Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty

Citation for published version (APA):
https://doi.org/10.1080/03081060.2019.1576384

DOI:
10.1080/03081060.2019.1576384

Document status and date:
Published: 03/04/2019

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty

Tao Feng & Harry J. P. Timmermans

To cite this article: Tao Feng & Harry J. P. Timmermans (2019) Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty, Transportation Planning and Technology, 42:3, 274-292, DOI: 10.1080/03081060.2019.1576384

To link to this article: https://doi.org/10.1080/03081060.2019.1576384

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 13 Feb 2019.

Submit your article to this journal

Article views: 125

View Crossmark data
Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty
Tao Feng and Harry J. P. Timmermans

Urban Planning Group, Department of the Built Environment, Eindhoven University of Technology, Eindhoven, The Netherlands

ABSTRACT
Procedures to transform GPS tracks into activity-travel diaries have been increasingly addressed due to their potential benefit to replace traditional methods used in travel surveys. Existing approaches for data annotation however are not sufficiently accurate, which normally involves a prompted recall survey for data validation. Imputation algorithms for transportation mode detection seem to be largely dependent on speed-related features, which may blur the quality of classification results, especially with transportation modes having similar speeds. Therefore, in this paper we propose an enhanced integrated imputation approach by incorporating the critical indicators related to trip patterns, reflecting the effects of uncertain travel environments, including bus stops and speed percentiles. A two-step procedure which embeds a segmentation model and a transportation mode inference model is designed and examined based on purified prompted recall data collected in a large-scale travel survey. Results show the superior performance of the proposed approach, where the overall accuracy at trip level reaches 93.2% and 88.1% for training and surveyed data, respectively.

KEYWORDS
Activity-travel diary; GPS; imputation; travel survey; trip patterns; uncertainty

1. Introduction
Transforming GPS data into a sequence of activity and travel data has been largely discussed in recent years because of the potential merits it offers for travel surveys, i.e. high accuracy, high resolution of spatial and temporal information and light burden for respondents. In comparison to conventional survey methods, GPS-based surveys need to extract diary data from GPS traces using imputation algorithms. As most algorithms/processes are not perfect, reporting an average accuracy between 70% and 85%, the instrument of prompted recall surveys to verify the imputed activity-travel data is normally required. However, the extra efforts involved in prompted recall surveys could blur the value of GPS surveys regarding the time gain and possible human errors. Developing an accurate and efficient imputation algorithm therefore is still of high importance for travel surveys.

The primary task in transforming GPS traces into a sequence of activity-trip data is the detection of transportation modes and activity types, which are essential for travel

CONTACT  Tao Feng  t.feng@tue.nl  Urban Planning Group, Department of the Built Environment, Eindhoven University of Technology, Vertigo 8.16, 5600MB, Eindhoven, The Netherlands

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.
behavior research. A common approach in existing literature adopts the sequential process where activity episodes (trip end or stationary) is identified first to divide the whole sequence into segments, and then the transportation mode for each trip segment is inferred using various features extracted from GPS traces (Wolf, Guensler, and Bachman 2001; Chung and Shalaby 2005; Du and Aultman-Hall 2007; Bohte and Maat 2009; Wolf 2004). To detect the transportation mode, algorithms largely depend on speed-based indicators, such as the average and/or the maximum values related to speed, acceleration, distance and duration (Schuessler and Axhausen 2009). These indicators are important in terms of the difference at epoch level, i.e. every second, or at trip level, i.e. a segment of GPS data. However, it suffers from the ambiguity between different transportation modes with similar features, such as cars and buses in urban areas. Similar difficulty also exists between trips using an e-bike and a conventional bike, and between a bike and a car in searching for parking locations. Especially in traffic congestion, these methods may not succeed differentiating between a car waiting in line or signals and slow walking. Thus, an integrated imputation method which incorporates all possible variations is needed.

Moreover, imputations algorithms for transportation mode classification also vary in the use of different features extracted from the GPS data. An improper selection/use of variables may introduce noise into the discrimination. In spite of the fact that the selection of variables should comply with the principle that they should be able to represent the unique features of specific transportation modes, it should be also considered in combination with the scale of application, e.g. at the trip or epoch level. For instance, the indicator of average speed may not be capable of differentiating between transportation modes at the trip level, especially in a complex urban setting.

Therefore, in this paper, we propose an integrated approach to enhance imputation accuracy by taking into account the effects of different urban contexts. A two-step method is designed to address these uncertain effects by combining the segmentation and transportation mode imputations. Using Bayesian network models, a segmentation model designed for epoch data imputation is built first to divide GPS traces into segments of trips or activities. The data merged at the trip level is then used to extract the variables for transportation mode imputation. These variables are selected by reflecting the unique features of the transportation modes and the effects of uncertain travel environments, using speed and spatial information which are popularly available in most GPS devices and smart phones. The validity of the proposed method is examined using the diary data collected in a large-scale travel survey.

The remainder of the paper is organized as follows. Section 2 will briefly introduce the relevant works in the data transformation of activity-travel diaries. Section 3 will present the selection of the variables for transportation mode imputation and the proposed approach. Section 4 will introduce the GPS data and the travel survey. The results of conditional probabilities and the validity of the proposed approach will be presented in Section 5. Section 6 summarizes and concludes the paper.

2. Relevant work

Earlier works to extract travel information from GPS data are largely dependent on the pre-defined ad hoc rules which are designed based on some general principles which
allow the classification between different transportation modes (Wolf 2001). For example, if the speed is higher than 50 km/h, it may be considered as a car. If the time gap between two observations of the GPS log points is longer than three minutes, it may be that people are inside a building. These rules have the advantage that they are straightforward in terms of reasonability and ease of use, and they have been demonstrated as useful in various applications (Stopher, Jiang, and FitzGerald 2005). However, such a method suffers from the possible misclassification of transportation modes with the increasing complexity of travel environments in that the rules can become very complicated and sensitive.

In order to avoid the possible misclassification induced by the combinations of multiple rules, machine learning algorithms have been used largely in recent years to identify transportation modes (Feng and Timmermans 2013a; Stopher, FitzGerald, and Zhang 2008; Tsui and Shalaby 2006). Relative to the predefined rules, machine learning algorithms are flexible to handle more complex situations in the sense that an automatic classification procedure can be accomplished by the inherent inference mechanism of the algorithm itself. In particular, the probability results reported by some algorithms are potentially valuable to further analysis on the reasoning. Popular algorithms of such kind which have been used to detect transportation modes include decision tables, neural networks and Bayesian networks.

Zheng and Xie applied a decision tree model to detect four types of modes using cellular phone data because of the superiority of the model relative to three other algorithms. However, selected a logistic regression model because it generates probabilities for each mode, even if its prediction accuracy is slightly lower than that of other methods. These studies have involved either a limited number of transportation modes or a limited number of input variables.

Furthermore, the difference in predicted accuracy depends not only on the algorithm but also on the number of identified transportation modes, type of input variables, urban setting and data used to validate the algorithms. Imputations have been conducted using speed-based indicators, such as speed and acceleration (Schuessler and Axhausen 2009), spatial location-based variables, such as distance to a road and/or to bus stops (Chung and Shalaby 2005; Bohte and Maat 2009), and/or personal profiles (Moiseeva, Jessuren, and Timmermans 2010). Many of these decisions depend on the specific philosophy underlying the approach. For example, the Trace Annotator system (Moiseeva, Jessuren, and Timmermans 2010) was built under the assumption of a minimum amount of information and fast, on-line processing of uploaded traces so that participants can wait for imputed activity-travel diaries and immediately change them if needed. In principle, one would assume that accuracy is improved by adding detailed specific information.

Studies also vary in terms of the number of transportation modes, ranging from three modes (Gonzalez et al. 2008) to a more complete list of 11 modes (Feng and Timmermans 2013a). Moreover, accuracy depends on how the imputation results were validated. Validation can be based on a comparison of imputed data and either individuals’ diaries or historical travel survey data. In the case of absence of so-called ground truth (mostly the prompted recall data of the same individual), comparisons with historical survey data at an aggregated level have been made (e.g. Schuessler and Axhausen 2009; Feng and Timmermans 2013a), although various sources of error exist in such data. More importantly, aggregate comparisons do not allow capturing accuracy at the individual
level. In this regard, Feng and Timmermans (2016) have systematically compared the performance of different machine learning algorithms in the application of GPS data imputation. A Bayesian network has been shown as superior to others.

In spite of the fact that the detection of transportation mode is vital, one cannot ignore the segmentation where the GPS tracks are divided into several portions of moving or static states. It is also related to the fundamental issue of how to detect the transportation mode. A common approach is to use a sequential method where GPS tracks are first segmented into trip or activity period, and then the portion of GPS logs labeled as trip period is used to detect transportation mode. Most of the applications implement rules to the segmentation, e.g. the three minutes rule. In many cases, this is reasonable regarding the actual situation. However, it may encounter problems when the GPS observations are influenced by the weak signal strength where some logs were not recorded.

To overcome the possible effects of signal strength on the GPS observations and the consequent detection of transportation modes, a mixed procedure based on the epoch level of the GPS data have been proposed (Moiseeva, Jessuren, and Timmermans 2010). Instead of the two-step procedure, the method emphasizes on the detection of transportation modes at the epoch level first and then applies a merging procedure to generate the sequence of activities and trips. This approach is considered to be superior in terms of the avoidance of the signal effects. However, the result of the segmentation is much sensitive to the design of the merging rules.

In addition to the approaches used in the detection of transportation mode and segmentation, researchers have proposed to use different critical indicators, including average speed, maximum speed, acceleration change, maximum acceleration and accumulated distances. In general, these indicators are extracted based on the whole or a portion of the GPS traces within a certain time period of the trip. However, there is no consensus about which critical indicators performs better than the others regarding the different limits and purpose in different applications. It is obvious that existing works still need to be improved in terms of the indicator selection, the design of imputation procedure and imputation algorithms. A more sophisticated imputation procedure seems to be important.

3. Segmentation and mode detection using aggregated pattern of trips

3.1. Selection of critical indicators

In general, the extent of differences in the critical indicators between different transportation modes depends on the traffic environment and urban context. The classification can be blurred especially for the transportation modes which have similar speed. Although one can commonly expect that a car has a higher speed than a bus, the difference may not be evident in a congested urban area. Similar to this is the differentiation between a bike and an e-bike where the latter is normally expected to have a relatively higher speed. To what extent the difference of the mode pairs is dependent on the urban and spatial environment needs to be investigated. Figure 1 represents an example of speed distribution of trips by different transportation modes. As one can see that patterns of the real-time speed are similar between car and bus and between bike and e-bike.

Therefore, to make better classification between the transportation modes, especially those with similar speed, it seems necessary to incorporate additional indicators which
can represent the spatial contexts in more details. For example, one may need to know the state of traffic flow at certain road segments during the time of the trip, and/or the number of signals and/or stops made on the route. Thus, it is necessary to specifically extract variables which can reflect the effects of uncertain environments.

3.1.1. Speed related indicators

Speed information is vital to differentiate different transportation modes. It is also the main indicator in which a large scale of variation exists, e.g. the normal driving speed of car vary according to different urban contexts. The speed of trip can vary from a value less than 1 km/h to a value larger than 130 km/h. Even for a same trip, different portions of the trip will have different speeds. One of the examples is that the parking search period when driving in a relatively low speed can be referred to a similar speed as a bike or walking.

In spite that existing applications have proposed the use of various indicators which are calculated based on the real speed measurement which is automatically recorded in the GPS traces, e.g. maximum and/or average speed, average acceleration, they may be not always helpful to identify transportation mode. Imagine a trip by car, for example, the distribution of real-time speed may be largely affected by the different urban context, e.g. traffic situation, quality of service of the road. At the trip level, a car trip with traffic congestion will lead to a longer tail than that with less congestion. The average speed of the trip in this case does not make much sense to the inference of transportation mode. Similarly, the acceleration of a walking trip may yield a larger value than that of a car and a bus.

Therefore, it is important to select properly the speed related indicators in practice. In fact, one can label the difference relatively well by use of the maximum speed of the trip. This is understandable because it can partly avoid the consequences of congestion, e.g. even a portion of the trip is with congestion. However, a single value, the maximum speed of the trip, may not be stable enough to avoid the uncertain effects in the

Figure 1. Example of similar speed patterns between transportation modes.
measurement, e.g. weak signals lead to a sudden change of real-time speed and distance. Therefore, it might be wise to use additionally variables to capture the speed in the maximum range.

Therefore, we propose to use the percentile of speed at the level of 85%. The 85th percentile itself represents the level which is higher than 85% of the other speed observation of a trip, representing the distribution of speed with larger values. Actually, the 85th percentile speed of traffic on a road is often used as a guideline in setting speed limits and assessing whether such a limit is too high or low (Robert and Patricia 2007).

3.1.2. Spatial indicators

Spatially, within a certain area, a set of objects can be found. Assume a radium of \( r \) is pre-defined and used to draw an area from the GPS log point, the object \( j \) in the set \( J \) found within the area can be used to calculate the distance to the point. The minimum distance to the object is then computed as

\[
d'_j = \begin{cases} 
\min(d_j, j \in J), & \text{if } J > 0 \\
r, & \text{if } J = 0 
\end{cases}
\]

This means the \( d'_j \) or \( d_j \) must be less than or equal to the radium \( r \). If there are no objects found, the default distance is set to the maximum, \( r \). If the spatial object is a point instead of a line (road), we apply the same method to calculate the distance to the searched starting point, i.e. bus stops. In this sense, the distance to bus stops calculated is the minimum distance of the point to all bus stops within the searching area.

Assume the distances from a GPS log point to the closest railway, motorway, cycleway and footway are \( d_r, d_m, d_c, d_f \), respectively. Then the minimum distance of the GPS point to the specific road category is \( d'_r, d'_m, d'_c, \) and \( d'_f \), respectively. It should be noted that, in actual case, the radius used to search the closed roads and/or points may be different according to the different types.

For a trip which has \( N \) log point in total, the number of points where \( d_j \) is less than \( r \) for a specific road category \( c_i \) is \( N_{c_i} \), where \( i \) refers to the category index, \( i \in [1, 4] \). For each GPS point, we have four values of distances for the four types of road categories. For a specific road category \( i \), we can compare it with the rest values. We use \( N'_{c_i} \) to label the number of GPS points for road type \( c_i \) which has the smallest distance value among the four categories. Then, the percentage of available distance for road category \( c_i \) is

\[
t_{c_i} = \frac{N'_{c_i}}{N_{c_i}}
\]

\( t_{c_i} \) indicates the level of consistency of the matching between GPS log point and the road networks. It represents exactly the same concept that the majority of the distribution of points in spatial should be consistent with the type of the roads. Moreover, it can avoid the possible effects induced by the missing geographical data. In this paper, we use \( t_{c_i} \) as the input variable to detect transportation modes. Four variables (\( t_{c1}, t_{c2}, t_{c3} \) and \( t_{c4} \)) will be generated based on the four road categories.

3.1.3. Bus stop and bus line information

One of the difficulties in the classification of transportation modes is the differentiation between cars and buses, due mainly to the large extent of variations in the two
transportation modes. In some cases where a bus is running on a bus priority road where the car is forbidden, the differentiation is less problematic. In many cases, however, depending on the different urban settings, a car and bus may have a similar running speed, leading to the complexity in the detection. Therefore, a further classification between the transportation modes which have similar speed should borrow additional input of other variables.

In general, buses need to anyway stop at the bus terminal and/or bus stops. Although the actual number of bus stops depends on the number of boarding passengers and waiting at the bus stops in different types of day, the location where buses stop for boarding/alighting should be close to the exact location of bus stops.

It is true that some stops detected in GPS tracks are because of a red signal at an intersection rather than a bus stop. However, it is reasonable to assume the majority of intersections with signals is not very close to the bus stop or the distance between signal location and bus stops should be larger than the distance between a stop location and a bus stop. Moreover, it seems plausible that the more bus stops detected, the more likely it is to be a bus trip.

Therefore, we propose to use bus stop information as a critical indicator. Here, we consider two types of information related to bus stops: the number of bus stops detected in the trip and whether the route is consistent with a bus line. Because the detection of bus stops depends on the stops of movement in the GPS data, we first explain how to extract such information.

For every GPS log point, we can use the speed data. Assume a normal stop \( P \) for a log point \( j \) is detected if the speed is less than 1 km/h, then

\[
P_j = \begin{cases} 
true, & \text{if } \text{speed}_j < 1 \\
false, & \text{if } \text{speed}_j \geq 1
\end{cases}
\]  

Then, for every stop point, we calculate the distance to the closest bus stop/terminals and get \( d'_j \).

Now, considering the duration of a real stop (dur), if we assume a bus should be stationary at a bus stop for at least five seconds, then the condition of a stop detected as a bus stop, \( p_b \), can be as follows

\[
p_b = \begin{cases} 
true, & \text{if } p_j = true \text{ and } \text{dur} \geq 5 
\end{cases}
\]

where \( \text{dur} \) is the summed time between adjacent GPS log points which are detected as a stop. The identification is implemented for every GPS log point of the trip. For the whole trip, the number of bus stops on the trip can be calculated, \( N_{\text{bstop}} \).

It is expected that the number of bus stops can be used to differentiate between a car and bus in the sense that car trips should include no stops or less than the number of bus stops. Of course, the number of bus stops depends on the total number of bus stops on the route. In some cases where none of the links on the route belongs to a part of bus line, there will be no bus stops detected, therefore, the bus mode should then be excluded. In general, the higher the percentage of the number of bus stops relative to the total number of bus stops on the route, the higher probability the transportation mode is a bus relative to a car. However, it can be the case that the total number of bus stops is zero for non-bus lines. Therefore, we propose to use the dummy variable on
whether the route is consistent with an existing bus line.

\[ S = \text{true, if consistent; false, if inconsistent} \]  

(5)

It should be noted that the inference procedure of transportation mode is implemented after the segmentation of activities and trips where the location information of start and end nodes is already known in advance. This allows us to detect to what extent the location is a bus stop/terminal.

To identify whether the route is a bus line, one may need a map matching process which identifies the road segment and paths on the real network. Here, to be simple, we consider a route as a bus route only if the start and end locations are detected as closed enough to bus stops/terminals, e.g. the distance is less than 10 meters.

### 3.2. Bayesian network model

A Bayesian network (BN) is a graphical representation of probabilistic causal information incorporating sets of conditional probability tables. It can be considered an enhanced naïve Bayesian model by relaxing the assumption of independent distributions in that BN consider the joint probability of an attribute with its parent attributes, while the naïve Bayesian assume all variables are independent. Thus, a BN represents all factors deemed potentially relevant for observing a particular outcome.

The model is described qualitatively by directed acyclic graphs where nodes and edges represent variables and dependencies between variables. The nodes where the edge originates and ends are called the parent and the child, respectively. Because of the statistical characteristics of BN for probabilistic inference, the probability of each value of a node can be computed when the values of the other variables are known. In a Bayesian network, each variable is conditionally independent of its non-descendant given the state of its parents. That is, if \( X_i \) is a variable with parents \( \text{parents}(X_i) \), all variables that are not descendants of \( X_i \) are conditionally independent of \( X_i \) given \( \text{parents}(X_i) \). Since independence among the variables is clearly defined, not all joint probabilities in the Bayesian system need to be calculated, which provides an efficient way to compute the posterior probabilities.

BN considers the joint probability of an attribute with its parent attributes. Suppose the set of variables in a BBN is \((X_1, X_2, \ldots, X_n)\) and that parents \((X_i)\) denotes the set of parents of the node \( X_i \) in the BBN. Then, the joint probability distribution for \((X_1, X_2, \ldots, X_n)\) can be calculated from the product of individual probabilities of the nodes:

\[ p(X_1, X_2, \ldots, X_n) = \prod_{n=1}^{N} p(X_i|\text{parents}(X_i)) \]

(6)

The network is represented as a directed graph, together with an associated set of probability tables. In our case, the Bayesian network measures the interrelationship between spatial and temporal factors (input), and activity-travel pattern (output), i.e. transportation modes. All the input variables which are selected as critical indicators are considered as child nodes of the transportation mode. The parameters are estimated using the maximum likelihood method when the network structure is determined.
3.3. Proposed approach

In order to make a better detection of transportation mode, we propose a two-step procedure: the segmentation of activities/trips, and transportation mode inference. A flowchart of the proposed approach is shown in Figure 2.

The segmentation module divides the GPS traces into different portions which can be recognized as a trip or an activity. Here, we adopt the same strategy as Moiseeva, Jessuren, and Timmermans (2010), where we use an improved BN model to predict the different patterns at the epoch level and employ a merging procedure to determine the sequence of activities and trips. Basically, the output of the segmentation imputation model is the state of movement and urban context. The category in the segmentation procedure includes not only the activity episode and trip, but also the state of devices and the surrounding environment, e.g. cold start, inside building and urban canyon. Note that the activity period can be very slow moving or standing outside where the GPS signal is

---

**Figure 2.** Proposed approach for segmentation and transportation mode detection.
still available. This is important in the sense that one needs to identify the difference between real walking and actual activities involving some slow walking.

The predicted pattern which includes five categories is further merged into two categories: activity episodes and trip episodes. In order to avoid the influences of the urban context, e.g. inaccurate observations, we have incorporated the variables related to the measurement accuracy, PDOP and HDOP, which refers to the level of the positional and horizontal accuracy. We also include the number of satellites in the imputation model, the number of used and viewed satellites. These variables can characterize the situation where a GPS log point is recorded. Other variables in the segmentation model are calculated based on the real-time speed data by setting a time window. For details on how these variables are manipulated within a time window and the design of merge rules, one can refer to the earlier work (Moiseeva, Jessuren, and Timmermans 2010). The imputation of transportation mode at the second stage is based on the segmented trip data applied with merge rules. The critical indicators used in the transportation mode imputation are extracted at the trip level, including the speed related indicators, spatial indicators and bus stop related indicators. Table 1 summarizes the full list of indicators used in the two models.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Explanation of the variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Segmentation model (indicators at the epoch level)</strong></td>
<td></td>
</tr>
<tr>
<td>PDOP</td>
<td>Position accuracy of 3-D coordinate</td>
</tr>
<tr>
<td>HDOP</td>
<td>Horizontal accuracy of 2-D coordinate</td>
</tr>
<tr>
<td>NUSED</td>
<td>Number of satellites used</td>
</tr>
<tr>
<td>NVIEWED</td>
<td>Number of satellites viewed</td>
</tr>
<tr>
<td>MAXSPEED</td>
<td>Maximum speed for the GPS logs within the time window</td>
</tr>
<tr>
<td>STDSSPEED</td>
<td>Standard deviation of speed for the GPS logs within the time window</td>
</tr>
<tr>
<td>AVGPEED</td>
<td>Average speed for the GPS logs within the time window</td>
</tr>
<tr>
<td>ACCUDIST</td>
<td>Accumulated distance for the GPS logs within the time window</td>
</tr>
<tr>
<td>MAXACC</td>
<td>Maximum acceleration for the GPS logs within the time window</td>
</tr>
<tr>
<td>AVGACC</td>
<td>Average acceleration for the GPS logs within the time window</td>
</tr>
<tr>
<td><strong>Transportation mode imputation (indicators at the trip level)</strong></td>
<td></td>
</tr>
<tr>
<td>MAXSPEED</td>
<td>The maximum speed of the trip</td>
</tr>
<tr>
<td>SPEEDPERC85</td>
<td>The percentile at the 85% level of the maximum speed of the trip</td>
</tr>
<tr>
<td>NBUSTOP</td>
<td>The number of stops which are identified as bus stops in the trip</td>
</tr>
<tr>
<td>ISBUSLINE</td>
<td>True, if the start and end location of the trip are bus stop/terminals; False, otherwise.</td>
</tr>
<tr>
<td>DRAIL</td>
<td>The percentage of logs which are most close to railways</td>
</tr>
<tr>
<td>DMOTORWAY</td>
<td>The percentage of logs which are most close to motorways</td>
</tr>
<tr>
<td>DCYCLEWAY</td>
<td>The percentage of logs which are most close to cycle ways</td>
</tr>
<tr>
<td>DFOOTWAY</td>
<td>The percentage of logs which are most close to footways</td>
</tr>
</tbody>
</table>

4. Data

4.1. Research area and spatial data

The data used were collected in the region of Eindhoven, The Netherlands. For spatial reference, we use the OpenStreetMap (OSM) data and the national road network data. The OSM data has varieties of information relating to road, land use and maps. It is a free map dataset with a massive amount of data. In this study, we use the OSM road network data and the point data of the whole Netherlands.

The road network has different categories including railways, motorways, cycle ways, footways, etc. For a motorway, there are also detailed classifications for different grades.
of road, i.e. highways or local roads. The point data includes all locational point of interest data with a full list of categories one can use for a variety of purposes. In the case of the Netherlands, the OSM data is detailed in the sense that the major locations are available in the data and the road network data is almost complete. In our study, we extracted the sub-networks by four categories (rail, motor, cycle and foot) from the national road network data file and the bus stop data from the point data file.

To generate the indicators related to the distance to different networks, e.g. motorways, cycle ways, etc., we developed a spatial searching function which can generate the distance from a particular GPS log point to a spatial object, e.g. a line or a node, which is the closest one among a set of objects. Therefore, for each GPS log point, the distances to the closest road by different road categories and bus stops are generated. A map showing the research area and the categorized OSM data is presented in Figure 3.

4.2. GPS data collection

The GPS data used in this study were collected in 2012 and 2013 in the Eindhoven region of The Netherlands. To collect the daily activity-travel data, portable GPS devices, 747pro, and a web-based survey system were used. Respondents were instructed to keep the GPS device active while traveling, upload the raw GPS traces using software that was provided to them, and verify and, if needed, correct the imputed GPS traces on a dedicated website through the Internet.

After GPS traces were uploaded to the server, a data processing program was automatically activated to transfer the GPS traces into a sequence of annotated activities and trips,
including start time, end time, activity location, activity type, transportation mode, duration and travel distance. After processing the data on-line, the imputed daily activity-travel diaries were shown on dedicated webpages of the prompted recall survey instrument. Respondents were asked to validate their activity-travel diaries and to provide some additional information that could be extracted from the GPS traces, such as travel party composition, extent and nature of multi-tasking, expenditures and judgments of travel experiences.

The data collection, running for a full year, consisted of four waves. In each wave, participants were asked to join the survey for three consecutive months. Invited participants who received a unique account name and password were directed to our webpage. Respondents could therefore log into the web survey system to manage their own activity-travel data. Respondents were compensated for their participation in the survey – a maximum amount of 40 Euro, conditional upon three months participation. The exact amount of compensation depended on the number of validated diaries, and whether the GPS devices were returned safely.

GPS traces were processed based on an improved version of TraceAnnotator (Moiseeva, Jessuren, and Timmermans 2010), which was embedded in the Web system as a background processing application. It incorporates an imputation model at the epoch level and a merge procedure to split activities and trips at the aggregated level. The resulting activity-travel diary was then shown to respondents in the prompted recall survey for validation. In this sense, it provided a way to obtain the so-called ground truth data of activities and trips. It should be realized, however, that prompted recall instruments are not necessarily error free.

After processing, the system immediately showed respondents their daily activity-travel diary data, derived from their uploaded GPS traces. Respondents could then check the correctness of the imputed diaries and revise them if necessary. They were able to change, remove or merge the activity and/or trip data as needed, and insert new data if any information was missing. A screenshot of the prompted recall page is shown in Figure 4.

The activity and trip data generated by the system includes a sequence of activities and trips. In addition, transportation modes, activity type, start and end time and travel distance formed output of the imputation model. If the transportation mode and/or activity type could not be detected, respondents were asked to select an option from a drop list of pre-defined options. Respondents were also given the opportunity to provide additional information about activities and trips in the text box. The activity and travel data were shown sequentially. For each event (either an activity or a trip), some additional questions were asked, such as travel party, degree and nature multitasking, degree of satisfaction with the trip and detailed classification of sub-activities. Only when all information was completed, the validation process proceeded. In designing these additional questions, the principle of balance was considered in the sense that the number of questions attached to activities and trips should not become too burdensome with respect to continued participation in this multi-week survey.

5. Results

The validity of the proposed approach was demonstrated by comparing the imputation results with the validated prompted recall data. The total GPS data collected in Eindhoven
involved the data from 314 respondents who reported 42,532 number of valid trips. The validated trips were then filtered by a control check manually to remove the unrealistic data which was induced by human error or ignorance in the validation procedure. The transportation modes which were not in the list of imputed modes were not considered. The total number of validated trip data used to examine the performance of the proposed model was 41,516 trips.

The comparison between the predicted transportation mode using the proposed model and the so-called ground truth is not so straightforward in that the prompted recall data involves various changes. Respondents may change the identified transportation mode in the prompted recall survey if it was recognized as incorrect. The changes can happen on the start and end time of the activities/trips, or the incorrect imputed transportation mode. In the process of comparison, one of the common issues therefore is to detect which two trips, the imputed and validated trip data, belong the same trip. This may be problematic if the validated trip data was removed and/or the trip start- or end-time was changed. Thus, we assume the segmentation is relatively accurate, where the majority of changes happened at the transportation mode correction. [This issue is also discussed in Feng and Timmermans 2014.]
Therefore, in order to accurately measure the accuracy of the proposed approach, we compare a section of the trips which are identified as consistent using start and end time. For every trip, the time information in the validated data will be compared with that in the imputed data. If the start and time are completely same in minutes, the trips are considered as the same portion of the journey. One can imagine, in this way, some validated data will not be used for comparison. However, it would not influence the result of the accuracy measurement.

5.1. The conditional probability of the transportation mode imputation

In order to obtain the conditional probabilities for the inference of segmentation and transportation mode, the models proposed in the paper were calibrated using the training samples. The conditional probabilities among the variables were fitted with a training dataset. The training data for segmentation model includes 53,030 samples at the epoch level and the transportation mode imputation model includes 372 samples at the trip level. Results of the conditional probability tables are shown in Tables 2 and 3.

It is clear from the conditional probabilities that all variables play a significant role in terms of the classification. In the segmentation model, for example, the category of ‘very small’ in the maximum speed (MAXSPEED) variable indicates the large share of activity (97.9%) and inside buildings (81.2%).

Similarly, the conditional probability for transportation mode imputation also represents the interdependencies among different variables. For example, the probability for a bus can be detected as high as 83.3% if the route is a bus line (ISBUSLINE equals true), assuming the rest does not change. With the number of bus stops increases (from the category of little to several), the probability to be a bus also increases (from 12.5% to 50%).

5.2. Results of the imputation model

Results of the imputations need to be examined comparatively with the true data. It is a common process that prompted recall data is taken as the ground truth to examine the validity of imputation algorithms. Because of the variation in the error of validation, however, the comparison between imputed data and validated data is not very straightforward.

Especially in the multi-week data collection, the error in validated data can be random, e.g. respondents may forget changing the transportation modes to the correct ones or changed to an incorrect one because of the temporal memory (Feng and Timmermans 2013b). The quality of the validation in general also depends on the closeness of the imputed data to actual data. The more closed to the actual trips of the imputed data, the less burden the respondents will have, thus resulting to a better data quality. Therefore, we additionally examine the validated data manually and visually in combination with using maps to filter out any unreasonable data.

To measure the accuracy of imputation models, the percentage of correctly identified instances is normally used in the literature. This approach is suitable to analyze the training samples. When applied to the real surveyed data, it might be biased because of the unequally distributed number of samples in terms of the participation days, number of
Table 2. Conditional probabilities of the segmentation model (values in percentages).

<table>
<thead>
<tr>
<th></th>
<th>PDOP</th>
<th>HDOP</th>
<th>USEDSAT</th>
<th>VIEWSAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very small</td>
<td>Small</td>
<td>Avg</td>
<td>Large</td>
</tr>
<tr>
<td>Activity</td>
<td>45.3</td>
<td>36.7</td>
<td>17.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Cold start</td>
<td>20.0</td>
<td>20.0</td>
<td>20.0</td>
<td>20.0</td>
</tr>
<tr>
<td>Trip</td>
<td>54.1</td>
<td>37.3</td>
<td>6.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Inside building</td>
<td>1.2</td>
<td>8.7</td>
<td>9.9</td>
<td>15.5</td>
</tr>
<tr>
<td>Urban canyon</td>
<td>3.1</td>
<td>3.1</td>
<td>3.5</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAXSPEED</td>
<td>MAXACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>97.9</td>
<td>2.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cold start</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>Trip</td>
<td>2.4</td>
<td>11.3</td>
<td>6.8</td>
<td>23.2</td>
</tr>
<tr>
<td>Inside building</td>
<td>81.2</td>
<td>14.5</td>
<td>2.1</td>
<td>1.2</td>
</tr>
<tr>
<td>Urban canyon</td>
<td>44.6</td>
<td>22.2</td>
<td>0.2</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>AVGSPED</td>
<td>AVGACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>99.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Trip</td>
<td>3.8</td>
<td>11.6</td>
<td>12.7</td>
<td>21.6</td>
</tr>
<tr>
<td>Inside building</td>
<td>87.2</td>
<td>10.9</td>
<td>1.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Urban canyon</td>
<td>75.7</td>
<td>18.1</td>
<td>2.3</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>ACCUMDISTANCE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>99.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cold start</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>Trip</td>
<td>15.0</td>
<td>18.4</td>
<td>18.4</td>
<td>12.2</td>
</tr>
<tr>
<td>Inside building</td>
<td>83.2</td>
<td>5.0</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Urban canyon</td>
<td>60.8</td>
<td>12.1</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 3. Conditional probabilities of transportation mode imputation (values in percentages).

<table>
<thead>
<tr>
<th>ISBUSLINE</th>
<th>NBUSTOP</th>
<th>MAXSPEED</th>
<th>SPEEDPERC85</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>False</td>
<td>No</td>
<td>Little</td>
</tr>
<tr>
<td>CAR</td>
<td>7.7</td>
<td>92.3</td>
<td>80.0</td>
</tr>
<tr>
<td>BUS</td>
<td>83.3</td>
<td>16.7</td>
<td>12.5</td>
</tr>
<tr>
<td>BIKE</td>
<td>9.1</td>
<td>90.9</td>
<td>76.9</td>
</tr>
<tr>
<td>WALK</td>
<td>6.7</td>
<td>93.3</td>
<td>82.4</td>
</tr>
<tr>
<td>TRAIN</td>
<td>12.5</td>
<td>87.5</td>
<td>70.0</td>
</tr>
</tbody>
</table>

DMOTORWAY | DRAIL

| CAR | 12.5 | 6.3 | 6.3 | 18.8 | 56.3 | 56.3 | 25.0 | 6.3 | 6.3 | 6.3 |
| BUS | 22.2 | 11.1 | 44.4 | 11.1 | 11.1 | 44.4 | 22.2 | 11.1 | 11.1 | 11.1 |
| BIKE | 21.4 | 35.7 | 14.3 | 14.3 | 14.3 | 64.3 | 7.1 | 14.3 | 7.1 | 7.1 |
| WALK | 27.8 | 5.6 | 27.8 | 22.2 | 16.7 | 66.7 | 5.6 | 5.6 | 11.1 | 11.1 |
| TRAIN | 45.5 | 27.3 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 9.1 | 63.6 |

DCYCLEWAY | DFOOTWAY

| CAR | 68.8 | 6.3 | 6.3 | 12.5 | 6.3 | 43.8 | 12.5 | 18.8 | 6.3 | 18.8 |
| BUS | 44.4 | 11.1 | 22.2 | 11.1 | 11.1 | 22.2 | 11.1 | 22.2 | 33.3 | 11.1 |
| BIKE | 7.1 | 7.1 | 21.4 | 50.0 | 14.3 | 28.6 | 7.1 | 35.7 | 21.4 | 7.1 |
| WALK | 22.2 | 16.7 | 33.3 | 22.2 | 5.6 | 11.1 | 11.1 | 5.6 | 38.9 | 33.3 |
| TRAIN | 54.5 | 9.1 | 18.2 | 9.1 | 9.1 | 54.5 | 9.1 | 18.2 | 9.1 | 9.1 |
trips of different participants, etc. One can imagine the absolute percentage of the same incorrect imputation might be high for a specific respondent if the respondent has a large proportion of such trips in the data, which influences the overall level of the accuracy measurement.

Therefore, a data filtering process is implemented in advance before the direct comparison between imputed and validated data. Basically, the filter ensures no duplicate instances in the sample in the sense that the possible misclassification between the same pairs of transportation modes can only appear once for each respondent. In addition to that, it is assumed that the originally imported data which participants did not change at all are all correct ones where the imputed and validated transportation modes are fully consistent. For the rest of the data which were labeled as updated, we examine how the imputation results of the proposed model match with the validated results.

The overall accuracy of the proposed imputation model is presented in the application for training samples and real surveyed trip data, respectively. This is to examine the generality of the proposed algorithm. As shown in Table 4, the accuracy for training samples is 93.2%, which is good relative to the accuracy in similar works of GPS data imputation.

When applied to the surveyed data, the accuracy yields 88.1% based on the filtered number of samples (813 trips in this case). This is satisfactory in terms of the validation in real data, especially considering that few existing works have reported the accuracy of the imputation algorithms in actual applications.

In order to further investigate the variation in the classification, we examine the detailed accuracy results for specific transportation modes. Figure 5 shows the accuracy of the proposed imputation model when applied to the survey data, represented by the hit ratio of different transportation modes. The hit ratio is an indicator which indicates the percentage

Table 4. Overall accuracy of the proposed imputation model.

<table>
<thead>
<tr>
<th></th>
<th>Correct instances</th>
<th>Total instances</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy for training samples</td>
<td>347</td>
<td>372</td>
<td>93.2%</td>
</tr>
<tr>
<td>Accuracy for surveyed data</td>
<td>716</td>
<td>813</td>
<td>88.1%</td>
</tr>
</tbody>
</table>

Note: ‘Instances’ includes filtered trip data without duplicated instances in terms of respondent and pairs of transportation modes.

Figure 5. Hit ratio of the transportation modes for surveyed data.
which is correctly predicted for each class based on the real classes. The higher the value, the more accurate the predictions.

As can be seen, the car and walk both achieve an accuracy which is higher than 90%. A satisfactory level of accuracy is also obtained for bike (89.2%) and train (83.2%). Among all transportation modes, bus results into the lowest level of accuracy, 77.9%, which is understandable regarding the mixed range of driving speeds with other transportation modes in cities.

6. Conclusions

One of the remaining challenges in GPS data imputation is whether and how much the algorithms are reliable and/or ready in the sense that they can be used in activity-travel data to replace traditional surveys, without the need for a prompted recall survey. The average accuracy reported in existing literature varies in terms of different contexts of the applications, which is not ideal yet. Further improving the imputation process and algorithms for the classification of transportation modes is still a highly demanding task. Moreover, the proposed imputation algorithms are mainly verified using well-managed training data. Reporting on the performance of imputation algorithms in real applications is rare. Therefore, it is highly necessary to develop a more sophisticated imputation algorithm/process and to examine their performance using actual data.

Therefore, in this paper, we have proposed an integrated imputation process and algorithm to enhance the imputation accuracy. The proposed approach involves a two-step procedure which incorporates a segmentation model to classify activities and trips and a transportation mode inference model. The segmentation model considers the imputation at the epoch level while the transportation mode imputation model emphasizes on the merged trip data. The transportation mode imputation considers the features related to the pattern of transportation modes specifically, e.g. number of bus stops, consistency with a bus line, and spatial indicators extracted using a geographical database, which are thought to be useful to further classify the pairs of transportation modes with similar speeds.

The performance of the proposed approach was examined using the training data and the prompted recall data collected in a large-scale travel survey project. Results show that the overall accuracy of the transportation mode imputation reaches 93.2% for training data and 88.1% for surveyed data. Further comparison between imputed and validated data shows that all transportation modes are detected with a satisfactory level of accuracy.

The integrated approach proposed in this paper can be extended to a broader scale of applications. For example, the list of transportation modes included in this paper considers those specific to the Eindhoven case. Larger cities may have more travel options, including trams, metros and light rail systems. Depending on the complexity of the traffic flow distribution and efficiency of system operations, the state in the variables may be adjusted based on the real observations collected from specific cities. We intend to consider the application of the proposed approach to other cities in the future.

Disclosure statement

No potential conflict of interest was reported by the authors.
References


Stopher, P. R., Q. Jiang, and C. FitzGerald. 2005. “Processing GPS Data from Travel Surveys.” Paper presented at the 2nd international colloquium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications, Toronto, June.

