EJTIR

Issue 15(4), 2015
pp. 662-674
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Detecting activity type from GPS traces using spatial and temporal information

Tao Feng[1]
Urban Planning Group, Department of the Built Environment, Eindhoven University of Technology, The Netherlands.

Harry J.P. Timmermans[2]
Urban Planning Group, Department of the Built Environment, Eindhoven University of Technology, The Netherlands.

Detecting activity types from GPS traces has been important topic in travel surveys. Compared to inferring transport mode, existing methods are still relatively inaccurate in detecting activity types due to the simplicity of their assumptions and/or lack of background information. To reduce this gap, this paper reports the results of an endeavour to infer activity type by incorporating both spatial information and aggregated temporal information. Three machine learning algorithms, Bayesian belief network, decision tree and random forest, are used to investigate the performance of these approaches in detecting activity types. The test is based on GPS traces and prompted recall data, collected in the Rijnmond region, The Netherlands. Results show that the random forest model has the highest accuracy. The model incorporating spatial and temporal information can predict activity types with an accuracy of 96.8% for the used dataset. These findings are expected to benefit research on the use of GPS technology to collect activity-travel diary data.

Keywords: *activity type, GPS, random forest.*

1. Introduction

Activity types (or trip purposes) are key aspects in modeling travel behavior. Traditionally, data related to activities and trips are collected using activity-travel diaries in the context of travel surveys. The information collected in this way mostly relies on what people reported. It has been widely recognized that the data collected with conventional survey methods can be erroneous as people may not be able to provide correct or accurate information about their activities and trips, i.e. duration and location information. As an alternative of traditional survey methods, GPS-based travel surveys may be more efficient in collecting accurate information about people's activities and trips in the sense that the data, which is automatically recorded, represents substantial details at both spatial and temporal scales (Wolf, 2001).

The imputed activity types and transportation modes from GPS traces provide inputs for transport modelling, accessibility analysis and discrete choice modelling (Neutens et al., 2007,

[1] A: Rondom 70, 5612AP Eindhoven, The Netherlands T: +31 40247 2301 F: +31 4024 38488 E: t.feng@tue.nl

[2] A: Rondom 70, 5612AP Eindhoven, The Netherlands T: +31 40247 3315 F: +31 4024 38488 E: h.j.p.timmermans@tue.nl

2008; Van Wee and Geurs, 2011). However, the detection is not straightforward as the activity and travel pattern need to be extracted from GPS traces. A variety of procedures to transform GPS traces into segmented activities and trips have been proposed in literature (Wolf, et al., 2004; Stopher, et al., 2005; Chung and Shalaby, 2005; Tsui and Shalaby, 2006; Du and Aultman-Hall, 2007; Deng and Ji, 2010). One common approach is a sequential procedure where GPS traces are divided into activities or trips using an empirical time gap value, e.g., an activity is recognized if the time gap is shorter than 120 seconds (Schuessler and Axhausen, 2009). The transportation modes and activity types are then detected respectively for each portion of the segmented GPS traces. Alternatively, Moiseeva, et al. (2010) and Feng and Timmermans (2013) treated trips and activities together considering the difference in patterns. A set of merge rules was incorporated to determine the most possible segmentation of trips and activities.

Relative to detecting transportation mode, the imputation of activity types from GPS traces has not been addressed sufficiently in literature. Basically, different activities have their own unique characteristics. The timing and duration of activities at the aggregate level may vary between different types. For example, home or work activities are generally considered to have a longer duration than other activities. Differences between activities may also relate to time of day. For instance, the work commute mostly happen during peak hours in weekdays while shopping activities are likely to happen during off-peak time and/or weekends (Schönfelder, et al., 2006).

In spite of the possible reasoning in the temporal perspective, the type of activity also depends on spatial information (Wolf, et al., 2004). In general, the data related to activities has different characteristics in the spatial representation from the data related to travel where the latter follows a sequence of movement. The GPS traces related to activities however are either missing or fluctuated to different locations around a certain area. The measurement is relatively less accurate because of effects of urban contexts. This results into higher complexity in the detection of activity types relative to transportation modes.

Due to the importance of locational information, existing studies commonly combine GPS traces with GIS data, i.e. land use data, point of interest (POI) data, to assist the identification of activity types. The rules developed based on such data are relatively simple. In principle, the shorter the distance from a trace point to a targeted location, the more likely the activity type matches the location type. For example, Wolf, et al. (2001) showed that activity types could be derived from GPS data and an extensive GIS land use database. Wolf, et al. (2004) derived the activity types for home activity if the activity location is within 200 meters from the home location. Other types of activities were derived by matching the GPS data with available points of interest and land use data. Although the results are promising, it also indicated that more GIS data are needed to increase the accuracy.

To get a better quality of activity type data, Stopher, et al. (2008) incorporated individual locations into the process to derive activity types. Four personal addresses of respondents were collected beforehand, including, workplace or school, and the address of the two most frequently used grocery stores. A set of heuristic rules were designed to deduce the activity types. Bohte and Maat (2009) adopted similar rules to derive activity types. The activity locations were determined from GIS maps listing points of interest data. For example, if a trip ends within a radius of 50 meters from a known location, it is assumed to be the location being visited. If an activity is located within 100 meters from home or work address, it is assumed as a home or work activity. This method is considered as efficient. However, it relies on the completeness of personal location information. In case that personal location data were not provided, activity types were then assigned as missing.

In order to improve the high rate of missing activity types, Moiseeva, et al. (2010) used multiple data sources to detect activity types. A searching procedure was designed to match the activity location with different data sources in a sequential order of personal address data, point of interest data, land use data and road network data. The method is treated as a kind of integration of existing approaches, while keeping the lowest level of missing activity types. However, such a deterministic approach is still limited in the sense that simply using spatial attributes only may not be able to capture the differences in temporal scales.

As a recent extension using the deterministic approach, Li and Stopher (2013) proposed an improved process by introducing additional aggregated information related to activity duration and tours to derive activity types. The rules were obtained by analyzing the basis information based on the NHTS data. Results showed the accuracy by incorporating the determined rules yielded an accuracy of 66.5%. Although using the activity duration and tour-related information is interesting, it seems there is still room to improve the efficiency of the imputation algorithm.

Machine learning algorithms are normally considered as be a better alternative of deterministic approaches regarding the intrinsic inference and learning capability. In general, deterministic approaches may perform not as good as machine learning algorithms, because the latter is relatively more flexible to handle complex problems. In a context of multi-functional land use areas, i.e. shopping centers, activity types cannot be easily detected using spatial information only because of the too much details or missing information of the locations. In this case, incorporating more relevant attributes into the identification procedure becomes extremely necessary.

In the perspective of methodological development, different approaches have been applied to detect activity locations and activity types. Griffin and Huang (2005) used a density based clustering algorithm to identify the most possible point of interest location related to the activity location. A decision tree model was then used to infer the activity type using the data such as time of day, duration and arrival time. The accuracy reported reached nearly 90%. However, it was only implemented in a small pilot trial. A lot of details were insufficiently presented, e.g., which activity types were inferred and how the input variables were used.

McGowen and McNally (2007) used two algorithms, the decision tree and the discriminant analysis algorithm, to detect activity types. They used 26 input variables to identify 22 types of activities. Accuracy of the two algorithms is similar, around 62%. However, authors also pointed out the potential problems which can be conducted in future, e.g. improve the accuracy of activity location detection, choose an optimal list of activity type categories, etc. In recent, Reumers, et al. (2013) used a decision tree-based model to infer activity types from GPS traces. The model considers only two indicators, the activity start time and activity durations, without considering the locational information. An accuracy of 74% and 76% for training and test data was achieved respectively. The models indicated the importance of time information in the semantic enrichment process.

In summary of existing studies, the method adopted to recognize the type of activities is still relatively simple. Researchers have been using locational or temporal information to detect activity types. Combined use of the two types of information is rare. Moreover, it seems there is still room to improve the imputation accuracy of the existing methods. To fill this research gap, we attempt to propose an advanced approach to infer activity types from GPS data. The paper will shed lights to the enhancement of the discrimination between different activity types for GPS-based travel surveys. More specifically, we incorporate both aggregated temporal information and spatial attributes into a learning-based model. The performance of this approach will be investigated based on three popular machine learning algorithms, Bayesian belief network, decision tree and random forest. The GPS traces and prompted recall data collected recently in the Rijnmond region, The Netherlands, will be used for model estimation and validation.

The remainder of the paper is organized as follows. Section 2 will briefly introduce the three algorithms. Section 3 will present our GPS data and the process to extract the attribute information for imputation. Section 4 will present the results. The paper will be concluded in Section 5.

2. Imputation algorithms

Because of the inherent difference in activity characteristics, detecting activity types can be treated as a kind of classification problems. Regardless of the ad hoc rule based approach, in this paper, we only investigate the feasibility of advanced machine learning algorithms. Regarding the key roles and popularity of these algorithms in application of GPS data imputation, we selected three representative methods, the Bayesian belief network (BBN), the decision tree (DT) and the random forest (RF). We will first briefly describe some fundamental elements of each algorithm in the following sections.

2.1 Bayesian belief network

A Bayesian (Belief) 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 naive 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-descendent given the state of its parents. That is, if X_i is a variable with parents *parents(X_i)*, all variables that are not descendants of X_i are conditionally independent of X_i given *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.

A 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:

$$\boldsymbol{p(X_1, X_2, \ldots, X_n) = \prod_{n=1}^{N} p(X_i|\mathbf{parents}(X_i))} \tag{1}$$

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 and activity episode. All the input variables are considered as child nodes of the activity type. The parameters are estimated using the maximum likelihood method when the network structure is determined.

2.2 Decision tree

Decision tree (DT) is a type of learning method for approximating discrete-valued target functions. It classifies instances by sorting them down the tree from the root to leaf nodes, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute. In general, an instance is classified by repeating a process which starts at the root node, tests the attribute and moves down the tree branch.

Among a variety of decision tree learning methods, C4.5 has been used popularly. C4.5 builds decision trees from a set of training data using the concept of entropy. The training data is a set

S $= s_1, s_2, \ldots, s_n$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, \ldots, x_{p,i})$, where the x_j represent attributes or features of the sample, as well as the class in which s_i falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy).

$$entropy(j|\overline{s}) = \frac{|s_j|}{|\overline{s}|} log \frac{|s_j|}{|\overline{s}|} \tag{2}$$

$$entropy(\overline{s}) = -\sum_{j=1}^{n} \frac{|s_j|}{|\overline{s}|} log \frac{|s_j|}{|\overline{s}|} \tag{3}$$

$$Gain(\overline{s}, j) = entropy(\overline{s} - entropy(j|\overline{s})) \tag{4}$$

Then, the attribute with the highest normalized information gain is chosen to make the decision.

C4.5 has been considered as a very efficient classifier in many other applications (Kotsiantis, 2007). We will compare C4.5 with other major algorithms in this paper in the context of transportation mode detection.

2.3 Random forest

Random forests (RF) are considered as an ensemble method for classification problems. It starts with a standard machine learning technique by constructing a forest of uncorrelated decision trees. The algorithm for inducing a random forest was first developed by Breiman (2001). The so called out-of-bag estimate was used to monitor the error of random forest. The accuracy of a random forest depends on the strength of the individual tree classifiers and a measure of the dependence between them. A tree with a low error rate is a strong classifier. Therefore, increasing the correlation and the strength of the individual trees increases and decreases the forest error rate, respectively.

The out-of-bag estimates normally mean the internal estimates of the generalization error, classifier strength and dependence between trees. The study of error estimates for bagged classifiers in Breiman (1996) gives empirical evidence to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out-of-bag error estimate removes the need for a set aside test set. In another words, using random forest does not need the separate test data set for cross-validation.

RF grows each tree based on a random selection made from the examples in the training set. All the trees are trained with the same parameters but on different training sets. These sets are generated from the original training set using the bootstrap procedure: for each training set, the same number of vectors is randomly selected as in the original set. The vectors are chosen with replacement. That is, some vectors will occur more than once and some will be absent. At each node of each trained tree, not all the variables are used to find the best split, but a random subset of them. With each node a new subset is generated. However, its size is fixed for all the nodes and all the trees. None of the built trees are pruned.

3. Data and processing

The GPS data used in this study was collected recently in a large-scale data collection project conducted in the Eindhoven and Rotterdam regions, The Netherlands. A Web-based data collection and data processing system which was developed before was improved and used in this study. The survey was divided into four waves. In each wave, participants were invited take the GPS logger for three months. The data were recorded for every three seconds. In addition to this, individuals also need to download data from the logger, upload their GPS traces and validate their daily activity-travel sequences online. The data regarding activities and trips were then derived using the Trace Annotator system from the GPS traces. In the follow-up online prompted recall survey, people were requested to fill out and/or correct missing information which they think is inaccurate or missing. The system allows changing, removing and merging the imputation data, and creating new activity/travel data. Confirmed daily activity-travel sequences are automatically saved into a background database. For details regarding the data collection, one can refer to the paper by Feng and Timmermans (2014).

In this paper, we use the data from Rotterdam because unlike Eindhoven the city has a broader set of available transportation modes, including metro and tram. The test of the algorithm is therefore more critical in this city. The Rijnmond region includes the city of Rotterdam and its surrounding suburbs and municipalities. Around 6 million inhabitants are living in the Rotterdam region. Data on daily activity-travel patterns were collected for approximately 430 respondents consecutively in the Rijnmond Region. As a result, around 329 respondents fully or partly completed the survey. A map representing the region is shown in Figure 1.

To detect activity types and transportation modes, different data were used in the imputation process. The GPS trace data were used in combination with the point of interest data to measure the spatial relationship between two locations. The prompted recall data (validated data) which are although not necessarily error free are considered to perfectly describe true activity-travel patterns. More specifically, the activity type data obtained from the prompted recall survey will be used to investigate the accuracy of our algorithms.

The categories of activity types in the survey cover both in-home activities and out-of-home activities. For each activity type, it may have multiple sub-categories. For example, a "home" activity can be just staying at home or a social activity, e.g., meet friends or relatives. Therefore, in this paper, the "home" activity was not considered. In addition, we only include the main activity types and exclude the "waiting", "other" and "unspecified" activities. In total, eleven activity types are used in this paper, as shown in Table 1. The data were filtered in advance regarding the categories of activity types where data related to missing activity information were ruled out. The final data used for imputation has 10,545 number of activity records.

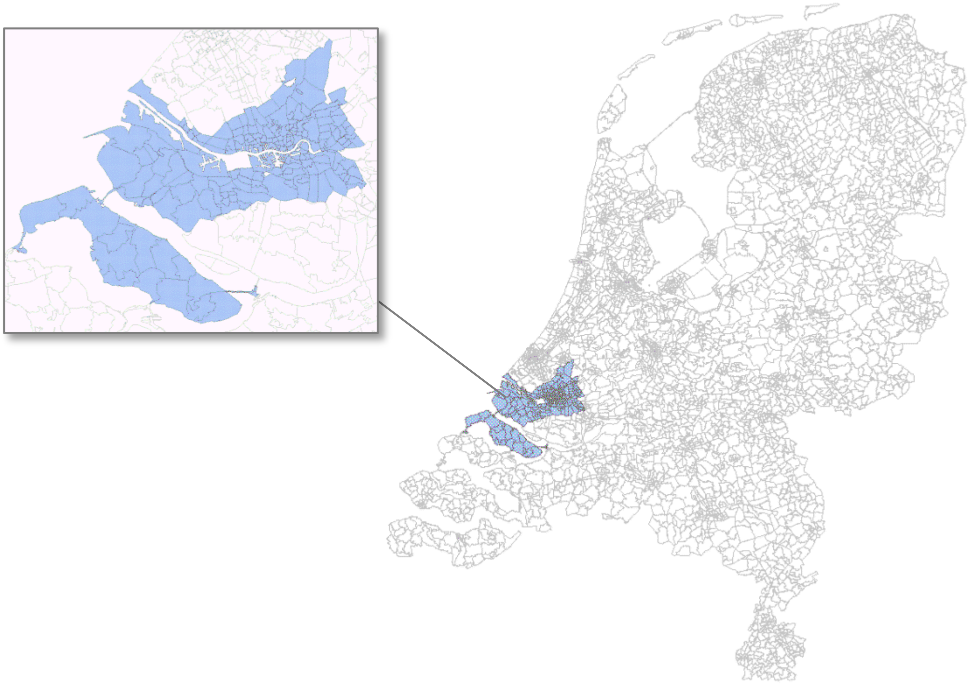

Figure 1. The Rijnmond region

Table 1. Categories of activity types and transportation modes

Items	Categories
Activity types	Paid work, daily shopping, non-daily shopping, help parents/children, recreation, social, voluntary work, service, leisure, picking-up people, study
Transportation modes	Walk, bike, bus, car, taxi, tram, metro and train

As shown in the table, there are two types of work activities (paid work and voluntary work). Some non-daily activities are grouped as service, e.g., getting money from ATM, cutting hair, visiting home doctor, etc. In the prompted recall survey, a sub-activity type was listed for selection if the main activity included sub categories. Because the type of activities may also be relevant to which transportation mode is used, Table 1 also lists the major 8 types of transportation modes used to conduct activities. The other two transportation modes, motorped and running, are not listed in the table because these modes are quite few and did not appear in the filtered dataset. In the following analyses, we do not simulate the effects of these two transportation modes.

3.1 *Characteristics of activity types*

Before we further move into the results of imputation algorithms, we first examine the differences in the patterns of different activities using the validated data. This might help us to understand the importance of contribution factors. Without loss of generality, we use the average activity duration and starting time as an example. These two indicators are mostly used in previous studies.

Figure 2 shows the average duration by different activity types. It is obvious that different activities have different level of average duration. As one can see the paid work and study have a longer duration than other activities. This is reasonable and understandable. In addition, it also seems that people pay slightly longer time for non-daily shopping than daily grocery shopping.

Figure 3 shows the distribution of starting time by different activity types. The starting time was classified into eight categories. It is evident that most work activities happen in the morning during 7:00~9:00, while social activities happen mostly during the afternoon time during 13:00~15:00 and after 7:00 at night. In addition, most daily shopping activities happen between 9:00 in the morning and 17:00 in the afternoon. Taking a special look at the picking-up activity, it happens mostly during the morning time period (7:00~9:00) and evening time period (15:00~17:00), while the frequency during morning time is higher than that during evening time.

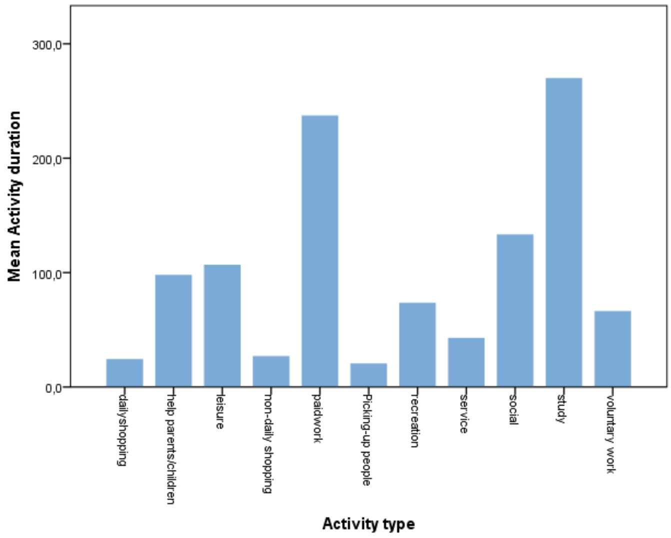

Figure 2. Average duration of different activities

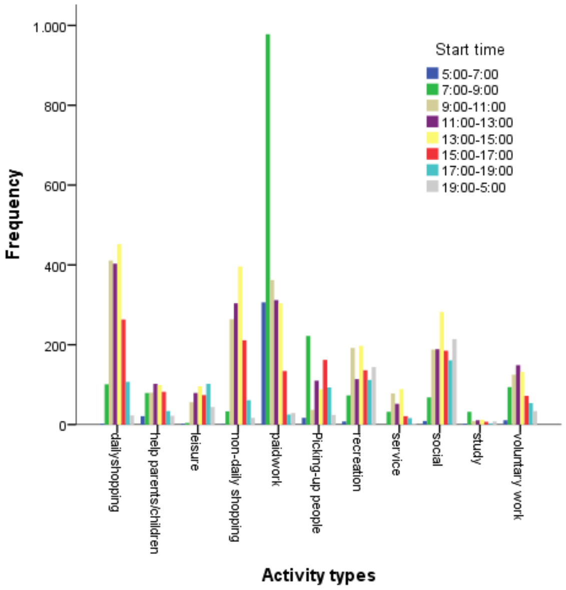

Figure 3. Start time of activities

3.2 *Feature extraction*

Because the activity types are related to both the type of the spatial location and the aggregate timing and duration information, we therefore consider the factors from both the spatial and temporal scales. To extract the spatial and temporal attributes, we process the GPS traces in combination of GIS data. The variables considered as the model input are listed in Table 2.

Except for the variables related to activities, i.e. activity duration and starting time, information regarding trips to activities are also considered here. The idea behind there is that the type of activities has a relation to the transportation mode. For example, people may not often use bus or train for grocery shopping activities.

Table 2. Variables used in the imputation algorithms

	Names
Input variables	Activity duration
	Starting time of activities
	Transportation mode to the activity location activities
	Travel time to the activities
	Have locations related to paid work activities, 1 yes and 0 otherwise
	Have locations related to daily shopping, 1 yes and 0 otherwise
	Have locations related to help parents/children activities, 1 yes and 0 otherwise
	Have locations related to non-daily shopping activities, 1 yes and 0 otherwise
	Have locations related to recreation activities, 1 yes and 0 otherwise
	Have locations related to social activities, 1 yes and 0 otherwise
	Have locations related to voluntary work activities, 1 yes and 0 otherwise
	Have locations related to service activities, 1 yes and 0 otherwise
	Have locations related to picking-up people activities, 1 yes and 0 otherwise
	Have locations related to study activities, 1 yes and 0 otherwise
Output variables	Activity types

Table 3. Correctly identified instances (CCI) among the three models

	BN	DT	RF
Correctly identified instances	46.2%	69.8%	96.8%

In addition, the variables related to the locations to certain activity are a kind of spatial information with dummy values. Basically, all points which are within 100 meters to the activity location were considered here. This method differs from the definite way where the activity type is same as the type of closest GIS data, but represents the spatial correlation in a flexible manner. It means that, within a certain area, the value will be 1 if the pre-defined area covers data belongs to the specific type, and 0, otherwise. The number of such variables considered here is same as the number of activity type categories. This is expected to enhance the capability to infer activity types by incorporating spatial information.

Such type of data is obtained in the way that the geocoded activity location data were matched with the personal location data and POI data. The personal location data were prepared with geocoded location information from the intake survey. The POI data, which are from the Openstreetmap data source, was processed in advance to keep consistent with our pre-defined activity categories.

Because the location data of activities collected in the prompted recall survey are in the format of street names, the first step is then to transform the text of street names into geo-coded locations, which can be used to measure the spatial correlation between activity location and GIS data. However, the street names can be also erroneous, which leads to the case that we cannot find any coordinates in practice. Therefore, a sequential process was designed. First, the street names were matched with an online geocoding program to fetch the information through Google service. For the street names which cannot be geo-coded, we used the original GPS traces. Extracting the coordinates from the original imputation/GPS traces is based on the validated data. The geo-coding process was designed like this: first, the starting time and ending time of activities are compared with the time in GPS traces. A block of epoch data were then recognized. The final coordinates attached to the activity location was recognized as the middle epoch in this block.

Because the data was validated by individuals, it is normal that some activities/trips were added by the respondents additionally. This means the portion in the original GPS trace does not exist because of the device was left somewhere or the battery runs out. In this case, matching the time information with the trace data becomes impossible because the geo-information of the activity location cannot be derived from missing data. In this case, we first compare the name with the personal locations. If it can be matched, the coordinates in personal data will be used as the coordinate of the activity location. Otherwise, the activity location is considered as missing. In validating the performance of different algorithms, the data where coordinates are not recognizable were excluded.

4. Results

4.1 Performance of the three algorithms

The three algorithms presented above were applied using the same dataset. Because the RF does not need the test dataset, we do not have to divide it into a training data and a test data. Here, we use the whole dataset to compare the performance of different algorithms. Table 3 shows the overall accuracy.

The CCI means the percentage which the number of data were recognized correctly divided by the total samples. The larger is the value, the better the accuracy. As shown in the table, RF results into the highest accuracy (96.8%) among the algorithms. In this sense, RF is the best algorithm to infer activity types. This is perhaps because of the inherent flexibility of RF models where the multiple decision trees can handle very complex problems. Therefore, in the next section to investigate the prediction accuracy for each activity type, we will use RF only.

The overall accuracy of DT yields 69.8%, which is relatively higher than BN. It should be noted that DT has been applied also in other studies (McGowen and McNally, 2007; Reumers, et al., 2013) to detect activity types. The accuracy obtained in existing studies using DT ranges from 62% to 75%. It means that the accuracy using our GPS data and the models is consistent with other findings. Moreover, the BN model only yields the accuracy of 46.2%, lower than 50%. It perhaps means that the BN model may not be suitable to detect activity types in such a way of combination of spatial and temporal information.

4.2 Confusion matrix of random forest model

Because of the better performance of RF relative to other algorithms, we use RF to further investigate the imputation accuracy for each activity type. The confusion matrix is used to examine the correctness and incorrectness of the imputation algorithm for each activity type. Confusion matrix is also known as a contingency table or an error matrix. It is a specific table layout to represent the performance of our algorithm. As shown in Table 4, each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The diagonal cells (in bold) represent the hit ratios which indicates the percentage which is correctly predicted for each class based on the real classes.

The table shows that the highest prediction accuracy is obtained for the paid work activity (99.0%). As presented in previous sections, the work activity normally has a longer duration then others. This might account for some reasons for the high accuracy. Furthermore, the spatial information related to the paid work variable is obtained based on the personal location data given by respondents, indicating that the work location data is more definitive than other POI data.

Table 4. Confusion matrix (predicted and real values by activity types) of RF model

		Predicted classes											
		A	B	C	D	E	F	G	H	I	J	K	Total
Real classes	A	99.0%	0.3%	0.0%	0.2%	0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	100%
	B	0.4%	97.7%	0.0%	0.9%	0.3%	0.1%	0.1%	0.1%	0.1%	0.3%	0.0%	100%
	C	0.6%	1.7%	95.8%	0.6%	0.2%	0.6%	0.4%	0.0%	0.0%	0.2%	0.0%	100%
	D	0.2%	3.2%	0.0%	95.5%	0.3%	0.2%	0.4%	0.1%	0.0%	0.2%	0.0%	100%
	E	0.2%	0.5%	0.2%	1.5%	96.8%	0.2%	0.3%	0.0%	0.0%	0.2%	0.0%	100%
	F	0.5%	0.5%	0.2%	0.4%	0.5%	97.5%	0.2%	0.0%	0.2%	0.2%	0.0%	100%
	G	1.2%	1.3%	0.3%	1.3%	1.2%	1.2%	93.1%	0.1%	0.0%	0.1%	0.0%	100%
	H	0.3%	2.1%	0.3%	1.0%	0.7%	0.7%	0.0%	94.8%	0.0%	0.0%	0.0%	100%
	I	0.4%	1.3%	0.0%	0.4%	0.7%	2.6%	0.7%	0.0%	93.7%	0.2%	0.0%	100%
	J	0.3%	0.9%	0.3%	1.5%	0.5%	0.5%	0.1%	0.1%	0.0%	95.8%	0.0%	100%
	K	7.1%	0.0%	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	91.7%	100%

Note: A- Paid work; B- Daily shopping; C- Help parents/children; D- Non-daily shopping; E- Recreation; F- Social; G- Voluntary work; H- Service; I- leisure; J- Picking-up people; K- Study

Apart from the paid work activity, the accuracy for voluntary work activity is relatively low (93.1%). Unlike the major paid work activity, voluntary work normally requires multiple different locations and sometimes the duration for each activity can vary as well. Moreover, people may also not be able to provide the exact multiple work locations by themselves, which makes the detection difficult. In this sense, the identification of voluntary work activity may be mainly attributed to the temporal information.

The study activity, which has a similar duration than paid work activity, results in the lowest accuracy (91.7%) among all activity types. Looking into the details, one can find that a substantial percentage of study activities are incorrectly recognized as work activities (7.1%). This is perhaps because the two types of activities have a similar pattern of duration. It might also indicate that the distinction between work and study activities may need additional information of individuals, e.g., if he/she is a student, or one uses the age of individuals.

Social activities are important since they account for a substantial percentage of daily trips. As presented in Table 4, the accuracy of imputing social activities is as high as 97.5%. The high accuracy may be attributed to the fact that different social activities, which can happen in different locations, may have consistent timing and duration characteristics. For example, people who work during the day mostly conduct social activities after work or during night time.

Moreover, the daily shopping and non-daily shopping activities have a high accuracy, 97.7% and 95.5%, respectively. Similarly, the Help parents/children activity, recreation activity and picking-up activity also obtain a good accuracy (95.8%, 96.8% and 95.8%, respectively). In summary, the RF demonstrates a good prediction capability.

5. Conclusion and discussion

Although activity types are considered as difficult to predict relative to transportation modes, it is potentially possible to discriminate activities in terms of their intrinsic features from the perspective of spatial and temporal scales. GPS data which has detailed spatial information potentially provides additional dimension in the identification process. Incorporating both spatial and temporal information into the activity type detection seems to be important to increase the prediction accuracy, which in turn benefits to the data imputation in travel surveys and follow-up travel behavior analysis.

As presented in this paper, the activity type can be better discriminated according to integration of aggregated temporal information and spatial attributes. The results using three modern algorithms showed that the random forest model is outstanding among the three machine algorithms. The overall accuracy yields 96.8%. The other two algorithms, although have been adopted in other studies, did not show the accuracy as good as the random forest model in this paper. The Bayesian belief model results into the accuracy of 46.2%, which means it may not be suitable to identify activity types. The decision tree model obtained the accuracy of 69.8% which is similar to the findings in other studies.

Although one can imagine the activity types may be highly dependent on the personal locations which individuals visited frequently, this paper assume an automatic imputation without personal information. Part of the reasons is because that, in reality, people may not intend to provide detailed information of all personal locations, i.e. the location of picking-up their children from school. Inferring activity types then will be mainly dependent on the spatial data such as POI and the intrinsic characteristics of different activities at the aggregate level. In our survey, most people did not provide their personal location information completely as we required in the intake survey. The missing information has an effect on the activity type detection. It is obvious that the more complete the personal addresses, the higher accuracy the final detection. However, our approach will bring additional value to avoid the possible privacy issues related to GPS-based travel surveys.

Alternatively, the imputation accuracy maybe improved by incorporating personal attributes. As showed in the above analysis, knowing the status of work or study of individuals might help the discrimination between work activities and study activities. In addition, the point of interest data used in this paper was filtered as consistent with the pre-defined category. The data is not sufficient enough considering the requirement of data imputation, which results into the case that points might be not obtainable within a diameter of 100 meters. Therefore, a better/rich geographical dataset should be more supportive to infer activity types. It should be noted that activities like service or leisure can include multiple sub-categories. To what details the activity types are necessary to be divided into sub-categories will depend on the purpose of the research questions. We leave the further detection of sub-activity types to the future.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 230517 (U4IA project). The views and opinions expressed in this publication represent those of the authors only. The ERC and European Community are not liable for any use that may be made of the information in this publication.

References

Bohte, W. and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C*, 17(3), 285-297.

Breiman, L. (1996) Out-of-bag estimation. ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5-32.

Chung, E.H. and Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, 28(5), 381-401.

Deng, Z. and Ji, M. (2010). Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach. *Traffic and Transportation Studies,* 2010, 768-777.

Du, J. and 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. and Timmermans, H.J.P. (2014). Travel survey using GPS devices: Experiences in The Netherlands In: *Mobile Technologies for Activity-Travel Data Collection and Analysis*. IGI Galobal.

Feng, T. and Timmermans, H.J.P. (2013). Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C*, 37, 118-130.

Griffin, T. and Huang, Y. (2005). A decision tree classification model to automate trip purpose derivation. In: *Proceedings of the ISCA 18th International Conference on Computer Applications in Industry and Engineering*, November 9-11, 2005, Sheraton Moana Surfrider, Honolulu, Hawaii, USA.

Kotsiantis, S.B. (2007). Supervised machine learning: A review of classification techniques. In: *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, The Netherlands.

Lin, L., Patterson, D.J., Fox, D. and Kautz, H. (2007). Learning and inferring transportation routines *Artificial Intelligence*, 171, 311-331.

McGowen, P. and McNally, M. (2007). Evaluating the potential to predict activity types from GPS and GIS data. In: *Proceeding of the 86th Annual Meeting of the Transportation Research Board*, January, Washington, D.C.

Moiseeva, A., Jessuren, J. and 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.

Neutens, T., Witlox, F. and De Maeyer, P. (2007). Individual accessibility and travel possibilities: A literature review on time geography. *European Journal of Transport and Infrastructure Research*, 7(4), 335-352.

Neutens, T., Schwanen, T., Witlox, F. and De Maeyer, P. (2008). My space or your space? Towards a measure of joint accessibility. *Computers, Environment and Urban Systems,* 32, 331-342.

Reumers, S., Liu, F., Janssens, D., Cools, M. and Wets, G. (2013). Semantic annotation of global positioning system traces: Activity type inference. *Transportation Research Record: Journal of the Transportation Research Board*, 2383, 35-43.

Schönfelder, S., Li, H., Guensler, R., Ogle, J. and Axhausen, K.W. (2006). Analysis of commute Atlanta instrumented vehicle GPS data: Destination choice behavior and activity spaces. Paper presented at the 85th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2006.

Schuessler, N. and Axhausen, K.W. (2009). Processing raw data from global positioning systems without addition information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105, 28-36.

Shen, L. and Stopher, P.R. (2013). A process for trip purpose imputation from global positioning system data. *Transportation Research Part C*, 36, 261-267.

Stopher, P.R., FitzGerald, C. and Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C*, 16(3), 350-369.

Stopher, P.R., Jiang, Q. and FitzGerald, C. (2005). Processing GPS data from travel surveys. Paper presented at 2nd International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, Toronto, June 2005.

Tsui, S.Y.A. and Shalaby, A. (2006). An enhanced system for link and mode identification for GPS-based personal travel surveys. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 38-45.

Van Wee, B. and Geurs, K. (2011). Discussing equity and social exclusion in accessibility evaluations. *European Journal of Transport and Infrastructure Research*, 11(4), 350-367.

Wolf, J., Guensler, R. and Bachman, W. (2001). Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. Paper Presented at 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M. and Axhausen, K.W., (2004). 80 weeks of GPS-traces: Approaches to enriching the trip information. *Transportation Research Record: Journal of the Transportation Research Board*, 1870, 46-54.