Chapter 5

Semantic Annotation of Mobile Phone Data Using Machine Learning Algorithms

Feng Liu, JianXun Cui, Davy Janssens,
Geert Wets and Mario Cools

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70255

Abstract

Cell phone call location data has been utilized for the study of travel patterns, but the underlying activities that originate the movement are still at a less explored stage. Resulted from routine and automated features of decision-making processes, human activity and travel behaviour exhibit a high level of spatial-temporal periodicities as well as a certain order of the activities. In this chapter, a method has been developed based on these regularities, which predicts activities being conducted at call locations. The method includes four steps: a set of comprehensive variables is defined; feature selection techniques are applied; a group of state-of-the-art machine learning algorithms and an ensemble of the above algorithms are employed; an additional enhancement algorithm is designed. Using data gathered from natural communication of 80 users over a period of 1 year, the proposed method is evaluated. Based on the ensemble of the models, prediction accuracy of 69.7% was achieved. Using the enhancement algorithm, the performance obtained 7.6% improvement. The experimental results demonstrate the potential to annotate call locations based on the integration between machine learning algorithms and the characteristics of underlying activity and travel behaviour, contributing towards the semantic interpretation and application of the massive data.

Keywords: cell phone location annotation, activity and travel behaviour, machine learning algorithms, feature selection techniques, sequential information

1. Introduction

1.1. Problem statement

Nowadays, cell phones are frequently used as an attractive means for sensing human behaviour on a large scale. They provide a source of real and reliable data, enabling automatic monitoring call and travel behaviour of users. Studies have been conducted to discover statistical laws that govern the key dimensions of human travel, e.g. travel distance and time spent at different locations [1]. These studies provide a modelling framework capable of describing general features of human mobility.

However, despite the discovery of these general features, previous studies do not provide further insights into the motivation or activities behind the identified mobility features. In general, most of the current research on cell phone data has focused on spatial-temporal dimensions. The behavioural aspects associated with the mobility features, e.g. travel mode and activities being conducted at the locations, are still at a less studied stage. Due to privacy concerns, cell phone data provided by phone operation companies usually does not have contextual information, leading to a wide gap between the raw data and the semantic interpretation of the traces. If a method can be found which helps to bridge this gap, the potential applications of the semantically enriched phone data are immense. They include inferring people's travel motivations in activity-based transportation modelling, mining individual life styles and activity preferences in urban planning, and providing activity tailored services in the cell phone environment [2].

1.2. Related state of the art

Methods have been developed to derive activities being conducted at a location from global positioning systems (GPS)-based data or from multi-modal data recorded by cell phones. The GPS-based methods first decompose continuous GPS points into a chain of *stops,* where the individual stays for a minimum period of time conducting activities, and *moves* that are the points between two consecutive stops. The stops are then compared with a geographic map by matching them in space, and interesting places that are relevant to the studies are subsequently found. The GPS-based methods have received much attention during the past years [3], but are still faced with a number of limitations. (1) The data collection process is expensive in terms of battery consumption of GPS devices. (2) Linking a GPS trajectory to detailed geographic information on all interesting places in a study area needs a lot of computational work. (3) The methods are location-specific, and the quality of the annotation process depends on the study area, making the process not transferable to other areas. (4) The matched location alone may not disclose a particular reason of why an individual travels there. A person could go to a place (e.g. a shopping mall) with different purposes (e.g. working, shopping or having a lunch). (5) The matching of exact GPS positions raises privacy concerns, as some of the places visited by an individual may be highly privacy-sensitive.

Some of the above-described limitations have been addressed by the annotation process based on multi-modal data recorded from sensors equipped on cell phones [4]. This process is composed of two steps. In the first step, data from GPS and other sensors (e.g. Wi-Fi and accelerometer) is collected from each individual. The data is then clustered into a number of visit places, each of which is represented by an ID number rather than geographic positions of the cluster points. In the second step, the obtained places are annotated based on contextual information from the sensors and phone applications, as opposed to GPS data. In this process, various machine learning methods are proposed, and different sets of features are defined [5]. These studies achieved good prediction performance without the need of additional geographic information and GPS data. Nevertheless, while the machine learning methods eliminate the need for a map, this entire annotation process still partly relies on GPS data for the identification of visit places in the first step. Thus, this process as a whole does not fully address the privacy concern. On top of that, while these studies mainly focus on selecting efficient classification models and relevant features, none of them have conducted post-processing analysis to examine how the predicted results are consistent with the sequential information that is embedded in daily activity and travel sequences. In-depth examination into the prediction errors is also lacking in these studies.

1.3. Research contributions

Extending the current research on annotating people's movement traces, our study proposes a new approach. The method utilizes data collected from simple cell phones, and it combines machine learning methods with the characteristics of underlying activity and travel behaviour that originates the traces. It has the following advantages over the existing studies. (1) The method is based on spatial-temporal regularities as well as sequential information intrinsic to human activity and travel behaviour. (2) It does not depend on additional sensor data and map information, reducing data collection costs and increasing transferability. (3) An enhancement algorithm has been developed to improve the prediction results by machine learning methods. (4) A set of extensive experiments and in-depth examination into the classification errors have been conducted. (5) Compared to GPS points, the wide coverage of a cell ID allows the process to reduce privacy concerns considerably.

The rest of this paper is organized as follows. Section 2 introduces the cell phone data and Section 3 elaborates on the annotation process. Experiments are conducted in Section 4 and examination into the experiment results is carried out in Section 5. Finally, Section 6 ends this chapter with major conclusions and discussions for future research.

2. Data

The cell phone data is composed of full mobile communication patterns of 80 users over a period of 1 year, collected by a European phone company for billing and operational purposes. The data records the location and time when each user performs a call activity, including initiating or receiving a voice call or message. The locations are represented with cell IDs, each of which has a coverage ranging from a few hundred square metres in cities to a few thousand in rural areas. The users along with their phone numbers and the corresponding cell IDs are all anonymized. **Table 1** illustrates typical call records of an individual identified as '10027534' on a day.

User ID	Cell ID	Time	Duration	Call type	Direction
10027534	10163	10:18	12	Voice call	Outgoing
10027534	10269	12:40	0	Message	Incoming

[a] The columns, respectively, denote the user, cell ID, time and duration (in minutes) of the call, the call type including 'voice call' and 'message' and the direction including 'incoming', 'outgoing' and 'missed calls'.

Table 1. Call records of a user.[a]

Among all the users, 9132 distinct call locations were detected and 259 (2.8% of the total identified locations) were labelled with activities conducted at these places. These labelled locations are used as the ground-truth data for training and validating our models. Activities are divided into five types, including 'work/school', 'home', 'social visit', 'leisure' and 'non-work obligatory', accounting for 30, 29, 15, 14 and 12% of the training data, respectively. The type of 'work/school' represents all work- or school-related activities outdoors; while 'home' accommodates all time spending at home. 'Social visit' refers to all visit activities, 'leisure' includes recreational activities outside home, e.g. sports and eating/drinking, and 'non-work obligatory' consists of activities like bringing/getting people, shopping and personalized services. If activities in multiple types are executed in the same location for a particular individual, the most frequent activity is selected, such that each location is uniquely linked to an activity type for the individual.

3. Methodology

3.1. Overview of the approach

The approach incorporates basic knowledge about human activity and travel decision-making processes and their resultant activity and travel behaviour. As Liu et al. [6] underlined, human activity and travel decision-making processes demonstrate routine and automated features. People do not generally schedule their activities on a daily basis; but rather depend on fixed routines or scripts executed during the day without much alteration. This leads to a high level of spatial-temporal regularities in activity and travel behaviour as well as a certain sequential order of the activities [6]. The spatial-temporal recurrences of the locations can be adequately reflected in the movement traces of cell phone users through a long period of call records. In addition, the spatial-temporal constraints of locations, stemming from the characteristics of various activities, which are performed in their own daily, weekly or monthly rhythms, can thus suggest the possible activities carried out at the locations. This enables the annotation for the third dimension, i.e. travel motives (activities). Furthermore, evidence also suggests that activity and travel behaviour differs across various time periods of a day, between weekdays and weekends, and between normal days and holidays [7].

The method consists of four major steps. (1) A set of variables characterizing call locations in the spatial-temporal dimensions is defined. (2) Feature selection techniques are applied to choose the most effective variables. (3) Upon the obtained variables, a set of classification models and an additional ensemble method to combine these prediction results are employed. (4) An enhancement algorithm is developed to improve the annotation performance based on sequential constraints of the activities.

3.2. Variable definition

For each user, all distinct locations, where the person has performed at least a call activity during the entire data collection period, are extracted. Let N as the total number of these locations. At each location Loc_i $(i = 1\ldots N)$, a set of variables is defined from two perspectives, including the *call behaviour* and the underlying *travel behaviour*. The *call behaviour* defines the variables that are directly related to call communication activities. Most of the variables are also used in the multi-modal data annotation process, as described in Section 1. The *travel behaviour*, however, approximates the spatial-temporal features of a location. The difference between these two perspectives can be illustrated by two groups of major variables. The first group includes the call frequency CFreqR and visit frequency VFreqR. CFreqR depicts how often calls are made at a location; by contrast, VFreqR reveals how often the location is reached, irrespective of the number of calls that are made at each visit. The second is the call duration CDur and visit duration VDur. CDur describes the duration of the call; while VDur is defined as the time interval between the first and last calls at the location. Apart from the different perspectives, all the variables are also divided based on spatial-temporal factors, including spatial repetition, temporal periodicity, day types and day segments. All the variables are listed in **Table 2**.

In terms of day segments, different definitions of time periods have been adopted, depending on the context of the study area [8]. Instead of making such an a priori assumption, a method that is proposed in this study estimates the splitting points of the day from empirical data. The resultant splitting points delimit the largest difference in the distribution of various activity types across these time intervals. Specifically, the segment process starts with a full day of 24 hours, and each hour is examined independently. An hour under investigation divides the day into two time intervals, e.g. 0–10 am and 10 am to 24 pm at 10 am. A contingency table is then constructed, in which these two time intervals and the five activity types are the row and column variables, respectively. The frequencies of the aggregated observations from the labelled call locations that fall into the corresponding time intervals and activity classes are the cell values. A chi-square statistics is subsequently calculated for this table. After chi-square statistics is obtained for each of the 24 hours, the hour with the largest statistics is chosen as the first splitting point, denoted as S_1. This point divides the day into two intervals between 0 and S_1 as well as between S_1 and 24. This process is repeated for each of the latest formed intervals, until further splitting does not generate substantial difference or until a pre-specified number of intervals is reached.

3.3. Feature selection

Due to the small size of the training dataset, particularly relative to the large number of defined variables, over-fitting is a potential problem. To address this issue, feature selection techniques are employed in order to decrease the number of predictors actually utilized by the

Travel behaviour

Spatial repetition. **(1) VFreqR**: the visit frequency at the location divided by the total visit frequencies to all locations by the individual.

Temporal variability. **(1) TotVDurR**: the total duration of all the visits to the location divided by the duration of visits to all locations by the individual. **(2) [Ear/Lat]VTime**: the earliest and latest call time of all calls at the location. **(3) AveV [StartT/ EndT], VarV[StartT/EndT]**: the average and variance of the first and last call time over all visits at the location. **(4) [Longest/Ave/Var]VDur**: the longest and average duration of all visits to the location, and the variance of the duration.

Day type. **(1) VFreqR[Week/Weekend/Sun/Sat/Hol],TotVDurR**

[Week/Weekend/Sun/Sat/Hol]: ‘VFreqR’ and ‘TotVDurR’ at weekdays, weekend, Sunday, Saturday, or public holidays.

Day segment. **(1) VFreqR[1/…/m], TotVDurR[1/…/m]:** ‘VFreqR’ and ‘TotVDurR’ are segmented during different time periods of a day.

Call behaviour

Spatial repetition. **(1) CFreqR**: the call frequency at the location divided by the total call frequencies at all locations by the individual. **(2) [VoiC/Mes]FreqR**: ‘CFreqR’ is segmented between voice calls and messages. **(3) [Inc/Mis/Out]CFreqR**: ‘VoiCFreqR’ is divided into incoming, missed and outgoing **(4) [Inc/Out]MesFreqR**: ‘MesFreqR’ is divided into incoming and outgoing.

Temporal variability. **(1) TotCDur’**: the total call duration of all calls at the location by the individual. **(2) CInt[Max/Ave]**: the maximum and average time interval between 2 consecutive calls at the location. **(3) [Ave/Var]CTime**: the average and variance of call time of all calls at the location. **(4) [Longest/Ave/Var]CDur’**: the longest, average and variance of duration of all calls at the location.

Day type. **(1) CFreqR[Week/Weekend/Sun/Sat/Hol], TotCDur’R**

[Week/Weekend/Sun/Sat/Hol],VoiCFreqR[Week/Weekend/Sun/Sat/Hol], MesFreqR[Week/Weekend/Sun/Sat/Hol]: ‘CFreqR’, ‘TotCDur”. ‘VoiCFreqR’ and ‘MesFreqR’ at weekdays, weekend, Sunday, Saturday, or public holidays.

Day segment. **(1) CFreqR[1/ …/ m], TotCDur’R[1/ …/ m], VoiCFreqR[1/ …/ m], MesFreqR[1/ …/ m]:** ‘CFreqR’, ‘TotCDur”, ‘VoiCFreqR’ and ‘MesFreqR’ are segmented during different time periods of a day.

[a] The symbol [] denotes different variables, e.g. [Ear/Lat]VTime for variables ‘EarVTime’ and ‘LatVTime’. Each day is divided into m segments, and m is decided by the method described as follows.

Table 2. Variable definition.[a]

classification models. Two methods including wrapper [9] and filter [10], which have shown effectiveness in the multi-modal data annotation process, are chosen for feature selection. Wrapper searches for an optimal feature subset using the classification model itself. In contrast, filter examines each feature separately and selects the feature that has high correlation with the target variable, but low relation with the features that have already been chosen.

3.4. Machine learning

A group of state-of-the-art machine learning algorithms, including decision trees (DTs) [11], random forests (RF) [12], multinomial logistic regression (MNL) [13] and multiclass support vector machines (SVMs)[14], are employed. These algorithms have demonstrated comparative performance for multi-category classification problems. These methods mainly differ in terms of the way the classification question is formulated, the learning function and the solution to deciding the optimal function parameters. As each learning algorithm has its strength and weakness, it is often challengeable to identify a single algorithm that performs best for a particular classification problem [15]. Thus, in this study, a fusion process is developed, which integrates the results of these algorithms, in order to utilize the strength of one while complementing the limitation of another. In this process, the four individual model prediction results (i.e. the probabilities of different possible activity types) for each call location are used as predictors, and the observed activity types are still as the dependent variable. The correlation between these predictors and the observed activity types can be built again by a classification model.

3.5. The enhancement algorithm

While machine learning methods provide an effective solution to annotating each single location, they disregard the activity orders and transitions embedded in daily activity and travel sequences. When the annotated locations on a day are linked according to the temporal order, they should follow a certain sequential constraint. The interdependencies of daily activities have been considered as a crucial factor in activity and travel decision making, as discussed in Section 3.1. By considering sequential information, the activity locations that are accessed by an individual on a day are viewed and tackled as a whole, rather than isolated participation in activities.

The enhancement algorithm takes the preliminary inference results as well as the sequential knowledge as inputs and aims to improve the prediction. The method is composed of two components: transition probability-based enhancement and prior probability-based enhancement. **Figure 1** illustrates how the prediction is improved using a daily location sequence of a user.

According to the training data of the user, he/she has conducted the chain of activities of 'work-social visit-work' at the respective call time on a day. But the prediction from the classification models is 'work-non-work obligatory-work'. A prediction error occurs at the second location. In this case, if a location (e.g. the second location) has a prediction probability P (0.443) smaller than a threshold T_1 (0.72 in our case study), it is assumed that the location is likely to be wrongly annotated. The enhancement algorithm is then applied to the false location to improve its prediction in the following steps. (1) If there is an additional location adjacent to the false one (including backwards and forwards) in the predicted sequence for that day and if this location has P larger than a threshold T_2 (0.9), it is considered as possibly correct prediction. The additional location is thus used to fix the prediction of the false one, using the

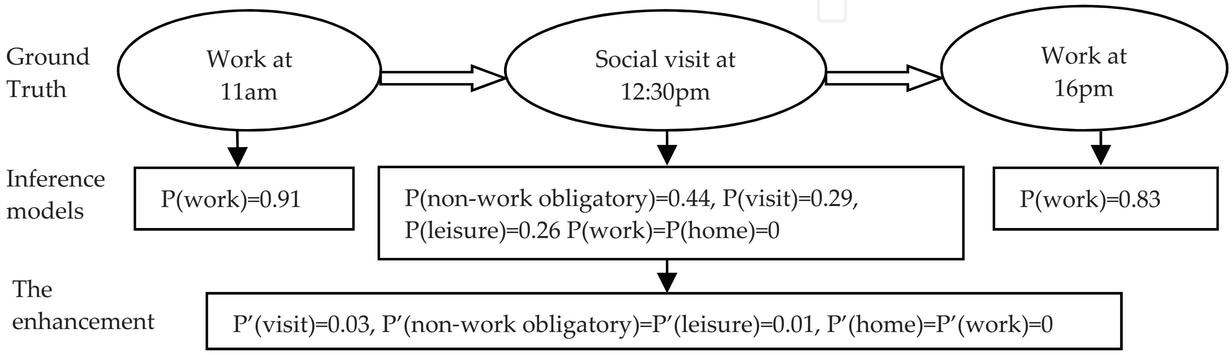
Ground Truth
Work at 11am
Social visit at 12:30pm
Work at 16pm
Inference models
P(work)=0.91
P(non-work obligatory)=0.44, P(visit)=0.29, P(leisure)=0.26 P(work)=P(home)=0
P(work)=0.83
The enhancement
P'(visit)=0.03, P'(non-work obligatory)=P'(leisure)=0.01, P'(home)=P'(work)=0

Figure 1. A daily call location sequence.

transition probability-based enhancement. (2) Otherwise, if no other locations in the neighbouring areas are predicted with a high probability, the *prior probability-based enhancement* method is employed to increase the prediction accuracy based on the call time at the false location. After recalculation, the activity type with the largest enhancement probability P' is chosen as the annotation result of the false location on that particular day. As a location may be repeatedly visited on multiple days, the multiple days' enhancement results are integrated by majority voting rules as the final annotation for the location. Under the appropriate parameters T_1 and T_2, the false prediction is likely to be corrected while accurate inference results are maintained. **Figure 2** demonstrates the details of the enhancement process.

3.5.1. Transition probability-based enhancement

The sequential information is represented in a 5 × 5 transition probability matrix between different activities. Let a_i and a_j (a_i, a_j = 1,…5) as the activities performed at the previous location i and current location j, respectively; $Tr(a_j|a_i)$ as the transition probability from a_i to a_j, calculated from the training data as follows:

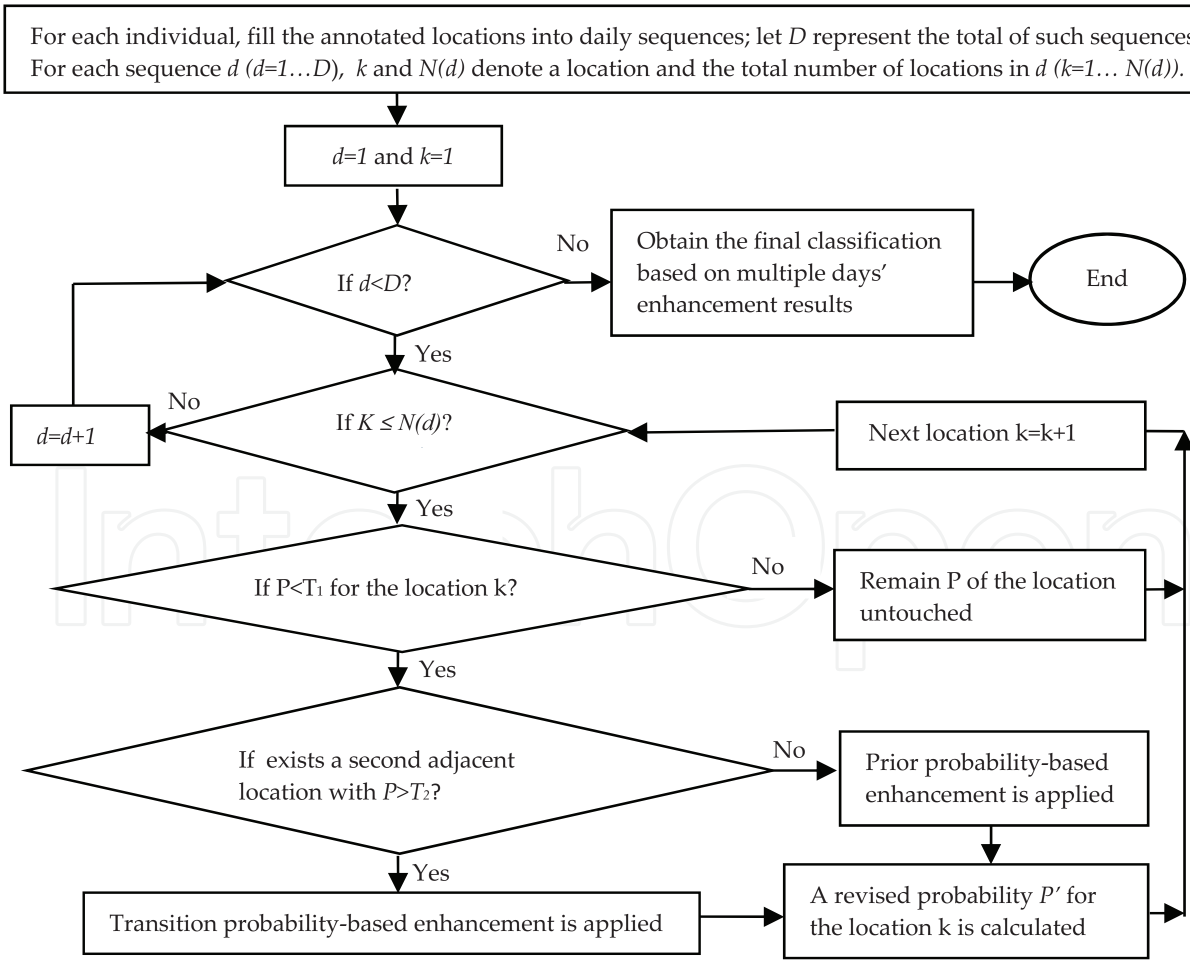
For each individual, fill the annotated locations into daily sequences; let D represent the total of such sequences.
For each sequence d (d=1…D), k and N(d) denote a location and the total number of locations in d (k=1… N(d)).
d=1 and k=1
If d<D?
No
Obtain the final classification based on multiple days' enhancement results
End
Yes
No
d=d+1
If K ≤ N(d)?
Next location k=k+1
Yes
If P<T1 for the location k?
No
Remain P of the location untouched
Yes
If exists a second adjacent location with P>T2?
No
Prior probability-based enhancement is applied
Yes
Transition probability-based enhancement is applied
A revised probability P' for the location k is calculated

Figure 2. The enhancement algorithm.

$$Tr\left(a_j|a_i\right) = \frac{F\left(a_j|a_i\right)}{\sum_{a_k=1}^{5} F\left(a_k|a_i\right)} \tag{1}$$

$F(a_j | a_i)$ is the frequency of a_j followed by a_i. The probability of the location j being annotated as a_j conditioned by a_i at the previous location i can be recalculated as $P^0(a_j | X)$ according to Eq. (2).

$$P^0\left(a_j|X\right) = P\left(a_j|X\right) \times Tr\left(a_j|a_i\right) \tag{2}$$

$P(a_j | X)$ is the result of the classification model. It is noted that $P^0(a_j | X)$ is biased towards frequently visited locations, e.g. home and work/school places, as transitions to these places are more likely than to other less visited locations. Consequently, most of the locations under Eq. (2) will be redirected to these two activity types. To overcome this, $Tr(a_j | a_i)$ is divided by the frequency of a_j, resulting in the probability $Qr(a_j | a_i)$.

$$Qr\left(a_j|a_i\right) = \frac{F\left(a_j|a_i\right)}{\sum_{a_k=1}^{5} F\left(a_k|a_i\right) \times \sum_{a_k=1}^{5} F\left(a_j|a_k\right)} \tag{3}$$

$P^0(a_j | X)$ can be revised as P' $(a_j | X)$.

$$P'\left(a_j|X\right) = P\left(a_j|X\right) \times Qr\left(a_j|a_i\right) \tag{4}$$

In the user's case, as shown in **Figure 1**, since the transition probability Qr from work to non-work obligatory activities is very small, after the enhancement, P' $(non - work - obligatory)$ (0.008) drops behind P' $(visit)$ (0.033), we get the visit activity as the revised annotation.

3.5.2. Prior probability-based enhancement

The above-described transition probability-based enhancement involves at least two locations, which are adjacent in time, and one of which has a prediction probability larger than T_2. However, such daily trajectories derived from the classification models are not always available for each day. For example, one of the neighbouring locations has a probability smaller than T_2. Or, in the case where people may stay at a location (e.g. home) during an entire day, engaging only in a single (home) activity. This is particularly true with cell phone data. People may not make calls when travelling to an activity location, resulting in the daily movement traces not being fully revealed by their call data. In these cases, we utilize the typical activity and travel behaviour at different time of a day through the prior probability distribution of the activity a_j at different call time t, i.e. $P(a_j | t)$. By applying Bayesian methods, we compute the posterior probability of a_j based on X and t, i.e. P' $(a_j | X, t)$. This probability can be computed as follows, with the assumption that X is independent of t.

$$\begin{aligned} P'\left(a_j|X,t\right) &= \frac{P\left(a_j,X,t\right)}{P(X,t)} = \frac{P\left(X,t|a_j\right)\times P\left(a_j\right)}{P(X)\times P(t)} \\ &= \frac{P\left(a_j|X\right)\times P(X)}{P\left(a_j\right)} \times \frac{P\left(a_j|t\right)\times P(t)}{P\left(a_j\right)} \times \frac{P\left(a_j\right)}{P(X)\times P(t)} \\ &= \frac{P\left(a_j|X\right)\times P\left(a_j|t\right)}{P\left(a_j\right)} \end{aligned} \tag{5}$$

$P(a_j|X)$ is the output of the classification model, i.e. the probability of a_j performed at the location j conditioned on the previously defined variables X. When $P(a_j|X)$ is compared with the new probability P' $(a_j|X, t)$, since t is added in the conditional part of P', the new probability is more discriminative and informative than P.

$P(a_j|t)$ and $P(a_j)$ can be derived from the training data as follows:

$$\begin{aligned} P\left(a_j|t\right) &= \frac{F\left(a_j|t\right)}{\sum_{a_k=1}^{5} F(a_k|t)} \\ P\left(a_j\right) &= \frac{F\left(a_j\right)}{\sum_{a_k=1}^{5} F(a_k)} \end{aligned} \tag{6}$$

Here, $F(a_j|t)$ refers as the occurrences of a_j at t and $F(a_j)$ refers as the occurrences of a_j at all time. It should be noted that from the theoretic perspective, the above enhancement process has two weak assumptions. One is the replacement of $P(a_j|X)$ with the result of the classification model and the other concerns the hypothesis of the independence between X and t. Nevertheless, based on Eq. (5), the preliminary prediction probability is complemented with the prior probability distribution.

4. Case study

In this section, adopting the proposed method and using the cell phone data described in Section 2, a set of experiments is presented. The results of these experiments are discussed and the performance of the annotation process is evaluated.

4.1. Day segments

Table 3 lists the optimal points for each of the intervals, based on the method described in Section 3.2. The first splitting point over an entire day was found at 9 am, generating two intervals of 0–9 am and 9 am to 24 pm. This process was iterated for each of the two newly obtained intervals. If the largest chi-square value over all potential points of an interval was lower than a predefined threshold, i.e. 200 in this experiment, this search stops.

Current interval	[0,24]	[0,9]	[9,24]	[9,19]	[19,24]	[9, 14]	[14,19]
S	9 am	7 am	19 pm	14 pm	20 pm	10 am	16 pm
Chi-square	3302	139	1603	855	75	194	30
If split?	Yes	No	Yes	Yes	No	No	No
New intervals	[0,9], [9, 24]	X	[9,19], [19,24]	[9, 14], [14,19]	X	X	X
Order	1	5	2	3	6	4	7

[a]The rows, respectively, denote the current interval (hour) under investigation, the optimal splitting point *S*, the chi-square value, the decision on whether or not the interval is split (if it is 'Yes' then two new intervals are formed and if it is 'No' then the symbol 'X' is used), and the order of the optimal points according to the chi-square values.

Table 3. The optimal points of a day.[a]

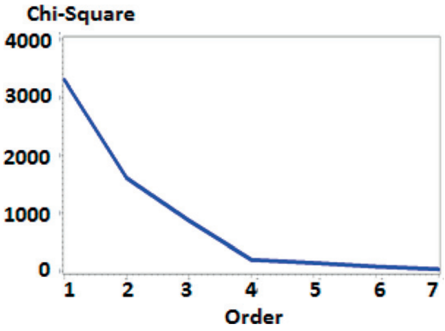

Figure 3. The evolution of chi-square statistics of the optimal points.

Figure 3 further shows the evolution of the chi-square statistics, in which the first 3 orders yield much higher values than the remaining ones. From the fourth order on, the statistics starts to decline sharply. Thus, the first 3 optimal points were extracted and 4 time periods were generated including 0–8:59 am, 9–13:59 am, 14–18:59 pm and 19–23.59 pm. After each day was segmented into the four periods, all the variables defined in **Table 2** were obtained and used as candidates for subsequent feature selection and machine learning. Weka, an open-source Java application consisting of a collection of machine learning algorithms for data mining tasks [16], was used for the implementation.

4.2. Results of individual classification models

The original training dataset is randomly divided into 10 subsets. In each model run, one of these subsets is used as the validation data and the remaining subsets combined as the training data. The number of correctly annotated locations in the validation subset is denoted as $C_i(i = 1\ldots10)$. Let *Num* as the total number of locations in the training dataset; the *prediction accuracy* can be defined as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{10} \mathrm{C_i}}{\mathrm{Num}} \tag{7}$$

The individual classification models are built on the features of locations drawn from the perspectives of both travel and call behaviour as well as on the features profiling only call behaviour, respectively. In addition, the models are also run separately on all candidate variables as well as on the variable subsets that are chosen by filter or wrapper. The prediction results with the best parameter setting in each case are presented in **Table 4**.

From the prediction results, the following observations can be drawn. (1) The models running on a subset of variables perform better than those operating on all predictors. The average improvement is 0.85% for wrapper and 2.13% for filter. This demonstrates the importance of feature selection techniques in dealing with a large number of predictors relative to a small training set. (2) There are no general conclusions on which feature selection methods are better, depending on specific classification models. SVM performs better with filter, DT and RF do not show much difference between these two feature selection techniques, while MNL gains remarkable improvement of 4.8% with wrapper. (3) When the different models are compared, it is noted that MNL produces the best results with 68.98% accuracy. This is followed by accuracy of 66.06% from RF, 65.69% from SVM and 60.95% from DT. (4) Variation is also exhibited between the variables drawn from different perspectives. In most cases, the prediction accuracy derived from the combination of both travel and call behaviour is higher than that from solely call behaviour. The average accuracy increases by 2.96 and 1.20% for filter and wrapper, and 2.09% for all variables included. This underlines the added value of the variables built based on underlying activity and travel behaviour.

Apart from different model performance, the feature selection techniques combined with various classification models also yield divergent optimal subsets of features. Eight variables are picked up by the multiple selection processes and they are regarded as important predictors, including VFreqRWeek, TotVDurRSun, VarVEndT, VarVStartT and AveVEndT describing activity and travel behaviour, and AveCallTime, IncMesFreqR and MesFreqR3 related to only call behaviour.

Classification models	**DT**	**RF**	**MNL**	**SVM-poly**	**SVM- RBF**
Parameters	*N* = 4	*N* = 0	*C* = 1	*c* = 100, degree = 1	*c* =100, Gamma = 0.01
Travel and call behaviour					
Filter	**60.95**	65.33	64.23	63.50	**65.69**
Wrapper	**1.1.** 60.58	**1.2. 66.06**	**1.3. 68.98**	**1.4.** 59.26	**1.5.** 56.57
1.6. All Variables	**1.7.** 59.12	**1.8.** 64.60	**1.9.** 63.50	56.93	**1.10.** 59.85
Call behaviour					
Filter	58.76	62.77	62.77	59.85	60.58
Wrapper	59.85	63.50	65.69	59.49	58.39
All variables	56.57	62.04	60.58	57.30	59.85

[a]The highest prediction accuracy for each model is in bold.

Table 4. Prediction accuracy of the individual classification models (%).[a]

4.3. Results of fusion models

In this fusion process, the four individual classification models are, respectively, employed as the fusion models to predict the activity types, while the results from each of the classifiers with the best parameter performance shown in **Table 4** are used as the predictors. The prediction with the two best performances for each fusion model is presented in **Table 5**. The results reveal that a fusion model does not necessarily outperform the individual models; the performance depends on the choice of the selected individual classifiers as the predictors. For instance, MNL obtains 68.98% accuracy as an individual classifier, while it achieves 69.71% when used as the fusion model built on the integration of all the four individual models' results. However, the accuracy drops to 61.68% when only DT and SVM-RBF are employed as the predictors.

4.4. Enhancement algorithm

4.4.1. Transition matrix

Similar to the temporal variables, the transition matrix is also built for weekdays, weekend and holidays separately as well as for different periods of a day. The identification of optimal cutting points for the matrix is the same as the previously described method, except the time intervals. For each potential dividing point, two intervals but three scenarios are obtained depending on the time of the two concerned activities in the transition. The first and second scenarios occur when both activities take place in the first interval or in the second. The third scenario is when the first activity takes place in the first interval and second activity in the second interval. Given the small size of the training set, only the first significant cutting point was identified, which is 18 pm. Under this time division, the largest difference in the distribution of activity transitions is among the three scenarios: transitions within 0–17:59 pm or 18–23:59 pm, and transitions from 0–17:59 pm to 18–23:59 pm. **Table 6** shows the transition matrix in the first scenario during weekdays.

Predictor	DT	RF	MNL	SVM - RBF	Accuracy
Fusion models					
DT			X	X	69.71
DT	X		X		67.15
RF		X	X		68.98
RF			X	X	68.24
MNL	X	X	X	X	69.71
MNL		X	X		68.98
SVM-RBF	X	X	X	X	67.52
SVM-RBF		X		X	67.15

[a]The rows represent the fusion models, and the columns include the individual classifiers and the prediction accuracy. X indicates the corresponding individual models being chosen as the predictors.

Table 5. Prediction accuracy of fusion models (%).[a]

Transition probability	Activity type	Home	Work/school	Non-work	Social visit	Leisure
Tr	**Home**	0.008	**0.546**	**0.700**	0.197	**0.797**
	Work/school	**0.883**	0.328	0.300	**0.701**	0.153
	Non-work	0.032	0.010	0.000	0.000	0.000
	Social visit	0.017	0.081	0.000	0.080	0.051
	Leisure	0.061	0.036	0.000	0.022	0.000
Transition probability	**Activity type**	**Home**	**Work/school**	**Non-work**	**Social visit**	**Leisure**
Qr	**Home**	0.002	**0.159**	**0.204**	0.057	**0.232**
	Work/school	0.060	0.022	0.020	0.047	0.010
	Non-work	**0.066**	0.019	0.000	0.000	0.000
	Social visit	0.023	0.114	0.000	**0.113**	0.072
	Leisure	0.059	0.035	0.000	0.021	0.000

[a] The row and column represent the current and previous activities respectively; the maximum probability for each column is in bold.

Table 6. Transition matrix.[a]

As expected, for the probability $Tr(a_j|a_i)$, the highest values are dominated by the transitions to either home or work/school activities. With $Qr(a_j|a_i)$, however, the dominance of these two activities is reduced by their high frequencies, and transitions to other less represented activities are exposed. This can be manifested by the high transitions from home to non-work activities and from social visit to second social visit locations.

4.4.2. Activity distribution at different time

The activity distribution is also differentiated between weekdays, weekend and holidays. The weekday distribution at each hour $P(a_j|t)$ is shown in **Figure 4(a)** and the distribution of the

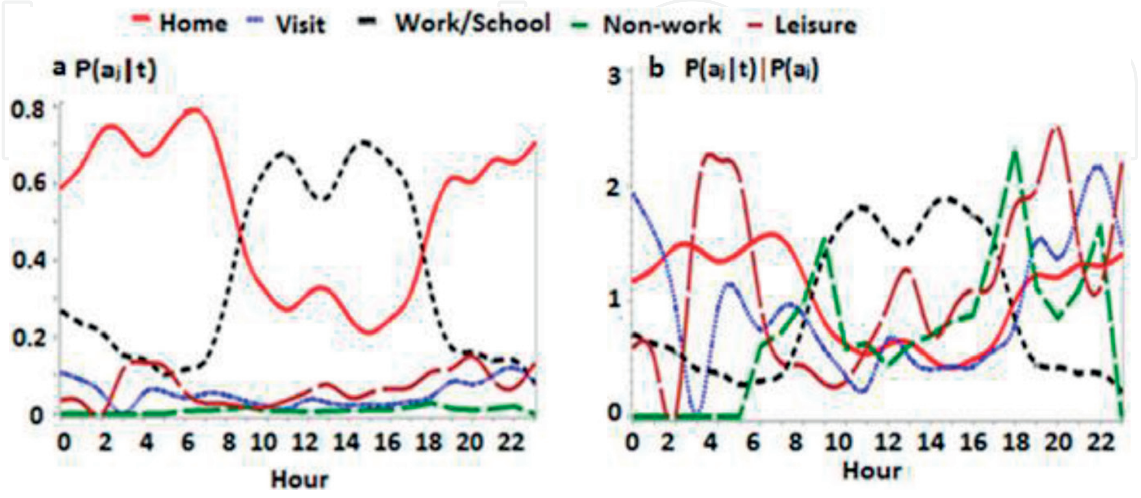

Figure 4. Absolute activity distribution (a) and relative activity distribution at each hour (b).

ratio between $P(a_j|t)$ and the overall probability of the activity $P(a_j)$ is depicted in **Figure 4(b)**. These two distributions show remarkable deviation: in **Figure 4(a)** either home or work/school types dominate the activities, whereas in **Figure 4(b)** the most likely activity shifts across various types as the day unfolds.

4.4.3. Selection of T_1 and T_2

Based on the previous results, two fusion models, including MNL built on all the four individual classifiers and RF on the combination between this model and MNL, are selected for the enhancement algorithm. To decide the threshold T_1, the correlation is examined between different values of T_1 and the prediction rates of the fusion models, as shown in **Figure 5**. It is observed that for both models, when T_1 is below the crossing point of 0.72 in **Figure 5(a)** and 0.8 in **Figure 5(b)**, the number of false prediction is higher than that of the correct one. Thus, 0.72 and 0.8 are selected as T_1 for MNL and RF, respectively. T_2 is set as 0.9, above which the prediction rate is 69.7 and 66.4% for these two models.

4.4.4. Enhancement results

Table 7 presents the prediction results by the enhancement algorithm (in the column 'After'), along with the results before the enhancement (in the column 'Before') as well as the difference between these two prediction results (in the column 'Difference'). Overall improvement of 4.4 and 7.6% for MNL and RF is achieved. The examination into the results across various activities discloses that the enhancement algorithm particularly performs better on less representative activity types, e.g. non-work obligatory, social visit and leisure activities. This could be originated from the fact that the machine learning algorithms usually favour majority types if the prediction accuracy is used as the evaluation criterion, while the enhancement algorithm puts equal weights on all activity types of the dependent variable (call locations).

The effectiveness of each of the two enhancement methods is also investigated, by running the RF fusion model using each of these methods independently to revise a weak prediction result.

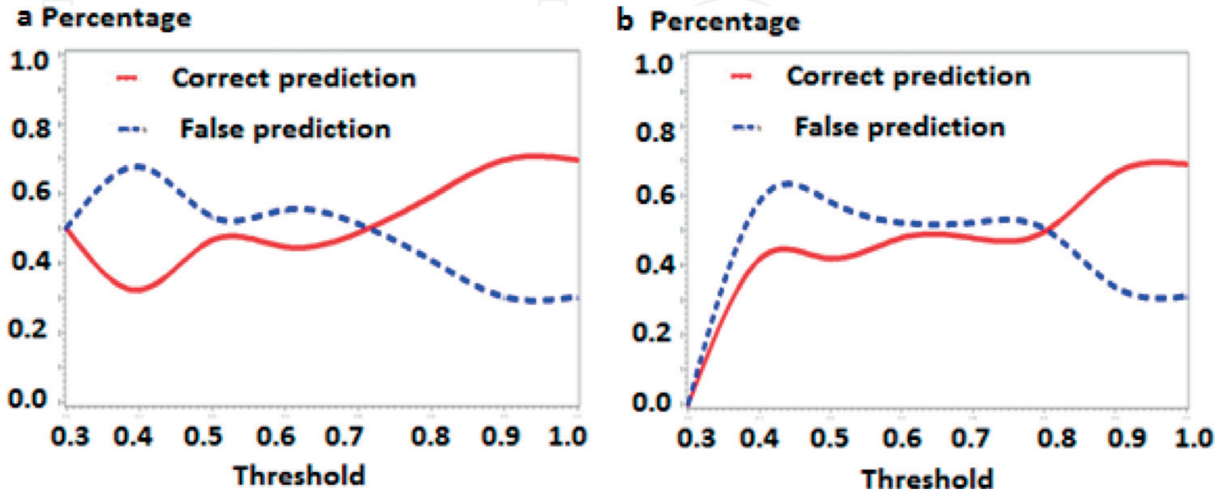

Figure 5. Relation between the prediction and the probabilities from MNL (a) and RF (b) fusion models, respectively.

Fusion model	MNL			RF		
Enhancement	**Before**	**After**	**Difference**	**Before**	**After**	**Difference**
Home	91.3	91.3	0	91.3	91.3	0
Work/school	80.9	82.0	1.1	74.2	79.8	5.6
Non-work	37.5	59.3	21.8	53.1	78.1	25.0
Visit	47.4	55.3	7.9	52.6	60.5	7.9
Leisure	45.7	48.6	2.9	37.1	51.4	14.3
Overall accuracy	69.7	74.1	4.4	69.0	76.6	7.6

Table 7. Prediction result comparison between before the enhancement algorithm and after that (%).

The prediction rates of 73.7 and 75.2% were obtained for the transition probability-based and prior probability-based enhancement methods, respectively. Due to the small size of the training set, many locations are labelled as one single known activity of a day, the sequential information is thus not available on these days. With a large dataset, the transition matrix would better represent typical activity and travel behaviour of users. This would lead to the transition probability-based method and the enhancement algorithm as a whole bringing greater improvement over the current experimental results.

5. Analysis on the prediction results

Table 8 presents the annotation results by the RF fusion model with the enhancement algorithm, showing a large variation in the prediction accuracy across different activity types. Home, work/school and non-work obligatory activities are better predictable, with the accuracy of 91.3, 79.8 and 78.1%, respectively. Social visit activities show a middle level of predictability of 60.5%. By contrast, leisure activities are only 51.4% recognizable. Overall, prediction accuracy of 76.6% is achieved. Despite the promising results, misclassification exists for each of the activity types, prompting for further examination into the potential reasons for the errors.

(1) Home. Homes are featured with high visit frequencies and spatial-temporal regularities. However, seven homes are misidentified, of which five have lower visit frequencies than 10% on weekdays, i.e. less than 1 in 10 trips on weekdays ending at home. The unusually less visited homes could be due to the fact that the corresponding users spend less time at home and/or they make fewer calls than expected at home. This results in the home visit frequencies less represented by their call records. Alternatively, some of the misclassified locations can be a second home for users who already have a home at different locations. Two in these five users have two labelled homes. While their second homes are occasionally accessed, their main homes are routinely visited and correctly annotated. (2) Work/school. Like homes, work/school locations are also characterized by a high level of routine visits, but these two types differ regarding the time of the visits. While most of the trips to homes are at night and weekends, trips to offices or schools occur during the daytime on weekdays. Of all the work/school

Annotated activity	Original activity				
	Home	Work/school	Non-work	Social visit	Leisure
Home	91.3	10.1	3.1	15.8	2.8
Work/school	1.2	79.8	9.3	7.9	14.3
Non-work	2.5	5.6	78.1	10.5	17.1
Social visit	3.8	4.5	6.2	60.5	14.3
Leisure	1.2	0	3.1	5.2	51.4

Table 8. Prediction results (%).

locations, 10.1% are wrongly predicted as non-work obligatory or social visit activities if they are accessed infrequently during weekdays. All the corresponding users work/study at multiple places, and the misidentified locations are their additional work/school places. Another 10.1% are mistaken as homes, if they have high visit frequencies at weekend. For instance, one of these users has two labelled work locations. They were visited at rates of 32% during weekdays and 42% on Sunday, respectively. While the first one was correctly identified, the second one was wrongly predicted as home. This suggests that the work regime plays an important role in distinguishing work locations from homes. While most people work during weekdays, certain minorities work on different shifts, especially to weekends or nights, generating distinct activity and travel patterns from the main stream of the population. (3) Non-work obligatory. The activities have low visit frequencies and short duration. The misclassification of the activities can be partially attributed to a combination of heterogeneity within this category. The various detailed types of the activities are likely performed at spatially independent locations and temporally varied preferences. For instance, shopping is mostly done in later time of the day than service or bringing/picking up activities. (4) Social visit. The activities are profiled with a middle level of visit frequencies during weekdays. If the locations are accessed less, they tend to be annotated as leisure or non-work obligatory activities; if more, they are considered as home or work/school places. The limited predictability could be caused by the underlying structure of an individual's social network, in which various degrees of relationship exist, ranging from closed one they visit routinely to the one they just meet occasionally. This generates variations in spatial-temporal features of the locations. (5) Leisure. Leisure activities are conducted in various places and at different time for an individual; they exhibit the lowest level of regularities and thus are the most challengeable to annotate. Apart from the spatial-temporal irregularities, the examination into two falsely predicted leisure locations reveals additional causes for the misclassification. The first one has a visit frequency of 36.3% in both the afternoon and evening on weekdays. It was the second most visited place for the corresponding user who has accessed this place 170 times over 337 days, such that 1 in 2 days he/her was observed there. This location is originally labelled as a restaurant; however, the call records suggest a high probability that he/her may work there instead of eating as a customer. The second location was ranked as the most visited place for the concerned user. He/she has in total conducted 383 visits over 442 days during both weekdays and weekends as well as at night. Nearly three in 4 days, he/she made calls there. Furthermore, the user has five locations collected in the training dataset, but none of which is labelled as home. This location is documented as sports; however, for this particular user, it is likely that this place is a home rather than a recreation site. While further investigation into the above two typical cases is needed before any definite conclusions are drawn, they nevertheless illustrate that our annotation method based on underlying activity and travel behaviour can effectively predict the activities, which are tailored to each individual. A location may have a single or multiple functions, but people visiting there could have different purposes. The match with geographic information alone is not able to identify this distinction. We shall call the location annotation at the individual level as *micro-location-annotation.*

6. Conclusions and future research

In this study, a cell phone location annotation method has been developed based on spatial-temporal regularities as well as sequential information intrinsic to activity and travel behaviour. The method does not depend on additional sensors and geographic details. The data requirement is simple and its collection cost is low. It is also generic to be transferable to other areas. On top of that, the method is independent of precisely geometric positions of individuals, thus considerably reducing privacy concerns.

Experiments on the annotation method using data collected from natural phone communication of users have achieved 76.6% prediction accuracy. With this probability, the activity conducted at a location for a user can be predicted by the spatial-temporal features of the visits disclosed by his/her call records. Furthermore, this study also shows the added value of the integration between machine learning methods and underlying activity and travel behaviour when annotating the location traces.

Nevertheless, despite the spatial-temporal regularities, activity locations still share commonalities in these two dimensions at a certain degree. Activity and travel behaviour is not solely decided by spatial-temporal elements, it is also affected by socio-economic conditions. The first improvement in future research should thus take this general background information into account. In particular, to address the potential causes for misclassifications of home and work/school locations, the annotation should be combined with the information on the number of home and work/school places of users as well as their work sectors and regimes. A broad picture of users' social networks, obtained from direct surveys and/or social networking sites, would strengthen the prediction of social visit activities. For non-work obligatory and leisure activities, the detailed types in each of these two categories should be handled separately, if a sufficient size of training data for the detailed types is available. The second improvement lies in finding an effective way of annotating locations, which are visited for multiple purposes for a particular user. While this study links the most frequent activity to a location, it dismisses additional activity types, which are performed by the user at different parts but within a same cell. In the training dataset, 5% of all the locations are visited for multiple purposes.

Today when simple phones are still prevalent constituting nearly 85% of total global handsets in use, this research makes undoubtedly an important contribution to the semantic explanation of the movement data. With the development of smart phones, the data from additional sensors installed on the phones will provide a third possibility of improvement by integrating the contextual information into the annotation process.

Author details

Feng Liu[1]*, JianXun Cui[2], Davy Janssens[1], Geert Wets[1] and Mario Cools[3]

*Address all correspondence to: feng.liu@uhasselt.be

1 Transportation Research Institute (IMOB), Hasselt University, Diepenbeek, Belgium

2 School of Transportation Science, Engineering, Harbin Institute of Technology, Harbin, China

3 LEMA, University of Liège, Liège, Belgium

References

[1] Song CM, Koren T, Wang P, Barabási AL. Modeling the scaling properties of human mobility. Nature Physics. 2010;**6**:818-823

[2] Liu F, Janssens D, Cui JX, Wang YP, Wets G, Cools M. Building a validation measure for activity-based transportation models based on cell phone data. Expert Systems with Applications. 2014;**41**(14):6174-6189

[3] Pink O, Hummel BA. Statistical approach to map matching using road network geometry, topology and vehicular motion constraints. In: The 11th International IEEE Conference on Intelligent Transportation Systems; 12–15 October 2008; Beijing, China. IEEE; 2008

[4] Laurila JK, Gatica-Perez D, Aad I, Blom J, Bornet O, Do TMT, Dousse O, Eberle J, Miettinen M. The mobile data challenge: Big data for mobile computing research. In: Proceeding Mobile Data Challenge (by Nokia) Workshop; 18–19 June 2012; Newcastle, UK. 2012

[5] Sae-Tang A, Catasta M, McDowell LK, Aberer K. Semantic place prediction using mobile data. In: Mobile Data Challenge (by Nokia) Workshop; 18–19 June; Newcastle, UK. 2012

[6] Liu F, Janssens D, Cui JX, Wets G, Cools M. Characterizing activity sequences using profile hidden Markov models. Expert Systems with Applications. 2015;**42**(13):5705-5722

[7] Cui JX, Liu F, Janssens D, An S, Wets G, Cools M. Detecting urban road network accessibility problems using taxi GPS data. Journal of Transport Geography. 2016b;**51**:147-157

[8] Cui JX, Liu F, Hu J, Janssens D, Wets G, Cools M. Identifying mismatch between urban travel demand and transport network services using GPS data: A case study in the fast growing Chinese city of Harbin. Neurocomputing. 2016a;**181**:4-18

[9] Kohavi R, John G. Wrappers for feature subset selection. Artificial Intelligence. 1997;**97**(1-2): 273-324

[10] Hall MA. Correlation-Based Feature Subset Selection for Machine Learning. New Zealand: Hamilton; 1998

[11] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufmann; 1993

[12] Breiman L. Random forests. Machine Learning. 2001;**45**(1):5-32

[13] Le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. Applied Statistics. 1992;**41**(1):191-201

[14] Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation. 2001;**13**(3):637-649

[15] Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications. 2012;**39**(12):11303-11311

[16] Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Elsevier Inc.; 2011