Transportation Research Part C 37 (2013) 118–130

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Transportation mode recognition using GPS and accelerometer data

Tao Feng *, Harry J.P. Timmermans [1]

Urban Planning Group, Department of the Built Environment, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 29 May 2012
Received in revised form 12 August 2013
Accepted 27 September 2013

Keywords:
Accelerometer
Activity type
Bayesian Belief Network
GPS
Transportation mode

ABSTRACT

Potential advantages of global positioning systems (GPS) in collecting travel behavior data have been discussed in several publications and evidenced in many recent studies. Most applications depend on GPS information only. However, transportation mode detection that relies only on GPS information may be erroneous due to variance in device performance and settings, and the environment in which measurements are made. Accelerometers, being used mainly for identifying peoples' physical activities, may offer new opportunities as these devices record data independent of exterior contexts. The purpose of this paper is therefore to examine the merits of employing accelerometer data in combination with GPS data in transportation mode identification. Three approaches (GPS data only, accelerometer data only and a combination of both accelerometer and GPS data) are examined. A Bayesian Belief Network model is used to infer transportation modes and activity episodes simultaneously. Results show that the use of accelerometer data can make a substantial contribution to successful imputation of transportation mode. The accelerometer only approach outperforms the GPS only approach in terms of the predictive accuracy. The approach which combines GPS and accelerometer data yields the best performance.

1. Introduction

During the past years, GPS-based technology has proven its usefulness in collecting activity-travel diary data. Professional organizations are now discussing possible replacement of traditional travel survey methods by GPS data collection. Compared to conventional paper-based or phone-based data collection methods, GPS technology has been argued to reduce respondent and researcher burden, while the accuracy of the data would be better than that of conventional survey methods. The latter, however, does not necessarily apply to all facets of activity-travel diaries because transportation modes and activity types need to be imputed from the GPS traces. Such traces may contain errors and moreover (semi)-automated imputation algorithms are not perfect.

Previous studies have adopted different imputation methods, including ad hoc rules (Du and Aultman-Hall, 2007; Bohte and Maat, 2009; Stopher and Wargelin, 2010), regression models (Rudloff and Ray, 2010) and learning-based algorithms (Moiseeva et al., 2010). These studies have shown that all these methods are only partially successful in correctly identifying transportation modes and activity purpose. All these approaches rely on speed and time information which are extracted from the GPS traces. These traces may contain errors in some situations. For example, when traveling underground or in urban canyons, GPS signals may not be received accurately or may be even completely lost. As a result, GPS data may be

* Corresponding author. Tel.: +31 40 247 2301; fax: +31 40 243 8488.
E-mail addresses: t.feng@tue.nl (T. Feng), h.j.p.timmermans@tue.nl (H.J.P. Timmermans).

[1] Tel.: +31 40 247 3315; fax: +31 40 243 8488.

0968-090X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2013.09.014

incomplete or inaccurate, causing problems for the correct imputation of particular facets of activity-travel patterns. In addition, the use of speed may in some cases result in misclassifications. When measured or calculated speed served in a range which is feasible for multiple transportation modes, no algorithm will be able to perfectly discriminate between transportation modes on this piece of information only. In practice, this often happens for fast walking and slow biking, for bus and car on a congested road and for tram, light rail, and bus. A prompted recall process may compensate for such incomplete or erroneous data, but it involves participation and additional respondent burden in confirming imputed activity-travel diaries. Moreover, prompted recall data are not necessarily error-free either (Feng and Timmermans, 2013). Therefore, a better imputation process, especially in contexts where speed information is not sufficiently discriminating between transportation modes and/or the GPS signal becomes a problem, would be beneficial to the data collection process.

Accelerometer data provide such an opportunity to enrich the data. This technology is not sensitive to the problems mentioned above and could therefore be used in a complementary fashion to GPS data. An accelerometer is a sensor that returns a real valued estimate of acceleration along the *x*, *y* and *z* axes from which velocity and displacement can be estimated. It can capture data independent of the exterior situation. Accelerometers have been used to identify the type of people's physical activity (Bao and Intille, 2004; Ravi et al., 2005), such as walking, running, sitting and relaxing, watching TV, scrubbing, brushing teeth and climbing. The technology has also been used as motion detectors for body-positioning and posture sensing (Ravi et al., 2005). Recent research has attempted to combine GPS and accelerometer data to recognize physical activities (Wolf et al., 2006; Troped et al., 2008; Cooper et al., 2010; Oliver et al., 2010). A few studies have also attempted to detect transportation modes using accelerometer data from smart phone sensors (Reddy et al., 2010; Wang et al., 2010; Xu et al., 2011). However, these studies either covered a limited number of physical activities or transportation modes (e.g. Troped et al., 2008; Cooper et al., 2010; Reddy et al., 2010), or have presented only illustrative findings (e.g. Oliver et al., 2010). Furthermore, none of these studies paid attention to the simultaneous detection of transportation modes and activity episodes.

Thus, this relatively scant knowledge about the potential advantage of the combined use of accelerometer and GPS data to infer transportation mode and activity episode suggests that research on this issue is timely and relevant. In this paper, therefore, we report the findings of a study which aimed at examining the identification of transportation mode by combining accelerometer and GPS data. We adopt a learning-based Bayesian Belief Network (BBN) model to investigate three different approaches: GPS data only, accelerometer data only and the combination of both types of data.

The remainder of the paper is organized as follows: Section 2 will briefly describe the data and the GPS device used in this study. Then, the possibility of accelerometer data to impute transportation mode will be discussed in Section 3. Next, the improved imputation model will be presented in Section 4. Section 5 will present the results. Finally, Section 6 will summarize and conclude this paper.

2. Data and the GPS device

The data was collected in the context of a project contracted between National Center for Social Research (Natcen) and Eindhoven University of Technology and funded by the Department of Transport (Dft) to assess the feasibility of GPS-only data for the new National Travel Survey (NTS) for Great Britain. According to our previous experience (Moiseeva et al., 2010), we adopted the Bayesian Belief Network approach to infer transportation modes and activity episodes. A small sample of volunteers working for these organizations was used for the present study. GPS data were recorded for every second for almost all transportation modes available in London area. Carefully recorded activity-travel diaries were considered as the ground truth. Activity-travel diaries are not necessarily error-free, but considering the critical importance of this pilot study for the organization, error is expected to be relatively small. In total, 80,670 data records were available as a training dataset, used for model development. We identify the conditional probability between input and output at the epoch level (in this case each second).

The device used for data collection was an accelerometer-enabled GPS device, named MobiTest GSL. The device is equipped with a broad range of sensors including three accelerometer sensors and a GPS sensor. The device has a battery life of up to 100 h, and its internal memory ranges between 128 Mb and 512 Mb for approximately 3 months storage of data (MGE, 2009). The GPS measurements include longitude, latitude, height, number of satellites, time, date, HACC and VACC. The accelerometer measurements represent accelerations along three axis (XACC, YACC and ZACC) and the state on whether the device is moving relatively (no acceleration detected in three directions). The accelerations are accurate in 10 Hz.

The speed information which is popularly available in many existing GPS devices however was not recorded by this device. On the other hand, the distance information was recorded only when certain threshold of the measurement accuracy was satisfied, which was rare in practice. Therefore, we propose an algorithm to approximate the instantaneous distance and speed for the imputation of transportation modes.

3. Accelerometer for transportation mode detection

In domains other than transportation, studies using accelerometer data have reported accuracies up to 80–90%, although most studies have been conducted in laboratory settings. Bao and Intille (2004) conducted an experiment using five accelerometers placed at different places of the body instantaneously to check the sensitivity of the accelerometer device for 20 activities. They found that the accelerometer attached to the thigh and wrist produced stable results. Accuracy ranged between 41% and 89% for activities involving movement.

Given these promising results, some researchers realized the potential of accelerometers to identify transportation modes. For example, Troped et al. (2008) used a combination of GPS and accelerometer data to discriminate between four types of what was called activity modes, but what would be called transportation modes in the transportation community (walking, jogging/running, bicycling, inline skating or driving a car). They found that imputation based on accelerometer data was correct in 89% of their cases, and this was increased to 93% when both types of data were used. However, only 61% of the driving minutes were correctly classified, and the combined used of these two data did not always result in better predictions, suggesting that the advantage of adding GPS to accelerometer monitoring, and vice versa, may depend on the type of analysis conducted and/or the purpose of the study. Similarly, Cooper et al. (2010) combined accelerometer and GPS data to investigate the level and location of physical activities of children walking to school. The mean values of the accelerometer data were used to identify differences between three transportation modes (walking, car and bus).

Basically, accelerometers record accelerations in three dimensions, which do not directly reflect differences in transportation modes. However, such differences may be inferred using statistical properties such as mean and variation, and correlations among three axis accelerations. Fig. 1 shows the different features of the accelerations in three dimensions for different transportation modes in a period of 10 min. It is evident that the profiles differ for different transportation modes. The most fluctuating profile is observed for running, followed by walking and biking. The least fluctuating profile with relatively stable values is observed for train and underground. Car and motorcycle seem to have similar average distributional values, especially for YACC. However, the XACC and ZACC of motorcycle have less vibration (fluctuating between 120 and 140) than car (fluctuating between 100 and 140). It can also be observed that tram and bus share a similar pattern in terms of the mean values along the three axes. Previous research has, however, shown that transportation modes can only be partly detected by representative statistics such as mean, standard deviation, energy, and correlations (Bao and Intille, 2004; Oliver et al., 2010). More advanced algorithms are needed to better differentiate between different transportation modes.

4. Improved method and model structure

The method we adopted here is the Bayesian Belief Network which replaces ad hoc rules with a dynamic structure, leading to improved classification if consistent evidence is obtained over time from more samples (more traces). A Bayesian Belief Network is a graphical representation of the conditional probability and causality relationships between variables. The model is described qualitatively by directed acyclic graphs where nodes and edges represent variables and the dependencies between variables. The nodes where the edge originates and ends are called the parent and the child, respectively. Bayesian Belief Networks allow probabilistic inference, indicating that the probability of each value of a node can be computed when the values of the other variables are known.

The nodes that can be reached from other nodes are called descendent. In a Bayesian network, each variable is independent of its non-descendent given the state of its parents. Since the independence among the variables is clearly defined, not all joint probabilities in the Bayesian system need to be calculated. This provides an efficient way to compute the posterior probabilities. Suppose the set of variables in a BBN is $\{A_1, A_2, \ldots, A_n\}$ and that $parents(A_i)$ denote the set of parents of the node A_i in the BBN. Then, the joint probability distribution for $\{A_1, A_2, \ldots, A_n\}$can be calculated from the product of individual probabilities of the nodes:

$$P(A_i, \ldots A_n) = \prod_{i=1}^{n} P(A_i | parents(A_i)) \tag{1}$$

4.1. Speed and distance calculation

Among these location-based variables, speed and distance are at the central position. The GPS device that was used for the data collection records distance only when the accuracy matches the required threshold (default setting is HACC < 10 m). This threshold can be captured only when sufficient satellite signals are received. In high density urban environments, this is often difficult because signals are disturbed or their strengths weakened. More importantly, the device does not record speed. This is problematic because internally recorded instantaneous speed is crucial in accurately detecting transportation mode. Missing information of speed and distance can be an obstacle to identify the activity-travel patterns. While in general the choice of another device would be preferably, in this case this limitation might be an advantage as the ease of transportation mode detection on the basis of GPS information might be more troublesome.

To calculate distance and speed, we developed an algorithm based on Haversine formula (Sinnott, 1984) for calculating great-circle distances between two points on a sphere from their longitudes and latitudes. The Haversine equation is often used in navigation, giving great-circle distances between two points on a sphere from their longitudes and latitudes. It is a special case of a more general formula in spherical trigonometry, the law of Haversines, relating the sides and angles of spherical “triangles”. Haversine’ formula calculates the shortest distance over the earth’s surface between the points,

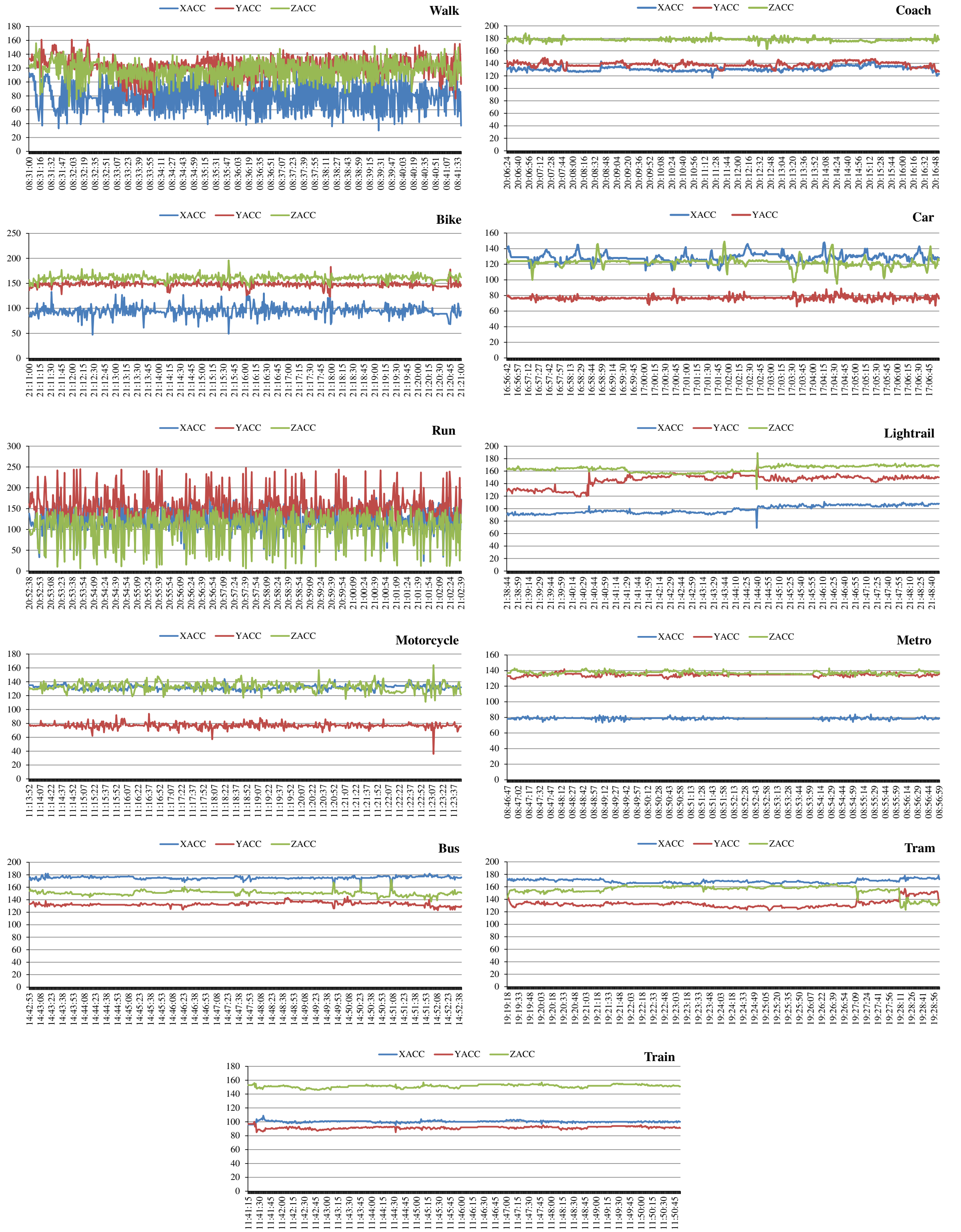

Fig. 1. Frequency of three-axis accelerations for different transportation modes.

ignoring any hills. Fig. 2 shows these inherent relations: the three points (M, N and P) are connected by the great circle (a for $M \sim N$, b for $M \sim P$ and c for $N \sim P$). The law of Harversines is formulated as follows:

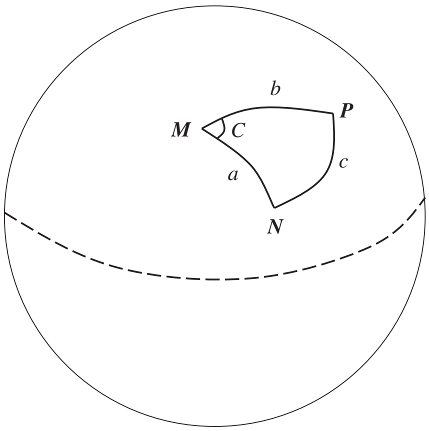

Fig. 2. The law of the Haversines.

$$haversin(c) = haversin(a - b) + \sin(a) + \sin(a) \cdot \sin(b) \cdot haversin \tag{2}$$

Based on the law of Haversines states, the equations to calculate the distance between two points are:

$$d = R \cdot haversin^{-1}(h) = 2R \cdot arcsin(\sqrt{h}) \tag{3}$$

$$haversin(\theta) = sin^2(\frac{\theta}{2}) = (1 - cos(\theta))/2 \tag{4}$$

$$h = haversin(\varphi_2 - \varphi_1) + \cos(\varphi_2) haversin(\Delta\lambda) \tag{5}$$

where *haversin* is the Haversin function; *d* is the distance between the two points; *R* is the radius of the sphere, here is set as 6371 km; φ_1 and φ_2 are the latitude of point 1 and point 2, and $\Delta\lambda$ is the longitude separation.

Substituting Eq. (3) into Eq. (4), we get

$$h = \sin^2(\Delta\varphi/2) + \cos(\varphi_1)\cos(\varphi_2)\sin^2(\Delta\lambda/2) \tag{6}$$

where $\Delta\varphi$ is the altitude separation.

The performance of this algorithm was examined using specific examples. Results suggest that the algorithm is quite robust in estimating distances from GPS data. Nevertheless, when the GPS data recorded are inaccurate, distances calculated this way may still be (highly) unrealistic. To correct for such values, any unrealistic distance calculation, based on typical speed of the transportation mode, was imputed by calculating the distance of the last reliable epochs before the epochs with unrealistic distances/speeds. For example, if the distance between time *i* to time *i* + 1 is larger than 50 m (around 180 km/h), we alternatively use the distance between time $i - 1$ to time *i* as a replacement if this value matches a feasible scale. For the starting point of a trace file, the first data record will be removed if the distance is extremely unrealistic.

Speed was estimated using the following approach. As trace data were recorded for every second, we designed an algorithm to calculate the speed for any particular time period as it is not immediately clear what the best epoch would be. Based on reported experiences (mainly on the Internet) and our own experimentation, (average) speed was calculated for every three seconds based on accumulated distance. Results indicated that the three second epoch seems a good choice to compensate to some extent for deviations from real instantaneous velocity.

The relevant equations for the calculation of speed are:

$$v_i = accudistance_{i,i-3}/(time_i - time_{i-3}) \tag{7}$$

where v_i is the speed at time point *i*; $accudistance_{i,i-3}$ is the accumulated distance from point $i - 3$ to *i*; $time_i$ and $time_{i-3}$ are the time at point *i* and point *i*–3.

The accumulated distance was calculated,

$$accudistance_{i,i-3} = dist_{i,i-1} + dist_{i-1,i-2} + dist_{i-2,i-3} \tag{8}$$

where $dist_{i,i-1}$, $dist_{i-1,i-2}$ and $dist_{i-2,i-3}$ are distances between two points in consecutive time.

Here, the calculated speed is taken as an approximation of the instantaneous speed which is available in most GPS devices. In the application of transportation mode imputation, we applied a time window of 60 s. This means the speed and/or the maximum speed data is averaged before taking them into the BBN model. Moiseeva et al. (2010) present details of the setting and implementation issues of the time window.

4.2. Modeling framework

To better understand the potential of this approach; first the ability of accelerometer data to detect differences in transportation modes was examined. To that effect we only incorporated a limited number of transportation modes. In particular, train journeys were excluded from this work as the GPS data was missing or incomplete for the majority of training data examples. This is likely due to overcrowding, the device being placed in different positions such as in a handbag under a table, the metal based coating applied to certain train operator's carriage windows which is known to block GPS signals.

Fig. 3 shows the network structure that was used to infer the type of transportation mode from the GPS traces. The output is the conditional probability that a particular type of transportation mode has been used as a function of the states of the variables included in the BBN. The node MODE considers 8 different transportation modes: walking, running, bicycle, motorcycle, bus, car, tram and metro.

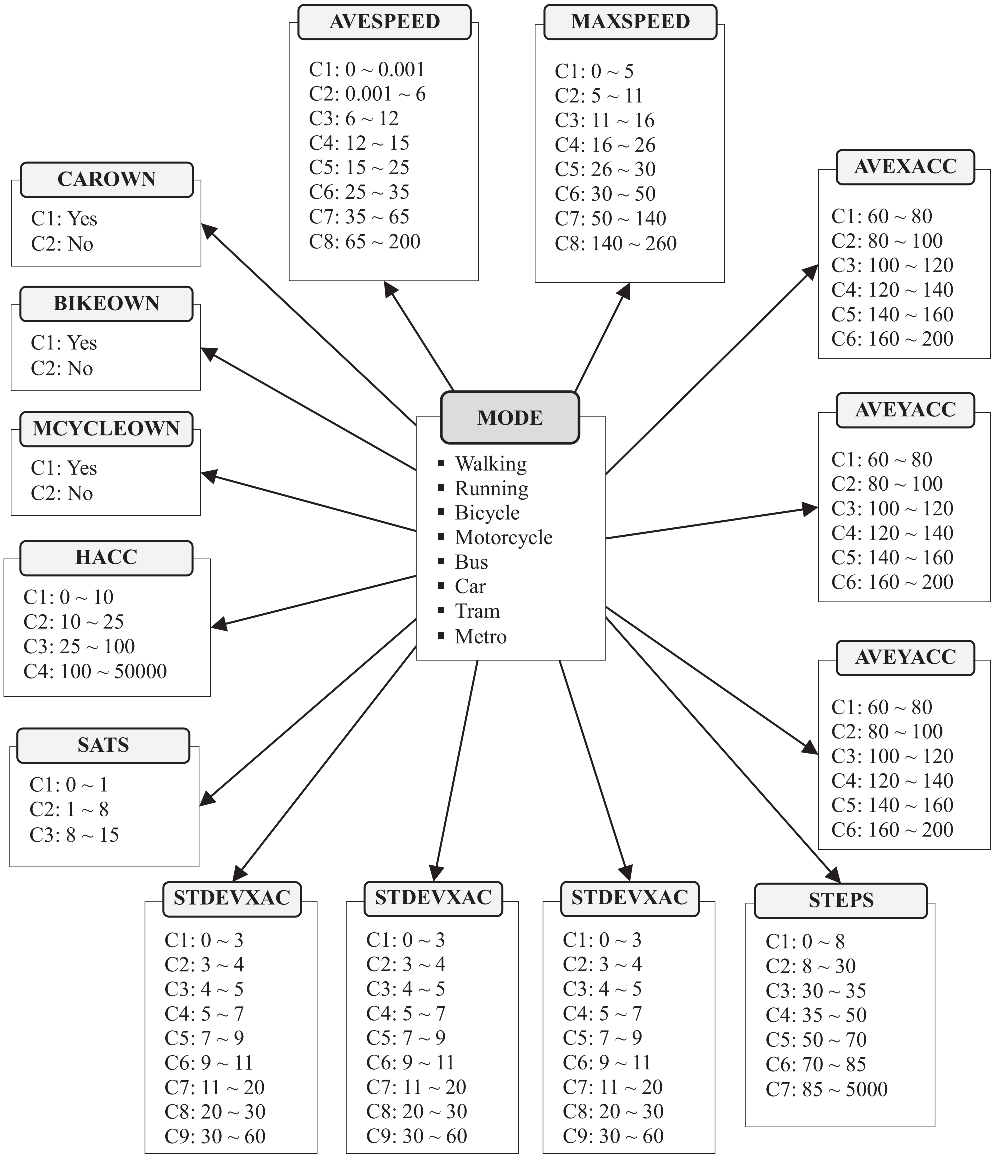

Fig. 3. Model structure and classification settings for the inference of mode of transportation in the comparative exercise. CAROWN: Yes if the respondent has a car, otherwise No; BYKEOWN: Yes if the respondent has a bicycle, otherwise No; MCYCLEOWN: Yes if the respondent has a motorcycle, otherwise No; HACC: estimated horizontal measurement error, *m*; SATS: number of satellites used for position calculation; AVEXACC: average value of *X*-axis acceleration change; AVEYACC: average value of *Y*-axis acceleration change; AVEZACC: average value of *Z*-axis acceleration change; STDEVXACC: standard deviation of *X*-axis acceleration change; STDEVYACC: standard deviation of *Y*-axis acceleration change; STDEVZACC: standard deviation of *Z*-axis acceleration change; STEPS: the average time duration of the device not moving in one minute; AVESPEED: the average speed in every 3 min, km/h; MAXSPEED: the maximum average speed, km/h.

The variables (nodes) included in the network were derived from the raw files extracted from the GPS device. These raw files consist of two types of data: GPS data and accelerometer data. The GPS data provide basic information about coordinates, date and time, accuracy measurements of the device and other information such as distance at every second (although as previously noted, this was only recorded where horizontal accuracy of a position was within a range of 10 m). The accelerometer data provides information about the change in acceleration on three-axis with respect to the device, moving or non-moving of the device, and state of the device (turned off, sleeping, etc.). To feed the imputation model, we first generated some statistical variables (averaged) on the scale of configured time window based on the second-by-second data. We used two variables for measuring the speed pattern, AVESPEED and MAXSPEED which are the average speed and maximum speed for the three minutes epoch. The additionally generated variables related to the accelerometer data are non-moving time duration (STEPS), average value and standard deviation of the three-axis acceleration change (AVEXACC, AVEYACC, AVEZACC, STDEVXACC, STDEVYACC and STDEVZACC).

Since the accelerometer records the information about movements by an inherent variable (NOMOVE) in the device, if the device does not sense movement, the value is automatically increased by one per second. This information is considered important to differentiate traveling activities with similar speeds but different motions, i.e. running and cycling. To use this variable properly, we created a variable (STEPS) equal to the average value of time duration. Seven states were specified after carefully examining the distributional frequencies of the sample, ranging from C1 (not much random movement) to C7 (high levels of random movement). The higher the STEPS value is, the more random the motion of traveling becomes.

In addition to the data extracted from the GPS device, the network also incorporates the effects of personal characteristics into the imputation. Three variables were used to describe whether the respondent has a car (CAROWN), a bicycle (BIKEOWN) or a motorcycle (MCYCLEOWN). In case of the training data, where no person information was collected, these data were set consistent with the transportation mode data. In case of the large sample, these input variables were explicitly collected from the personal profiles.

In addition, two precision related variables of the device were included: HACC and SATS. HACC is the key variable of the device for accuracy control. It means specific data such as distance from last recorded GPS point (DIST) will not be recorded if the measured HACC value is greater than the threshold. In our data collection, the threshold value was set as 10 m by default.

The BBN also requires the states of the input and output variables to be classified. In this case, the initial conditional probability (*C* values in Fig. 3) and the assumed interdependency between the variables are based on either expert judgments or on the processing of a small set of GPS traces. More specifically, we check on the frequency of input variables for each value of the transportation mode, and examine the impact of classifications on prediction accuracy in terms of the true transportation modes. Special emphasis on the setting of classifications is also given to process these 'invalid' data (the number of satellites, quite large value of HACC, etc.). For instance, the samples without any satellites should in general have a zero value for speed, but still have the instantaneous accelerometer data.

5. Results

5.1. Conditional probabilities

In the system, the BBN is a computational object able to represent compactly joint probability distributions, which denote dependencies and independencies among the variables as well as the conditional probability distributions of each variable, given its parents in the graph. The example shown in Table 1 is the conditional probability of the variable 'Mode' by the seven levels of the variable 'STEPS'. One can see that running behavior has the highest score (96%) associated with the top level 'C7' – which indicates that the person carrying the device is moving in a highly random or 'jerky' fashion – followed by walking (70%) and cycling (66%). This is understandable in that running involves continuous movement with many random variations. On the other hand, the tram has the highest probability at the minimum level of movement 'C1' (49%), followed by metro (28%) and bus (11%). These are notably modes of transportation where the individual is most likely to be sitting passively. It should be noted that the distribution of car mode seems flat. This might be accounted for by the different places where people put their device.

Table 1
Conditional probability table of STEPS variable (number of sample: 52,421).

%	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C6 (%)	C7 (%)
Walking	0	0	0	2	9	18	70
Cycling	0	1	0	1	6	26	66
Running	0	0	0	0	0	3	96
Motorcycle	0	0	0	7	34	53	6
Bus	11	37	8	16	9	17	2
Car	4	10	3	18	35	23	6
Metro	28	46	7	9	7	2	1
Tram	50	19	2	12	17	1	0

Table 2
Conditional table of AVGSPEED variable (%).

	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C6 (%)	C7 (%)	C8 (%)	C9 (%)
Activity episode	11.18	81.61	5.87	0.68	0.24	0.38	0.02	0.02	0.02
Walking	0.00	5.04	92.08	2.64	0.21	0.03	0.00	0.00	0.00
Running	0.03	0.03	0.03	98.30	0.82	0.70	0.03	0.03	0.03
Biking	0.01	0.01	0.29	0.50	99.14	0.01	0.01	0.01	0.01
Bus	0.09	0.09	0.09	1.43	0.09	97.58	0.45	0.09	0.09
Motorcycle	0.01	0.01	0.01	0.01	0.01	9.84	88.53	1.58	0.01
Car	0.00	0.00	0.00	0.00	0.03	1.70	34.18	63.44	0.63
Train	54.76	0.67	0.70	0.35	0.34	0.54	0.66	15.10	26.88
metro	99.06	0.80	0.00	0.00	0.12	0.00	0.00	0.00	0.00
Tram	0.08	0.08	0.08	0.66	0.08	89.09	8.60	1.24	0.08
Light rail	0.13	0.13	0.13	0.13	0.76	18.64	69.14	10.83	0.13

Table 2 shows the conditional probabilities related to the variable of average speed. One can see that the traveling underground has the highest score (99.06%) associated with the level C1, followed by train (54.76%) and activity (11.18%). This is because the underground traveling will have no GPS data recorded where the average speed is considered as zero. The signal cannot be well received when traveling by train and when people are doing activities.

Walking will be mostly identified (92.08%) when speed is at the level of C3 and running will be mostly identified (98.30%) when speed is at the level of C4. Car will be identified with a probability of 63.14% when speed is at the level of C8. In addition, taking the state of C6 and C7 specifically, there are similar probabilities between bus and tram for C6 and between motorcycle and light rail for C7. This is true at least based on the sample data that each of the pairs of transportation modes has in general the similar average speed. It should be noted that the conditional probabilities for one variable reflects in part the mode predictability for transportation mode, however, the overall performance of the BBN model will rely on all input variables. Here as shown in Table 1, the similar speeds between different transportation modes are potentially difficult for a GPS only model. Therefore, it indicates the potential merits to incorporate accelerometer data to further identify transportation modes.

5.2. Assessment of imputation models

Table 3 sets out the variables used in each of the three models being assessed. The GPS-only model includes the nodes only related to the GPS traces (AVESPEED and MAXSPEED), while the Accelerometer-only model excludes the speed related data and only incorporates accelerometer variables. To assess the model performance with combined data, all related variables are combined in the last model.

The data used in this section to assess the three models exclude the trips for which modes were not included in the comparative model (for example, rail). This dataset was further divided into 65% and 35% respectively for the purpose of model calibration and validation. Thus, the number of data records used for calibration and validation were respectively 52,424 and 27,630. The performance of the models was assessed in terms of accuracy (hit ratio, i.e. the percentage of corrected predicted classes). In addition, a confusion matrix, describing how misclassified cases were assigned, was constructed. Thus, the main diagonal of this matrix lists the proportion of correctly classified cases, while off-diagonal elements describe the proportion of incorrect imputations. It should be noted that the comparison required some operational decisions. For example, typically the start and end times in the verbal descriptions of the training data often were rounded-off. Comparisons were based on most closely matching times.

Table 4 present the results for the first comparison, which focuses on transportation modes only and excludes rail travel as there were too many missing GPS data for that mode. It illustrates that the accuracy of all models for the calibration data is higher than for the validation data. Taking the calibration-based models as an example, the accelerometer-only model achieves a higher precision (96%) than the GPS-only model (81%). When comparing the GPS and accelerometer model to the accelerometer-only model the accuracy is increased by less than one percentage point for the calibration data models and three percentage points for the less accurate validation data based models.

Table 5 shows the confusion matrix for the model based on Accelerometer-data-only. It shows that the percentage of correctly predicted transportation mode is higher than 90% for all modes except for bus (89%) and tram (84%). In case of the bus,

Table 3
Model structure and input variables for the three models.

Models	Contents	Variables
1	GPS-only	AVESPEED, MAXSPEED, CAROWN, BIKEOWN, MCYCLEOWN, HACC, SATS, MODE
2	Accelerometer-only	STEPS, STDEVXACC, STDEVYACC, STDEVZACC, AVEXACC, AVEYACC, AVEZACC, CAROWN, BIKEOWN, MCYCLEOWN, HACC, SATS, MODE
3	GPS-and-accelerometer	AVESPEED, MAXSPEED, STEPS, STDEVXACC, STDEVYACC, STDEVZACC, AVEXACC, AVEYACC, AVEZACC, CAROWN, BIKEOWN, MCYCLEOWN, HACC, SATS, MODE

Table 4
Results for hit ratios of the three models (proportion of epochs for which mode was correctly identified).

Model	Calibration data (%)	Validation data (%)
GPS-only	81	75
Accelerometer-only	96	82
GPS-and-accelerometer	96	85

Table 5
Confusion matrix of model Accelerometer-only for validation data.

Actual	Predicted							
	Walking (%)	Bicycle (%)	Running (%)	Motorcycle (%)	Bus (%)	Car (%)	Metro (%)	Tram (%)
Walking	96	0	0	0	1	3	1	0
Bicycle	2	94	0	0	0	4	0	0
Running	0	2	98	0	0	0	0	0
Motorcycle	0	0	0	100	0	0	0	0
Bus	8	0	0	0	89	0	1	2
Car	2	1	0	0	0	95	1	0
Metro	2	0	0	0	0	0	97	0
Tram	10	0	0	0	6	0	0	84

eight percent of the cases (epochs) is misclassified as walking and two percent as tram. In case of the tram, six percent and ten percent of tram modes are misclassified as bus and walking.

The results of the combined accelerometer-GPS data model are reported in Table 6. The main difference from the previous model is the inclusion of speed related data into the model. It is expected that speed information may allow further discrimination between transportation modes. This seems to be supported by the results. Table 6 shows a substantial improvement in hit rates for all transportation modes. Moreover, the misclassifications between walking, bus and tram in this model also show significant improvements. The percentage of tram epochs misclassified as walking decreased from 10% to 6%, while the misclassification of bus epochs as walking fell from eight percent to four percent.

These results pertain only to the application of the BBN to infer transportation modes. The ultimate application is, however, more complex in the sense that the network is also used to detect trip purpose. This is done by detecting activity episodes and differentiating these from travel episodes. Consequently, activity episodes may be misclassified as travel episodes and vice versa and in turn this may affect the inference of transportation modes. Hence, to examine the performance of the BBN in this more complex situation, an extended network, including activity type (trip purpose) and train, was developed and assessed. Fig. 4 shows the structure of this model. Moreover, some unrealistic data were filtered out.

Compared with Fig. 3, this comprehensive model incorporates more input variables, the activity episode and more transportation modes. Here, VACC is the estimated vertical measurement error, m; AVGACC is the average estimated horizontal acceleration, km/h^2; MAXACC is the maximum horizontal acceleration, km/h^2; ACCUMDIST is the accumulated distance, m; RRDIST is the distance to road, m; RMDIST is the distance to metro line, m; RLRDIST is the distance to light rail line, m.

Table 7 suggests that the performance of all three models for the calibration data has been slightly reduced (this compares to Table 4). On the other hand, the performance of the models has been improved to the extent that there is now little difference between the models using calibration data and validation data; as one might expect. The accelerometer-only model correctly identifies transportation modes for 89% of the calibration data and 89% of the validation data. Table 7 shows that over all data, the GPS-only model has a lower level of prediction accuracy than the accelerometer-only model.

Table 6
Confusion matrix of model GPS-and-accelerometer for validation data.

Actual	Predicted							
	Walking (%)	Bicycle (%)	Running (%)	Motorcycle (%)	Bus (%)	Car (%)	Metro (%)	Tram (%)
Walking	98	0	0	0	1	1	0	0
Bicycle	2	95	0	0	0	3	0	0
Running	0	2	98	0	0	0	0	0
Motorcycle	1	0	0	99	0	0	0	0
Bus	4	0	0	0	91	0	0	5
Car	2	0	0	0	0	97	1	0
Metro	0	0	0	0	3	0	95	2
Tram	6	0	0	0	6	0	0	88

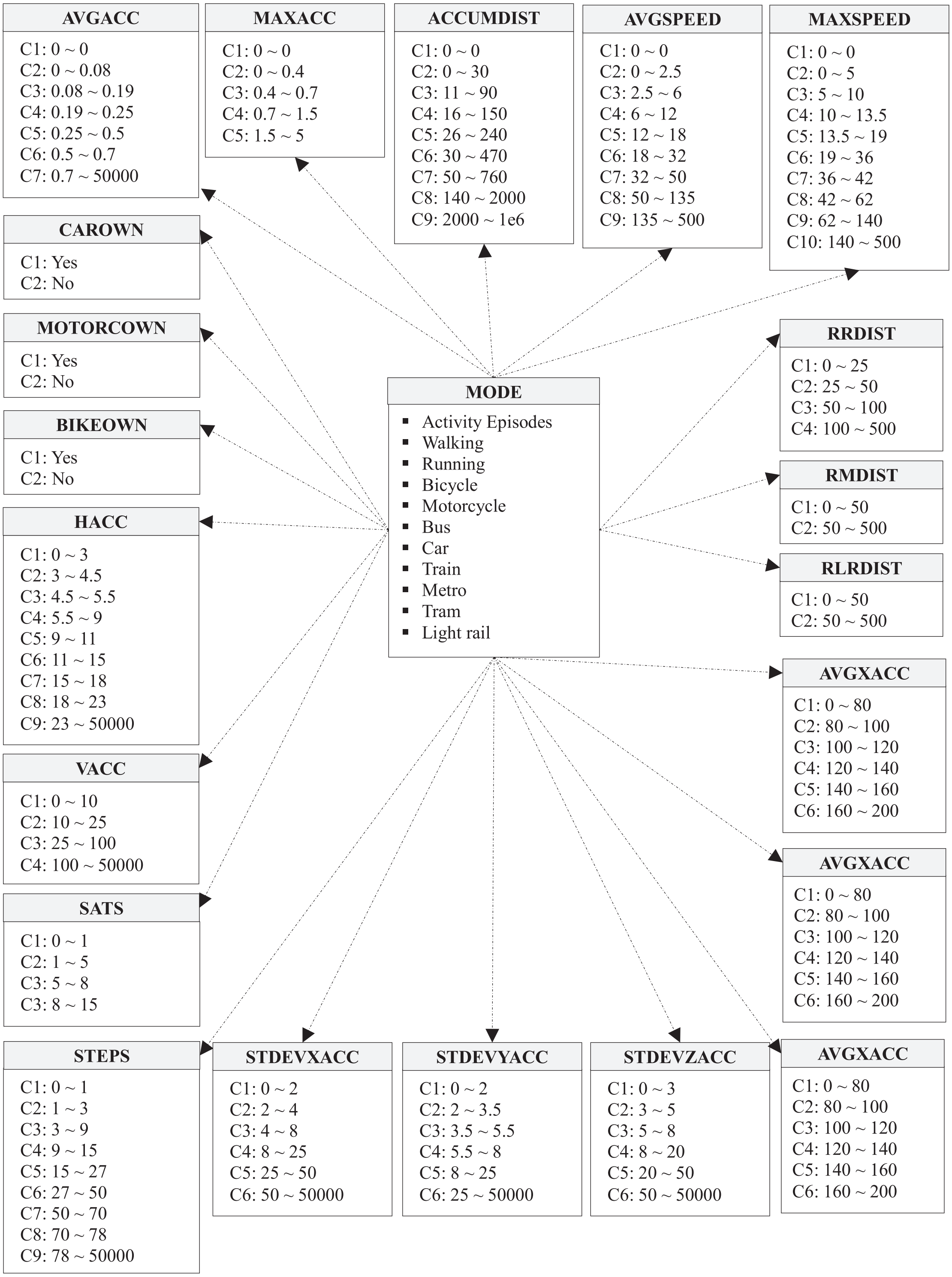

Fig. 4. Revised model structure and classification settings for the inference of conditional table.

Interestingly, classifying both activity episodes and travel episodes seems to have positively influenced the transportation mode-specific hit ratios (Table 8), all are higher than they were before (compared to Tables 5 and 6). Train is a relatively difficult transportation mode to predict: only 83% of epochs were classified correctly.

Table 7
Results of error rate of the three models based on filtered data (proportion of epochs for which mode was correctly identified).

Model	Calibration data (%)	Validation data (%)
GPS-only	78.5	78.4
Accelerometer-only	88.9	88.8
GPS-and-accelerometer	91.7	91.7

Table 8
Correctly identified hit ratios by activity type based on filtered data.

	Accelerometer only (%)	GPS only (%)	Combined Accelerometer and GPS (%)
Activity	33	84	83
Walking	92	97	98
Running	97	98	100
Cycling	88	100	100
Bus	78	87	98
Motorcycle	100	100	100
Car	93	98	99
Train	89	58	83
Metro	86	98	99
Tram	83	98	99
Light rail	98	98	99

Table 9
Confusion matrix of the Accelerometer-only model based on filtered data.

	Activity (%)	Walking (%)	Running (%)	Cycling (%)	Bus (%)	Motorcycle (%)	Car (%)	Train (%)	Metro (%)	Tram (%)	Light rail (%)
Accelerometer-only											
Activity	33	18	0	0	8	0	4	35	0	2	0
Walking	4	92	0	1	1	0	2	1	0	0	0
Running	0	2	97	0	0	0	0	1	0	0	0
Cycling	1	4	0	88	0	0	6	0	0	0	0
Bus	7	15	0	0	78	0	1	1	0	0	0
Motorcycle	0	0	0	0	0	100	0	0	0	0	0
Car	1	3	0	2	0	0	93	1	0	0	0
Train	3	1	0	0	0	0	0	89	7	0	0
Metro	0	0	0	0	0	0	0	14	86	0	0
Tram	3	1	0	0	0	0	0	13	0	83	0
Light rail	1	0	0	0	0	0	0	2	0	0	98
GPS-only model											
Activity	84	4	0	0	0	0	1	9	2	0	0
Walking	2	97	0	0	1	0	0	0	0	0	0
Running	0	0	98	0	1	0	1	0	0	0	0
Cycling	0	0	0	100	0	0	0	0	0	0	0
Bus	1	0	0	0	87	0	0	0	0	12	0
Motorcycle	0	0	0	0	0	100	0	0	0	0	0
Car	0	0	0	0	1	0	98	0	0	0	1
Train	0	0	0	0	0	0	5	58	36	0	0
Metro	1	0	0	0	0	0	0	1	98	0	0
Tram	0	0	0	0	0	0	2	0	0	98	0
Light rail	0	0	0	0	2	0	0	0	0	0	98
GPS–accelerometer											
Activity	83	5	0	0	0	0	1	3	8	0	0
Walking	1	98	0	0	1	0	0	0	0	0	0
Running	0	0	100	0	0	0	0	0	0	0	0
Cycling	0	0	0	100	0	0	0	0	0	0	0
Bus	1	0	0	0	98	0	0	1	0	0	0
Motorcycle	0	0	0	0	0	100	0	0	0	0	0
Car	0	0	0	0	1	0	99	0	0	1	0
Train	0	0	0	0	0	0	2	83	15	0	0
Metro	1	0	0	0	0	0	0	1	99	0	0
Tram	0	0	0	0	0	0	0	1	0	99	0
Light rail	0	0	0	0	0	0	0	2	0	0	99

It should be emphasized that these overall error rates (hit ratios) depend strongly on the frequency of observed transportation modes in the training data. Because in the present context, these data were collected for specific modes, the training data are not a random sample of trips and are therefore not representative. Table 9 shows that the hit ratio of the BBN based on the GPS traces is higher than that for the accelerometer only model for all transportation modes, except for train. However, the over-representation of train trips in the training data and the fact that GPS element was missing for many of the train data due to signal issues, the results of the accelerometer only model are better than the GPS only model.

6. Conclusions and discussion

Identification of transportation mode and activity types is extremely important in collecting travel data. The use of modern technology has been considered very promising in increasing data accuracy and reducing respondent burden. Although successful GPS data processing tools have been suggested and applied in academic and applied research, there is still space to improve and overcome some of the outstanding problems associated with GPS data processing. Accelerometer data provide such an opportunity.

This paper examined the potential advantages of accelerometer data in imputing transportation mode. We presented some findings for three approaches: GPS data only, Accelerometer data-only and the combination of GPS and accelerometer data based on a Bayesian Belief Network model. Results indicate that the accelerometer only model is better than the GPS only model in case of missing GPS signals. The combined use accelerometer and GPS data outperforms the other two approaches in inferring transportation modes.

As presented that the activity diary data used in this paper was considered as the ground truth. This is a common approach adopted in many existing studies, e.g. Stopher and Wargelin (2010), to validate the efficiency of an imputation algorithm. However, the diary data or the prompted recall data is erroneous in itself (Bonsall et al., 2011; Feng and Timmermans, 2013). As evidenced by Feng and Timmermans (2013), people are likely to have a wrong memory in validating the timing data in the sense that the real starting time of a trip might be changed wrongly in the process of prompted recall surveys to a time of an activity. Therefore, future research may incorporate the uncertainty in prompted recall data to enhance the validation process. In addition, the great-circle distances and the speed calculated based on Haversine formula shows a complementary alternative in case that distance and/or speed data are missing. However, additional work may need to investigate the accuracy by using the internally recorded real-time speed and distance.

In addition, regarding the characteristics of the accelerometer sensitivity, it is important to investigate in more detail the effects of where the device is put. Different profiles may exist in the accelerometer data between holding the device and putting the device into the pocket when running. Future research may also consider a more comprehensive list of modes and might include some targeted motion activities, such as reading, social communication and listening during the travel. This could be of interest for collecting data about multi-tasking behavior. Moreover, future research may incorporate more data to investigate the performance of the proposed model.

Acknowledgements

The data used in this study was collected by employees in the context of the 2011 National Travel Survey (NTS) GPS pilot conducted by the National Centre for Social Research (Natcen) on behalf of the United Kingdom Department for Transport (Dft). Details on the data collection and further background to this pilot project can be found in the National Travel Survey 2011 GPS Pilot Field Report, by Josi Rofique, Alun Humphrey and Caroline Killpack (NatCen, 2011). The views expressed in this paper are those of the authors only.

References

Bao, L., Intille, S.S., 2004. Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (Eds.), PERVASIVE 2004, LNCS 3001. Springer-Verlag, Berlin, Heidelberg, pp. 1–17.

Bohte, W., Maat, K., 2009. Deriving and validating trip purpose and travel modes for multiday GPS-based surveys: a large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies 17 (3), 285–297.

Bonsall, P., Schade, J., Roessger, L., Lythgoe, B., 2011. Can we believe what they tell us? Factors affecting people's engagement with survey tasks. In: 9th International Conference on Transport Survey Methods.

Cooper, A.R., Page, A.S., Wheeler, B.W., Griew, P., Davis, L., Hillsdon, M., Jago, R., 2010. Mapping the walk to school using accelerometry combined with a global positioning system. American Journal of Preventive Medicine 38 (2), 178–183.

Du, J., Aultman-Hall, L., 2007. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: automatic trip end identification issues. Transportation Research Part A: Policy and Practice 41 (3), 220–232.

Feng, T., Timmermans, H.J.P., 2013. Analysis of error in prompted recall surveys. In: The XII NECTAR International Conference.

MGE, 2009. Marketing Geographics Environment. <http://www.mgedata.com/>.

Moiseeva, A., Jessuren, J., Timmermans, H.J.P., 2010. Semiautomatic imputation of activity travel diaries: use of global positioning system traces, prompted recall, and context-sensitive learning algorithms. Transportation Research Record: Journal of the Transportation Research Board 2183, 60–68.

Oliver, M., Badland, H., Mavoa, S., Duncan, M.J., Duncan, S., 2010. Combining GPS, GIS and accelerometry: methodological issues in the assessment of location and intensity of travel behaviors. Journal of Physical Activity and Health 7 (1), 102–108.

Ravi, N., Dandekar, N., Mysore, P., Littman, M.L., 2005. Activity recognition from accelerometer data. In: The Seventeenth Conference on Innovative Applications of Artificial Intelligence, IAAI, American Association for Artificial Intelligence, Pittsburgh, Pennsylvania.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. ACM Transactions on Networking 6 (2), 1–27.

Rudloff, C., Ray, M., 2010. Detecting travel modes and profiling commuter habits solely based on GPS data. Transportation Research Board 89th Annual Meeting.

Sinnott, R.W., 1984. Virtues of the Haversine. Sky and Telescope 68 (2), 159.

Stopher, P.R., Wargelin, L., 2010. Conducting a household travel survey with GPS: reports on a pilot study. In: The 12th World Conference on Transportation Research.

Troped, P.J., Oliveira, M.S., Matthews, C.E., Cromley, E.K., 2008. Prediction of activity mode with global positioning system and accelerometer data. Medicine & Science in Sports & Exercise 40 (5), 972–978.

Wang, S., Chen, C., Ma, J., 2010. Accelerometer based transportation mode recognition on mobile phones. In: Proceedings of the 2010 Asia–Pacific Conference on Wearable Computing Systems.

Wolf, J.L., Oliveira, M.G.S., Troped, P., Mathews, C.E., Cromley, E.K., Melly, S.J., 2006. Mode and activity identification using GPS and accelerometer data. In: Transportation Research Board 85th Annual Meeting, Washington, DC, USA.

Xu, D., Song, G., Gao, P., Cao, R., Nie, X., Xie, K., 2011. Transportation modes identification from mobile phone data using probabilistic models. In: Tang, J., King, I., Chen, L., Wang, J. (Eds.), Advanced Data Mining and Applications, vol. 7121. Springer, Berlin, Heidelberg, pp. 359–371.