TIAS computes the circuitry of the paths between these directional change points and marks the higher circuitry ones (set at  $> 7$  for this analysis) as candidates for high-circuitry trip ends. TIAS searches for trip end placement options using each point's dwell time and the partial circuitry based on the path start location to each of the segment's points. A 'high-circuitry' trip end is placed on the point that features the highest circuitry and is a 'suspicious delay'. The list of habitual trip destinations performs a 'tiebreaker' if more than one trip end point candidate exists. If no potential trip end is found after this search, an abandoned trip end event is placed at the point with the highest circuitry. Figure 5 illustrates a TIAS-identified case of high-circuitry / overlap.

Figure 5 Examples of high-circuitry and overlaps

high-circuitry with overlap



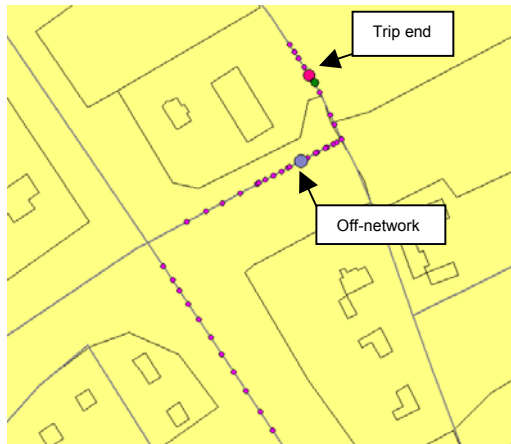
high-circuitry without overlap (real GPS data)



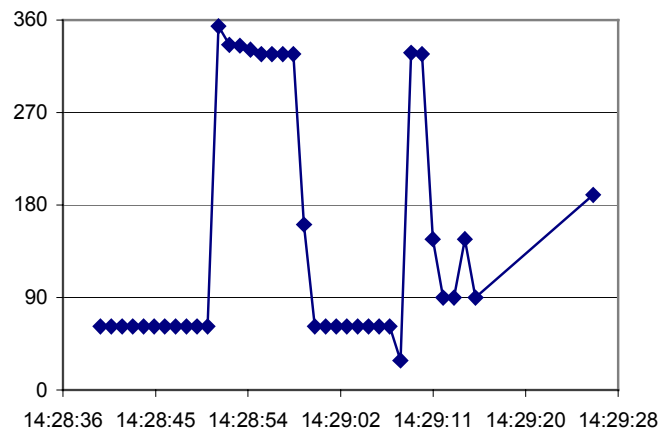
Off-network events are added when a 'confident' or an 'engine-off' stop occurs within the search area of the overlap. This last event happens quite often in the Borlänge data set due to the link-matched nature of the positional data. For example, when a vehicle leaves the road network and enters a parking area close to the road, its movements are still captured and snapped to the adjacent roads. As a result, sudden changes in heading (derived from the direction of travel on the link) over short distances appear, but are quickly followed by a trip end. See Figure 6 for an example.

Figure 6 Examples of off-network travel and corresponding vehicle headings

off-network point example



corresponding heading shifts over time



#### **Step 4: Signal-Loss, Congestion Delays and 'Habitual Destination' Trip Ends**

During this step, TIAS checks all dwell-time classified trip ends, and 'suspicious delay' points that remain in the dataset for GPS signal loss events. If signal loss characteristics are found, the trip end or delay point is reclassified as a signal loss (i.e., not a trip end). Next, TIAS places the all 'probable trip ends' and 'suspicious delay' points over the GIS road network to evaluate the possibility of congestion. Any of these points that occur upstream of an intersection (in the final 1/3 of the road segment) are changed to 'probable congestion' events and are not considered as trip ends.

Finally, the 'habitual destination' list is compared to the location of the remaining 'suspicious delay' points and the recently set 'probable congestion' points. 'Suspicious delay' and 'probable congestion' points that are near (within 50 meters) to an 'habitual destination' are changed to a 'derived location' trip end.

After completing trip end detection, TIAS examines the identified trips with respect to their sequence of places and generates warnings for the following situations: trip starts and ends in the same place, skipped places in the sequence of trips, and first and last trips of the day do not share a common location. At the moment, these warnings are logged into an exception table for later review. If a trip is found with a length less than 25 meters, it is combined with the adjacent trip based on time intervals.

## 6.2 Trip Processing Results

Each travel day took approximately 3.5 seconds to get loaded from the database, processed for trip end identification, and saved, with an average processing speed of approximately 5000 points per minute on a 2.4 GHz P4 Windows 2000 machine. The 24 vehicle-log files with valid data can be processed in about five hours using two machines; further processing time reductions can be achieved by adding more machines. However processing time gains taper off as the database server and network bandwidth become the TIAS performance bottlenecks.

Of the original 13,375 trips provided for the 24 vehicles, 411 trips were tagged with out of area end points and the remainder with in-area end points. Next, the trip identification algorithms were applied to the GPS point data.

There were 4 engine-off trip ends that were reclassified as engine stalls based on a dwell time of five seconds or less and combined into adjacent trips. The trip end identification algorithms found 3006 missing trip ends within the original trip files.

This result indicates a net trip increase of 22.5 %, with the breakdown of the 3006 new TIAS trips as follows:

- 235 trips were added by TIAS based on dwell times greater than 5 minutes ('confident' trip ends)
- 420 trips were added by TIAS based on dwell times greater than 2 minutes and less than or equal to 5 minutes (probable trip ends)
- 1751 trips based on high-circuitry for suspicious delays (between 5 and 120 seconds)
- 600 trips based on habitual destination (reclassification of congestion delays and delays between 5 and 120 seconds based on proximity to existing trip end)

It should be noted that these new trip ends do not include any TIAS-identified trip ends or delay points that were reclassified as GPS signal losses. In addition, TIAS classified 324 points as abandoned trip ends and 5517 points as 'probable congestion'. A total of 1309 off-network events were found among the 'engine off' and 'confident' trip ends.

## 7 Conclusions and Next Steps

The initial tests with a sub-sample of the Borlänge cars/drivers are promising. Some of these encouraging findings include:

- 1) The post-processing of the passively monitored trip data which aims to provide essential information for travel behavior analysis has so far provided mainly consistent but also few ambiguous results. There is need for a refinement of the methodology and especially for an enlargement of the reference database for identifying the trip purposes (e.g. further point-of-interest data).
- 2) Although the database is restricted to local car travel only and there is space for methodological improvements, the resulting database already shows potentials for intense spatial and temporal investigation of daily life travel. As *Mobidrive*, the data set reveals the variability and stability in trip making and activity performance over time.
- 3) The trip end identification results show clearly that a significant number of short duration stops exist within the dataset that occurred without an engine off event. Even if only the ‘confident’ trip ends (i.e., those exceeding five minutes in total dwell time) and those found based on highly circuitous / overlapping travel characteristics, the ‘missing trip’ percentage is still 14.8%.

Clearly, there is scope for methodological improvement on many fronts. The heuristics used for the trip purpose assignments are ad-hoc. A framework that is consistent with either a Bayesian approach to updating the probabilities or with a maximum expectation approach should be developed to make maximum use of the available information within and without the dataset.

There are a variety of sensitivity analyses that can be performed within TIAS to more fully examine the existence of missing trip ends, including increasing the ‘probable’ trip end threshold to 5 minutes, which would effectively increase the ‘suspicious delay’ range to 5 seconds through 5 minutes, allowing the algorithms to more closely examine the previously classified trip ends in the 120 to 300 second range. In addition, using only 1-second frequency vehicle files would eliminate any noise introduced by the 10-second logging intervals.

It would also be beneficial to run all analyses presented here on the same dataset – with the trip end identification process performed prior to the trip purpose imputation. Task schedules

and database extraction difficulties prevented this during these initial analyses, but an end-to-end analysis of a fixed dataset is scheduled soon.

## 8 Acknowledgements

The authors wish to thank our colleagues in Sweden, who gave us the kind permission to work with their data, generously included additional questions from us in their debriefing interviews and finally made the data set available: Lars Åberg (University of Uppsala and University of Dalarna), Niclas Brus (Columna AB, Borlänge). Numerous other Swedish colleagues gave support with supplementary information, including networks, land use data, travel diary data etc.

The research reported here is being supported by ETH Zürich through two internal grants. This is gratefully acknowledged. Uta Samaga's contribution forms part of her MSc dissertation at the Technische Universität Dresden.

## 9 References

- Biding, T. and G. Lind (2002) Intelligent Stöd för Anpassning av hastighet (ISA), Resultat av storskalig försöksverksamhet i Borlänge, Lidköping, Lund och Umeå under perioden 1999-2002, Publikation 2002:89, Vägverket, Borlänge.
- Jensen, F.V. (1996) *An Introduction to Bayesian Networks*, UCL Press, London.
- Little, R., and D. Rubin (1987) *Statistical Analysis with Missing Data*, John Wiley and Sons, New-York
- Pearson, D. (2001) Global Positioning System (GPS) and travel surveys: Results from the 1997 Austin Household Survey, Paper presented at the Eighth Conference on the Application of Transportation Planning Methods, Corpus Christi, Texas, April 2001.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
- Samaga, U. (2003) Entwicklung von GIS-Funktionen zur Analyse und Visualisierung GPS-basierter Mobilitätsdaten. Diplomarbeit an der Technische Universität Dresden, Technische Universität, Dresden.
- Schlich, R., B. Kluge, S. Lehmann and K. W. Axhausen (2002) Durchführung einer 12-wöchigen Langzeitbefragung, *Stadt Region Land*, **73**, Institut für Stadtbauwesen und Stadtverkehr, RWTH, Aachen, 141-154.
- Schönfelder, S. and U. Samaga (2003) Where do you want to go today – More observations on daily mobility, *paper presented the 3rd Swiss Transport Research Conference*, Ascona, March 2003.

- Sika Institut (2001) RES 2000, Den nationella resundersökningen, Statistiska Centralbyrån, Stockholm.
- Stopher, P., J. Bullock, and F. Horst (2003) Conducting a GPS Survey with a Time-Use Diary, Presented at 82<sup>nd</sup> Annual Meeting of the Transportation Research Board, January.
- Vägverket (2000) ISA Intelligent Speed Adaptation, Vägverket, unpublished, Vägverket, Borlänge.
- Wagner, D. P. (1997) Lexington Area Travel Data Collection Test; GPS for Personal Travel Surveys, Final Report for OHIM, OTA, and FHWA, Battelle, Columbus.
- Waßmuth, Volker (2001) Modellierung der Wirkungen verkehrsreduzierender Siedlungskonzepte, *Schriftenreihe des Instituts für Verkehrswesen*, **60**, Institut für Verkehrswesen, Universität, Karlsruhe.
- Wolf, J. (2000) Using GPS data loggers to replace travel diaries in the collection of travel data, Dissertation, Georgia Institute of Technology, School of Civil and Environmental Engineering, Atlanta, Georgia.
- Wolf, J. (2003) Tracing people and cars with GPS and diaries: Current experiences and tools, Presentation at the IVT-Seminar, ETH, Zürich, February 2003.
- Wolf, J., M. Loechl, M. Thompson and C. Arce (2003) Trip Rate Analysis in GPS-Enhanced Personal Travel Surveys, in P. Stopher and P.M. Jones (eds.) *Transport Survey Quality and Innovation*, 483-498, Pergamon, Oxford.
- Wolf, J., R. Guensler and W. Bachman. (2001) Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data, *Transportation Research Record*, **1768**, 125-134.

Die *Arbeitsberichte Verkehrs- und Raumplanung* dienen der schnellen Verbreitung der Ergebnisse der Arbeit der Mitarbeitenden und Gäste des Instituts. Die Verantwortung für Inhalt und Gestaltung liegt alleine bei den Autor/innen.

The *Working Papers Traffic and Spatial Planning* are intended for the quick dissemination of the results of the members and guests of the Institute. Their content is the sole responsibility of the authors.

Eine vollständige Liste der Berichte kann vom Institut angefordert werden:

A complete catalogue of the papers can be obtained from:

IVT ETHZ  
ETH Hönggerberg (HIL)  
CH - 8093 Zürich

Telephon: +41 1 633 31 05

Telefax: +41 1 633 10 57

E-Mail: [sekretariat@ivt.baug.ethz.ch](mailto:sekretariat@ivt.baug.ethz.ch)

WWW: [www.ivt.baug.ethz.ch](http://www.ivt.baug.ethz.ch)

Der Katalog kann auch abgerufen werden von:

The catalogue can also be obtained from:

[http://www.ivt.baug.ethz.ch/vrp/arbeitsberichte\\_d.html](http://www.ivt.baug.ethz.ch/vrp/arbeitsberichte_d.html)