Modelling Heterogeneous Location Habits in Human Populations for Location Prediction Under Data Sparsity

James McInerney[1], **Jiangchuan Zheng**[2], **Alex Rogers**[1], **Nicholas R. Jennings**[1]

[1]University of Southampton, Southampton, SO17 1BJ, UK
{jem1c10,acr,nrj}@ecs.soton.ac.uk
[2]Hong Kong University of Science and Technology, Hong Kong
jczheng@cse.ust.hk

ABSTRACT

In recent years, researchers have sought to capture the daily life location behaviour of groups of people for exploratory, inference, and predictive purposes. However, development of such approaches has been limited by the requirement of personal semantic labels for locations or social/spatial overlap between individuals in the group. To address this shortcoming, we present a Bayesian model of mobility in *populations* (i.e., groups without spatial or social interconnections) that is not subject to any of these requirements. The model intelligently shares temporal parameters between people, but keeps the spatial parameters specific to individuals. To illustrate the advantages of population modelling, we apply our model to the difficult problem of overcoming data sparsity in location prediction systems, using the Nokia dataset comprising 38 individuals, and find a factor of 2.4 improvement in location prediction performance against a state-of-the-art model when training on only 20 hours of observations.

Author Keywords

Human Behavior Learning, Mobile Phone Sensing, Human Activity Inference, Graphical Models

ACM Classification Keywords

I.5 Pattern Recognition: H.5.2User/Machine Systems

General Terms

Algorithms, Experimentation, Performance.

INTRODUCTION

Modelling routine human mobility has long been a topic of research interest, with traditional applications in epidemiology, urban planning, and emergency response planning [11]. In recent years, as the increasing adoption of GPS-enabled mobile devices has provided highly granular location data of ever greater numbers of people, researchers have sought to introduce more powerful models of group behaviour. Such group models represent the behaviour of multiple individuals in a unified way, enabling exploration (e.g., understanding which behaviours are prominent across a group [4]), inference (e.g., determining social structure [15]), and prediction (e.g., enhancing prediction [3]) that would not be possible with a set of individual models.

To date, existing approaches to group mobility modelling involve either discrete labels of location or continuous latitude/longitude data. In the former case, Eagle and Pentland used hidden Markov models on cell tower identifiers (i.e., the nearby cell tower to the user's current position) to infer whether an individual was at home, work, or elsewhere [4]. These labels, which we refer to as *semantic labels*[1], were then used (with principle component analysis) to place individuals in a group space, in order to discover similarities between people's mobility habits. Gao used the weighted sum of an individual's historical model and those of her friends to incorporate group behaviours [7]. In the continuous case, De Domenico *et al.* and Sadilek *et al.* both leveraged group data, specifically social links, to boost location prediction, by assuming spatial correlation between people's positions [3, 15].

However, progress in the development of highly detailed models of population mobility (i.e., in groups without spatial or social overlap) has been limited in three important respects. First, semantic labelling of locations is usually necessary before it is even possible to discover common structure in mobility data across a population. Moving beyond the basic home and work discrete location labels (in order to include a wider range of locations in an individual's daily life) is non-trivial, despite the availability of highly granular spatial information. This is because home and work locations are special cases that are easily identifiable, since they are usually the top most visited states in daily life (due to power laws in human mobility [8]), and can often be distinguished from each other by their time of day features. Generalising beyond these locations (i.e., to induce semantic overlap between individuals) might be possible with the use of additional location databases indicating the type of business or

[1]We distinguish between semantic labels and *significant locations* in the following way: the former gives information about the function of the location (e.g., home, work) while the latter does not. Both may be obtained either manually or automatically.

UbiComp '13, Sep 8-Sep 12, 2013, Zurich, Switzerland.

function of a location, but these are not always available or reliable. Second, in the absence of such semantic labels, spatial overlap between the mobility of individuals, or, third, a social connection between individuals, is required to be able to model multiple individuals together. This requires that the population contains friends, family, people who work together, or at the very least, people who live near one another. If there is no such spatial or social overlap, it is hard to discover any commonalities between people, as location behaviour is highly personal (with respect to favourite restaurants, gyms, shops, parks etc.).

The aforementioned limitations make it hard to perform exploration, inference, and prediction on large numbers of users. Yet, similar activities have, for several years, been possible with groups of text documents (corpora) using hierarchical Bayesian models, specifically, latent Dirichlet allocation (LDA) and the hierarhical Dirichlet process (HDP) [1, 18]. Both models are powerful representations of text corpora that work on the assumptions that (1) there exists a set of global latent topics and (2) these topics are represented heterogeneously amongst individual documents. The two approaches differ in that LDA represents (and requires the specification of) finite numbers of topics, while HDP can represent an unbounded number of topics, using the non-parametric Dirichlet process (DP). This makes the HDP the state-of-the-art in topic modelling, and the point of departure for our work.

Now, the ability of these topic models to capture corpora of text documents is clear. However, their interpretation and extension to deal with the behaviour of *populations of people* is not. This is for two reasons.

First, the initial work in using LDA for human location behaviour has focused on interpreting topics as features of *individual* behaviour. Specifically, both [6] and [5] present LDA as a model of individual behaviour that assumes that each day in a person's dataset is assigned a latent topic. This initial interpretation makes it difficult to see how populations may be modelled with the same probabilistic architecture. To overcome this, we provide such an interpretation that frames topics as *habits* that may be present among multiple individuals in a population, and are allowed to be expressed heterogeneously in different people. For example, the habit of visiting a recreational location (e.g., park or shopping mall) may have a characteristic temporal pattern that is seen in many people (e.g., weekend afternoons, but hardly ever at midday on weekdays).

Second, LDA and HDP require non-trivial extensions to deal with habitual spatio-temporal behaviour. Specifically, we address the issue that spatial behaviour is highly personal to the individual, while still allowing the sharing of temporal parameters that makes a population model useful. Sharing is important because it allows the generalisation of models across users. For example, a subset of individuals may share a tendency to go to work on weekends, while another subset may have the opposite habits. But there is obviously a limit to sharing in populations of people who may have no connection other than having similar routines to daily life. For this reason, our approach maintains an estimate of the key locations in the daily life of each individual in the model.

As an illustration of what is possible with our approach that is otherwise very hard to achieve, we tackle the problem of predicting the future locations of users under *data sparsity* (also known as the *cold start* problem) in location datasets [13][2]. In more detail, we define the *depth* of a dataset to refer to the number of observations associated with each individual and the *width* of the dataset to refer to the number of individuals present in the population being modelled. Typical approaches to predicting human mobility usually assume deep datasets, in which every person in the dataset has at least a few weeks' worth of data, in order to train statistically accurate models. However, predictions are poor for those individuals who have not yet amassed such a volume of data. Moreover, extant models are not able to take advantage of the increasing width of datasets, due to the aforementioned restrictions of generalised labelling, and spatial or social overlap. This problem, which we tackle in this work, is important to address if mobility prediction is to find broad applicability in location-based services. This is because there are always likely to be a significant number of new users, with fewer than several days of observations (i.e., where the userbase has a long and shallow tail).

The reason why we think an appropriate population model is applicable to overcoming data sparsity is highlighted in the following example. Alice visits the same location every weekday morning, as does Bob (to a different location). Charlie is a new user whose mobility we have only observed for one day, say, Wednesday. A model that shares habits appropriately will be able to extrapolate the pattern from Alice and Bob, and predict Charlie's location to be the same for all weekday mornings as it was for him on Wednesday (but still give a different prediction on the weekend). Crucially, the model can give *personalised* predictions for each user, while simultaneously considering the behaviour of the population as a whole. Clearly, even in this simple example there are reasonable objections (e.g., what if Charlie behaves fundamentally differently to Alice and Bob? What if Charlie was not where he usually is on a Wednesday morning? How to deal with multiple contradictory habits present in a population?). The model we present here is designed to overcome these issues in a general and principled way.

In more detail, our contributions are the following:

- We present the first approach to modelling group routine location behaviour across a population without the existing restrictions that currently apply. Specifically, without requiring semantic labels of locations or spatial/social overlap between individuals. To do this, we develop an extension of the HDP, called *LocHDP*, to deal with spatio-temporal behaviour in populations.

[2]The same conceptual apparatus could equally be used for exploration and inference in the mobility domain. We focus on data sparsity because it is an often overlooked problem that a population model could be particularly effective at addressing.

- We derive the inference process for LocHDP using Markov chain Monte Carlo (MCMC) sampling, and show how prediction using the shared parameters of the population may be performed.
- Using the Nokia Lausanne dataset containing detailed mobility observations of 38 people for 1 year, we explore LocHDP's ability to overcome data sparsity in location prediction by varying the width and depth of the data seen during training. For our experiments, the extensive depth of the Nokia dataset enables us to consider a wide range of training sizes, while still having a large number of observations that were unseen during training that allow us to get statistically significant measures of predictive accuracy. We find that our model outperforms the state of the art (an approach by [2]) by a factor of 2.4 in held-out data likelihood when given only 20 hours of training data. We find that this advantage holds all the way up to 100 hours of training data, after which point there is no advantage in prediction accuracy to using a population model in comparison to a set of models of individual behaviour.

The rest of the paper is organised as follows. First, we describe existing work related to the problems of modelling the location behaviour of groups and overcoming data sparsity in location prediction. Next, we give the full details of our approach, explaining our extension to the HDP. Then we derive the MCMC sampling process for LocHDP, and give the equations for prediction under data sparsity after training on population data. We test our approach on the Nokia dataset by varying the depth and width of the data, finding the held out data likelihood in each case. Finally, we conclude.

RELATED WORK

Existing approaches modelling the routine location behaviour of groups of people place restrictions on the data that limit their wider applicability. De Domenico *et al.* propose the use of dependencies, specifically mutual information, between the mobility patterns of friends (or people who live/work together) to improve prediction of future location [3]. However, as they acknowledge, this requires that users' movements be both spatially and temporally correlated, which is not true in general. In addition, their method scales $\mathrm{O}(N^n)$, where N is the number of people in the group and n is the number of connections for each person to others. In contrast, our approach scales $\mathrm{O}(N)$ per iteration during sampling and simultaneously considers all behavioural overlaps. Based on the same intuition about the mobility of friends, Sadilek *et al.* proposed a probabilistic approach (a variant of a hidden Markov model) that treats the location of friends as noisy observations of a person's current location [15]. Although they present their method as being predictive, it is actually inferential in nature (of a person's current location given her friends' current locations). In addition, they also present an approach to inferring social structure from location data, taking into account also user messages and existing links (therefore, they present an approach to the *missing link* problem in social networks using location).

Gao *et al.* consider discrete location prediction using both the historical behaviour of an individual, as well as the historical behaviour of his friends [7]. The historical aspect of their model is based on the hierarchical Pitman-Yor process (HPY) which is very close to the hierarchical Dirichlet process but which can also capture power laws that tend to arise in human behaviour. However, to clarify, they used the HPY for sequential prediction of individual mobility, in which the parent Pitman-Yor process at each level represents the recent history for that individual. Our interpretation and extension of the HDP is significantly different in aims and assumptions. Most significantly, ours allows temporal predictions (from several hours to several weeks into the future), whilst theirs is an individual sequential prediction method that places probabilities over future locations given very recent observations. This is an important distinction; temporal prediction (which we address) is useful for knowing where someone will be given the time context (e.g., 2pm next Tuesday, or, one year from now) while sequential prediction is useful for knowing where someone will be given their most recent locations (e.g., where will the individual go after work?). In [7], the role of groups is more of a final addition, since they do not provide a unified approach to group modelling, and instead opt for a weighted sum of probabilities of friends' mobility based on cosine similarity.

Cho *et al.* also build an individual model then use a heuristic approach to do *post hoc* correction based on social and spatial ties [2]. Their individual model does temporal prediction and is based on a mixture model, so is an extension to Nurmi *et al.*, who applied mixture models to spatial information only (for the purposes of identifying significant locations rather than prediction) [14]. Cho *et al.* tested their individual model against several benchmarks and found that it predicted locations almost as well as the social-historical model (which had privileged access to friends' locations). In our evaluation section, we compare LocHDP to their individual approach.

The exception to the trend of requiring semantic, social, or spatial information is provided by [19], who use hierarchical clustering to find groups based on discrete observations of location behaviour. The drawback of their approach is that they assume three levels of latent variable: a set of latent variables at the day level, summarising each day of behaviour, another set explaining individual observations contained within a day, and yet another set to capture multimodalities in the temporal observations. As a result, parameter inference is particularly challenging, and means that sampling has to proceed piecemeal (i.e., training one level of parameters before proceeding to the next) which has unexplored consequences on model accuracy. Another consequence is that they can only represent a small number of behaviours in their model, and in any case, only consider discrete location observations and ignore longer scale periodicities (which we consider in our work).

In addition to population modelling, a key contribution of our work is in showing how such models can be used to overcome data sparsity in location prediction. The accuracy of prediction results for existing approaches (e.g., by [16] and [4]) rely on the assumption that significant observation

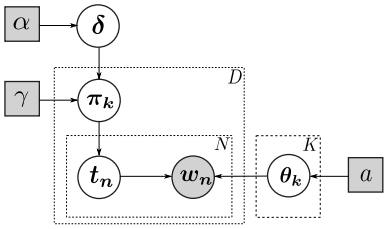

Figure 1. The graphical structure of standard HDP using plate notation, in which each random variable within a plate is repeated. Shaded nodes indicate observed variables and square nodes indicate hyperparameters.

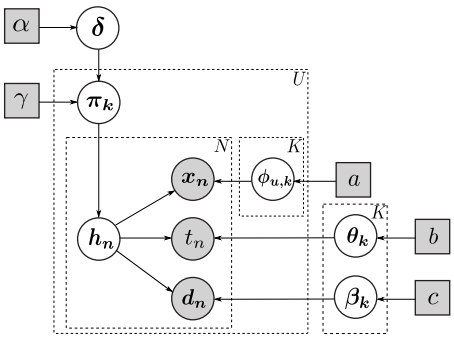

Figure 2. The graphical structure of `LocHDP` using plate notation, in which each random variable within a plate is repeated. Shaded nodes indicate observed variables and square nodes indicate hyperparameters.

histories (i.e., at least several weeks) are available. In our work, we explicitly consider and test for performance under data sparsity.

In the only work that directly addresses data sparsity in location prediction, we have previously presented a pairwise probabilistic model that matches the location of one (established user) to the new user to obtain a mapping between the significant locations of individuals [13]. A separate prediction algorithm is then required to actually generate location predictions. Due to its pairwise nature, this method is not suitable for multiple users, and retraining the model for every possible pairing scales prohibitively in the number of users.

Finally, the problem of overcoming data sparsity (or cold starts) is prominent in collaborative filtering [17]. In collaborative filtering for recommender systems, the degree of preference a user has for an *item* (e.g., book, movie) must be predicted under conditions of high data sparsity, because users typically only ever rate a few items. An important strategy in overcoming this exploits similarities between the preferences of users (e.g., two users who share a preference for romantic comedies might also share other movie tastes). Naïvely, we would expect to be able to apply the same tools used in collaborative filtering to overcome sparsity in location prediction. However, collaborative filtering assumes either an overlap between a common set of items (even if there are a vast number of such items) or the ability to categorise items (i.e., semantic labelling) whereas mobility habits are highly personal and hard to categorise in general (beyond the basic "work" and "home" categories). In short, the significant locations for each individual, such as the home, workplace, favourite restaurant or gym, vary from person to person and overlaps between locations are the exception (e.g., people who live together, or people who are close friends). Crucially, our model of multiple individuals' mobility does not assume or require any such connection between individuals. They could, in fact, live on opposite sides of the globe.

HABITUAL LOCATION BEHAVIOUR MODEL

In this section we briefly give an overview of the hierarchical Dirichlet process (HDP) before proceeding to describe our extension (LocHDP). We then derive the sampling process for LocHDP, and give the relevant equations for location prediction given population data.

The Standard Hierarchical Dirichlet Process (HDP)

We start with an overview of the standard HDP. For more details, see [18]. The key feature of the HDP is that a global set of topic coefficients is selected once, then a local set of topic coefficients is drawn from this global distribution for each document:

$$\begin{aligned}&\boldsymbol{\delta} \sim DP(\alpha, B)\\ &\text{for each document } i \in [1..D]:\\ &\quad \boldsymbol{\pi_i} \sim DP(\gamma, \boldsymbol{\delta}) \end{aligned} \tag{1}$$

where the global topics ($\boldsymbol{\delta}$) and the local coefficients of topics ($\boldsymbol{\pi_i}$) are unknown parameters in the model. The bigger the topic coefficient, the more the topic is expressed in the document, resulting in a higher frequency words associated with that topic in the document. We next explain in more detail the connection between words and topics.

In the HDP, each document, i, consists of a set of words represented as a bag of discrete tokens $\boldsymbol{X_i}$, and each word $\boldsymbol{w_n} \in \boldsymbol{X_i}$ has a latent (i.e, hidden) assignment to a topic t_n which is drawn from the local multinomial distribution $t_n \sim \mathcal{M}(\boldsymbol{\pi_i})$. Then, given the topic, the word is assumed to be drawn from another discrete multinomial distribution that is selected by this topic $\boldsymbol{w_n} \sim \mathcal{M}(\boldsymbol{\omega_{t_n}})$. Note that we adopt the standard notation of using bold symbols for vector random variables, and of using binary vectors for discrete variables, so that observation $\boldsymbol{w_n} = [0, 0, ..., 1, ..., 0]$ has a single 1 to indicate the word token for observation n (so that $w_{n,v} = 1$ iff word v was observed). This ensures that the equations for noisy discrete observations are identical to those of noiseless observations throughout. In addition, random variables without subscripts indicate the whole dataset (e.g., $\boldsymbol{w}$ indicates all word observations comprising $\boldsymbol{w_n}$).

The aforementioned assumptions for the HDP are summarised in the graphical model shown in Figure 1. As it is, the standard HDP is not directly applicable to spatio-temporal be-

Symbol	Description
N	Total number of observations (for all users)
$\boldsymbol{x_n}$	Latitude/longitude of observation n (in degrees)
t_n	Time of day for observation n (in hours)
$\boldsymbol{d_n}$	Day of week for observation n (Mon-Sun)
$\boldsymbol{h_n}$	Latent habit assignment for observation n
U	Number of users in the data
$\boldsymbol{\pi_u}$	Discrete distribution over habits for user u
K	Number of habits (*not* pre-specified with HDP)
$\boldsymbol{\phi_{u,k}}$	Parameters for Gaussian spatial distribution for user u and habit k
$\boldsymbol{\theta_k}$	Parameters for Gaussian temporal distribution for hour of day observations of habit k
$\boldsymbol{\beta_k}$	Parameters for multinomial temporal distribution for day of week observations of habit k
$\boldsymbol{\delta}$	Global (parent) distribution over habits
α, γ, a, b, c	Hyperparameters

Table 1. Summary table of symbols for LocHDP.

haviour data in populations. We now present our extension for human spatio-temporal data.

An Extension to Location HDP (LocHDP)

LocHDP models a set of N location data points $(\boldsymbol{x_n}, t_n, \boldsymbol{d_n})$, where $\boldsymbol{x_n}$ is the location of a person (continuous, in degrees of latitude and longitude), t_n is the time of day that the observation was recorded (continuous, in hours), and $\boldsymbol{d_n}$ is the day of the week (discrete).

Our approach starts with the assumption that there exists a set of *habits* that explain daily life mobility. Similar to topics, habits are considered as discrete random variables $\boldsymbol{h_n}$ (for $0 \leq n < N$) that have a latent assignment to each observation $(\boldsymbol{x_n}, t_n, \boldsymbol{d_n})$ explaining both the spatial and temporal aspects of the data. The marginal likelihood of spatio-temporal data can therefore be expressed as a mixture of habits, meaning, intuitively, that each person can have multiple habits that can be expressed to varying degrees in their behaviour (depending on the probability $p(h_n = k|\pi_{u,k})$):

$$p(\boldsymbol{x_n}, t_n, \boldsymbol{d_n}) = \sum_k p(\boldsymbol{x_n}, t_n, \boldsymbol{d_n}|h_n = k)p(h_n = k|\pi_{u,k})$$

where we have marginalised out the uncertainty over latent habit assignments, and introduced parameter $\boldsymbol{\pi_u}$ to represent the local mixture of habits. This implies that there is a one-to-one mapping between documents in the standard HDP and users in LocHDP. But the standard HDP assumes globally shared parameters (i.e., $\boldsymbol{\omega}$) between observations which does not take account of the fact that the spatial aspect of location behaviour is highly personal (i.e., we do not generally live, work, and relax at the same locations).

To address this, we assume separate spatial parameters for each user u. Specifically, we assume that spatial observations follow a Gaussian distribution, representing both sensor and behavioural noise that is present in the data. The former is caused by fluctuations in the GPS (e.g., random error from measurements taken when the GPS is still warming up, or when nearby buildings obscure satellite communication), while the latter refers to the tendency for a person's position to change slightly even if they are in the same significant place (e.g., walking around a big building). By the central limit theorem, the sum of independent random noise caused by such factors can be approximated with a Gaussian:

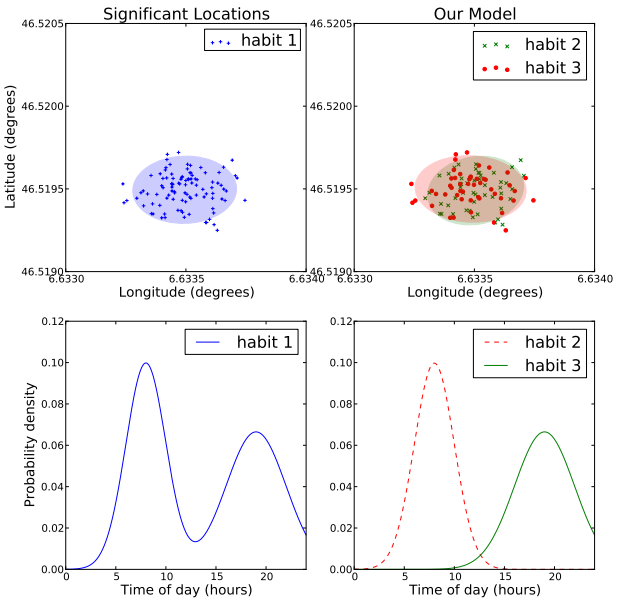

Figure 3. An illustration of two alternate interpretations of the same routine spatio-temporal data: the left-hand side panels assume that a *significant location* may be visited at many different times, the right-hand side panels (our model) find separate habits in the data. Note that the discrete weekday temporal distribution at the location is omitted from this figure.

$$p(\boldsymbol{x}_n|u, \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\phi_{h,u}}) \tag{2}$$

where we have introduced the parameters $\boldsymbol{\phi_{h,u}}$ indicating the mean and variance of the spatial information for user u given the habit h.

However, if we take the same approach to the temporal aspect of observations, then there can be no parameter sharing between individual mobility models. Sharing is important as it allows generalisation across users, which enables exploration, inference, and possibly improved prediction.

Given the periodic nature of routine behaviour, we model the time of day and day of the week of each observation:

$$p(t_n|h, \boldsymbol{\beta}) = \mathcal{N}(t_n|\boldsymbol{\beta_h}) \tag{3}$$

$$p(\boldsymbol{d_n}|h, \boldsymbol{\theta}) = \prod_w^W \theta_h^{d_{nw}} \tag{4}$$

where t_n represents the local time of day (in hours) of the observation and is a continuous 1-dimensional Gaussian, so that variations around arrival times at locations are smoothed.

Smoothing is usually preferable to temporal binning [9] because it allows missing periods to be filled in sensibly. The distribution over arrival times is parameterised by $\boldsymbol{\beta_h}$, which is selected by habit h alone (and not the user). The day of the week observation is discrete, so is therefore given a multinomial distribution with parameter θ_h for habit h. In this case, $W = 7$, the number of days of the week (though this can be easily changed for other discrete information, e.g., weather, national holidays).

The unimodal distribution over the time of day implies that each habit is assumed to be active for only one period on any given day (if we ignore the Gaussian noise). This might seem counterintuitive, since it is common to spend several different periods at a single location per day (e.g., at home or commuting route). To clarify, a *significant location* may be visited at many different times of the day, but a *habit* is only active around a single period of each day. To illustrate this point, see Figure 3, in which we show the two different interpretations of the same synthetic data. The left-hand side panel assumes that complex temporal patterns can be assigned to a single (normal) spatial distribution while the right-hand side requires that each temporal mode be represented separately. Therefore, in making this assumption, we are shifting the burden of rediscovering existing significant locations to the spatial element (i.e., the home significant location may be discovered once for morning time periods, and then again for evening periods). But there are considerable advantages to this approach. Firstly, habits are more modular, making them more likely to be shared amongst users (e.g., leaving home at around 8:30am may be common to lots of people, but leaving home at 8:30am and arriving back home at 9pm is more idiosyncratic). Secondly, a multimodal temporal distribution would require an extra latent variable in the model, creating additional complexity.

This extension from HDP to LocHDP results in a modified generative process. The generative process describes how new observations can be generated from the parameters alone and is important because it fully specifies the model:

$$
\begin{aligned}
&\boldsymbol{\delta} \sim DP(\alpha) \;\text{ (draw global habits)}\\
&\text{for each latent habit } h \in [1..\,|\boldsymbol{\delta}|]:\\
&\quad \boldsymbol{\theta_h} \sim IG(d) \;\text{ (draw hour of day parameter)}\\
&\quad \boldsymbol{\beta_h} \sim Dir(e) \;\text{ (draw day of week parameter)}\\
&\text{for each user } u \in [1..U]:\\
&\quad \boldsymbol{\pi_u} \sim DP(\gamma, \boldsymbol{\delta}) \;\text{ (draw user habits from global set)}\\
&\quad \text{for each latent habit } h \in [1..\,|\boldsymbol{\delta}|]:\\
&\quad\quad \boldsymbol{\phi_{u,h}} \sim IG(c), IG(c) \;\text{ (draw spatial params)}\\
&\quad \text{for each observation } n \in [1..N]:\\
&\quad\quad \boldsymbol{h_n} \sim \mathcal{M}(\boldsymbol{\pi_u}) \;\text{ (draw habit)}\\
&\quad\quad \boldsymbol{x_n} \sim \mathcal{N}(\boldsymbol{\phi_{u,h_n}}) \;\text{ (draw spatial obs.)}\\
&\quad\quad t_n \sim \mathcal{N}(\boldsymbol{\beta_{h_n}}) \;\text{ (draw time of day obs.)}\\
&\quad\quad \boldsymbol{d_n} \sim \mathcal{M}(\boldsymbol{\theta_{h_n}}) \;\text{ (draw day of week obs.)}
\end{aligned}
\tag{5}
$$

For comparison with the standard HDP, we also provide a graphical depiction of the model in Figure 2.

Parameter Inference

There is no tractable closed-form solution for the parameters for LocHDP (nor for standard HDP). We therefore derive the Markov chain Monte Carlo sampling process for LocHDP from the generative specification in Equation 5. As a basis, we use one of the most efficient sampling methods for HDP, which is collapsed Gibbs sampling by [18], the Chinese restaurant franchise representation.

The collapsed version converges more quickly than other variants of MCMC because it requires sampling fewer random variables ("collapsing" out most of the parameters by integration). Here, we need to sample from only two distributions, $p(\boldsymbol{h_n}|\boldsymbol{h_{-n}}, \boldsymbol{\delta}, \boldsymbol{x_n}, t_n, \boldsymbol{d_n})$, the posterior over the habit assignment for observation n, and, $p(\boldsymbol{\delta}|\boldsymbol{h})$, the posterior over the coefficients of the parent DP. The former can be expanded using Bayes' theorem and the conditional independence of $\boldsymbol{x_n}$, t_n, and $\boldsymbol{d_n}$:

$$
\begin{aligned}
& p(\boldsymbol{h_n}|\boldsymbol{h^{-n}}, \boldsymbol{\delta}, \boldsymbol{x_n}, t_n, \boldsymbol{d_n}, \gamma, a, b, c)\\
\propto\;& \boldsymbol{p(h_n}|\boldsymbol{h^{-n}}, \gamma) p(\boldsymbol{x_n}|\boldsymbol{h_n}, a) p(t_n|\boldsymbol{h_n}, b) p(\boldsymbol{d_n}|\boldsymbol{h_n}, c)
\end{aligned}
$$

We substitute collapsed versions of the prior and observation likelihoods to get the first equation required for MCMC:

$$
\begin{aligned}
& p(h_{n,k} = 1|\boldsymbol{h^{-n}}, \gamma) p(\boldsymbol{x_n}|h_{n,k} = 1, a) p(t_n|h_{n,k} = 1, b)\\
&\qquad \times\; p(\boldsymbol{d_n}|h_{n,k} = 1, c)\\
\propto\;& \left(\gamma\delta_k + v_{u,k}^{-n}\right) f(\boldsymbol{x_n}|\boldsymbol{X_{x,k,u}^{-n}}, a) f(t_n|\boldsymbol{X_{t,k}^{-n}}, b)\\
&\qquad \times \left(c\beta_k + \boldsymbol{X_{d,k}^{-n}}\right)
\end{aligned}
\tag{6}
$$

where f is the Student's t distribution, which is the result of collapsing a normal distribution, giving the predictive likelihood given just the hyperparameters and *sufficient statistics*[3]. In Equation 6, the day of week Dirichlet distribution has also been collapsed. The sufficient statistics for the prior is v_k, simply the total number of data points assigned to habit k. $\boldsymbol{X_x}$,$\boldsymbol{X_t}$, and $\boldsymbol{X_d}$ are the sufficient statistics for the spatio-temporal likelihoods and can be calculated as:

$$
\begin{aligned}
\boldsymbol{X_{x,k,u,0}} &= \sum_{n=1}^{N} \boldsymbol{x_{n,k}} I[h_{n,k} = 1] I[U(n,u)]\\
\boldsymbol{X_{x,k,u,1}} &= \sum_{n=1}^{N} \boldsymbol{x_{n,k}}^2 I[h_{n,k} = 1] I[U(n,u)]\\
X_{\boldsymbol{t},k,0} &= \sum_{n=1}^{N} t_n I[h_{n,k} = 1]\\
X_{\boldsymbol{t},k,1} &= \sum_{n=1}^{N} t_n^2 I[h_{n,k} = 1]\\
X_{\boldsymbol{d},k,w} &= \sum_{n=1}^{N} d_{n,w} I[h_{n,k} = 1]
\end{aligned}
\tag{7}
$$

[3]That is, the information that is alone sufficient to calculate a distribution, allowing improved efficiency during inference.

where $I[B]$ is the indicator function that is 1 if B is True, and 0 otherwise, and $U(n, u)$ is evaluated True iff data point n came from user u. Given initial values of $\boldsymbol{X_x}$, $\boldsymbol{X_t}$, $\boldsymbol{X_d}$, and $\boldsymbol{v}$, it is then trivial to update them as each $\boldsymbol{h_n}$ is reassigned during sampling. Intuitively, Equation 7 can be understood as first partitioning each observation n according to its current habit assignation $\boldsymbol{h_n}$, then aggregating each partition through summation. A simple example of aggregation is the calculation of the mean and variance of a single Gaussian distribution, which requires both $\sum_n x_n$ and $\sum_n x_n^2$ as sufficient statistics. A similar calculation is required here, albeit in a full Bayesian framework involving both continuous and discrete observations.

The MCMC algorithm is given in Algorithm 1. In practice, we found that 50 iterations were enough to reach convergence for all users, after which point, we took every 3rd sample until 10 samples could be collected. We now consider how to derive location predictions given such samples.

Algorithm 1 LocHDP sampling process

procedure SAMPLE-LocHDP($\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{d}$)
  Randomly initialise $\boldsymbol{h}, \boldsymbol{\delta}$ ▷ Eq$^{\text{n}}$ 5
  Initialise sufficient statistics $\boldsymbol{X_x}, \boldsymbol{X_t}, \boldsymbol{X_d}, \boldsymbol{v}$ ▷ Eq$^{\text{n}}$ 7
  for $s \leftarrow 1, S$ **do** ▷ For each sample
    for $n \leftarrow 1, N$ **do**
      $\boldsymbol{h_n^{(s)}} \sim p(\boldsymbol{h_n}|\boldsymbol{h_{-n}}, \boldsymbol{\delta}, \boldsymbol{x_n}, t_n, \boldsymbol{d_n})$ ▷ Eq$^{\text{n}}$ 6
      Update $\boldsymbol{X_x}, \boldsymbol{X_t}, \boldsymbol{X_d}, \boldsymbol{v}$
    end for
    $\boldsymbol{\delta^{(s)}} \sim p(\boldsymbol{\delta}|\boldsymbol{h})$ ▷ See [18]
  end for
  return $\boldsymbol{h}, \boldsymbol{\delta}$
end procedure

Location Prediction

After obatining samples of $\boldsymbol{h}$ and $\boldsymbol{\delta}$, using historical data, we consider the scenario of predicting an individual's location given an arbitrary temporal query, specifically the time of day, and day of week for the prediction. LocHDP can produce a 2-dimensional predictive probability density over continuous space in the world given this query. After sampling is performed, this density may be found by averaging over the samples [12], and marginalising over the hidden habit for the temporal query point:

$$p(\mathbf{x}^*|t, d, u, \mathbf{X}) = \int p(\mathbf{x}^*|h, u, \mathbf{\Omega}) p(h|t, d, u, \mathbf{\Omega}) p(\mathbf{\Omega}|\mathbf{X}) \mathrm{d}h \mathrm{d}\mathbf{\Omega} \quad (8)$$

$$\approx \frac{1}{|R|} \sum_{r \in R} \sum_{h=1}^{K} \mathcal{N}(\mathbf{x}^*|\boldsymbol{h}^{(\boldsymbol{r})}, a) \frac{\pi_{u,h,r} \beta_{\mathbf{h},\mathbf{r}} \mathcal{N}(t|\theta_{\mathbf{h},\mathbf{r}})}{\sum_{h'=1}^{K} \pi_{u,h',r} \beta_{\mathbf{h}',\mathbf{r}} \mathcal{N}(t|\theta_{\mathbf{h}',\mathbf{r}})}$$

where $\mathbf{x}^*$ is the location, t is the query hour of the day, $\boldsymbol{d}$ is the query day of the week, u is the individual that we want to predict, $\boldsymbol{h}$ is the habit responsible for this observation, $\mathbf{X}$ represents the full dataset (for all individuals), $\mathbf{\Omega}$ is the whole set of parameters in the model[4], and R is the number of samples kept during MCMC. To derive Equation 8, we used Bayes' theorem to expand the posterior distribution over habits:

$$p(h|t, d, u, \mathbf{\Omega}) \propto p(h|\pi) p(d|h, \beta) p(t|h, \theta) \quad (9)$$

$$= \pi_{u,h} \beta_{\mathbf{h}} \mathcal{N}(t|\theta_{\mathbf{h}}) \quad (10)$$

In the next section, we evaluate the quality of the model's predictions on real world data of human location behaviour.

[4]The hyperparameters are assumed given, and are omitted from these equations for clarity.

EMPIRICAL EVALUATION

In this section, we first describe our choice of experimental setup, specifically, how we test our approach under the constraints outlined in the introduction. Then, we give the empirical results for the evaluation of our approach against two benchmarks.

Experimental Setup

At a high level, our experiments are designed to investigate the effects of limited training data on location prediction. Our main hypothesis is that our approach will be less affected by such conditions. Evaluating this hypothesis requires the consideration of four experimental elements. Firstly, we introduce all the approaches that we evaluated (our model, plus two other approaches for comparison). Secondly, we describe the dataset of real human location data that we used for our experiments. Thirdly, we briefly discuss the metric we use for evaluation of model and prediction quality. Finally, the details of exactly how we test under data sparsity are discussed.

Approaches

The approaches we consider include LocHDP, as well as an existing state-of-the-art approach and a benchmark that performs only basic learning.

1. **Location HDP Model (`LocHDP`)** our approach.

2. **Individual Mixture Model (`Mixture`)** in which each individual is modelled separately. This represents an extended version of the approach proposed by [2], a state-of-the-art prediction method that assumes that location behaviour follows a spatio-temporal mixture model. This is similar to `LocHDP` with the important difference that spatio-temporal clusters cannot be shared between users or generalised across a population. There also exist other state-of-the-art prediction sequential methods that take advantage of very recent history (e.g., with a variable-order Markov model, or Hierarchical Pitman-Yor process [7]), but these are not appropriate when predicting potentially several days into the future.

3. **Single Gaussian (`Single`)** places probability mass fairly uniformly over the whole area that an individual occupies. Formally, we use a single Gaussian $\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the maximum likelihood estimates of the whole training data for an individual. This approach therefore provides a lower bound on performance that any learning method is expected to beat.

Nokia Dataset

All our experiments were conducted using the Nokia dataset, which was chosen because it comprises highly granular (both in time and space) GPS data taken in a longitudinal study involving 38 people moving about in their daily lives, in Lausanne, Switzerland [10]. Furthermore, this data was captured using commodity mobile devices, giving us greater confidence of repeatability in practice.

The Nokia dataset contains 1,553,154 continuous location readings (with a mean of 40,872 and standard deviation of 36,848 per user) comprising latitude, longitude, and timestamp taken over a period of a year. The timestamp is represented in seconds, which we processed into the periodic measures of day of the week and hour of the day. No further pre-processing was required to make the data work with our model. However, we remain mindful of the potential biases introduced by the data collection method. Specifically, GPS sensor activation was optimised by Nokia so that the power requirements of data collection did not drain the batteries of the participants' mobile phones during the day (important for user compliance). For example, if a user was stationary for a long time, the rate of location recordings was reduced significantly, to respond to the fact that not much extra information would be gained from a higher time granularity. The implication for our model (or, indeed, any model that learns spatio-temporal structure in unprocessed behaviour data) is that sensor behaviour is also learnt, introducing some artefacts that may bias our results concerning human behaviour. We discuss in the next section how we mitigate this concern though the choice of evaluation metric.

Evaluation Metric

The purpose of our evaluation method is to determine the prediction quality of all the approaches as they relate to human behaviour. Given this, we treat any statistical structure related to sensor behaviour (i.e., optimising energy efficiency in data collection) as noise. A common metric for model quality is the joint data likelihood of the test data. However, we wish to mitigate the concerns of bias introduced by sensor behaviour (which we discussed in the previous subsection). We do this by evaluating with the conditional likelihood of the test locations, $p(x_{test}|t, d, u, \boldsymbol{X_{train}})$, which indicates how much probability density the model gives at the test location given information about the time that this location was recorded. Since we are providing the time context from the ground truth, the focus is on the performance of the model to give an accurate *location* distribution, not on the ability of the model to learn *when* the GPS was most likely to be activated.

Testing Data Sparsity

Given the aforementioned elements, we now detail the experiments we performed. The condition of data sparsity may be recreated (i.e., simulated using behaviour data) by varying which data points are seen by the model during training, while keeping the test dataset fixed. The reason why we need to intentionally recreate sparsity is because we need the ground truth (data that was held back during training) to verify the accuracy of each model's predictions. This requires access to a deep dataset for our experiments, to obtain statistical significance, though this would obviously not be a requirement in deployment. We now detail how we vary the depth and width of the dataset for our experiments.

- **Experiment 1: Varying Dataset Depth** To recreate the case of a new user arriving to a system, we train the model on the first H non-empty hours of location observations of the individual user. Thus, the independent variable in this experiment is the number of fixed time units (hours) observed during training. We emphasise that H is always shorter than the total observation period for the user due to missing data (either because their phone was switched off, or because of a lack of GPS satellite reachability). Since the number of missing hours varies from user to user, there is no universal conversion factor between number of non-empty hours and total observation period. However, H corresponds approximately to $1.5H$ hours of observation (e.g., $H = 100$ corresponds to over a week's worth of location observations).

  As we increase H, we observe the effect on predictive ability of all the approaches. However, our model (`LocHDP`) is also able to see the full mobility patterns of all other users in the dataset, who represent the established users in a location prediction system (this set obviously excludes the individual we are testing). For this experiment, we randomly selected 10 such auxiliary users. Comparing the performance of our approach to the other approaches, we can evaluate how much benefit is derived from having this set of auxiliary established users' data. In order to get a full evaluation for different values of H (for $0 < H \leq 200$), we select the 9 individuals in the Nokia data set who had more than 1000 non-empty hours associated with them as the test individuals, testing with the location observations that were not seen during training (i.e., held-out testing).

- **Experiment 2: Varying Dataset Width** While Experiment 1 deals with the number of training hours provided (i.e., the depth of the dataset), here, we investigate the benefit of having a wider dataset for prediction by varying the number of *auxiliary* users (i.e., users in the population, but not directly evaluated for prediction). We randomly sample the size of the auxiliary set *aux*, for sizes in the range [0..20], and test predictive performance of the 9 individuals in turn, fixing the number of hours for each of the 9 users to 20 hours.

Experiments Using Human Behaviour Data

The results for Experiment 1 can be seen in Figure 4, with errorbars indicating the 95% confidence interval for each result. The `Single` baseline performs consistently badly because it places probability mass at many locations that the user never visits[5]. The `Mixture` does much better because the temporal context is used to refine the prediction. However, `LocHDP` outperforms `Mixture` by a factor of 2.4 (since the liklihoods are presented in $\log_e$ space) with only 20 hours of training data, rising to a factor of 6.4 with 40

[5]The values of `Single` are so low for $H < 190$ that they do not fit on the plot in Figure 4.

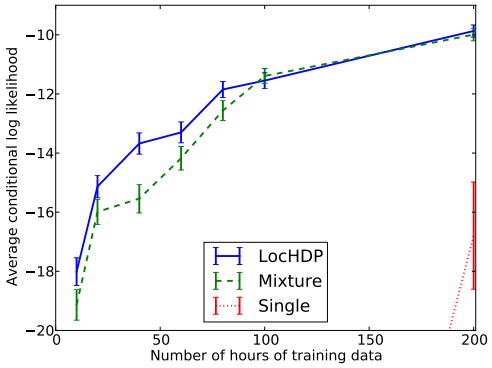

Figure 4. The results of Experiment 1. We see that the model quality of the population model is significantly better than all other approaches, up to 100 (non-zero) hours of training data. Error bars represent the 95% confidence range (using the empirical standard error measure).

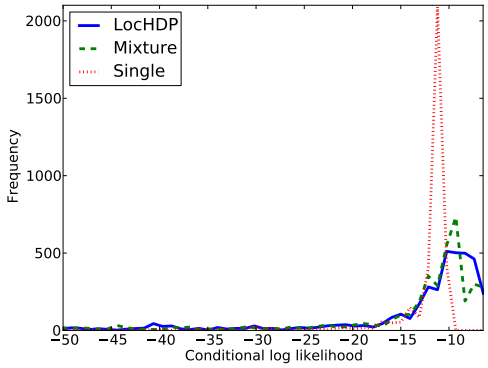

Figure 5. The results of Experiment 1, showing the distribution of conditional data likelihood for all 9 target users for each approach when training on 20 hours of data.

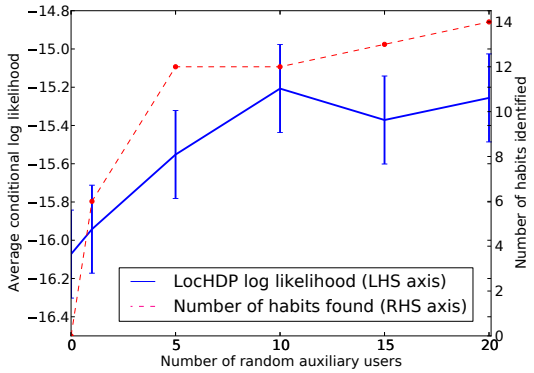

Figure 6. The results of Experiment 2, showing the model quality of LocHDP (left-hand axis) and the number of habits discovered (right-hand axis) when training with varying numbers of auxiliary users. Error bars represent the 95% confidence range (using the empirical standard error measure).

hours of data. However, this advantage disappears once a significant amount of training data is available (100 hours, corresponding to a period of observation of over a week). This is to be expected, as the most useful information about a person's future location behaviour is their own past behaviour. Our experiments suggest that 100 hours of training data is the point beyond which past behaviour of an individual alone is enough to predict their future behaviour. In the absence of such information, sophisticated assumptions about user similarity provide a boost to prediction. Therefore, modelling multiple individuals with our approach provides a considerable benefit in behaviour prediction (under sparsity) compared to considering only individuals. We attribute this to the fact that parameter sharing (specifically of $\boldsymbol{\delta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$) gives our model of the new user a head start over the purely individual approach.

To understand more deeply the difference in performance between the three approaches, we plot the distribution of error for all the test points after training on just 20 hours of observations (Figure 5). For all approaches there is a concentration of density in the range [-14,-10] and a heavy tail. This corresponds to the well known power laws in human routine mobility, in which a few locations are visited very often, while many locations are visited infrequently (causing low accuracy in the prediction results) [8]. The difference between the approaches is in the key [-14,-10] range. We see that the benchmark, `Single`, has strongly concentrated mass around -12. This is because it places fairly even probability density around the whole area where the individual lives and works, ensuring a fairly reasonable accuracy with most predictions, but foregoing the possibility of higher precision. The reason why `Single` performs so badly on average (see Figure 4) is because of its longer tail of very bad predictions (not shown in Figure 5). `Mixture`, on the other hand, is able to achieve highly accurate predictions for a large number of time contexts, and also avoids very low accuracy predictions. However, `LocHDP` is the only approach that has many results in the highest accuracy range [-9,-6], which we attribute to better model quality.

We now consider Experiment 2, in which we kept the number of training hours fixed at 20, but varied the number of auxiliary users (i.e., users in the model that may help with predictions of the new user). The results for this experiment are shown in Figure 6. Since this is only applicable to the population approach, the results for `Single` and `Mixture` are omitted (they stay constant for all group sizes). The general trend is an increase in prediction accuracy as the number of auxiliary users increases. This is due to the fact that data describing more individuals is likely to result in `LocHDP` learning new habits that may be applied for better prediction. However, there appears to be no additional benefit in this dataset of having more than 10 auxiliary users, after which the log likelihood remains in the range [-15.0,-15.5]. Against this background, the small dip in performance for 15 auxiliary users can be explained by random variation alone.

We verified the assertion that `LocHDP` learns more habits with increased number of auxiliary users by also plotting the number of habits discovered for varying numbers of people in the group (using the right-hand axis in Figure 6). We see that the model does indeed identify new habits as the group size increases, but that this rate of increase declines in a similar way to that of prediction accuracy (after 10 auxiliary users). This evidence supports the hypothesis that `LocHDP` overcomes the shallowness of a dataset using its width.

To see how these habits are shared amongst users, we also plotted the distribution over habits for 9 randomly selected users in Figure 7. The most prevalent was habit 12, which the model inferred was responsible for more than 30% of the observations in 5 of the users displayed (as indicated by the y-axis in Figure 7). Manual verification[6] indicates that these were the home locations of the 5 users. On the other hand, habit 12 did not have as clear an interpretation for the other 4 users in Figure 7. This is an expected feature of heterogeneous habit modelling, in which the behaviour parameters of some users will exhibit strong commonalities.

CONCLUSIONS AND FUTURE WORK

We have presented an extension to the HDP, called LocHDP, that probabilistically captures routine location behaviour in populations, without the limiting assumptions present in existing group models. As a demonstration of the usefulness of LocHDP, we identified data sparsity as an often overlooked problem in existing location prediction work, to which we applied the model. In experiments on a human mobility dataset we found a significant improvement in prediction.

One possible limitation of our sparsity result comes from the dataset we used. The Nokia dataset may contain habitual behaviour overlaps that may not always be present in any randomly selected subset of the whole population. This is due to the recruitment campaign for the study, which was referral based [10]. On the other hand, the dataset did contain many users who do not know each other directly. In future work, we would like to perform further experiments with a larger (i.e., wider) dataset. Furthermore, we want to investigate strategies for boosting completely new users to a mobility model using a judiciously selected set of questions to ask the user (e.g., direct input of key locations).

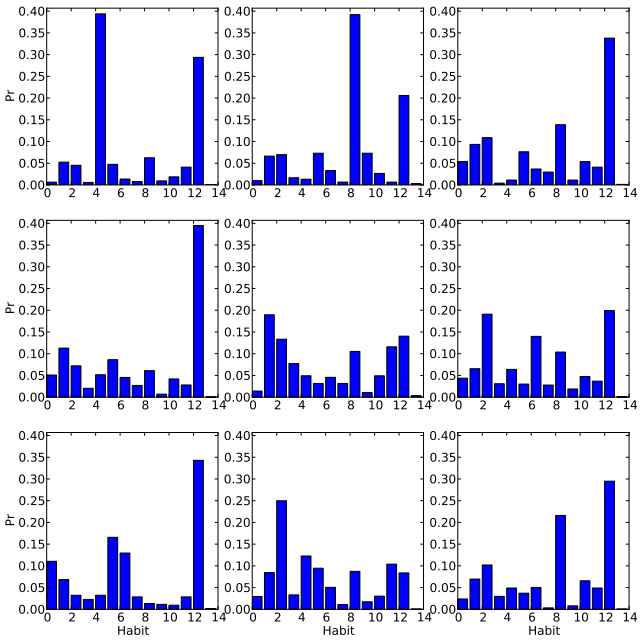

Figure 7. The distribution over habits for 9 randomly selected users (out of 20 who were modelled together).

REFERENCES

1. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.
2. E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. 17th ACM SIGKDD*, pages 1082–1090, 2011.
3. M. De Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. In *Mobile Data Challenge by Nokia Workshop, in conjunction with Pervasive*, 2012.
4. N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
5. K. Farrahi and D. Gatica-Perez. Extracting mobile behavioral patterns with the distant n-gram topic model. In *Proc. ISWC*, 2012.
6. L. Ferrari and M. Mamei. Discovering daily routines from google latitude with topic models. In *PerCom Workshops*, pages 432–437, Mar. 2011.
7. H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *6th ICWSM*, 2012.
8. M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.
9. J. Krumm and A. Brush. Learning time-based presence probabilities. *Pervasive*, 6696:79–96, 2011.
10. J. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge by Nokia Workshop, in conjunction with Pervasive*, 2012.
11. D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Life in the network. *Science*, 323(5915):721–723, 2009.
12. D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.
13. J. McInerney, A. Rogers, and N. R. Jennings. Improving location prediction services for new users with probabilistic latent semantic analysis. In *Proc. 4th Workshop on Location-Based Social Networks*, UbiComp, 2012.
14. P. Nurmi and S. Bhattacharya. Identifying meaningful places: The non-parametric way. In *Pervasive*, pages 111–127, 2008.
15. A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. WSDM, pages 723–732, New York, NY, USA, 2012.
16. S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169, San Francisco, CA, USA, 2011.
17. X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:1–19, 2009.
18. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *J. of the American Statistical Association*, 101(476):1566–1581, 2006.
19. J. Zheng and L. M. Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proc. UbiComp*, pages 153–162, 2012.

[6]by inspecting an animation of each user's location history and finding that the user was at this location mostly in late evenings and weekends.