Enhancing Travel Data Collection with an Emphasis on Active Transport

Hamid Safi, Mahmoud Mesbah, and Luis Ferreira
School of Civil Engineering, University of Queensland, Brisbane, Australia
Email: {h.safi, mahmoud.mesbah}@uq.edu.au

***Abstract*—The promotion of active transport has been considered as a fundamental objective of transport planning policies since it can improve passenger's health, relieve traffic congestion, and reduce air pollution. However, the share of active transport in planning studies has not traditionally been estimated accurate enough. One of the main reasons for this problem is the inability of conventional travel survey methods to identify trips made by active modes in the travel behaviour of individuals. Nowadays, the employment of smartphones/GPS devices and elaborated algorithms has improved the accuracy of travel data collections. Yet, the accuracy of current algorithms in detecting active modes of transport is questionable. This paper addresses this concern by proposing a comprehensive and practical framework for detecting trips especially those made by an active mode of transport. In this study, a smartphone application has been developed in conjunction with an improved post-processing analysis framework. It is suggested to revise the conventional method of trip detection and employ smoothing techniques after trip detection instead of employment on GPS raw data to improve the accuracy of data collection. In addition, important attributes for trip detection are applied in a rule-based model. The results demonstrate that the model can detect trips more accurately compared to an active travel data collection approach, and increase the number of detected active-mode trips by 22%. The proposed approach can be employed in all GPS-assisted travel surveys, thereby improving their accuracy and reduce the under-reporting rate, specifically for trips made by active modes of transport.**

***Index Terms*—active transport, travel survey, smartphone, GPS, framework**

I. Introduction

The dominance of motorized modes in Transport systems has led to increased congestion, emissions, pollution and health problems and obesity. On the other hand, active transport modes, walking and cycling, are clean and sustainable alternatives to deal with these concerns (Beckx, Broekx *et al*. 2013). One of the initial requirements for increasing the share of active transport is providing appropriate facilities for active transporters based on their real travel behaviour and considering them in future planning. However, literature shows that active transport modes have been under-reported in most of travel surveys, especially by conventional data collection methods, where participants are requested to report their travel behaviour, relying on their memory and judgment (Yang and Diez-Roux 2012; Millward, Spinney *et al*. 2013). In fact, participants who are responsible to recall, detect and report their travel behaviour are not able to detect their active transport trips effectively. Generally, during data collection, participants go through a three-step procedure to report their trips, as presented in Fig. 1. They have to remember their travel behaviour, review it completely and recognize each trip and its attributes and finally report them. The critical steps in this procedure are the first and second steps, where participant have to recall their travel behaviour and judge about it in order to detect trips and their attributes before reporting.

In the meanwhile, trips which are made by active modes of transport can be easily ignored because, firstly these trips are easily forgotten after a short period of time (Gleave 2004; Forrest and Pearson 2005; Ohmori, Nakazato *et al*. 2005), and secondly most of connection trips, which are made by active modes to access to public transport or parking, are ignored by participants and considered as a part of another trip (Wolf, Hallmark et al. 1999; Duncan and Mummery 2007; Gonzalez, Weinstein *et al*. 2010). In order to cover some part of this shortcoming, an interview or phone-assisted recalling procedure is undertaken in some household travel surveys, in which reported trips are evaluated by experts following the main data collection and if there is any ambiguity in reported data, participants are contacted and invited to revise their reported travel behaviour (TRB Committee on Travel Survey Methods 2010). However, this procedure requires significant human and financial resources, and cannot efficiently detect all active-transport trips, especially those which occurred before or after a private mode trip.

The introduction of GPS devices has supported participants to more accurately recall their travel behaviour and therefore improve the quality of collected data (Wolf, Schöunfelder *et al*. 2004; Itsubo and Hato 2006; Duncan and Mummery 2007; Stopher, FitzGerald *et al*. 2008; Gonzalez, Weinstein *et al*. 2010). Yet, the problem of under-reporting of active transport trips still exists even in improved GPS-assisted methods to some extent, since it is fundamentally assumed that respondents are able to detect their trips completely, while this assumption has been evaluated in this paper and results show that the assumption should be reconsidered.

Manuscript received January 15, 2016; revised June 4, 2016.

©2016 Journal of Traffic and Logistics Engineering
doi: 10.18178/jtle.4.2.117-127

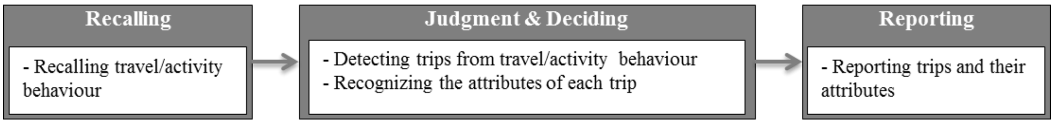


Figure 1. The procedure of travel data recording.

TABLE I. A BRIEF COMPARISON BETWEEN CONVENTIONAL TRAVEL DATA COLLECTION APPROACHES

		Commencement	Implementation requirements	Participants' involvement	Under-reporting rate	Accuracy of collected data	Participant's memory roll	participants judgment
Self-reporting methods	Paper & pencil methods	1950's	high	high	high	low	high	high
	Telephone assisted methods							
	Computer-assisted methods							
	Interview-assisted methods							
GPS-assisted methods	Passive data collection	1994	low	low	Questionable	N/A	-	-
	Active data collection	1999	moderate	high	low	high	low	high
	Prompted recall data collection	2001	moderate	moderate	moderate	moderate	moderate	high

This paper focuses on proposing an enhanced smartphone-assisted travel data collection and post-processing approach, which can specially detect active mode trips and improve the quality of collected data. After a brief review of the literature on recent studies for collecting travel data using smartphones and also trip detection models, significant attributes for trip detection are discussed. Section 3 provides a comprehensive framework for travel data collection and trip detection which is accompanied by a post-processing procedure. In Section 4, the proposed framework and model is calibrated and tested based on a dataset collected through an active data collection approach. Finally, some conclusions and practical recommendation are presented, which will be applicable to all GPS-assisted travel survey studies.

II. LITERATURE REVIEW

A. *General*

Travel data collection methods can be divided into two main categories, namely self-reporting and GPS-assisted methods. As presented in Table I, all conventional travel data collections belong to the first category, where respondents are requested to recall and report their travel behaviour in a specific period of time. The quality of collected data using self-reporting approach relies on the memory and judgment of respondent, which imposes significant burden on them and adversely impacts the quality and quantity of collected data, as far as the under-reporting ratio of these methods is estimated up to 45 percent (Ashley, Richardson *et al.* 2009). With technological improvement in positioning systems and the introduction of GPS technology in transport industry, GPS devices have been used widely to improve the accuracy of travel surveys (Pierce, Casas *et al*. 2003).

GPS-assisted data collection approaches have provided the possibility of collecting travel data more accurately in regards to time, geographic location and route of trips. Yet some other valuable travel attributes, such as transfer points, trip mode and purpose, cannot be directly derived from GPS loggers and need to be detected and labelled by participants. In order to deal with this concern, three specific categories of approaches have been proposed, ‘passive data collection’, ‘active data collection’ and ‘prompted-recall data collection’ (See Table I).

In ‘passive data collection’, which is introduced in the earliest attempts to employ GPS devices in travel surveys, travel data is collected without any labelling procedure from participants (Tsui and Shalaby 2006; McGowen and McNally 2007; Griffin, Huang *et al*. 2008; Schüssler and Axhausen 2008; Srinivasan, Bricka *et al*. 2009; Chen, Gong *et al*. 2010). Although, the implementation of this approach is easier compared to other approaches, the accuracy and quality of collected data cannot be evaluated based on a labelled data. In the other approach ,‘active data collection’ approach, participants are invited to actively label data during data collection (Asakura and Hato 2004; Ohmori, Nakazato *et al*. 2005; Rehrl, Göll *et al*. 2007; Barbeau, Labrador *et al*. 2009; Niu, Zhang *et al*. 2012). This approach can be considered as the most accurate data collection method since the impact of recalling travel behaviour is eliminated. However, the method might impose an extra burden on participants, thereby causing difficulties in the procedure itself (Auld, Williams et al. 2008; Li and Shalaby 2008; Safi, Mesbah *et al*. 2013). The ‘prompted-recall data collection’ approach has been proposed as an alternative to providing the possibility of labelling collected data without imposing significant burden on participants (Auld, Williams et al. 2009). In this approach, travel data is collected passively, and trips are detected using some post-processing algorithms, then participants are invited to label the detected trips, supported by the depiction of their travel behaviour. In fact, if the post-processing algorithm cannot detect a trip, it will be missed in the follow-up prompted-recall survey. Based on the literature, there is a tendency in previously proposed trip detection models to ignore trips which are made by active modes of transport trips. For instance, Zheng, Chen *et al*. (2010) and Stopher and Collins (2005) ignore walking trips systematically, assuming that each trip is undertaken with a single mode.

TABLE II. SOME RULES WHICH ARE CONSIDERED FOR DATA CLEANSING IN RECENT STUDIES

Study	Speed	Acceleration	Time difference	Distance difference	Altitude
Auld, Williams et al. (2009)	less than 160 km/hr	-	Less than 15 second in a nine-log interval	-	-
Bohte and Maat (2009)	less than 200 km/hr (not less than 5 km/hr for more than 1 min)	-	-	More than 10 m between two adjacent points	-
Schüssler and Axhausen (2008)	less than 180 km/hr (50 m/s)	less than 10 m/s^2	-	10 meter	between 200 & 42060 m

B. Data Preparation

Generally, the procedure of data preparation for a trip detection based on GPS navigational raw data includes two steps, namely ‘data cleansing’, and ‘data smoothing’ (Schüssler and Axhausen 2008; Auld, Williams *et al*. 2009; Bohte and Maat 2009). Data cleansing take cares of systematic errors while data smoothing tries to detect and remove random errors. Systematic errors mainly happen due to the lack of in-view satellites or insufficient signal strength (depending on the positioning algorithm of data collection medium). In early GPS-based studies (e.g. Wolf, Hallmark *et al*. (1999); Ogle, Guensler *et al*. (2002)), data cleansing process mainly focused on checking the number of in-view satellites and also the PDOP[1] values to remove obviously incorrect entries, while in more recent studies (e.g. Tsui and Shalaby (2006); Schüssler and Axhausen (2008); Auld, Williams *et al*. (2009); Bohte and Maat (2009)) more rules are added for data cleaning (See Table 2).

The next step for data preparation is ‘data smoothing’, where statistical techniques are employed to remove random errors of the collected data, which are caused by factors such as signal blocking, inaccurate probabilistic errors, or receiver problems (Schüssler and Axhausen 2008). It is necessary to mention that the employment of suitable data smoothing method is crucial as an inappropriate data smoothing eliminate informative logs and reduce the accuracy of trip detection.

The next step for data preparation is ‘data smoothing’, where statistical techniques are employed to remove random errors of the collected data, which are caused by factors such as signal blocking, inaccurate probabilistic errors, or receiver problems (Schüssler and Axhausen 2008). It is necessary to mention that the employment of suitable data smoothing method is crucial as an inappropriate data smoothing eliminate informative logs and reduce the accuracy of trip detection.

C. Trip Detection Model

Based on the literature and considering the complexity of the behaviour of the collected data, rule-based approach is used in all previous trip detection models. Table III demonstrates significant attributes which have been employed for trip detection models in previous studies. ‘Dwell time’, ‘bundle of GPS points’, ‘average speed’ and ‘transfer points’ are the most important attributes which are incorporated in trip detection models. ‘Dwell time’ is the minimum time-difference between two consecutive GPS points after which it can be assumed that the activity took place. Several values have been considered for dwell-time, ranged from 45 seconds (Pearson 2001) to 900 seconds (Schüssler and Axhausen 2008), depending on the specifications of study area and the accuracy of data collection medium. ‘Bundle of GPS points’ is another important attribute which is mainly employed for passive or prompted recall data collection approaches to recognize trips from activities. In an active data collection, where a discontinuous stream of GPS log is collected for each trip, the bundle of GPS points is replaced by trip-id, which is assigned to each recording automatically.

Speed of participants is another important attribute, which is always considered in developing trip detection models. As can be seen in Table III, some researchers assume that before any trip/activity change, there should be an inactive period (e.g. Schönfelder, Axhausen et al. (2002); Tsui and Shalaby (2006); Schüssler and Axhausen (2008)), while some others assume that there should be a walking section for transition between different modes of transport. The second group try to initially retrieve walking segments and then partition the trajectory into several portions based on detected walking segments (Chen, Gong *et al*. (2010); Schuessler and Axhausen (2009); Zheng, Li *et al*. (2008); Zheng, Chen *et al*. (2010)). In another study, Auld, Williams *et al*. (2009) define a low-speed threshold and classify logs into slow and fast movements, and assume that when a point is found in low-speed class, the probability of modal change or trip/activity transfer would be increased substantially. In addition, some researchers employ the GIS

[1] Position Dilution of Precision (PDOP) is a term used to specify the additional multiplicative effect of GPS satellite geometry on GPS precision.

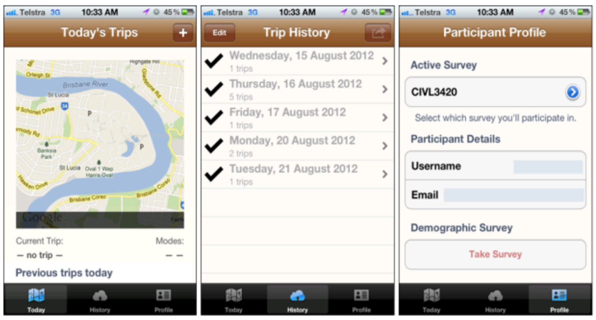
database of study area to recognize mode changes points and separate trips into trip-legs (e.g. Bohte and Maat (2009); Chen *et al*. (2010)), or merge adjacent trips in a same land-use polygon into one trip (Schüssler and Axhausen 2008).

TABLE III. RULES AND ATTRIBUTES EMPLOYED FOR TRIP DETECTION

Study	Dwell time	Bundle of GPS points or Trip-id	Speed	Geographic Criterion
Nitsche, Widhalm et al. (2012)	-	Trip id (from the smartphone app)	-	-
Chen, Gong et al. (2010)	120 sec or 250 meters	Trip id (from a supplementary survey)	low-speed transitions (walking - less than 10 km/h)	mode transfer points
Zheng, Chen et al. (2010)	-	-	low-speed transitions (walking)	-
Bohte and Maat (2009)	180 sec or 10 meters	Loss of satellite reception	Stopped intervals between each mode change	mode transfer points
Auld, Williams et al. (2009)		Distance and time thresholds	low-speed trip-legs (less than 16 km/h)	-
Schuessler and Axhausen (2009)	900 sec	Distance thresholds (3rd times of the standard deviation of tracker's accuracy)	speed less than 0.01 m/s for more than 120 sec	mode transfer points
Flamm and Kaufmann (2007)	-	Spatial density algorithm & Trip id (supplementary survey)	-	-
Tsui and Shalaby (2006)	120 sec	Signal-loss related activities	No movement interval between each mode change	-
Wolf, Schönfelder et al. (2004)	300 sec	Trip id (engine on & off)	-	-
Ashbrook and Starner (2003)	-	K-means clustering algorithm	-	-
De Jong and Mensonides (2003)	120 sec	-	No movement interval between each mode change	-
Schönfelder, Axhausen et al. (2002)	-	Trip id (engine on & off)	No movement intervals	-
Pearson (2001)	45 sec	-	No movement intervals	-
Wolf, Guensler et al. (2001)	120 sec	Trip id (engine on & off)	No movement detections	-
Doherty, Noël et al. (2001)	300 sec	Distance and time thresholds method	No movement detections	-

III. METHODOLOGY

In this paper an enhanced data collection and analysis framework is proposed, which includes the use of a smartphone and a post-processing analysis which can detect trips more accurately compared to previous methods.



a. Travel-related questions b. Trip history - Upload c. Socioeconomic Questions

Figure 2. Three tabs of the smartphone application (ATLAS-I).

A. Data Collection

A smartphone-based travel data collection method has been designed, developed and implemented in this study. ATLAS-I (Advanced Travel Logging Application for Smartphones), is an iPhone application designed for performing travel surveys. It can record individuals' trips when it is running on their phones through an active data collection approach in which participants can label recorded data during the survey through the app interface.

ATLAS-I requires a minimum level of participants' cooperation to start recording and specify the mode and purpose of their trips and also uploading them to the research server. Fig. 2 presents three tabs of the application, namely ‘Today’, ’History’ and ‘Profile’.

‘Today’ tab is the main tab of the application and shows the trajectory, duration and travelled distance of the current trip of participants as well as list of previous trips of the day. One can define new trips during the survey by tapping on the plus (+) button on the top right corner of this tab. All recorded trips are accessible in the ‘History’ tab, which are categorized based on their date and whether they are uploaded or not. In addition, each user creates a profile during installation where they can choose the survey they want to participate and also choose a user-name, password for their profile.

ATLAS-I is designed to work in background while recording a trip in order to allow participants to use the phone services as usual. This is important since any ambiguity in the procedure of data collection or limitation in data collection medium can change the real travel behaviour of participants (Asakura and Hato 2004; Ohmori, Nakazato *et al*. 2005; Itsubo and Hato 2006) and adversely impact the quality of collected data. The only anticipated limitation is an extra battery usage of ATLAS-I in the background. This problem has been addressed in the next generation of the proposed system. More information can be found in Safi, Assemi *et al*. (2015).

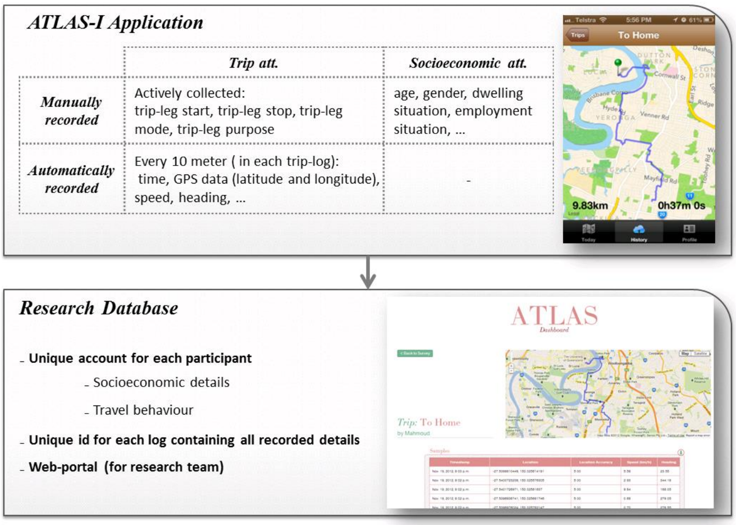


Figure 3. Data-flow in the proposed data collection approach using ATLAS-I

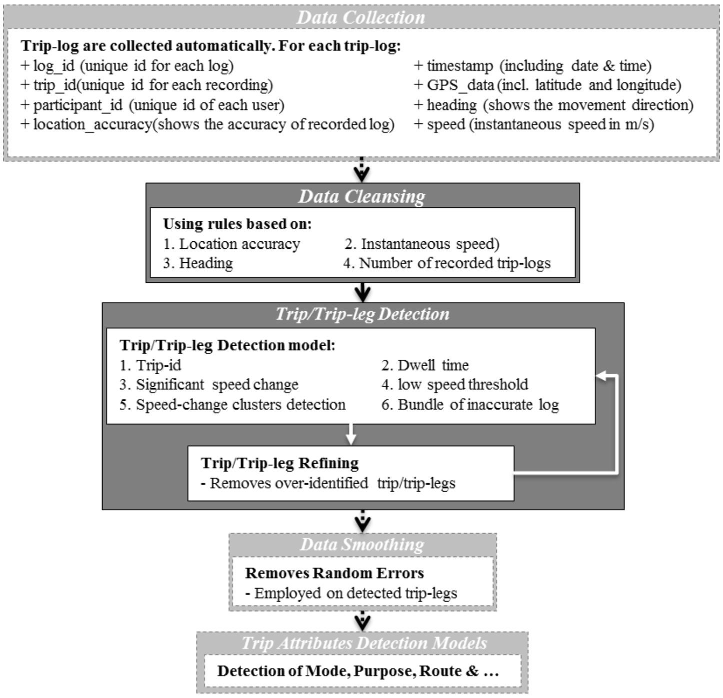


Figure 4. Proposed framework for trip detection based on collected data.

Participants can see their current location and the path they took on a map, and also the attributes of their current trip, as well as the details of their previous trips which were recorded and saved automatically by the application. Considering the importance of socioeconomic details in future modeling, a questionnaire is designed and developed within the application, and participants are invited to complete the questionnaire during the profile creation. ATLAS-I invites participants to define and label their trips by specifying mode and purpose, while ATLAS-I automatically records several other attributes of their travel behaviour, such as time, speed, GPS position, heading and positioning accuracy for every ten meter of their significant movement. This set of recorded attributes is called trip-log. As can be seen in Fig. 3, which demonstrates the flow of data in the proposed data collection approach, the uploaded travel data is stored in a database, which is accessible through a web-portal. This web-portal enables the research team to have a graphical interface to the collected data of participants which facilitates the post-processing analysis.

IV. Data Analysis

Fig. 4 presents the data analysis framework for extracting the travel attributes of participants based on the data collected using ATLAS-I. The proposed framework includes three specific steps, namely 'data cleansing', 'trip detection', and 'trip refining'.

Unlike previously proposed trip detection methods, data smoothing is postponed in the proposed trip detection framework. In fact, all previous studies employed data smoothing on GPS raw data while this lead to removing some informative trip logs which can be used in next steps of trip attribute extraction, specifically trip detection. However, considering the necessity of data smoothing for removing random errors in the collected data, it is postponed to after trip detection. In addition, taking into account the tendency of most of the previously proposed trip detection models to over-identify trips, some rules are suggested to employ on the output of trip detection model and remove the over-identified trips. More explanation regarding the procedure of developing and calibration of 'data cleansing', 'trip detection', and 'trip refining' models is presented in Section 4.2.

V. Application of Proposed Methodology

A. Data Collection

The proposed travel data collection and analysis framework calibrated and evaluated in a multi-day pilot survey, where 6 respondents installed ATLAS-I and work with it periodically for a period of three months. They became familiar with the survey, data collection medium and technical concepts of trip, trip-leg, trip mode and purpose in induction sessions before the data collection, and they were requested to report their travel behaviour accurately by reporting all modes that they use for making a trip. For instance, if they use public transport, they are asked to report the walking trips before and after the public transport trip. The details of collected data are presented in Table IV.

TABLE IV. Total Trips and Trip-Logs which are Collected on the Server

Number of trips	89
Number of trip-logs	73483
Time Span	12/11/2012 - 02/03/2013
Recorded travel time	52:58 Hours
Recorded travel distance	1273.495 Km

Fig. 5 demonstrates the share of each mode based on the data manually labelled by participants. As can be seen, around 85% of trips were made by motorized modes while the share of reported non-motorized trips is around 15%.

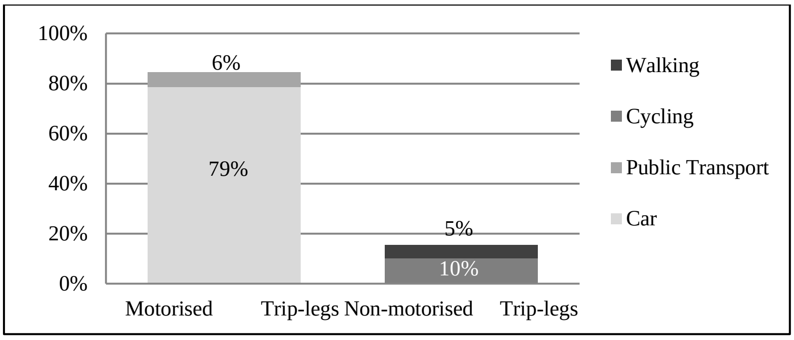


Figure 5. Share of each mode based on users' labelled data.

TABLE V. Proposed Data Cleansing Rules for Removing Inaccurate Logs

#	Rule	Explanation
1	If (Location_accuracy < 55) then Delete trip-log	inaccurate logs
2	If (speed < 0 or speed > 34 m/s) then Delete trip-log	recorded speed (less than 0 or more than 120 km/h)
3	If (speed =0 & heading = 0) then Delete trip-log	unrealistic speed & heading
4	If (number of trip-logs < 5) then Delete trip	too short trips

TABLE VI. List of Rules Employed in Each Model for Developing the Trip Detection Model

Rules (see Table 7)	Model 1	Model 2	Model 3	Model 4	Model 5
Dwell time	✓	✓	✓	✓	✓
Trip-id	-	-	✓	✓	✓
Significant speed change	-	✓	✓	✓	✓
Speed-change clusters detection	-	-	-	✓	✓
Low-speed threshold	-	-	-	-	✓
Bundle of inaccurate logs	✓	✓	✓	✓	✓

©2016 Journal of Traffic and Logistics Engineering

B. Data Analysis

As mentioned in Section 3.2, a three-step framework is proposed for processing the collected data.

1) Data cleansing

A set of simple rules is employed to deal with obviously incorrect logs since the employment of complex cleansing rules may eliminate informative trip-logs and reduce the accuracy of a leg-detection model.

Table V shows the rules and thresholds which are employed for data cleansing.

2) Trip detection model

A step-wise procedure is employed to develop a trip detection model. As presented in Table VI, several models are developed based on the attributes recognized in the literature as significant attributes for trip detection. In each step, it is tried to calibrate the rules and thresholds in order to maximize the number of correctly detected trips, while minimizing incorrect detections. As can be seen in the final model (Model 5), ‘speed change cluster detection’ and ‘bundle of inaccurate logs’, are introduced as two new rules and implemented in the model and significantly improved the accuracy of trip detection model.

Fig. 6 presents the highest obtained level of accuracy (maximized correct detection and minimized incorrect (false) detections) for each model based on the employed rules. As presented in Fig. 6, reducing the number of incorrect detections is as much important as increasing the number of correct detections. For instance, the rate of correct detection is very high in models 3 to 5 but the improvement is in the number of incorrect detections.

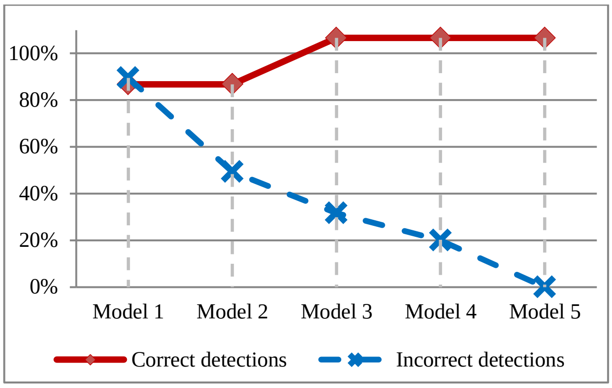


Figure 6. Procedure of developing and calibrating the trip detection model.

As can be seen in Fig. 6, the rate of correct detections for the last three models is more than 100%, which means that the number of detected trips is more than the number of reported trips. Initially it might seem counter intuitive, but after a manual check on the origin, destination and chosen route of each trip on the GIS database of study area, it is revealed that all newly detected trips are real trips which are made by participants but were not reported as independent trips.

For the final proposed trip detection model ‘dwell-time’, ‘trip-id, ‘instantaneous speed’, and ‘location accuracy’ are recognized as the significant attributes, and six rules are proposed and calibrated based on the literature and the behaviour of collected data in an empirical procedure (See Table VII).

TABLE VII. PROPOSED TRIP DETECTION MODEL

#	Rule	Explanation
1	Dwell time (second)	$if\,(timestamp_i - timestamp_{i-1} > 300)$ then logi is the start of a new trip
2	Trip-id	$if\ (trip-id_i \neq trip-id_{i-1})$ then logi is the start of a new trip
3.1	Significant speed change (m/s)	$if\left(\left\lvert\frac{\sum_{i-10}^{i} speed_i}{10} - \frac{\sum_{i}^{i+10} speed_i}{10}\right\rvert > 10\right)$ then logi can be the start of a new trip
3.2	Speed-change clusters detection	Choose the first log if Significant speed change happened for 5 consecutive logs
4	Low-speed threshold	$if\left(\frac{\sum_{i-10}^{i} speed_i}{10}\ or\ \frac{\sum_{i}^{i+10} speed_i}{10} < 2\right)$ then logi is mode-transfer point
5	Bundle of inaccurate logs	If $location\ accuracy > 10$ and speed is not available for 5 consecutive logs

Dwell time: Choosing a suitable value for the dwell-time is crucial, as a small value of dwell-time would lead to many incorrect trip/activity detections (Schüssler and Axhausen 2008). For instance, being in traffic congestion or waiting at a red-light can significantly increase the time difference between two consecutive trip-logs, but dwell-time has to be chosen appropriately to ignore all of these interim points and just detect real trip start or finish points. In this study, several values have been tested for dwell-time in a trial and error process, and it is concluded that the value of 300 seconds (5 minutes) for dwell-time returns the most accurate results.

Trip-id: Trip-id is used to partially detect trips from activities, yet other attributes such as instantaneous speed, location accuracy and timestamp are incorporated to detect those activities which were happened between two consecutive trips (usually short activities), such as waiting in a bus stop or a short shopping activity.

Significant speed change: In order to detect those non-motorized trips which are made immediately before or after another motorized trip, a speed-change threshold (10 m/s) is defined to differentiate these trips. For reducing the number of incorrect detections, another rule is added to detect the clusters of speed change.

Low-speed threshold: In order to detect those mode-transfer points where a motorized mode changed to a none-motorized (or contrariwise), a low-speed threshold (2 m/s) is defined. This rule is employed on those points which are chosen based on the previous rule (significant speed change) and candidates those points that there is a none-motorized mode on one side of them.

Bundle of inaccurate logs: It is empirically found that recorded logs have lower level of accuracy when participant start their trip (warm start/cold start problem) and also when they are in an under-covered area (usually at the end of their trip). This rule specifically looks for inaccurate logs by checking their location accuracy in a 5-log interval.

3) Trip refining

Based on the literature, most of the previously proposed trip detection models have a tendency to over-identify trips and need further data reduction algorithms. One of the common rules to remove over-identified trips is removing too short trips (which is defined as trips with less than 50 meter). Another rule is merging two consecutive trips with a similar mode, which requires the output of mode detection model. In fact, there should be an interaction between trip detection and mode detection steps, which is a time-consuming procedure and requires manual audits (Auld, Williams *et al*. 2009).

A new approach is proposed in this paper to deal with this concern, in which detected trips are classified into two classes based on their mode, motorized and active-transport trips. Then consecutive trips are controlled based on their mode-class and merged if they have similar mode-classes. Therefore, two refining rules are suggested to employ on the output of trip detection model and remove the over-identified trips.

TABLE VIII. EMPIRICAL RULES FOR DIFFERENTIATING MOTORIZED AND NON-MOTORIZED TRIPS

#	Rule	Explanation
1	If (length of trip < 50m) then Delete trip-log	Remove too short detected trips
2	If(mode-class$_i$ = mode-class$_{i+1}$) then Merge trips	Merge two consecutive trips with similar mode class

In order to classify detected trips based on their mode class, an empirical simple rule-based model developed using the data collected in the pilot. Table IX presents the developed model for differentiating motorized trips from active-modes of transport (non-motorized) trips. Both of these rules, which are respectively defined based on the average speed and the average acceleration of the detected trip, have to be satisfied for specifying the mode class of the detected trip.

TABLE IX. EMPIRICAL RULES FOR DIFFERENTIATING MOTORIZED AND NON-MOTORIZED TRIPS

Mode	Mode-detection rules
Active Modes of Transport (Cycling + Walking)	$0 < average\ speed \leq 8.5$
	$-0.1 < average\ acceleration \leq +0.1$
Motorized Modes (Public Transport + Passenger Car)	$8.3 < average\ speed < 34$
	$0.0 < average\ acceleration \leq +2.2$

VI. DISCUSSION

A. Model Evaluation

In order to validate the proposed model and also evaluate the significance of each attribute, and its corresponding rule, in the accuracy of trip detection, a sensitivity analysis procedure is employed. In fact, the rules of the final models (Table VII) is eliminated from the model in a step-wise procedure and it is tried to get to the best level of accuracy with the remaining rules, then the accuracy of the developed model is evaluated through comparing with the total number of reported. The performance of each model (or significance of the eliminated rule) is compared by three indices namely, "Correct Detection Rate (CDR)", "Incorrect Detection Rate (IDR)" and "Significance Rate (SR)", as bellow:

$$CDR_i = \frac{number\ of\ correct\ detection_i}{total\ number\ of\ reported\ trips} \quad (1)$$

$$IDR_i = \frac{number\ of\ incorrect\ detection_i}{total\ number\ of\ detection_i} \quad (2)$$

$$SR_i = (CDR - IDR)_{main\ model} - (CDR - IDR)_i \quad (3)$$

CDR compares the number of correct detections with the number of reported trips, while IDR shows the number of incorrect detection versus total detections (correct and incorrect detections). In addition, SR is defined based on the two previously defined indices, which compares the success rate of the developed model with the success rate of the final model. In fact, a high SR indicates the more significance of the eliminated rule and its related attribute.

Table X presents a comparison between different rules and their relative significance on the accuracy of the proposed trip detection model. As can be seen, 'trip-id' is recognized as the most important attribute for trip detection, followed by 'bundle of inaccurate logs', 'significant speed-change clusters' and 'dwell-time'.

TABLE X. THE SIGNIFICANCE OF EACH RULE IN PROPOSED TRIP DETECTION MODEL

Rank	Eliminated Rule	All detections	Correct	Incorrect	CDR	IDR	SR
1	Trip-id	71	65	6	70%	8%	39%
2	Bundle of inaccurate logs	127	93	34	100%	27%	27%
3	Significant speed change clusters	110	93	17	100%	15%	15%
4	Dwell time	98	90	8	97%	8%	11%
5	Low-speed threshold	93	91	2	98%	2%	4%

Although ‘trip-id’ and ‘dwell time’ are reported in the literature as the most significant attributes in trip detection models, they couldn’t significantly improve the accuracy of the model and detected a considerable portion of incorrect trips (See Model 3 in Fig. 6). While, the ‘bundle of inaccurate logs’, which has been neglected in previous studies, is recognized as the second significant attribute for improving the accuracy of trip detection. This finding emphasizes the necessity of postponing data smoothing to after trip detection as well as employing appropriate data cleansing rules.

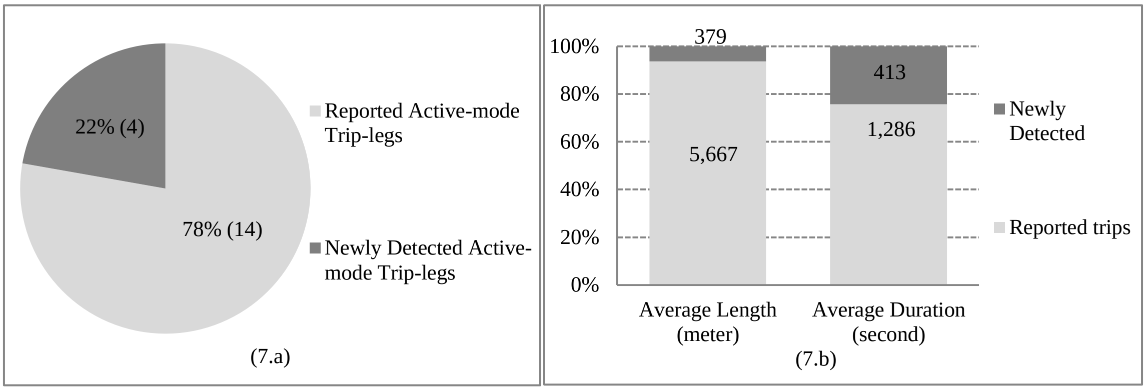


Figure 7. Comparison between reported trips made by active modes and newly detected trips.

B. Active Trip Detection

The employment of rules presented in Table IX, demonstrated that all detected trips are made using active-modes of transport, which means that active-transport trips are under reported up to 22% in the pilot active data collection survey (See Fig. 7a.). Moreover, the employment of post-processing analysis framework could successfully detect 4.5% more trips which were not reported by the participants. Fig. 7b shows the average distance and duration of detected trips made by active-modes of transport, which indicates that a considerable portion of those trips which were made by active modes of transport are missed in terms of length and duration.

The performed pilot survey demonstrates the success of the proposed approach for travel data collection. The participants could record and label data during the survey through Atlas-I. In addition, the possibility of employing post-processing analysis on collected data is another advantage of the proposed data collection method, which can improve the accuracy and quality of collected data.

VII. Conclusions

A comprehensive framework for performing travel surveys using smartphones has been proposed with an emphasis on detecting active modes of transport (walking and cycling). A post-processing analysis procedure is developed in which the traditional procedure of trip detection is revised by incorporating inaccurate logs in trip detection modelling which improved the accuracy of final results. It is suggested to revise the framework of trip detection with postponing data smoothing to after trip detection, in order to improve the accuracy of travel attributes extraction, specifically trip detection.

A rule-based trip detection model has been deployed based on an active data collection approach. ‘Trip-id’, ‘bundle of inaccurate logs’, ‘significant speed-change clusters’ and ‘dwell-time’ are recognized as the most significant attributes for improving the accuracy of trip detection. This paper confirms that walking and cycling trips are under-reported in travel surveys, and suggests a post-processing analysis to increase the accuracy of GPS-assisted travel data collections in extracting the active transport trips. The under-reporting rate of active transport trips has been estimated at up to 22%, compared with an active data collection survey.

Finally, the paper points to the necessity for developing more accurate trip detection models for GPS assisted travel surveys, since current trip detection methods are not able to detect trips accurately. This is a serious shortcoming adversely impacting on the quality and quantity of collected data.

References

[1] Y. Asakura and E. Hato, “Tracking survey for individual travel behaviour using mobile communication instruments,” *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 3, pp. 273-291, 2004.

[2] D. Ashbrook and T. Starner, “Using GPS to learn significant locations and predict movement across multiple users,” *Personal and Ubiquitous Computing*, vol. 7, no. 5, pp. 275-286, 2003.

[3] D. Ashley, T. Richardson, *et al.*, “Recent information on the under-reporting of trips in household travel surveys,” in *Proc. 32nd Australasian Transport Research Forum*, Auckland, New Zealand, 2009.

[4] J. Auld, C. Williams, *et al.*, “An automated GPS-based prompted recall survey with learning algorithms,” *Transportation Letters: The International Journal of Transportation Research*, vol. 1, no. 1, pp. 59-79, 2009.

[5] J. Auld, C. Williams, *et al*., Prompted Recall Travel Surveying with GPS, 2008.

[6] S. Barbeau, M. Labrador, *et al*., “TRAC-IT: A software architecture supporting simultaneous travel behavior data collection and real-time location-based services for GPS-enabled mobile phones,” in *Proc. Transportation Research Board 88th Annual Meeting*, Washington, D.C., USA, 2009.

[7] C. Beckx, S. Broekx, *et al*., “Limits to active transport substitution of short car trips,” *Transportation Research Part D: Transport and Environment*, vol. 22, pp. 10-13, 2013.

[8] W. Bohte and K. Maat, “Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-

scale application in the Netherlands," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 3, pp. 285-297, 2009.
[9] C. Chen, H. Gong, *et al*., "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 10, pp. 830-840, 2010.
[10] R. De Jong and W. Mensonides, "Wearable GPS device as a data collection method for travel research," Institute of Transport Studies Working Paper (ITS-WP-03-02), 2003.
[11] S. T. Doherty, N. Noël, *et al*., "Moving beyond observed outcomes: integrating global positioning systems and interactive computer-based travel behavior surveys," *Transportation Research Board*, 2001.
[12] M. J. Duncan and W. K. Mummery, "GIS or GPS? A comparison of two methods for assessing route taken during active transport," *American Journal of Preventive Medicine*, vol. 33, no. 1, pp. 51-53, 2007.
[13] M. Flamm and V. Kaufmann, "Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions," in *Proc. 11th World Conference on Transport Research*, 2007.
[14] T. L. Forrest and D. F. Pearson, "Comparison of trip determination methods in household travel surveys enhanced by a global positioning system," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1917, pp. 63-71, 2005.
[15] S. D. Gleave, The Use of GPS to Improve Travel Data: Use of GPS in Travel Surveys, Study report prepared for DTLR New Horizon Program, 2004.
[16] P. Gonzalez, J. Weinstein, *et al*., "Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks," *Intelligent Transport Systems, IET*, vol. 4, no. 1, pp. 37-49, 2010.
[17] T. Griffin, Y. Huang, *et al.*, "Computerized trip classification of GPS data Extended Abstract," *Relation*, vol. 10, p. 9220, 2008.
[18] S. Itsubo and E. Hato, "A study of the effectiveness of a household travel survey using GPS-equipped cell phones and a WEB diary through a comparative study with a paper-based travel survey," in *Proc. Transportation Research Board 85th Annual Meeting*, 2006.
[19] Z. J. Li and A. S. Shalaby, "Web-based GIS System for prompted recall of GPS-assisted personal travel surveys: System development and experimental study," in *Proc. Transportation Research Board 87th Annual Meeting*, 2008.
[20] P. McGowen and M. McNally, "Evaluating the potential to predict activity types from GPS and GIS data," in *Proc. Transportation Research Board 86th Annual Meeting*, Washington, 2007.
[21] H. Millward, J. Spinney, *et al*., "Active-transport walking behavior: destinations, durations, distances," *Journal of Transport Geography*, vol. 28, pp. 101-110, 2013.
[22] P. Nitsche, P. Widhalm, *et al*., "A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys," *Procedia-Social and Behavioral Sciences*, vol. 48, pp. 1033-1046, 2012.
[23] X. Niu, Q. Zhang, *et al*., "Using inertial sensors of iPhone 4 for car navigation," in *Proc. Position Location and Navigation Symposium*, 2012.
[24] J. Ogle, R. Guensler, *et al*., "Accuracy of global positioning system for determining driver performance parameters," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1818, pp. 12-24, 2002.
[25] N. Ohmori, M. Nakazato, *et al*., "GPS mobile phone-based activity diary survey," in *Proc Eastern Asia Society for Transportation Studies*, 2005.
[26] D. Pearson, "Global Positioning System (GPS) and travel surveys: Results from the 1997 Austin household survey," presented at 8th Conference on the Application of Transportation Planning Methods, Corpus Christi, April 2001.
[27] B. Pierce, J. Casas, *et al*., "Estimating trip rate under-reporting: preliminary results from the Ohio Household Travel Survey," in *Proc. Transportation Research Board 82nd Annual Meeting Pre-Print CD–ROM*, 2003.
[28] K. Rehrl, N. Göll, *et al*., "Smartphone-based information and navigation aids for public transport travelers," *Location Based Services and TeleCartography*, pp. 525-544, 2007.
[29] H. Safi, B. Assemi, *et al*., "Design and implementation of a smartphone-based system for personal travel survey: Case study from New Zealand," in *Proc. Transportation Research Board 94th Annual Meeting*, Washington, D.C., USA, 2015.
[30] H. Safi, M. Mesbah, *et al*., "ATLAS Project – Developing a mobile-based travel survey," in *Proc. Australasian Transport Research Forum*, Brisbane, Australia, 2013.
[31] S. Schönfelder, K. W. Axhausen, *et al*., "Exploring the potentials of automatically collected GPS data for travel behaviour analysis: A Swedish data source," ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau IVT, 2002.
[32] N. Schuessler and K. W. Axhausen, "Processing raw data from global positioning systems without additional information," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2105, pp. 28-36, 2009.
[33] N. Schüssler and K. W. Axhausen, "Identifying trips and activities and their characteristics from GPS raw data without further information," ETH, Eidgenössische Technische Hochschule Zürich, IVT, 2008.
[34] S. Srinivasan, S. Bricka, *et al*., "Methodology for converting GPS navigational streams to the travel-diary data format," Department of Civil and Coastal Engineering, University of Florida, Technical Report, 2009.
[35] P. Stopher and A. Collins, "Conducting a GPS prompted recall survey over the internet," in *Proc. Transportation Research Board Annual Meeting, 84th*, Washington, DC, USA, 2005.
[36] P. Stopher, C. FitzGerald, *et al*., "Search for a global positioning system device to measure person travel," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 3, pp. 350-369, 2008.
[37] TRB Committee on Travel Survey Methods. (2010). The on-line travel survey manual: A dynamic document for transportation professionals. [Online]. Available: http://www.travelsurveymanual.org/.
[38] S. Tsui and A. Shalaby, "Enhanced system for link and mode identification for personal travel surveys based on global positioning systems," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1972, pp. 38-45, 2006.
[39] J. Wolf, R. Guensler, *et al*., "Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1768, pp. 125-134, 2001.
[40] J. Wolf, S. Hallmark, *et al*., "Accuracy issues with route choice data collection by using global positioning system," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1660, pp. 66-74, 1999.
[41] J. Wolf, S. Schöunfelder, *et al*., "Eighty weeks of global positioning system traces: Approaches to enriching trip information," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1870, pp. 46-54, 2004.
[42] Y. Yang and A. V. Diez-Roux, "Walking distance by trip purpose and population subgroups," *American Journal of Preventive Medicine*, vol. 43, no. 1, pp. 11-19, 2012.
[43] Y. Zheng, Y. Chen, *et al*., "Understanding transportation modes based on GPS data for web applications," *ACM Transactions on the Web (TWEB)*, vol. 4, no. 1, p. 1, 2010.
[44] Y. Zheng, Q. Li, *et al*., "Understanding mobility based on GPS data," in *Proc. 10th International Conference on Ubiquitous Computing*, ACM, 2008.

Hamid Safi has received the M. Sc. Degree in Transport Planning and Engineering from Sharif University of Technology, Iran, in 2005. He is currently a Ph.D. candidate in the area of Transportation Engineering at the University of Queensland. His main research focus is on proposing a comprehensive framework for smartphone-based travel data collection methods. He has multiple papers published and presented in peer-reviewed transport and traffic related journals and conferences. His research interests include travel surveys, travel behavior analysis and machine learning.

Mahmoud Mesbah joined the School of Civil Engineering, the University of Queensland as a Lecturer in 2011. He has a B.Sc. in Civil Engineering from the University of Tehran, Iran, and an M.Sc. in Transportation Planning from Iran University of Science and Technology. He received his Ph.D. from Monash University, Melbourne, Australia in Transportation Engineering. His research interests are transport modelling and planning, optimization of transport systems, public transport, planning for cycling facilities, and application of advanced technologies in transport.

Luis Ferreira received the M. Sc. degree in Transport Planning and Management from The University of Westminster, UK, in 1976, and the Ph.D. degree in Transport Modelling from the University of Leeds, UK, in 1984. He is currently a Professor with the School of Civil Engineering, The University of Queensland, Australia.

He has a strong multi-modal research and consulting background encompassing road and rail, freight and passenger transport. He has worked for 30 years in technical and managerial roles covering transport planning, research, management and consultancy. He has been closely involved with transport and traffic planning and modeling, evaluation and performance measurement of transport programs and projects as a practitioner, researcher and trainer. His research interests include transport evaluation, intelligent transportation systems, safety, incident management, and driver behavior. He has an extensive and distinguished road and rail related publications record in international journals and conferences.