Discovering urban activity patterns in cell phone data

Transportation (2015) 42:597–623
DOI 10.1007/s11116-015-9598-x

Discovering urban activity patterns in cell phone data

Peter Widhalm[1] · Yingxiang Yang[2] · Michael Ulm[1] · Shounak Athavale[3] · Marta C. González[2]

Published online: 27 March 2015

Abstract Massive and passive data such as cell phone traces provide samples of the whereabouts and movements of individuals. These are a potential source of information for models of daily activities in a city. The main challenge is that phone traces have low spatial precision and are sparsely sampled in time, which requires a precise set of techniques for mining hidden valuable information they contain. Here we propose a method to reveal activity patterns that emerge from cell phone data by analyzing relational signatures of activity time, duration, and land use. First, we present a method of how to detect stays and extract a robust set of geolocated time stamps that represent trip chains. Second, we show how to cluster activities by combining the detected trip chains with land use data. This is accomplished by modeling the dependencies between activity type, trip scheduling, and land use types via a Relational Markov Network. We apply the method to two different kinds of mobile phone datasets from the metropolitan areas of Vienna, Austria and Boston, USA. The former data includes information from mobility management signals, while the latter are usual Call Detail Records. The resulting trip sequence patterns and activity scheduling from both datasets agree well with their respective city surveys, and we show that the inferred activity clusters are stable across different days and both cities. This method to infer activity patterns from cell phone data allows us to use these as a novel and cheaper data source for activity-based modeling and travel behavior studies.

Keywords Cell phone data · Mobility patterns · Activity-based models · Activity recognition · Unsupervised learning · Relational Markov network

✉ Peter Widhalm
peter.widhalm@ait.ac.at

[1] Austrian Institute of Technology, Giefinggasse 2, Vienna, Austria

[2] Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, USA

[3] Enterprise Technology Research & Innovation, Ford Motor Company, One American Road, Suite 1824 FMCC, Dearborn, MI 48176, USA

Introduction

The field of transportation modeling keeps evolving to adapt to the needs of transportation development. The required data collection and analysis procedures also change over time. The first generation of travel demand models, the four step models, were developed in the 1950s characterized by major investments in road infrastructure and rapid increase in car usage, which called for travel forecast models that are able to predict aggregate level demand in the long run (Jovicic 2001). The input data of these models are usually zoning level aggregate statistics collected through surveys. Trip attraction rates are modeled as functions of land use, which allows to forecast travel demand based on planned or anticipated changes of land use. The second generation of travel demand models, characterised by disaggregate trip based or tour based demand models, emerged when the focus of transportation modeling turned to more detailed individual level travel behavior modeling. In these models individual level travel diary surveys began to play a more important role, but trips/tours of the same individual were still analyzed independently. The Travel Model Improvement Program in the early 1990s marked the boom of the third generation travel demand models, activity based models. Many of them are based on theoretical foundations proposed by Hägerstraand (1970) and Chapin (1974). Activity based modeling treats travel as being derived from the demand for activity participation. Activities are motivated by economical, physiological and sociological needs of an individual. Travel is therefore viewed in a broader context of activity scheduling in time and space (Rasouli and Timmermans 2014). Activity based models are supposed to capture all the interconnections between activities and trips and to avoid shortcomings of trip based models such as lack of behavioral realism, strong aggregate nature, and assumption of independency between the four steps of the traditional urban transportation planning procedure. Activity-based models imply a shift from aggregate quantities and relationships to disaggregate models and micro-simulations. The focus on individual and household level decision making process poses higher requirement on individual level data collection. The ideal input of such models would include detailed activity time, location, mode choice and interaction with other family members of a large sample of the population over a long observation period. This requirement is unattainable for traditional manual surveys, which leads to the more recent focus on automated data collection methods.

GPS technology has been proposed to enhance travel surveys for more than 15 years (Casas and Arce 1999; Wolf et al. 2003a, b; Bachman et al. 2011; Nitsche et al. 2013). The application of GPS data in transportation modeling includes: generation of trip rate correction factors (Wolf et al. 2003a; Bricka and Bhat 2006), travel mode detection (Tsui and Shalaby 2006; Reddy et al. 2010; Widhalm et al. 2012), trip end/activity location detection (Wolf et al. 2001; Stopher et al. 2005; Ashbrook and Starner 2003), assessment of transportation network conditions (Hackney 2005; Stopher and Swann 2007), and route choice analysis (Jan et al. 2000; Li et al. 2005; Hood et al. 2011; Quddus et al. 2003). GPS traces can provide accurate spatial and temporal information of individuals, but generally the attainable sample size and observation period of GPS-assisted surveys are still limited. In contrast, the cellular networks of mobile phone operators act as an ubiquitous sensor that provides an immense amount of information about the movement patterns of almost the entire population. In recent years this has inspired intensive research, such as the analysis of human mobility behavior, e.g. (Gonzalez et al. 2008; Schneider et al. 2013; Ratti et al. 2006; Ratti et al. 2007; Sevtsuk and Ratti 2010; Calabrese et al. 2013; Hoteit et al. 2014), origin–destination flows, e.g. (Tettamanti and Varga 2014; Wang et al. 2013; Caceres et al.

2012; Calabrese et al. 2011; Friedrich et al. 2010), and road usage patterns (Wang et al. 2012). More detailed reviews of the use of cell phone data in traveler information systems and travel behavior studies can be found in (Qiu and Cheng 2007; Yue et al. 2014; Wang et al. 2014). While automatically collected mobile phone records have the advantages of large sample size and long observation periods, they also have obvious weaknesses: cell phone traces are sparsely sampled in time, provide only a low spatial resolution and include noise stemming from pure signal movement. Therefore the data have to be carefully processed to extract trip origins and destinations as well as starting and ending times of activities.

A simple method to extract trips from cell phone records was described in Wang et al. (2010) where consecutive location measurements are clustered according to their geographical distance and the resulting clusters are then used as origins and destinations of trips. While the clustering method accounts for moderate noise in the cell phone track, it does not include any trajectory filtering to cope with outliers resulting from occasional large positioning errors. Nor does it filter out "passing-by" points which do not indicate a trip origin or destination but instead mark positions along the route of a trip. Several suitable methods to filter cell phone trajectories were compared in Horn et al. (2014), including Kalman filtering and Recursive filtering. To filter out passing-by points some of the stay point extraction methods previously applied to GPS traces were also applied to cell phone datasets with minor adjustments (Zheng et al. 2010; Zheng et al. 2009; Hariharan and Toyama 2004; Jiang et al. 2013). These methods require at least two position records at each stay location which accurately define the stay's beginning and end. However, cell phone traces often do not fulfill this requirement because of their sparse and irregular sampling, and stays cannot be easily delimited in time.

Trip extraction is the basis for observing origin–destination traffic flows and estimating the current travel demand. But to derive information for an activity-based travel demand model, the trips need to be related to activities and trip attractors. Activities are typically categorized into classes such as "home", "work", "education", "recreation", "shopping" and so on, which are supposed to reflect basic personal and family needs. Each individual is assumed to follow a weekly and daily activity schedule and to optimize trips so as to perform all activities with a required daily or weekly frequency, taking into account constraints on time and duration of each activity as well as the transportation and activity location infrastructure, which they share with other individuals. Therefore, another interesting challenge in terms of the research proposed here is that cell phone tracks are semantically poor: they do not include activity type labels and therefore do not reveal the purposes of the trips, which are the key determinant for trip scheduling and destination choice. This shortcoming hampers their use in activity-based modeling and travel behavior studies.

For GPS data several approaches to automatically infer the type of activity conducted at each visited location have been proposed in both the transportation and computer science community. Trip end locations where matched to land use data to derive trip purposes and good agreement was reported for "go to home" and "go to work" trips (Wolf et al. 2001). A multistage hierarchical matching procedure was designed to infer trip purposes (Schönfelder et al. 2003). Other types of statistical approaches include Bayesian frameworks (Hurtubia et al. 2006; Moiseeva et al. 2010) and decision trees (McGowen and McNally 2007; Reumers et al. 2013). GPS data, GIS data, and individual and household demographic data were usually combined (Bohte and Maat 2009; Stopher et al. 2008). The relationship between consecutive trips was introduced as tour based corrections (Shen and Stopher 2013). A combination of the approaches proposed in Schönfelder et al. (2003) with probabilistic multinomial logit models was proposed in Chen et al. (2010). A framework for activity recognition using Relational Markov Networks was proposed in Liao et al. (2005). However, it remains unclear if the methods proposed for activity recognition using GPS tracks can directly be applied to mobile data, since the spatial and temporal accuracy is not comparable to that of GPS data. Moreover, the methods require manually labeled training data for the recognition of activity classes which is often difficult to obtain. On the other hand, the vast amount of data allows to analyze mobility behavior in whole new ways and opens the way towards data-driven approaches, where the required information for activity-based simulation models can be learned from the cellular data itself. To the best of our knowledge no previous study has investigated activity-behavioral clusters in cellular data with the goal to endow more semantic structure to the extracted trips, even when manually labeled training data are not available.

The contribution of this paper is twofold. First, we propose a method for robustly detecting stays and converting the raw cell phone tracks into a sequence of trips and activity locations including estimates of arrival times and stay durations. This allows to study travel patterns at intra-urban scales and forms the basis for the analysis of urban travel behavior. Different from previous approaches the method accounts for positioning errors and sparse sampling of cell phone tracks by combining properties of a low-pass filter with an incremental clustering algorithm. In order to delimit stays in time and to detect passing-by points the lower and upper bounds of arrival time and stay duration are computed based on minimum feasible travel times between the visited locations. Moreover, the method considers the geometry of travel trajectory to detect activity locations. Second, we propose an unsupervised learning method to reveal activity patterns in the cell phone tracks. Mobile phone data do not include activity labels of a predefined categorization scheme. We therefore propose a data-driven approach where the conventionally predefined activity classes are replaced by activity clusters that emerge from the data without the need of manually labeled training data. An activity cluster is defined as a set of activities that show similar properties in activity start time, duration, nearby landuse types around the activity location, frequency of similar activities and the sequence of activity locations. These features allow activity clusters to emerge naturally without predefined class labels such as “home” and “work”. Each activity cluster defines a daily frequency, attraction rates by each land use type, and properties regarding the trip chain patterns such as the number of times an activity location is visited during a day or the number of different locations where a particular activity is performed. Although the resulting activity clusters are not equivalent to activity classes in conventional surveys, they can be related by their characteristic spatial and temporal features. For example, if a cluster represents activities which start around 9 am, end around 5 pm, and the surrounding landuse are mainly office buildings, then this cluster can be interpreted as “work” activities. Once these activity clusters are revealed, the frequencies of daily activity chains as well as characteristic trip length or travel time distributions can also be easily computed. This way the discovered activity clusters can be used in activity based simulation models as a substitution for conventional activity categorization schemes. In this study we model trip attraction as a function of land use, although the proposed method can easily be extended to use other attraction factors such as points-of-interest. Similar to Liao et al. (2005) we model the dependencies between activity type, trip scheduling and land use types with a Relational Markov Network. Inference is done by sampling from the joint probability expressed by the RMN. Instead of relying on labeled training data to recognize predetermined activity types the RMN is trained in an unsupervised way, following an Expectation–Maximization scheme.

We demonstrate the proposed approach with two different datasets of the metropolitan areas of Vienna, Austria and Boston, USA, and we compare the results. For Boston the analysis was based on anonymized Call Detail Records (CDRs), whereas for Vienna mobile signaling traffic was used for analysis, including network communications of devices in idle mode. We show that the resulting activity time scheduling and trip sequence patterns agree well with data obtained with traditional surveys. The proposed method yields similar activity clusters in both cities. The clusters are stable across different workdays, while work days and weekends show different patterns, corresponding to the well-known differences in travel behavior between these types of days.

The remainder of this paper is organized as follows: the proposed method to extract trips and visited places is introduced in "Reconstruction of trips and visited places" section, and in "Activity patterns" section we explain our approach to revealing activity patterns in cell phone data. In "Results" section we describe the data sources and parameter settings and present the empirical results. "Conclusion" section concludes this paper and identifies directions for future research.

Reconstruction of trips and visited places

The goal of the first processing step is to extract from the raw cell phone records the times and locations where the cell phone user stayed to perform some activity. The challenges one faces when reconstructing activity times and locations from cell phone traces are to filter out noise while preserving the best possible spatial resolution and to interpolate the sparsely sampled trajectories to estimate arrival times and stay durations. A common approach to detecting stays in travel trajectories recorded with technologies such as GPS, is to define a radius corresponding to the positioning error and a minimum dwell-time (Hariharan and Toyama 2004). A stay is detected if the position estimates stay within the given radius for at least the predetermined time. A similar method to extract trips from cell phone records was described in Wang et al. (2010). Trajectories of moving objects can be filtered and interpolated by assuming constraints on velocity or acceleration. Several such methods to filter cell phone trajectories were compared in Horn et al. (2014), including Kalman filtering and Recursive filtering. The best results in terms of reduction of the positioning error were achieved with the Recursive Look-Ahead Filter. This filter considers the geographical distance and the time difference between consecutive records to calculate a speed, which is then compared to a threshold to detect and remove outliers.

The method proposed here combines properties of a low-pass filter with an incremental clustering algorithm to robustly detect stays and to convert the raw cell phone track into a sequence of visited places. Figure 1 illustrates the individual steps in our algorithm by an example of a cell phone track depicted in a space–time diagram. The red dots indicate cell phone records with timestamps represented on the horizontal axis and location estimates represented on the vertical axis, where, for illustration only, the coordinates are projected onto a single dimension. The short black bars extending vertically from the red dots represent a radius ρ around the location which is defined according to a desired and feasible spatial resolution. In this illustration the axes are scaled such that given an assumed maximum travel speed one spatial unit can be traversed in one unit of time. The light blue area along the cell phone track shows the resulting spatial uncertainty of the devices' location over time.

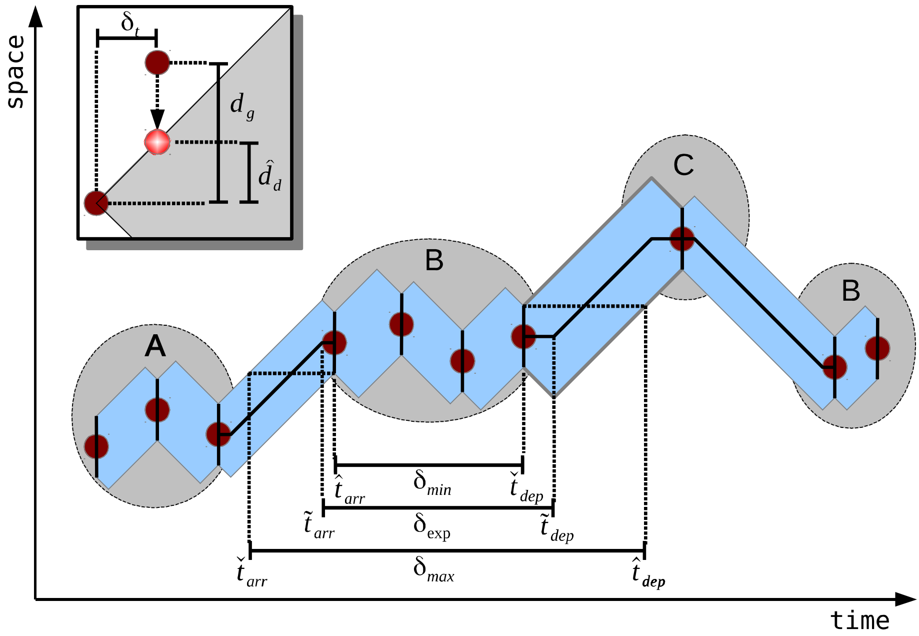


Fig. 1 Illustration of the algorithm to convert cell phone tracks into sequences of visited places

Positioning errors caused by rapid jumps to distant cells are smoothed out by introducing constraints on the travel speed. This is illustrated in the upper left corner of Fig. 1. The basic idea is to estimate an upper bound on the distance a mobile device has traveled without needing to know the devices' actual locations. The measure of distance between two cell phone records used in the clustering algorithm is therefore defined as $d^\star = \min(d_g, \hat{d}_d)$, where d_g is the geographical distance between the cell coordinates and $\hat{d}_d = \hat{v}\delta_\mathrm{t}$ is the upper bound of the distance the mobile device was able to travel, which is computed with an assumed travel speed $\hat{v}$ and the time span δ_t between the two records. In the Vienna data set it is possible to identify records belonging to the same cell in the mobile network via their location coordinates. This allows to calculate a measure of distance between pairs (Γ_a, Γ_b) of cells rather than pairs (r_a, r_b) of individual records by defining $\hat{d}_d(r_a, r_b) := \hat{d}_d(\Gamma_a, \Gamma_b) = \hat{v}\min_{r_{a'}\in\Gamma_a, r_{b'}\in\Gamma_b}\delta_t(r_{a'}, r_{b'})$, which increases the effectiveness of the approach.

The clustering procedure is described in Algorithm 1. It consecutively examines the cell phone records of a mobile device in their temporal order and incrementally creates and appends clusters of phone records with small distances d*. The distance between a new record r and an existing cluster C is computed as average distance to all the records already in the cluster:

$$D(r, C) = \frac{1}{|C|}\sum_{r_C \in C} d^\star(r, r_C).$$

The resulting clusters are shown as grey ellipses in Fig. 1 and the letters *A*, *B* and *C* are the cluster labels. In this example the sequence of clusters is therefore *A*-*B*-*C*-*B*. Each cluster represents a “virtual location” with coordinates calculated by averaging the location estimates of all records in the cluster.

Algorithm 1 Clustering of cell phone records into virtual locations.

```
Input:
    records $r \in T$ of cell phone track $T$;
Output:
    virtual locations $\mathbf{L}$;
    $\mathbf{L} \leftarrow \{\}$;
    create new cluster $C_{new}$ and add $r_1$ to $C_{new}$;
    add $C_{new}$ to $\mathbf{L}$;
    $C_{current} \leftarrow C_{new}$;
    for $i \leftarrow 2$ to $|T|$ do
        if $D(r_i, C_{current}) < \rho$ then
            add $r_i$ to $C_{current}$;
        else
            $C_{current} \leftarrow$ none;
            for all $C \in \mathbf{L}$ do
                if $D(r_i, C) < \rho$ then
                    add $r_i$ to $C$;
                    $C_{current} \leftarrow C$;
                    break;
                end if
            end for
            if $C_{current} =$ none then
                create new cluster $C_{new}$ and add $r_1$ to $C_{new}$;
                add $C_{new}$ to $\mathbf{L}$;
                $C_{current} \leftarrow C_{new}$;
            end if
        end if
    end for
```

The arrival times and stay durations at the virtual locations are estimated based on the timestamps of the cell phone records and by assuming constraints on the travel speed. An upper bound of the arrival time at location B is given by $\hat{t}_{arr}(B) = \min_{r_B \in B} t(r_B)$, the earliest transaction time at B. Likewise, a lower bound of the departure time from location B is estimated by $\check{t}_{dep}(B) = \max_{r_B \in B} t(r_B)$. Let us now consider a stay at location B, arriving from location A and continuing to location C. We assume $\check{\delta}_t(A,B) = d_g(A,B)/\hat{v}$ to be a lower bound for the travel time between A and B and estim.te thexpected arrival time

$$\tilde{t}_{arr}(B) = \frac{1}{2}\left(\hat{t}_{arr}(B) + \check{t}_{dep}(A) + \check{\delta}_t(A,B)\right)$$

Here the parameter $\frac{1}{2}$ means we assume the user would arrive randomly between the lower bound arrival time $\check{t}_{dep}(A) + \check{\delta}_t(A,B)$ and the upper bound arrival time $\hat{t}_{arr}(B)$. For the same reason, the expected time of departure is given by

$$\tilde{t}_{dep}(B) = \frac{1}{2}\left(\check{t}_{arr}(C) + \hat{t}_{dep}(B) + \check{\delta}_t(B,C)\right).$$

The stay duration is bounded by $\delta_{min}(B) = \check{t}_{dep}(B) - \hat{t}_{arr}(B)$ and $\delta_{max}(B) = \hat{t}_{dep}(B) - \check{t}_{arr}(B)$, the time span between the earliest possible time of arrival and the latest possible departure time. The expected duration of stay is estimated by

$$\delta_{exp}(B) = \tilde{t}_{dep}(B) - \tilde{t}_{arr}(B).$$

In the Vienna data set the user IDs are rotated daily at midnight which requires to introduce assumptions about the arrival time at the first location and the stay duration at the last location of a day. If the first and last virtual locations are identical, then we assume their arrival and departure times of day to be also identical. This assumption means that on the following day the phone users leave the location at the same time of day as they did on the previous day. For tracks where the first and the last virtual location differ, we simply set $\tilde{t}_{arr} = \hat{t}_{arr}$ and $\tilde{t}_{dep} = \check{t}_{dep}$

In order to identify *activity locations* and to filter out passing-by points we examine the upper and lower bounds of the stay duration as well as the geometry of the travel trajectory. Let us again consider a sequence *A*-*B*-*C* of virtual locations. Location *B* is identified as activity location if at least one of the following two criteria is met:

1. $\delta_{min} > \tau$, or
2. $\delta_{max} > \tau$ and $\left(d_g(A,B) + d_g(B,C)\right)/d_g(A,C) > \iota$,

where τ and ι are thresholds. The rationale behind the second criterion is the assumption that significant extra distances travelled are motivated by an activity, as illustrated in Fig. 2. These two criteria are applied to all the triplet sequences of virtual locations.

For illustration, an exemplary cell phorack and the reconstructed sequence of activity locations is shown in Fig. 3. In the further analysis the arrival times and stay durations at activity locations are estimated by $\tilde{t}_{arr}$. and δ_{exp}, and for easier notation we will simply write t_i and δ_i for the arrival time and stay duration of the *i*-th stay of a cell phone track.

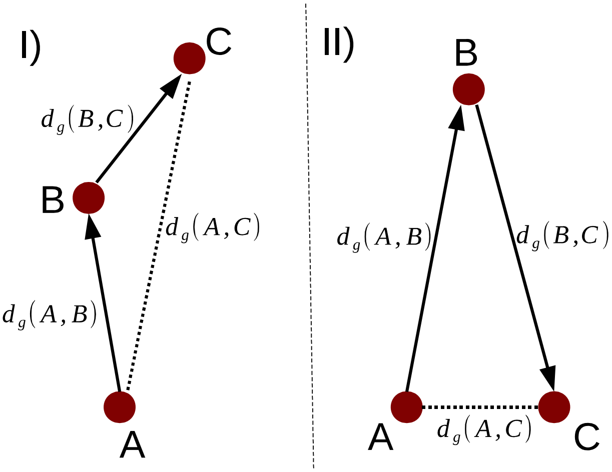


Fig. 2 Detection of activity locations by the geometry of the trajectory: in I) *B* is not detected as activity location, while in II) *B* is probably an activity location, assuming that significant extra distances travelled are motivated by an activity

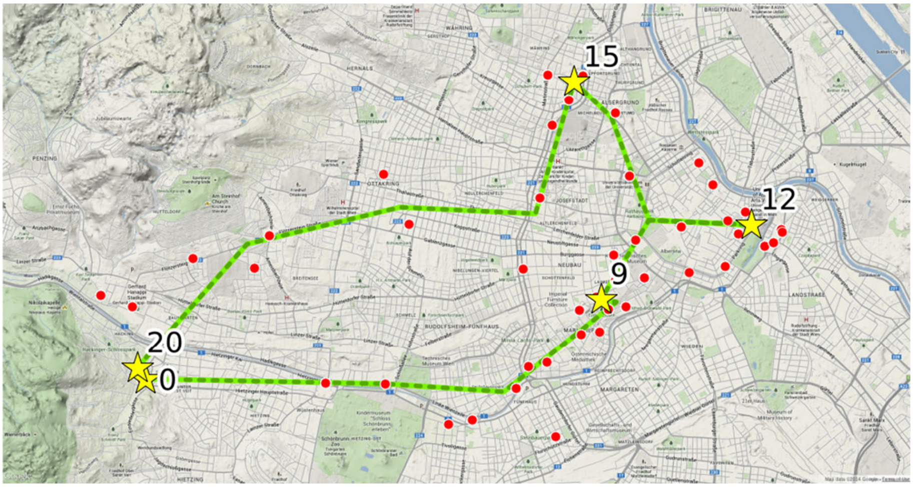


Fig. 3 Reconstruction of trips and visited places. *Red dots* represent the raw cell locations and the *green line* is the filtered trajectory. *Yellow stars* mark visited places and the numbers indicate the time of day in hours since midnight

Activity patterns

The reconstructed sequences of activity locations are suitable for analyzing trip time scheduling and to compute Origin–Destination flows between traffic assignment zones. The resulting flows will show the observed traffic without explaining any of the behavioral mechanisms behind the travel decisions. In order to allow modeling the underlying activity time and and location choice behavior, the reconstructed activity location sequences are enriched with an activity type and a land use type which serves as attraction factor in our study. However, due to the low precision of the location estimate, the land use type of the activity location cannot be unambiguously observed. Instead, we can only compute the land use shares within a buffer area around the location estimate. Moreover, a given land use type does not unambiguously determine the type of activity performed at that location.

In our approach we view the activity types as patterns defining activity time scheduling and the attractiveness of destinations. For example, the activity type "working" is often constrained by given working hours, e.g. 9 am to 5 pm, and is attracted by working locations indicated by certain land use types. In addition, there are dependencies between the sequence of activity locations and the types of activity performed at these places: some activity types, such as "working" or "being at home", are usually attracted by only one location while other activity types, such as "shopping" or "leisure" can be attracted by multiple distinct locations. Some activity types, in particular "being at home", typically involve returning to a specific location after performing some other activities elsewhere. The activity types are initially unknown and have to be inferred from the data.

We approach the problem by estimating the joint posterior probability distribution $\Pr(\mathbf{l}, \mathbf{a} | \mathbf{P}, \mathbf{t}, \boldsymbol{\delta}, \mathbf{i})$ of the land use types $\mathbf{l} = (l_1, \ldots, l_n) \in \mathcal{L}^n$ and the activity labels $\mathbf{a} = (a_1, \ldots, a_n) \in \mathcal{A}^{\mathrm{n}}$ of a track with n activity locations, given

- vectors $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_{\mathrm{n}})$ of land use shares in proximity of the estimated locations,
- arrival times $\mathbf{t} = (t_1, \ldots, t_n)$,

- stay durations $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$, and
- the sequence of activity location indexes $\mathbf{i} = (i_1, \ldots, i_n)$ which number consecutively the distinct locations visited during a day.

The activity sequences can be conveniently represented in a relational schema as shown in Fig. 4, and the joint distribution Pr(L, A|**P**, **t**, **δ**, **i**) can be modeled with a Relational Markov Network (RMN). Before we elaborate on the specific details of our probability model we provide a short introduction to RMNs in the following section.

Relational Markov networks

Relational Markov Networks (Taskar et al. 2002; Getoor and Taskar 2007) are an extension of undirected graphical models known as Markov Random Fields or Markov Networks. These models define a joint distribution over a set **V** of random variables and comprise a graph G = (**V**, *E*). and a set $\Phi = \{\phi_c(\mathbf{V}_c)\}_{c \in C(G)}$ of *clique potentials*. The links *E* of the graph indicate dependencies between the connected variables. A set $\mathbf{V}_c \subseteq \mathbf{V}$ where each $V_i, V_j \in \mathbf{V}_c$ is connected by an edge $\{V_i, V_j\} \in E$. is called a *clique*. The *clique potentials* $\phi_c(\mathbf{V}_c)$ are non-negative functions such that

$$\Pr(\mathrm{V} = \mathrm{v}) = \frac{1}{Z} \prod_{c \in C(G)} \phi_c(\mathbf{v}_c)$$

is a factorization of the joint density of **V** over the cliques $c \in C(G)$ of graph G with the normalization constant

$$Z = \sum_{\mathbf{v}'} \prod_{c \in C(G)} \phi_c(\mathbf{v}_c{}').$$

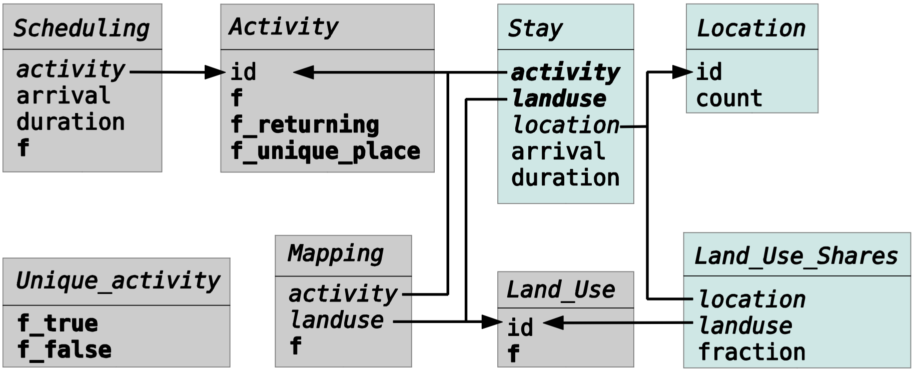


Fig. 4 Relational schema used for inference of activity clusters. The blue tables represent an activity sequence consisting of a number of stays at locations with certain land use shares. The attribute "count" in table "Location" gives the number of times the location is visited during the day. The attributes "activity" and "landuse" of table "Stay" are the label attributes to predict. The gray tables contain frequency counts *f* used for calculating the potential functions. Attributes in *bold letters* are automatically learnt from the cell phone data

If the random variables are partitioned into target variables $\mathbf{Y}$ and predictor variables $\mathbf{X}$, the conditional distribution of the target variables given the predictors is given by

$$\Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{\mathrm{c} \in \mathrm{C(G)}} \phi_c(\mathbf{x}_c, \mathbf{y}_c)$$

where

$$\mathrm{Z}(\mathbf{x}) = \sum_{\mathbf{y}'} \prod_{\mathrm{c} \in \mathrm{C(G)}} \phi_c(\mathbf{x}_c, \mathbf{y}'_c).$$

The model calibration task consists of optimizing parameters of the potential functions ϕ_c.

RMNs extend Markov Random Fields by defining them over a relational dataset. A RMN consists of a schema defining entity types and their attributes and a set of relational clique templates with their respective potentials. *Relational clique templates* are a means to specify cliques and potentials at a template level. Using a relational query language they specify the cliques to be constructed in an instantiation of the schema. The query consists of the same three parts as SQL queries:

1. a set of entitiy variables, corresponding to the *FROM*-clause,
2. a boolean condition corresponding to the *WHERE*-clause, and
3. a subset of selected attributes corresponding to the *SELECT*-clause.

Inspired by Liao et al. (2005), we extend the definition of the clique templates to allow a template to select aggregations of attributes, such as counts or sums. This makes it necessary to also include *GROUP-BY*-clauses in the queries.

For each data row retrieved by the query a clique is constructed, where the nodes correspond to the selected attributes. All cliques constructed from the same template share the same potential functions ϕ_c. Together, the generated cliques of all clique templates form an *unrolled* Markov network defining the distribution of all the *label attributes* in the instantiation conditioned on *content attributes* and on *reference attributes* which specify the relational structure.

Model specification

In our model we discretize arrival times and durations of stay to 15 min intervals. The land use shares at each activity location are computed according to the area covered by each of the land use types within a buffer area around the estimated locations. They serve as priors for the true land use type at the user's actual position before any context information is taken into account. The label attributes $\mathbf{y}_{\mathrm{c}}$. in our model are the activity and land use types at each activity location. Based on the schema shown in Fig. 4 we define the following clique templates:

- C_1: the activity type itself, to represent its prior probability:
  SELECT S.activity
  FROM Stay S
- C_2: activity type, land use type and the fraction of the buffer area covered by the land use type:
  SELECT S.activity, S.landuse, L.fraction

FROM Stay S, Land_Use_Shares L
WHERE L.location = S.location

- C_3: type, starting time and duration of the activity:
  SELECT S.activity, S.arrival, S.duration
  FROM Stay S
- C_4: the activity type and an indicator that the activity location is visited more than once during the day:
  SELECT S.activity, L.count > 1 AS is_returning
  FROM Stay S, Location L
  WHERE S.location = L.id
- C_5: the activity type and an indicator that the activity is performed at only one unique location:
  SELECT S.activity, COUNT(DISTINCT S.location) = 1 AS unique_place
  FROM Stay S
  GROUP BY S.activity
- C_6: an indicator that only one unique actity is performed at each location
  SELECT COUNT(DISTINCT S.activity) = 1 AS unique_actvitiy
  FROM Stay S
  GROUP BY S.location

Note that the structure of the unrolled Markov network can change during inference because in C_5 the label attribute "activity" appears in the GROUP BY-clause. Such *label specific cliques* have been introduced in Liao et al. (2005) and require a specific inference method which will be discussed in "Inference" section.

The potential functions are usually represented as log-linear combinations of real-valued feature functions of the variables in the clique, and the coefficients are optimized to fit the training data. In our model however, all variables—except for the land use shares—are discrete and we simply represent the potential functions with histograms which can be linked to the target entities "Activity" and "Landuse" as shown in Fig. 4. In detail, we define the potential functions ϕ_1, …, ϕ_6 for the corresponding clique temates C_1, …, C_6 described above as follows: let $f(\mathrm{X} = x)$ denote the frequency that a random variable X assumes the value x., then

$$\begin{aligned}\phi_1(a) &= f(\text{S.activity} = \text{a})\\ &\propto \Pr(\text{S.activity} = \text{a})\end{aligned}$$

$$\begin{aligned}\phi_2(l, a, p_l) &= p_l \times \frac{f(\text{S.landuse} = l, \text{S.activity} = a)}{f(\text{S.landuse} = l) f(\text{S.activity} = a)}\\ &\propto p_l \times \frac{\Pr(\text{S.landuse} = l, \text{S.activity} = a)}{\Pr(\text{S.landuse} = l)\Pr(\text{S.activity} = a)}\\ &\approx \Pr(\text{S.landuse} = l | \text{L.fraction} = p_l) \frac{\Pr(\text{S.activity} = a | \text{S.landuse} = l)}{\Pr(\text{S.activity} = a)}\\ &= \Pr(\text{S.landuse} = l | \text{L.fraction} = p_l, \text{S.activity} = a)\end{aligned}$$

$$\begin{aligned}\phi_3(a, t, \delta) &= \frac{f(\text{S.activity} = a, \text{S.arrival} = t, \text{S.duration} = \delta)}{f(\text{S.activity} = a)}\\ &= \Pr(\text{S.arrival} = t, \text{S.duration} = \delta | \text{S.activity} = a)\end{aligned}$$

$$\begin{aligned}\phi_4(a,r) &= f(\text{S.activity} = a, \text{S.returning} = r)/f(\text{S.activity} = a)\\ &= \Pr(\text{S.is\_returning} = r|\text{S.activity} = a)\end{aligned}$$

$$\begin{aligned}\phi_5(a,u_p) &= \frac{f(\text{S.activity} = a, \text{S.unique\_place} = u_p)}{f(\text{S.activity} = a)}\\ &= \Pr(\text{S.unique\_place} = u_p|\text{S.activity} = a)\end{aligned}$$

$$\begin{aligned}\phi_6(u_a) &= f(\text{S.unique\_activity} = u_a)\\ &\propto \Pr(\text{S.unique\_activity} = u_a)\end{aligned}$$

The land use share L.fraction is the only non-discrete variable. It serves as approximation of the probability $\Pr(\text{S.landuse} = l|\text{L.fraction} = p_l) \approx p_l$ that l is the land use type at the user's actual position given the share p_l of land use type l within the buffer area around the location.

Given these definitions we can write

$$\begin{aligned}\Pr(\mathbf{l},\mathbf{a}|\mathbf{P},\mathbf{t},\boldsymbol{\delta},\mathbf{i}) \propto &\left[\prod_{c\in C_1(G)} \phi_1(a_c)\right] \times \left[\prod_{c\in C_2(G)} \phi_2(l_c,a_c,p_{l,c})\right]\\ &\times \left[\prod_{c\in C_3(G)} \phi_3(a_c,t_c,\delta_c)\right] \times \left[\prod_{c\in C_4(G)} \phi_4(a_c,r_c)\right]\\ &\times \left[\prod_{c\in C_5(G)} \phi_5(a_c,u_{p,c})\right] \times \left[\prod_{c\in C_6(G)} \phi_6(u_{a,c})\right]\end{aligned}$$

where G is the graph of the unrolled Markov Network.

Inference

Computing the joint posterior distribution of land use and activity types exactly would require a summation over all possible combinations of land use and activity types that can be assigned to the activity locations of a cell phone track. However, since the number of such combinations increases exponentially with the number of activity locations and due to the vast number of tracks to analyse this is not practicable. A method commonly used in graphical models to compute the marginal sums in a much more efficient way is Belief Propagation (Pearl 1982). However, Liao et al. (2005) pointed out that because of the *label specific cliques*, standard Belief Propagation cannot be used and proposed to sample from the posterior distribution using a Markov chain Monte Carlo (MCMC) method. For our model it is difficult to create a Markov chain that rapidly converges to the target distribution and yields uncorrelated samples without requiring a large number of iterations. Instead we sample from the posterior distribution using Rejection Sampling which is another well-known technique to draw samples from an arbitrary distribution $p(x)$. The basic idea is to define a proposal distribution $q(x)$ for which we know how to efficiently generate samples and define an upper bound M. on $p(x)/q(x)$. A sample from $q(x)$ is accepted with probability $p(x)/(Mq(x))$., otherwise a new sample is drawn from $q(x)$. until the sample is accepted. A low upper bound M has the advantage that a smaller number of samples from $q(x)$ will be rejected, which results in faster sampling from $p(x)$. Rejection Sampling does not require normalization: it suffices to know functions $p'(x) = Zp(x)$ and $q'(x) = Zq(x)$ with some arbitrary normazation factor Z. We can therefore simply split the potential products into two parts and define

$$q'(\mathbf{l},\mathbf{a}|\mathbf{P},\mathbf{t},\boldsymbol{\delta},\mathbf{i}) = \left[\prod_{c\in C_1(G)} \phi_1(a_c)\right] \times \left[\prod_{c\in C_2(G)} \phi_2(l_c, a_c, p_{l,c})\right] \times \left[\prod_{c\in C_3(G)} \phi_3(a_c, t_c, \delta_c)\right] \times \left[\prod_{c\in C_4(G)} \phi_4(a_c, r_c)\right],$$

and as a consequence

$$p'(\mathbf{l},\mathbf{a}|\mathbf{P},\mathbf{t},\boldsymbol{\delta},\mathbf{i})/q'(\mathbf{l},\mathbf{a}|\mathbf{P},\mathbf{t},\boldsymbol{\delta},\mathbf{i}) = \left[\prod_{c\in C_5(G)} \phi_5(a_c, u_{p,c})\right] \times \left[\prod_{c\in C_6(G)} \phi_6(u_{a,c})\right].$$

Sampling from $q'(\mathbf{l}, \mathbf{a}|\mathbf{P}, \mathbf{t}, \boldsymbol{\delta}, \mathbf{i})$ is straightforward because the clique potentials ϕ_1, …, ϕ_4 do not include dependencies between the individual activity locations. A pair (l_i, a_i) can be sampled independently for each location and summing over all combinations of $(l_i, a_i) \in \mathcal{L} \times \mathcal{A}$ for a single location is feasible. An upper bound M is given by

$$M = \left[\prod_{a\in\mathcal{A}} \max_{u_p\in\{T,F\}} \phi_5(a, u_p)\right] \times \left[\max_{u_a\in\{T,F\}} \phi_6(u_a)\right]^m,$$

where m is the number of distinct activity locations.

We use an EM (expectation–maximization) based learning scheme summarized in Algorithm 2 to discover behavioral patterns in an unsupervised way. However, our approach allows using prior background knowledge about certain activity clusters to initialize some or all of the potential functions before fitting the model to the data. In particular, initializing ϕ_2 allows defining an initial probabilistic mapping between land use and activity types, and the clique potentials ϕ_4 and ϕ_5 allow formulating global constraints on the sequence of activity types and their locations. For example, it can be defined that there is only one "home" location, and that individuals have to return to that location, after performing some activity at a different place.

Algorithm 2 EM-based learning of activity clusters

```
Input:
    attributes $\mathbf{P}, \mathbf{t}, \boldsymbol{\delta}, \mathbf{i}$ of enriched cell phone tracks;
Output:
    activity labels $\mathbf{a}$, land use labels $\mathbf{l}$ and potential functions $\boldsymbol{\Phi}$;
    initialize potential functions;
    repeat
        for all cell phone tracks do
            draw $n$ samples from $\mathrm{p}(\mathbf{l},\mathbf{a}|\mathbf{P},\mathbf{t},\boldsymbol{\delta},\mathbf{i})$ using the current potential functions $\boldsymbol{\Phi}$;
        end for
        update the potential functions $\boldsymbol{\Phi}$ according to the observed frequencies in the samples;
    until potential functions $\boldsymbol{\Phi}$ have convergenced;
```

Results

We applied the proposed methods to two different datasets of the metropolitan areas of Vienna, Austria and Boston, USA, and we compared the results. In the following we describe the data sources and parameter settings, and report the achieved results.

Data sources

The data sources used in this study comprise cell phone and land use data of Vienna and Boston. We use travel surveys of Massachusetts and Austria for comparing the results with travel data obtained with traditional survey methods.

The Boston mobile dataset comprises CDRs collected from different US mobile carriers and preprocessed by a wireless data provider company. The data used for this study included approximately 600,000 users in the Greater Boston area for a period of 2 weeks in February 2010. Each record contains anonymous user id, longitude, latitude, and time stamp of the phone activity. The location coordinates of the records are estimated by the data provider using proprietary algorithms. The location accuracy is claimed to be about 200–300 m, which improves the resolution obtained by simply approximating locations with cell tower positions (Candia et al. 2008; Song et al. 2010).

The Vienna mobile signaling dataset was provided by an Austrian mobile carrier. We used data from approximately one million users in Vienna and surrounding areas. For this study we used cell phone records from a period of two weeks in September 2012. The records include an anonymous user id, the time the signaling event was generated, estimated latitude and longitude coordinates, and the type of event in the mobile communications. The coordinates do not directly correspond to the cell tower location but instead represent an estimate of the mobile subscriber's position, which is computed by the mobile operator using undisclosed algorithms. The Vienna cell phone data contains all events in the mobile communications protocol, e.g. when an outgoing or incoming call is started or ended, an SMS is sent or received or when data packages are sent via the mobile network. In addition to these examples the protocol events also includes Mobility Management signals of all devices even when they are in idle mode. The cells of a mobile network are organized in groups which are called Location Areas (LA). When a device is moved between LAs the network updates its position by registering the device with a LA Update event. The size of LAs determines the ratio of registration and paging costs and is chosen to minimize the total signaling traffic. Mobile carriers usually do not disclose the organization of their networks, and in practice the spatial extent of Location Areas varies and is difficult to quantify. In our experiments the spatial resolution provided by LAs seemed to be in the order of several kilometers. In addition to LA Update events, each device's location is automatically updated when a set time interval since the last signaling event has elapsed. The update intervals are determined by the operator and are typically in the order of a few hours. In summary this means that in addition to the information contained in the CDR data, Mobility Management signals provide location data when an active mobile device is moved between cells, when a device in standby mode is moved between Location Areas, or the device has been in standby mode for a set time interval.

The user IDs in the Vienna dataset are shuffled every day while the IDs for the Boston users remain the same during the two week observation period. The datasets include both local residents and foreign roaming clients, but we only consider records made when the users were physically in the Greater Boston area and Vienna Metro area, respectively.

The temporal sparsity of the cell phone traces differ between Vienna and Boston. In order to quantify sampling sparsity we devide the day into n time intervals of equal length and define the *sampling frequency level* S_n as the number of time intervals, where the location of the device is revealed at least once. The comparison of the sampling frequency levels in Vienna and Boston is shown in Fig. 5. In particular between midnight and noon a

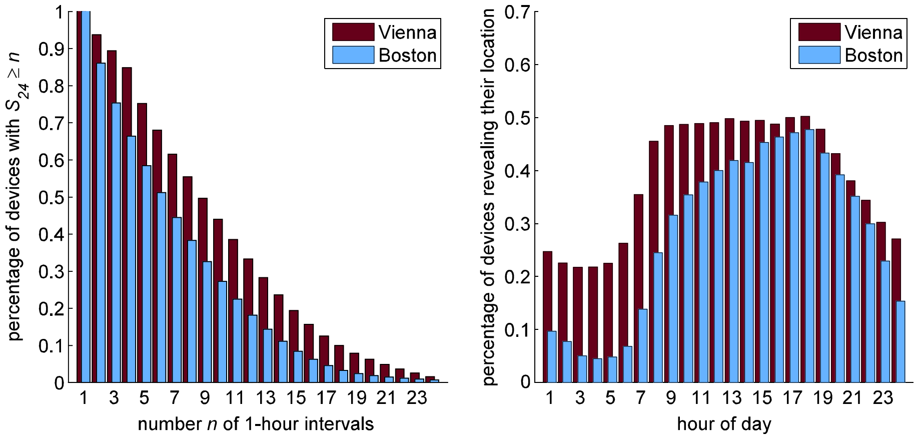


Fig. 5 Comparison of the sparsity of the two cell phone data sources. *Left* distribution of sampling frequency level S_{24}. *Right* percentage of devices with at least one record as a function of the hour of a day

significantly higher percentage of devices reveals its location at least once per hour in the Vienna dataset.

We used publicly available land use data provided by the city governments of Vienna and Boston, respectively. The Vienna dataset defines 32 distinct categories representing the factual land use, which was identified manually based on aerial photographs and on-site inspectations. The dataset was last updated in 2009. The Boston land use dataset defines 33 distinct categories and was created in 2005 with semi-automated methods based on digital ortho imagery.

The Vienna travel survey data was collected in October and November 1995 from 12,564 households representing the population of all nine federal states of Austria. We only used data from households located in Vienna to represent approximately the same population as the cell phone dataset, which results in a sample size of 727 households. Travel days were randomly selected from all weekdays, excluding public holidays. The Massachusetts travel survey dataset contains information for 15,033 households, which were randomly selected following a stratified sampling approach. Data collection activities for the full-study began in May 2010 and continued through October 2011, with a break during the summer. Travel days were evenly distributed among each weekday. The survey population represents all households residing in the thirteen MPO regions in the Commonwealth of Massachusetts.

Parameter settings

For our experiments we chose a location clustering radius $\rho = 1000$ m because of the spatial measurement accuracies in the cell phone data, and the maximum straight-line travel speed was assumed to be $v = 40$ km/h. The buffer size for land use analysis around each activity location was set to $\rho/2 = 500$ m. Since the minimum stay duration τ and the detour ratio ι control the filtering of passing-by points the values of τ a ι have a direct effect on the resulting trip chain lengths. How the average trip chain length changes as the two parameters change is presented in Table 1. The minimum stay duration τ is set to 15 min for reconstructing visited places based on our previous studies (Jiang et al. 2013).

The decline of trip chain length is steepest between $\iota = 1$ and $\iota = 1.5$, whereas for higher values of ι the effect on the average trip chain length abates rapidly. Based on these results we set the detour ratio $\iota = 1.5$.

For the inference of activity clusters we initialized ϕ_2 by assuming a uniform distribution over the land use types such that

$$f(\text{S.landuse} = l, \text{S.activity} = a) \propto f(\text{S.activity} = a | \text{S.landuse} = l)$$

and defined conditional probabilities of activity types given the land use category. The land use types were mapped to 5 different activity types: "home", "working", "shopping", "leisure" and "other" activities. The mapping was often ambiguous, which was accounted for by assigning one of three different weight levels proportional to 1 (low), 10 (medium) and 1000 (high) to each of the activity types and converting the weights into probabilities by normalizing their sum to 1. The activity type "other" was initialized with medium weights for all land use types. Potential ϕ_4 was initialized such that

$$\Pr(\text{S.is\_returning} = \text{true} | \text{S.activity} = \text{``home''}) = 1$$

and

$$\Pr(\text{S.is\_returning} = \text{true} | \text{S.activity} \neq \text{``home''}) = 0.5.$$

For initialization of potential ϕ_5 we assumed that

$$\Pr(\text{S.unique\_place} = \text{true} | \text{S.activity} = \text{``home''} \cup \text{S.activity} = \text{``working''}) = 1$$

and

$$\Pr(\text{S.unique\_place} = \text{true} | \text{S.activity} \neq \text{``home''} \cap \text{S.activity} \neq \text{``working''}) = 0.5.$$

Potential ϕ_6 was initialized such that

$$\Pr(\text{S.unique\_activity} = \text{true}) = 1.$$

The potentials ϕ_1 and ϕ_3 were initialized assuming uniform distributions.

Activity locations and trips

We used cell phone traces of a single work day to estimate the distribution of trip sequences between activity locations as well as the joint distribution of arrival time and duration of stay at each visited place. For comparison we also computed these distributions using travel survey data of Vienna and Boston. The cell phone data include devices with very few location records per day resulting in an extremely sparse sampling in time. Including these devices in the analysis leads to an overestimation of individuals with no or very few trips. We therefore discarded observation days with sampling frequency level

Table 1 The change of average trip length with parameter τ and ι

ι						
		1	1.5	2	2.5	3
τ	300 s	5.29	4.82	4.72	4.67	4.63
	600 s	5.27	4.77	4.65	4.58	4.55
	900 s	5.26	4.73	4.61	4.54	4.49

$S_{24} < 6$, according to previous studies [Schneider et al. 2013]. The comparison of the resulting trip sequences shown in Fig. 6 shows a good general agreement between the estimations derived from cell phone data and the distributions found in the survey datasets. However, the results obtained from cell phone data still tend to overestimate the number of patterns with only two trips in comparison to the survey data, while complex patterns with a larger number of trips tend to be underestimated. The reason could be that very short trips cannot be detected due to the low spatial resolution of cell phone data. The distribution of arrival times and durations of stay shown in Fig. 7 agree well between the different data sources, although data sparsity in the Boston CDR and Vienna survey data results in more noise in the distribution estimates. We calculated the Kolmogorov–Smirnov test statistic on the marginal distributions of activity start time and duration from the survey data and cell phone data to quantitatively characterize the difference between sub-figures in the left column and right column of Fig. 7. In the Boston data, the test statistic of activity start time distribution is 0.10, the test statistic of activity duration distribution is 0.08. In the Vienna data, these two values are 0.12 and 0.10 respectively. These results show that while statistically speaking the distributions extracted from the survey data and cell phone data are not the same, they still share resemblance. The difference might be explained by the difficulty to detected short trips in cell phone data, which is confirmed by Fig. 6. Another reason might be that the detectability of stays with short duration and/or early starting time is strongly influenced by the variations of phone usage and data sparsity at different times of a day (see right side of Fig. 5). As a result, trips and activities starting in the morning ours are underrepresented, especially if their duration is short.

Activity patterns

We used cell phone data of single work days and weekend days for the inference of activity patterns. Only intra-urban tracks which start and end at a location within the metropolitan area have been used for analysis. The temporal characteristics of the activity clusters of a work day and the ten most frequent activity chains are shown in Figs. 8 and 9. The cell phone data of the two cities reveal similar activity clusters, and also the resulting activity chain distributions are in high agreement. The activity type "home" peaks at an arrival time at about 6pm and a duration of approximately 14 h which is the typical pattern of people coming home after work and staying at home over night. This pattern also includes short activities in the afternoon corresponding to people coming home and leaving again to

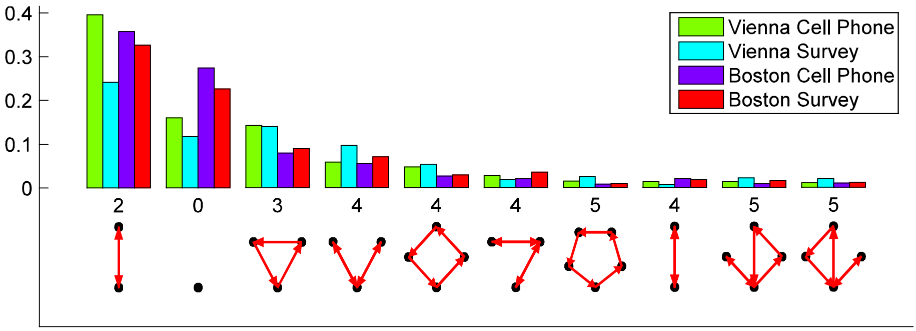
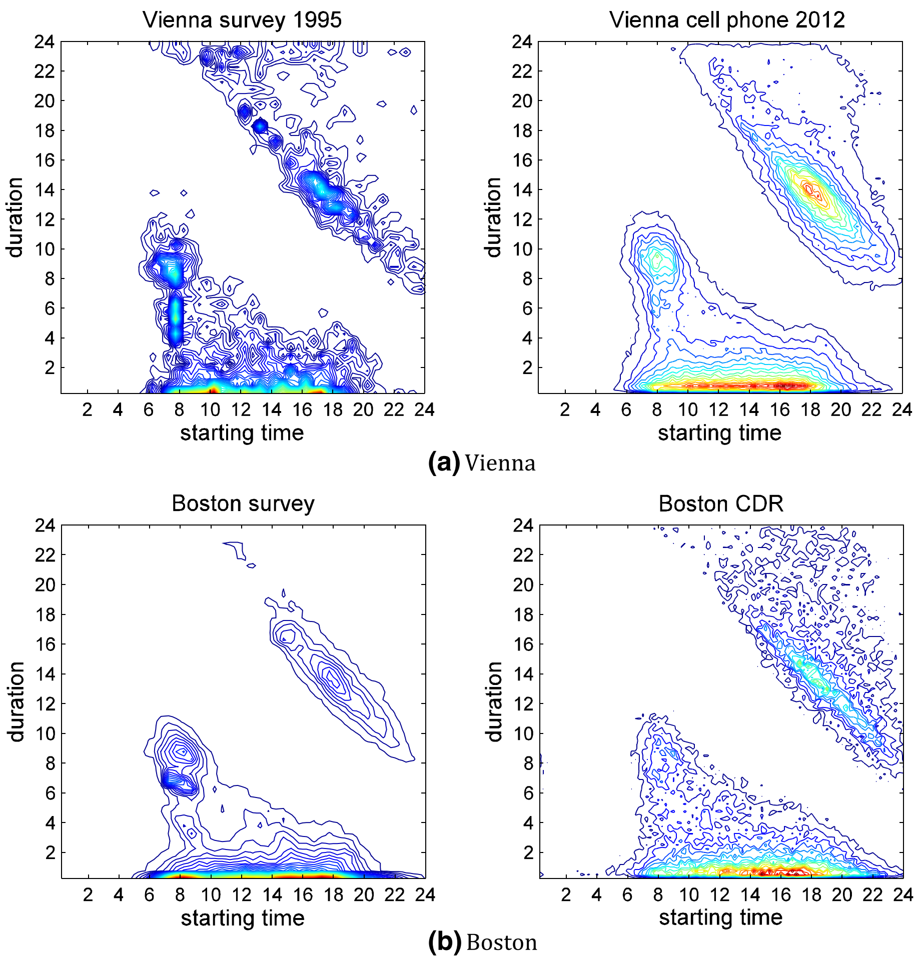


Fig. 6 Comparison of the distributions of activity location sequences in the cell phone and survey datasets of Vienna and Boston. The *numbers* above the location sequences represent the number of trips



Vienna survey 1995

Vienna cell phone 2012

(a) Vienna

Boston survey

Boston CDR

(b) Boston

Fig. 7 Comparison of the distribution of reconstructed starting time and duration of activities in the CDR and survey dataset in Vienna and Boston

carry out another activity in the afternoon or evening. The activity cluster “work” peaks at a starting time at approximately 9am and a duration of 9 h, corresponding to typical working hours. The pattern also includes activities with shorter durations and has a secondary local maximum at around noon and a duration of approximately 1 h. It is interesting to note that while these short activities obviously do not correspond to work behavior they are attracted by the same land use types. They could be removed from this pattern by introducing prior knowledge about the stay duration. However, the goal of this study was not to recognize predefined activity types but to discover activity clusters based on activity scheduling and attraction by land use. The “shopping” activity spans the time range between 6 am and 9 pm, extending until midnight in Boston. It peaks in both cities at about 5 pm and a duration of 1 h. In the Boston dataset this pattern also includes some longer activities starting at about 8 am, which potentially result from part-time workers in shopping areas. The peak time of the activity cluster “leisure” ranges from 3 pm to 6 pm.

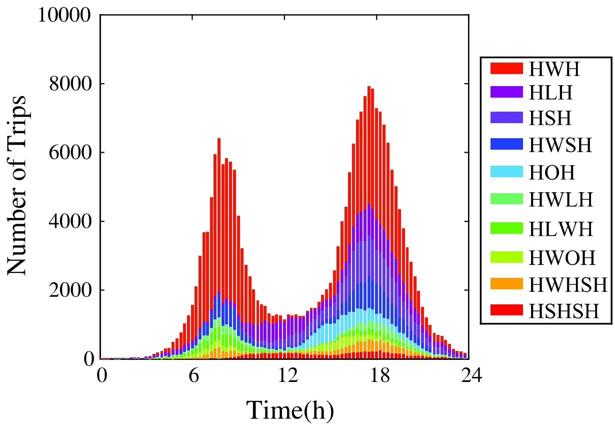
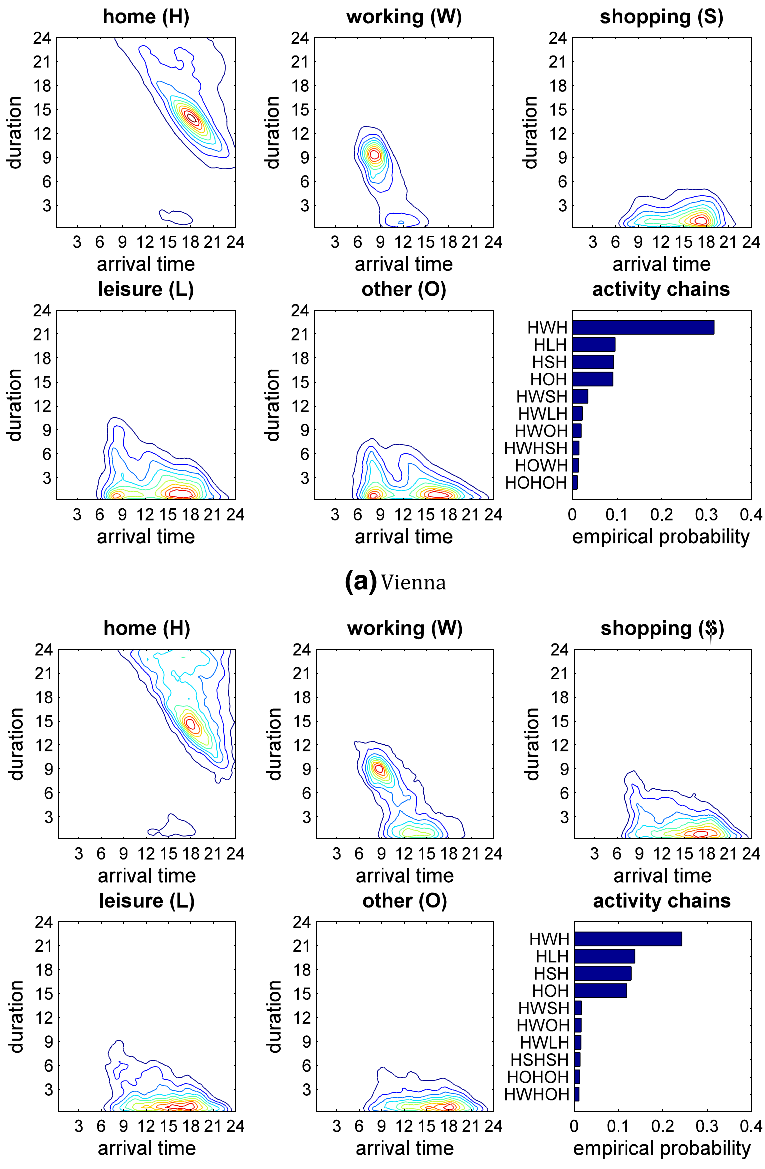


Fig. 8 The traffic flow in an example day in Vienna for the top 10 frequent activity chains

In both data sets this pattern also includes activities with long durations, starting in the morning hours. In Vienna the "leisure" pattern has a secondary peak at 8am, presumably corresponding to activities performed before work. The remaining activity type "other" shows start times ranging from 6am to midnight, again peaking between 3pm and 6pm. This activity type differs the most between the two cities. In Vienna the pattern shows a morning peak and an afternoon peak and drops at noon, and most of the morning activities in this category end before noon. Considering the temporal pattern and the activity sequence "home-other-home-other-home" one may speculate that this activity cluster includes taking and fetching of children.

We can divide the population into different groups according to their activity chains, and then observe how each group contributes to the daily activity flow as is shown in Fig. 8. In Vienna the daily flow has two clear peaks. the "HWH" type takes up over 70 % of the morning peak flow while during the evening peak this percentage drops to less than 50 %. The second and third most frequent chain, "HLH" and "HSH" are most active during evening.

To illustrate the revealed dependencies between land use type and activity we plotted the posterior distribution over the land use types for given activity clusters in Fig. 10. The values in each column sum up to 1 and show the distribution of trip destinations over types of land use. For example, in the Vienna dataset the activity cluster "Home" is mostly attracted by residential land use types, while the "Work" pattern is predominantly attracted by the land use types "office and administration", "industry, manufacturing and wholesale" and "education". In an agent-based simulation model these correlations can be used to define the attractiveness of land use types for each activity cluster. Together with activity-specific trip lengths or travel times these dependencies are the determinants for destination choice, and the temporal distributions in Fig. 9 determine activity time scheduling. Some examples of simulations using such data can be found in (Janssens et al. 2007; Bellemans et al. 2010; Yang et al. 2014). The results in Fig. 10 also show ambiguities of the activity clusters with respect to the land use of the activity location. This means that the same activity cluster can be found across different land use types, and likewise, many land use types are composed of mixed activity clusters. Examining horizontal rows shows that a significant amount of "work", "shop", "leisure", and "other" trips are conducted on land use type "dense residential(mixed) area" since this



(a) Vienna

(b) Boston

Fig. 9 Distributions of arrival times and durations for different activity classes. The distributions are learnt from the data based on an initial probabilistic mapping between land use types and activity classes

land use type takes up a large percentage in the entire land area. This suggests that a more detailed classification in this land use type could help better distinguish between different activities. In the Boston dataset these ambiguities seem to be stronger than in the Vienna dataset and the correlations between land use and activity clusters are weaker, which might result from partially inaccurate land use labelling. Further discussion of the interaction

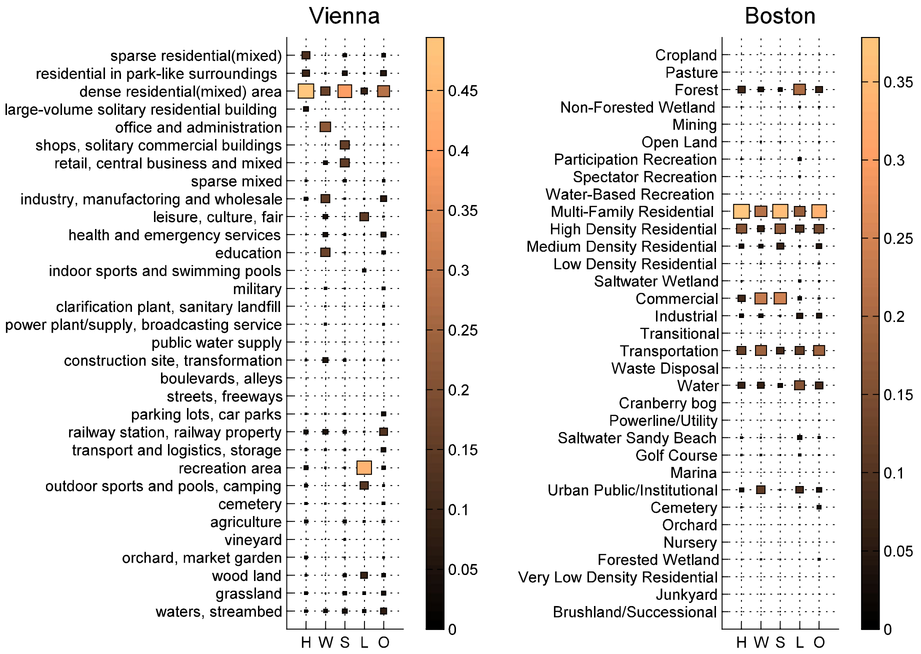


Fig. 10 Posteriors Pr(S.landuse $= l|$S.activity $= a$) of land use types given the activity type learnt from cell phone traces

between land use and travel behavior is an interesting future research topic and is covered in more detail in the conclusion section.

To test the stability of the cluster model, we compared the activity clusters discovered in cell phone data of different days. Similarity was measured by computing the correlation coefficients between the frequency counts f(S.activity $= a$, S.arrival $= t$, S.duration $= \delta$), which are shown in Fig. 11. The activity clusters are very similar across different work days and as expected, weekends show different activity clusters then workdays. Also, Friday is more different from other work days. One weekday in the Boston dataset deviated from the other days: this day was Wednesday, Feb. 10th 2010, where a lot of people seem to return to their homes at noon. Tracing back to previous news reports and we found that a major blizzard took place between Feb. 9th and 11th 2010. That blizzard influenced the entire Northeastern U.S. It is interesting to see that the proposed method is able to detect the influence of major incidents on the population's travel behavior.

Conclusion

In this paper we proposed a method to reveal activity behavioral patterns in cell phone traces that copes with the sparse sampling and low spatial precision of the location estimates. The presented approach consists of a trip extraction method that robustly detects stays and converts the raw cell phone track into a sequence of trips and visited places, and a method to reveal activity patterns by combining the reconstructed activity locations with land use data and modeling the dependencies between activity type, trip scheduling, and land use types with a Relational Markov Network (RMN).

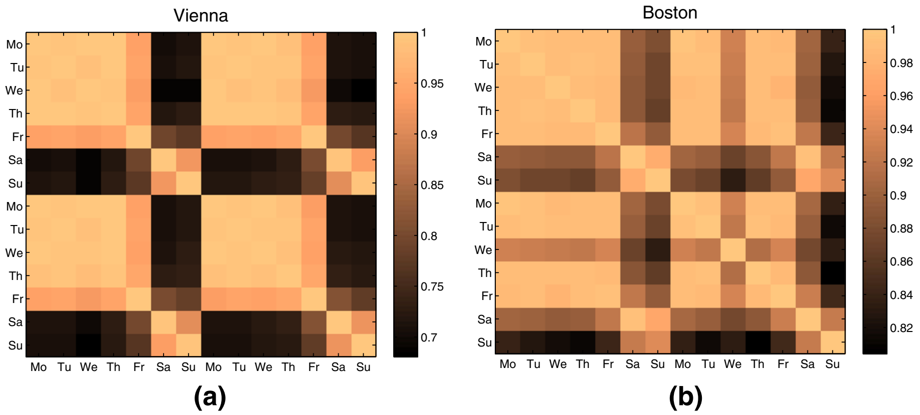


Fig. 11 Correlation between $\Pr(\text{S.arrival} = t, \text{S.duration} = \delta | \text{S.activity} = a)$ calculated for 14 consecutive days

We showed that the resulting trip chains and activity scheduling patterns agree well with data obtained with traditional surveys. The method yields robust results across different days, while work days and weekends show different patterns corresponding to the well-known differences in travel behavior between these types of days. The comparison between Vienna and Boston showed similar patterns in both cities. The inferred activity classes were not shown to directly correspond to activity types used in traditional surveys, but they determine activity scheduling and destination attractiveness in a similar way, so that by adding the inferred activity classes to the travel patterns our method opens up cell phone data as a new data source for activity-based modeling and travel behavior studies.

In this study we combined surrounding land use types with extracted activity start time and duration in the clustering procedure of activities. But we realize there is a deeper interaction between land use and travel behavior (Litman 2004; Litman 2005; McNally and Ryan 1992; McNally and Kulkarni 1997). Various land use factors, such as density, regional accessibility, land use mix, and roadway connectivity, together with travel behaviors such as mode choice, trip length, and activity location choice play a role on travel behavior. Existing studies give mixed results on the strength of connection between land use and travel behavior because they are based on different hypothesis and modeling approaches. The modeling approaches include descriptive studies, multivariate statistical studies, simulation models, choice models, among others (Crane 2000; Boarnet and Crane 2001; Handy 1996; Maat et al. 2005). Many of these modeling approaches require individual level travel behavior characteristics such as trip length, activity chain, activity location choice, etc. Therefore the output result of our proposed model is an ideal input for these studies. The proposed method turns raw mobile phone records into long term observations of individual activity patterns. The labeled trips resulting from the proposed methodology open the way to further interesting research questions that examine travel behavior and land use.

Future improvements of the presented method can add strategies to counter biases such as the underrepresentation of activities subject to time, duration and activity type. Allowing a flexible number of activity clusters and including points-of-interest databases in addition to land use data can further improve the results. We also plan to analyze the relationship between the automatically discovered activity clusters to the conventional activity types used in traditional surveys. In order to evaluate the utility of the presented methods for transportation forecasting, we plan to use the discovered activity patterns in a simulation model and compare the resulting traffic flows to actual traffic measurements.

Acknowledgments This work was partially funded the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the strategic program FIT-IT under grant 835946 (SEMAPHORE), by the Ford-MIT alliance, by the Accenture-MIT alliance and the Center for Complex Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris.[1] The authors would like to thank Airsage for providing the mobile phone data used in this study.

References

Ashbrook, D., Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. Pers. Ubiquitous Comput. **7**(5), 275–286 (2003)

Bachman, W., GeoStats L.P., Oliveira, M., Xu, J.: Using household-level gps travel data to measure regional traffic congestion. In 91st Annual Meeting of the Transportation Research Board, Washington, DC

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., Timmermans, H.: Implementation framework and development trajectory of FEATHERS activity-based simulation platform. Transp. Res. Rec. **2175**(1), 111–119 (2010)

Boarnet, M., Crane, R.: The influence of land use on travel behavior: specification and estimation strategies. Transp. Res. A **35**(9), 823–845 (2001)

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

Bricka, S., Bhat, C.R.: Comparative analysis of global positioning system-based and travel survey-based data. Transp. Res. Rec. **1972**(1), 9–20 (2006)

Caceres, N., Romero, L.M., Benitez, F.G., Del Castillo, J.M.: Traffic flow estimation models using cellular phone data. IEEE Trans. Transp. Syst. **13**(3), 1430–1441 (2012)

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. C **26**, 301–313 (2013)

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput. **10**(4), 36–44 (2011). doi:10.1109/MPRV.2011.41

Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L.: Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A **41**(22), 224015 (2008)

Casas, J., Arce, C.H.: Trip reporting in household travel diaries: a comparison to gps-collected data. In: 78th Annual Meeting of the Transportation Research Board, Washington, DC, vol. 428 (1999)

Chapin, F.S.: Human activity patterns in the city: things people do in time and in space. Wiley, New York (1974)

Chen, C., Gong, H., Lawson, C., Bialostozky, E.: Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. Transp. Res. A **44**(10), 830–840 (2010)

Crane, R.: The influence of urban form on travel: an interpretive review. J. Plan. Lit. **15**(1), 3–23 (2000)

Friedrich, M, Immisch, K., Jehlicka, P., Otterstätter, T., Schlaich, J.: Generating OD matrices from mobile phone trajectories. In: Transportation Research Board, 89th Annual Meeting (2010)

Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)

Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. Nature **453**, 779–782 (2008)

Hackney, J., Marchal, F., Axhausen, K.W.: Monitoring a road system's level of service: the canton zurich floating car study 2003. In: 84th Annual Meetings of the Transportation Research Board, Washington, DC (2005)

Handy, S.: Methodologies for exploring the link between urban form and travel behavior. Transp. Res. D **1**(2), 151–165 (1996)

Hariharan, R., Toyama, K.: Project lachesis: parsing and modeling location histories. In: Freksa, C., Miller, H.J. (eds.) Geographic Information Science, pp. 106–124. Springer, New York (2004)

[1] anas@mit.edu.

Hägerstraand, T.: What about people in regional science? Pap. Reg. Sci. **24**(1), 7–24 (1970)

Hood, J., Sall, E., Charlton, B.: A GPS-based bicycle route choice model for San Francisco. Calif. Transp. Lett. **3**(1), 63–75 (2011)

Horn, C., Klampfl, S., Cik, M., Reiter, T.: Detecting outliers in cell phone data: correcting trajectories to improve traffic modeling. In: Transportation Research Board 93rd Annual Meeting, 14-3690 (2014)

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., Pujolle, G.: Estimating human trajectories and hotspots through mobile phone data. Comput. Netw. **64**, 296–307 (2014)

Hurtubia, R., Flötteröd, G., Bierlaire, M.: Inferring the activities of smartphone users from context measurements using Bayesian inference and random utility models. In: European Transport Conference. EPFL-CONF-152362 (2006)

Jan, O., Horowitz, A.J., Peng, Z.-R.: Using global positioning system data to understand variations in path choice. Transp. Res. Rec. **1725**(1), 37–44 (2000)

Janssens, D., Lan, Y., Wets, G., Chen, G.: Allocating time and location information to activity–travel patterns through reinforcement learning. Knowl. Based Syst. **20**(5), 466–477 (2007)

Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J. Jr, Frazzoli, E., González, M.C.: A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, vol. 2. ACM, Beijing (2013)

Jovicic, G.: Activity based travel demand modelling. Danmarks Transp. Skn. (2001)

Li, H., Guensler, R., Ogle, J.: Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data. Transp. Res. Rec. **1926**(1), 162–170 (2005)

Liao, L., Fox, D., Kautz, H.: Location-based activity recognition using relational markov networks. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), pp. 773–778. Morgan Kaufmann Publishers Inc., San Francisco (2005). http://dl.acm.org/citation.cfm?id=1642293.1642417

Litman, T.: Evaluating transportation land use impacts (2004)

Litman, T.: Land use impacts on transport. Victoria Transport Policy Institute (http://www.Vtpi.Org) (2005)

Maat, K., Van Wee, B., Stead, D.: Land use and travel behaviour: expected effects from the perspective of utility theory and activity-based theories. Environ. Plan. B **32**(1), 33–46 (2005)

McGowen, P., McNally, M.: Evaluating the potential to predict activity types from GPS and GIS data. In: Transportation Research Board 86th Meeting, Jan 21 (2007)

McNally, M.G., Kulkarni, A.: Assessment of influence of land use-transportation system on travel behavior. Transp. Res. Rec. **1607**(1), 105–115 (1997)

McNally, M.G., Ryan, S.: A Comparative Assessment of Travel Characteristics for Neo-Traditional Developments. University of California Transportation Center, Berkeley (1992)

Moiseeva, A., Jessurun, J., Timmermans, H.: Semiautomatic imputation of activity travel diaries. Transp. Res. Rec. **2183**(1), 60–68 (2010)

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P.: Supporting large-scale travel surveys with smartphones—a practical approach. Transp. Res. C **43**, 212–221 (2013)

Pearl, J.: Reverend bayes on inference engines: a distributed hierarchical approach. In AAAI, pp. 133–136 (1982)

Qiu, Z., Cheng, P.: State of the art and practice: cellular probe technology applied in advanced traveler information system. In: 86th Annual Meeting of the Transportation Research Board, Washington, DC, p. 223 (2007)

Quddus, M.A., Ochieng, W.Y., Zhao, L., Noland, R.B.: A general map matching algorithm for transport telematics applications. GPS Solut. **7**(3), 157–167 (2003)

Rasouli, S., Timmermans, H.: Activity-based models of travel demand: promises, progress and prospects. Int. J. Urban Sci. **18**(1), 31–60 (2014)

Ratti, C., Sevtsuk, A., Huang, S., Pailer, R.: Mobile Landscapes: Graz in Real Time. Springer, Heidelberg (2007)

Ratti, C., Williams, S., Frenchman, D., Pulselli, R.M.: Mobile landscapes: using location data from cell phones for urban analysis. Environ. Plan. B **33**(5), 727 (2006)

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using mobile phones to determine transportation modes. ACM Trans. Sens. Netw. **6**(2), 13 (2010)

Reumers, S., Liu, F., Janssens, D., Cools, M., Wets, G.: Semantic annotation of global positioning system traces. Transp. Res. Rec. **2383**(1), 35–43 (2013)

Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. J. R. Soc. Interface **10**(84), 20130246 (2013)

Schönfelder, S., Ethz, I., Samaga, U.: Where Do You Want to Go Today?–More Observations on Daily Mobility. Citeseer, New York (2003)

Sevtsuk, A., Ratti, C.: Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. J. Urban Technol. **17**(1), 41–60 (2010)

Shen, L., Stopher, P.R.: A process for trip purpose imputation from global positioning system data. Transp. Res. C. **36**, 261–267 (2013)

Song, C., Koren, T., Wang, P., Barabási, A.-L.: Modelling the scaling properties of human mobility. Nat. Phys. **6**(10), 818–823 (2010)

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

Stopher, P., Clifford, E., Zhang, J., FitzGerald, C.: Deducing Mode and Purpose from GPS Data. Institute of Transport, Logistics Studies, Sydney (2008)

Stopher, P., Swann, N., FitzGerald, C.: Using an odometer and a GPS panel to evaluate travel behaviour changes. TRB Transp. Appl. (2007)

Taskar, B., Abbeel, P., Koller, D.: Discriminative probabilistic models for relational data. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, pp. 485–492. Morgan Kaufmann Publishers Inc., San Francisco (2002)

Tettamanti, T., Varga, I.: Mobile phone location area based traffic flow estimation in urban road traffic. Adv. Civil Environ. Eng. **1**(1), 1–15 (2014)

Tsui, S.Y.A., Shalaby, A.S.: Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. Transp. Res. Rec. **1972**(1), 38–45 (2006)

Wang, H., Calabrese, F., Di Lorenzo G., Ratti, C.: Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2010, pp. 318–323 (2010)

Wang, M.-H., Schrock, S.D., Vander Broek, N., Mulinazzi, T.: Estimating dynamic origin-destination data and travel demand using cell phone network data. Int. J. Intell. Transp. Syst. Res. **11**(2), 76–86 (2013)

Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C.: Understanding road usage patterns in urban areas. Sci. Rep. **2** (2012)

Wang, T., Chen, C., Ma, J.: Mobile phone data as an alternative data source for travel behavior studies. In: Transportation Research Board 93rd Annual Meeting, 14-2887 (2014)

Widhalm, P., Nitsche, P., Brandie, N.: Transport mode detection with realistic smartphone sensor data. In: 21st International Conference on Pattern Recognition (ICPR), 2012, pp. 573–576 (2012)

Wolf, J., Guensler, R., Bachman, W.: Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. Transp. Res. Rec. **1768**(1), 125–134 (2001)

Wolf, J., Loechl, M., Thompson, T., Arce, C.: Trip rate analysis in GPS-enhanced personal travel surveys. Transp. Surv. Qual. Innov. **28**, 483–498 (2003)

Wolf, J., Oliveira, M., Thompson, M.: The impact of trip underreporting on VMT and travel time estimates: preliminary findings from the California Statewide Household Travel Survey GPS study. In: 83rd Annual Meetings of the Transportation Research Board, Washington, DC (2003)

Yang, M., Yang, Y., Wang W., Ding, H., Chen, J.: Multiagent-based simulation of temporal-spatial characteristics of activity-travel patterns using interactive reinforcement learning. Math. Probl. Eng. (2014)

Yue, Y., Lan, T., Yeh, A.G.O., Li, Q.-Q.: Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. Travel Behav. Soc. **1**(2), 69–78 (2014)

Zheng, V.W., Zheng, Y., Xie, X., Yang, Q.: Collaborative location and activity recommendations with GPS history data. In: Proceedings of the 19th International Conference on World Wide Web, pp. 1029–1038. ACM, New York (2010)

Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y.: Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th International Conference on World Wide Web, pp. 791–800. ACM, New York (2009)

Peter Widhalm studied Computer Science at the Vienna University of Technology and specialized in Machine Learning and Pattern Recognition. Since 2009 he is researcher at the Mobility department of the Austrian Institute of Technology and focuses on novel methods of collecting and combining mobility data from various sources, and mining information for mobility behavior analysis and transport demand modeling.

Yingxiang Yang is a transportation engineering PhD candidate in the Department of Civil & Environmental Engineering at MIT. His research focus is on utilizing different types of digital traces, such as mobile phone data and gps traces, to facilitate transportation modeling.

Michael Ulm holds a PhD in Mathematics from the University of Ulm, Germany. He works at the Austrian Institute of Technology as a data scientist for transport science, where he focuses on mining hidden information in highly noisy data.

Shounak Athavale is Senior Process Methodologist at Ford Motor Company. Currently he leads two work streams, namely, Application Enablement and Urban Mobility. He holds a PhD in Mechanical Engineering from North Carolina State University, Raleigh, NC and an MBA from University of Michigan, Ann Arbor, MI.

Marta C. González is Associate Professor at the Department of Civil & Environmental Engineering at MIT and holds a PhD in Physics from the University of Stuttgart, Germany. She works in the area of urban computing, with a focus on the analysis of vast data collections gathered from different human-driven activities and the formulation of models that elucidate the underlying principles of the observed scenarios.