Inferring the activities of smartphone users from context measurements using Bayesian inference and random utility models

Ricardo Hurtubia, Gunnar Flötteröd, and Michel Bierlaire
Transport and Mobility Laboratory (TRANSP-OR), EPFL

Abstract

Smartphones collect a wealth of information about their users' environment and activities. This includes GPS (global positioning system) tracks and the MAC (media access control) addresses of devices around the user, and it can go as far as taking visual and acoustic samples of the user's environment. We present a Bayesian framework for the identification of the current activity type of a smartphone user. As the prior information, we use a random utility model that predicts the type of activity a user is likely to perform given (i) the user's socioeconomic features, (ii) the land use of the user's current location, and (iii) the time of day. This model is estimated using data from the 2005 Swiss transport microcensus. The smartphone measurements come from a experimental survey, where one user carried around a phone programmed to constantly record his GPS location and other context variables, including the MAC addresses of nearby Bluetooth devices. In addition to this, the user answered a daily survey, where he described and geo-located all the activities performed during this period. An analysis of the recorded data shows that the information about nearby Bluetooth devices can be related to particular activities that are conducted jointly with the owners of that devices. The likelihood function is therefore specified as the probability of observing particular Bluetooth devices when conducting particular activities. Due to the limited amount of available data, only exemplary results are given, which, however, clearly indicate that the accuracy of the prior model can be greatly improved by using Bluetooth data.

1 Introduction

Smartphones have access to an enormous amount of information about their user. They measure the user's location, speed, process every communication activity of the user, scan nearby wireless networks and connect to other devices, and they are equipped with both a microphone and a camera.

"Context-aware computing" infers from this type of data what information to provide to the user in a given situation (Chen & Kotz, 2000). For example, a user who has searched a particular item on the web a while ago might appreciate being informed about nearby shopping opportunities that provide this or a similar item. Context aware devices also bear a great potential for travel demand modeling in that they reveal the user's mobility and activity patterns.

Most research in the field of "context awareness" focuses on using the measurements that come from smartphones or other devices to infer the location of the user (Roos & Tirri, 2002), the transport mode he is using (Patterson et al., 2003) or the activity he is performing (Wilson & Atkeson, 2005). In most of the literature, the "activity" is understood as a movement-related state of the user (walking, driving, static) (Anderson & Muller, 2006) or is understood as a frequent place that can be associated with an activity, mostly because of the frequency and time of the day at which the location is visited (Papliatseyeu & Mayora, 2008). This implies that the forecast is difficult or inaccurate for locations where there are no previous measurements. This strain of research is strongly founded on "automatic learning processes" (Liao et al., 2007) which are based on data mining techniques and are strongly dependent on the availability of abundant data.

The approach presented in this paper is different from the previous ones because it mixes knowledge of the general behavior of the population with measurements for a specific user. The general information is represented in terms of a random utility model (Manski, 1977) that maps the socioeconomic characteristics of a user, the land use pattern of the user's current location, and the time of the day on a prior choice distribution of the user's current activity. This model is estimated from observations of a large number of people in the analysis zone, which is typically available in form of a transport census or an activity survey. The user-specific information considered in this work consists of the physical (MAC, media access control) addresses of nearby Bluetooth devices, which belong to, e.g., friends, working colleagues or relatives. Bluetooth is a data-transfer and communication technology that has become a standard in most medium to high-end mobile phones, being also frequently present in laptop computers and other mobile and stationary devices. Every Bluetooth device has a unique MAC address that is assigned by the manufacturer. The basic hypothesis that underlies in this work is that users conduct certain activities together with certain people, and hence there is a strong correlation between the presence of these people's Bluetooth devices and the user's current activity. This is consistent with the findings of Eagle & Pentland (2006), who observe that a person's social network is

related to the interactions between Bluetooth devices around that person. These two sources of information are combined in a Bayesian framework that updates the activity hypotheses generated by the random utility model based on observations of nearby Bluetooth devices.

The remainder of this article is organized as follows. Section 2 explains the structure and construction of our framework, explaining the prior model, the selection and processing of the measurements and the construction of the likelihood function. Section 3 shows an example of the application of the framework and the corresponding result analysis. Finally, Section 4 concludes the paper and identifies further work.

2 Framework

The aim of this work is to develop a framework to infer the activities an individual (in this case a smartphone-user) will perform. Activities are characterized by their purpose or type (work, shopping, leisure, etc.), denoted in our formulation by the index "a", by the time of the day at which they are performed (t) and by the location at which they are performed (a specific address or a zone, denoted by i).

We want to combine general knowledge of the user's activity-choice behavior (which we consider to be our "prior knowledge") with the measurements of context variables coming from the smartphone. This is done through Bayes' law, which, for our specific problem, can be essentially expressed as:

$$P(\text{activity} \mid \text{measurements}) \propto P(\text{activity}) \cdot P(\text{measurements} \mid \text{activity}). \quad (1)$$

In our specific formulation, $P(\text{activity})$ is generated by a random utility model that, more specifically, predicts the probability $P(a|i,t)$ of performing an activity of type a given the land use attributes of the current location i and the current time t. This model is general and estimated for a whole population. The likelihood $P(\text{measurements} \mid \text{activity})$ indicates how probable it is to observe a user-specific measurement Y at time t given that the user currently performs activity a, which we write as $P(Y|a,t)$. Using this we can write the general inference equation:

$$P(a|Y,i,t) = \frac{P(Y|a,t) \cdot P(a|i,t)}{P(Y|i,t)} \quad (2)$$

where

$$P(Y|i,t) = \sum_{a'} P(Y|a',t) \cdot P(a'|i,t) \quad (3)$$

is a normalizing term which ensures that the posterior probabilities of all possible activities sum up to one.

This specification accounts for time but it does not account for the temporal structure of the user's activities (Bowman & Ben-Akiva, 1998). While the Bayesian techniques for the recursive tracking of an activity sequence based on a dynamic model of user behavior are well-known in principle (Arulampalam et al., 2002), the limited amount of data available for this study does not allow to account for dynamics in this work.

The following subsections detail each element of (2).

2.1 Prior model

We assume that there is a relation between the activity a user performs and the land use characteristics of the zone were the activity is performed. For example the presence of a school indicates a high probability of performing education-related activities (if the user is a student) and the presence of commerce implies a high probability of performing shopping. We also assume that the time has an important effect in this relationship since some activities are more likely to be performed at specific periods of the day (e.g. work in the morning or leisure at night in weekdays).

Speaking in terms of a random utility model, we assume that the utility U_{na} an individual n perceives from performing an activity a is a function of the land use, the time of the day, and the individual's socioeconomic characteristics. Assuming a multinomial logit structure ((Ben-Akiva & Lerman, 1985)) the probability that person n performs activity type a at zone i and time period t is

$$P_n(a \mid i,t) = \frac{\exp(U_{na}(z_i, z_n, \delta_t))}{\sum_{a'} \exp(U_{na'}(z_i, z_n, \delta_t))} \tag{4}$$

where z_i is the vector of land use attributes of zone i, z_n is the vector of attributes of individual n and $\delta_t = (\delta_{tp})$ is a vector of indicators that relate the time of the day with a certain period of the day (night, morning, noon and afternoon). The possible activity types are work, study, shopping, consumption of services, leisure and other; these categories come from the types of activities reported in the Swiss Transport Microcensus (ARE/BfS, 2007).

The time periods p are also derived from the 2005 Swiss Transport Microcensus, which shows that different activities are concentrated in different periods of the day. Figure 1 shows the temporal distribution of the most frequently observed activities (work, study, shopping and leisure) and the

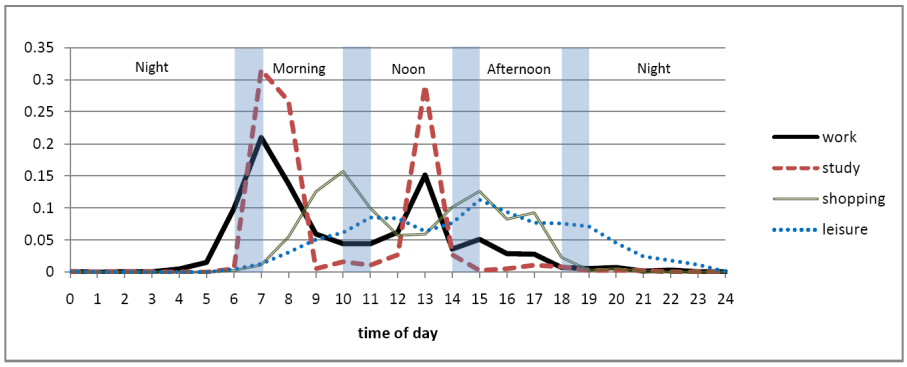

Figure 1: Distribution of starting times of activities over the day

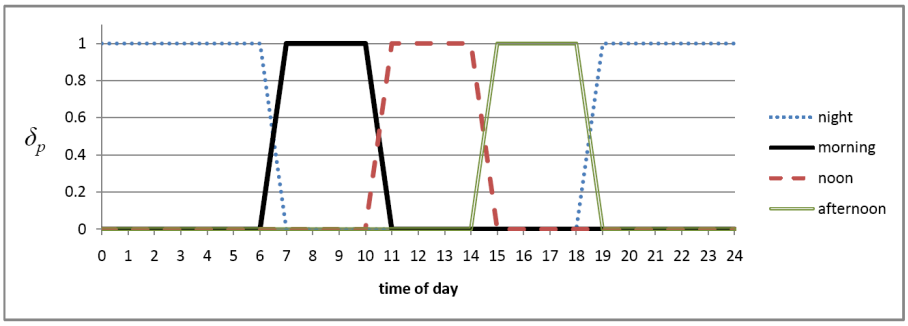

Figure 2: Membership function for periods of the day

chosen thresholds for the different periods of the day (night, morning, noon and afternoon). These thresholds are chosen such that they group the peaks of observed activity-changes while maintaining some consistency with the common understanding of these periods for easier interpretation.

The piecewise linear membership functions shown in Figure 2 are used to assign each time of the day to a period while maintaining "soft" transitions between the periods. These membership functions constitute the indicators δ_{tp} of which δ_t in (4) is comprised. The membership functions sum up to one at any point in time:

$$\sum_p \delta_{tp} = 1 \; \forall t, p \in \{\text{night, morning, noon, afternoon}\} \tag{5}$$

The model (4) is estimated using the observed performed activities reported in the travel diaries of the Swiss transport microcensus 2005. The chosen study area is the canton of Vaud, such that only activities performed in this region are accounted for. The land use attributes are calculated for grid cells of 100x100 meters from the 2005 Swiss Population and Enterprise Census.

Table 1: Prior model estimation results

Variable	Work	Study	Shopping	Service	Leisure	Other
constant	-	-0.532	2.031	2.311	3.522	0.656
male	0.713	-	-0.377	-0.278	-	-
employed	2.132	-	-	-	-	-
children	-	-	-	-	-	0.379*
industry	0.025	-	-	-	-	-
commerce	-	-	0.077	-	-	-
services	0.046	-	-	0.055	0.024	-
other	0.032	-	-	-	0.053	0.065*
retail	-	-	1.074	-	-	-
long_term_retail	-	-	0.554	-	-	-
restaurant	-	-	-	-	0.109	-
school·age<19	-	1.694	-	-	-	-
high_educ·student	-	1.328	-	-	-	-
δ_{morning}	2.720	-	0.887	1.341	-	-
δ_{noon}	1.001	-	-	-	-	-
δ_{morning}·student	-	6.516	-	-	-	-
δ_{nooon}·student	-	4.212	-	-	-	-
δ_{morning}·age>60	-	-	1.114	-	0.836	-
$\delta_{\text{afternoon}}$·age<19	-	-	-	-	0.813	-
$\delta_{\text{afternoon}}$·age>60	-	-	-	-	-0.242	-
δ_{night}·age19_25	-	-	-	-	1.683	-

* significance $< 95\%$

For each activity type, the utility is specified as a linear function of user specific attributes, land use variables and period-membership indicators. The model was estimated with Biogeme (Bierlaire, 2003) and results are shown in Table 1. Estimation results are considered to be good, since all parameters have the expected signs and are significant, with the exception of activity-type "other". As expected, activities are related with the land use of the zone where they are performed. For example, industry and services establishments have a positive effect in the probability of working; commerce and retail increase the probability of shopping, and the presence of service establishments and restaurants makes it more likely to perform leisure activities. The periods of the day also prove to be relevant: work is likely to be performed earlier in the day and leisure activities are also concentrated at specific periods of the day that depend on the age of the individual. (The model was also estimated for the canton of Zurich, where similar parameters are obtained. This indicates a certain robustness.)

2.2 Measurements

The context measurements available for this study come from one smartphone (Nokia N95) that was carried by one subject while performing his usual activity schedule over two months. The subject also answered a daily online activity survey, where each activity of the day is reported indicating type of activity, starting and finishing time, and location. Figure 3 shows a screenshot of the survey.

The smartphone measured the following information:

- GPS track
- speed and acceleration
- nearby Wi-Fi networks
- cellphone network tower (antenna ID)
- nearby Bluetooth devices

Figure 4 shows an example of how the number of detected Bluetooth devices changes between activity types. The empirical distributions (calculated over approximately 8700 observations) show that activities like leisure and service consumption have similar patterns, being most likely to observe zero devices. Shopping also exhibits a high probability of observing few devices, but in

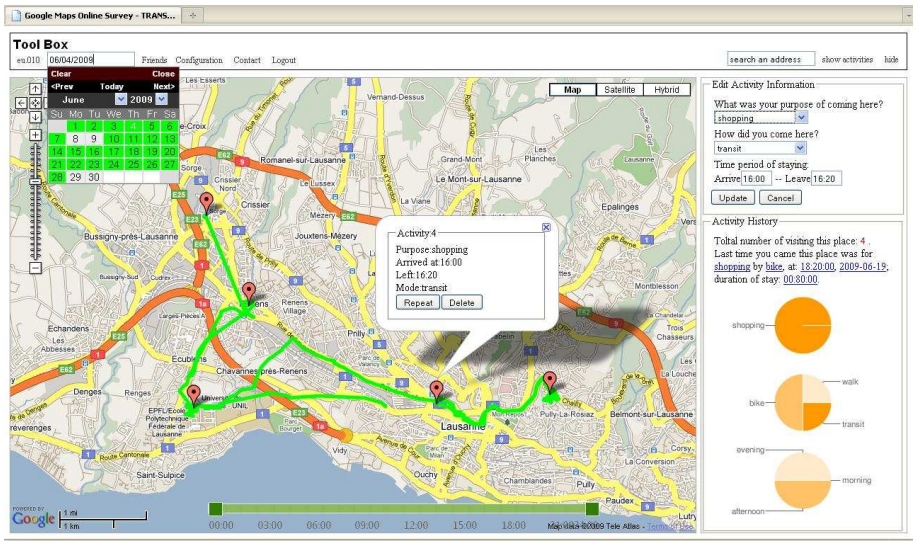

Figure 3: On-line activity survey screenshot

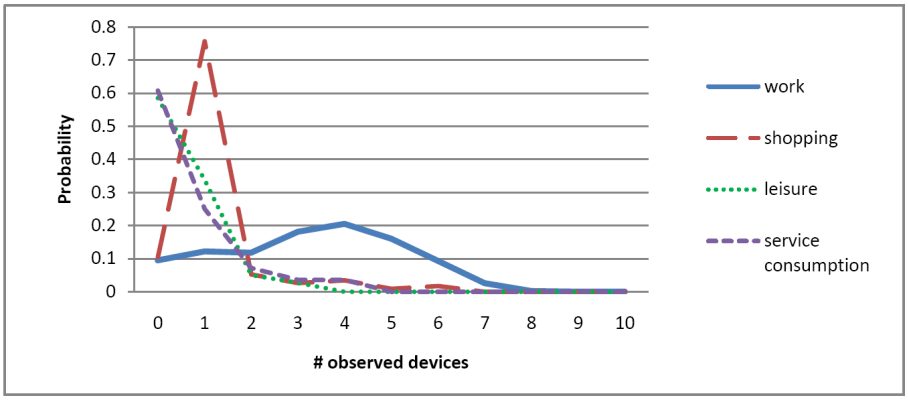

Figure 4: Empirical distribution of the number of detected Bluetooth devices by activity type

Table 2: Counts of frequent devices by activity type

Device	Work	Leisure	Shopping	Total count
A	26	0	0	26
B	24	0	0	24
C	24	0	0	24
D	23	0	0	23
E	21	0	0	21
F	19	1	0	20
G	0	9	7	16
H	15	0	0	15
I	15	0	0	15
J	8	0	0	8
K	8	0	0	8
L	7	1	0	8

this case the empirical mean is near one device, which indicates a structural difference from leisure and service consumption. In the case of work, there is a clear tendency to detect a much higher number of devices but, at the same time, this is a much flatter probability distribution. The cause of these differences in the observed patterns is not obvious, and it is unlikely that these observations can be generalized to other phone users. This motivates a more disaggregated analysis of the Bluetooth data.

An analysis of the available data reveals that some MAC addresses are frequently observed while performing some activities. Table 2 shows the counts by activity type for the 12 most frequently detected MAC addresses (detected more than 3 times).The letters A to L are assigned as identifiers to each device. Repeated observations of devices are only made while performing activities of type work, leisure or other. Most of them are observed only at work, with the exception of devices F and L, which are also observed at leisure activities. Device G constitutes a noticeable exception in that it is observed only at shopping and leisure activities.

When building a likelihood function, correlations of the measurements need to be accounted for. It is possible that some devices are only observed together; this means that even when each one of them can be strongly related to some specific activity type, observing all of them does not provide any additional information. This indicates that such devices should be grouped and treated as one independent device. Considering this, a pairwise comparison analysis was performed, and devices were grouped according to their correlation level. The critical threshold for grouping was roughly a joint

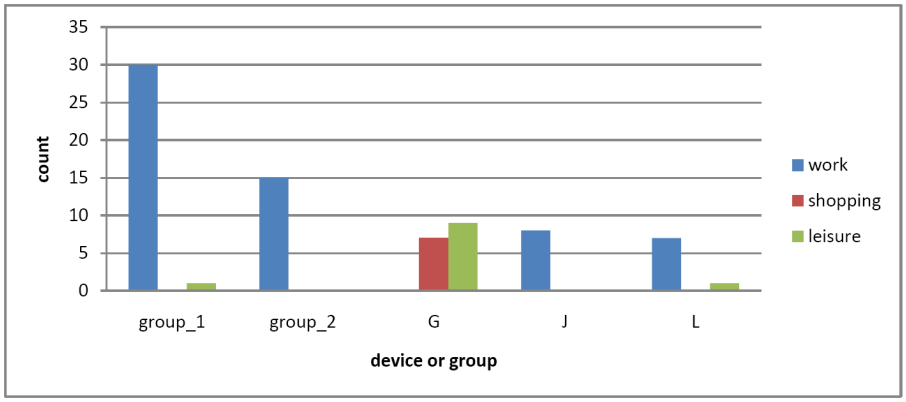

Figure 5: Counts of independent devices by activity type

detection rate of 70%. After grouping, the set of "effectively independent devices" consists of 2 groups and 3 single devices:

- group_1 = {A,B,C,D,E,F,I}
- group_2 = {H,K}
- devices G, J, L

Figure 5 shows the frequency by activity type for each independent device.

Based on this analysis, a measurement Y is defined as a vector of indicators y_j, where each indicator corresponds to one independent device:

$$Y = (y_j) \tag{6}$$

with $j \in$ {group_1,group_2,G,J,L} and

$$y_j = \begin{cases} 1 & \text{if device } j \text{ is observed} \\ 0 & \text{if not.} \end{cases} \tag{7}$$

As from now, the notion of a "device" refers to an "independent device group".

2.3 Likelihood function

The likelihood function is the probability of observing a combination Y of independent device groups given that a specific type of activity a is performed

at time t:

$$P(Y|a,t) = \prod_j (P(y_j = 1|a,t) \cdot y_j + (1 - P(y_j = 1|a,t)) \cdot (1 - y_j)) \quad (8)$$

where $P(y_j = 1|a,t)$ is the probability of detecting device j while performing activity a at time t , which is estimated from the user-specific survey data in the following way:

$$P(y_j = 1 \mid a,p) = \frac{N_{jap} + \varepsilon_a \cdot \alpha}{N_{ap} + \alpha} \quad (9)$$

where N_{ap} is the number of times an activity of type a is conducted during period p and N_{jap} is the number of times an activity of type a is conducted during period p while device j is detected.

The parameter ε_a is the expected probability of observing any device while performing an activity of type a if no measurements are available, and the α parameter weights this uninformed prior knowledge against the data that is obtained from the survey. These parameters account for the uncertainty in the survey data (e.g., not observing something does not mean it will never happen) and must be defined by the analyst. This is particularly relevant for our dataset where there are no counts for activities of type study, consumption of services, and other.

Using the membership functions defined in Figure 2, the likelihood can be applied at any time t of the day according to

$$P(y_j = 1 \mid a,t) = \sum_p \delta_{tp} P(y_j = 1 \mid a,p). \quad (10)$$

A substitution of this and (9) in (8) fully specifies the empirical likelihood function.

The final inference mechanism consists of a straightforward evaluation of the Bayesian posterior distribution given in (2). The following section shows an exemplary application of this method.

3 Application and analysis

Since the presented work is of experimental character, we apply the proposed method to a particular event registered by the user in the on-line survey. This event was selected because of its peculiarity since it is a leisure activity

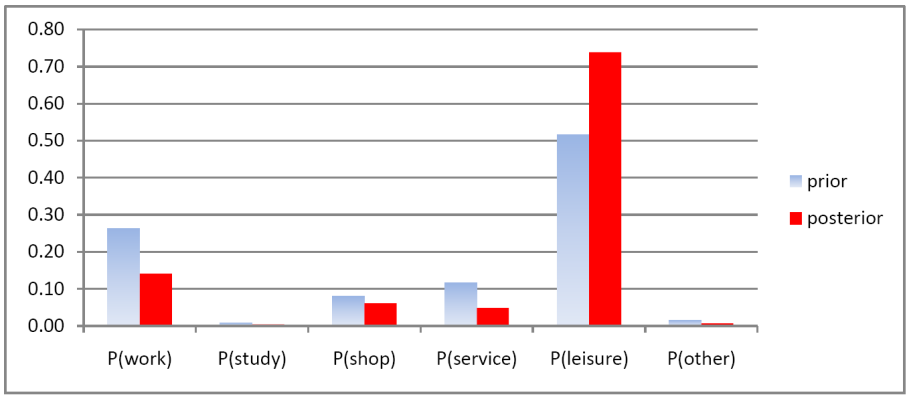

Figure 6: Prior and posterior probability distribution

performed at a location that was systematically identified as a work location in the survey. The activity was performed between 18:00 and 23:00 hrs and, during this period, 3 of the independent devices were detected (group_1 and devices G and J). The detection of group_1 and device J should increase the likelihood for work, since they are mostly observed while performing that type of activity; however, the detection of device G should decrease the probability of work, since it is never observed while performing that activity. Instead, device G should serve as a strong indicator of shopping and leisure activities.

Figure 6 shows both the prior and the posterior activity type distribution for this event, where the empirical likelihood is estimated using the parameters $\alpha = 10$ and $\varepsilon_a = 0.1$ $(\forall a)$, which are defined in the previous section.

As expected, the posterior distribution is significantly better in predicting a leisure activity than the prior distribution. However, the prior is already generating a high probability for leisure. This is owed to the late time at which the activity is performed, which is less consistent with a work activity than with a leisure activity (see the coefficients for the time indicators in Table 1).

The values used for α and ε_a were defined arbitrarily and therefore is interesting to see how results are affected when moving this parameters. Figure 7 shows the value of the probability for leisure in the analyzed event for different values of α and ε. The posterior leisure probability is systematically higher than the prior for every value of the parameter, with the exception of unreasonably small α values. This is consistent with the fact that the amount of available survey data is fairly limited. The results become almost insensitive with respect to α for larger α values, which indicates some robustness

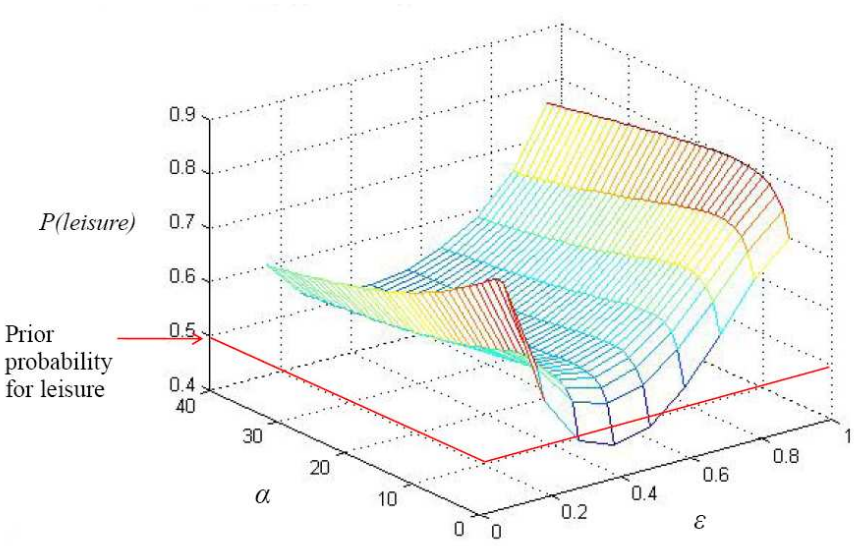

Figure 7: Sensibility of posterior probability to α and ε

of the method with respect to this parameter.

On the other hand, the posterior leisure probability is very sensible and even non-monotonous with respect to ε. It is hypothesized that this effect results from particular data sets in the survey that have too much of an effect because of limited amount of data. This is confirmed by the observation that the distribution flattens with respect to ε as α increases. In general, the reproduction of all registered activities in the survey decreases with ε. This can be seen in Figure 8, which shows the value of the joint log-likelihood for all activities in the survey against ε. This figure also shows that, for small values of ε, the posterior distribution has a much better overall-fit to the observations than the prior.

4 Conclusions and further work

This work presents a Bayesian framework for the identification of activity types from smartphone sensor-data and a prior random utility model that is estimated from readily available census data. In particular, the occurrence of certain Bluetooth devices is found to correlate strongly with the activity a user is performing. The use of a general model prior avoids the extensive data gathering procedures typically necessary for "machine learning" approaches. Due to limited data availability, only exemplary results could be presented.

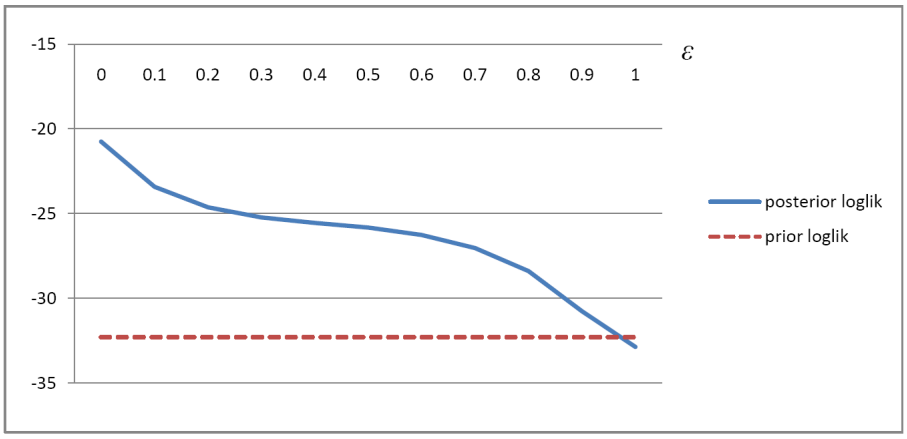

Figure 8: Overall fit to activities reported in the survey

Future work will concentrate on the following items:

- Collection of more data from more users. This is a joint project with Nokia, where the data collection starts in the end of 2009.
- Investigation of further ambient sensor information that is related to an activity, including the modeling of an according likelihood function.
- Regarding the Bluetooth data described in this article, a technique is needed that identifies the link between activities and detected devices from additional sensor data because the on-line survey is infeasible in a real-world application.
- Making the estimation procedure dynamic in that it accounts for the dynamics of activity scheduling and physical movement. Clearly, these modeling efforts will also benefit from the upcoming data collection campaign.

References

I. Anderson & H. Muller (2006). 'Practical Context Awareness for GSM Cell Phones'. In *Proceedings of the 10th IEEE International Symposium on Wearable Computers*, pp. 127–128.

ARE/BfS (2007). 'Mobilität in der Schweiz, Ergebnisse des Mikrozensus 2005 zum Verkehrsverhalten'. Tech. rep., Federal Office for Spatial Development and Swiss Federal Statistical Office, Bern and Neuenburg, Switzerland.

S. Arulampalam, et al. (2002). 'A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking'. *IEEE Transactions on Signal Processing* **50**(2):174–188.

M. E. Ben-Akiva & S. R. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, Ma.

M. Bierlaire (2003). 'BIOGEME: a free package for the estimation of discrete choice models'. In *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland. www.strc.ch.

J. Bowman & M. Ben-Akiva (1998). 'Activity based travel demand model systems'. In P. Marcotte & S. Nguyen (eds.), *Equilibrium and advanced transportation modelling*, pp. 27–46. Kluwer.

G. Chen & D. Kotz (2000). 'A Survey of Context-Aware Mobile Computing Research'. Tech. rep., Dartmouth College, Hanover, NH, USA.

N. Eagle & A. S. Pentland (2006). 'Reality mining: sensing complex social systems'. *Personal Ubiquitous Computing* **10**(4):255–268.

L. Liao, et al. (2007). 'Learning and inferring transportation routines'. *Artificial Intelligence* **171**(5-6):311 – 331.

C. Manski (1977). 'The structure of random utility models'. *Theory and Decision* **8**:229–254.

A. Papliatseyeu & O. Mayora (2008). 'Simultaneous Tracking and Activity Recognition (STAR) Using Many Anonymous, Binary Sensors'. In *3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008*, pp. 343–352.

D. Patterson, et al. (2003). 'Inferring High-Level Behavior from Low-Level Sensors'. In *UbiComp 2003: Ubiquitous Computing*, pp. 73–89.

M.-P. Roos, T. & H. Tirri (2002). 'A Statistical Modeling Approach to Location Estimation'. *IEEE Transactions on Mobile Computing* **1**(1):59–69.

D. H. Wilson & C. Atkeson (2005). 'Simultaneous Tracking and Activity Recognition (STAR) Using Many Anonymous, Binary Sensors'. In *Pervasive Computing*, pp. 62–79.