On Using Existing Time-Use Study Data for Ubiquitous Computing Applications

Kurt Partridge
Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
kurt@parc.com

Philippe Golle
Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
pgolle@parc.com

ABSTRACT
Governments and commercial institutions have conducted detailed time-use studies for several decades. In these studies, participants give a detailed record of their activities, locations, and other data over a day, week, or longer period. These studies are particularly valuable for the ubicomp community because of the large number of participants (often the tens of thousands), and because of their public availability. In this paper, we show how to use the data from these studies to provide validated and cheap (although noisy) classifiers, baseline metrics, and other benefits for activity inference applications.

Author Keywords
Time-use studies, diary studies, activity inference, evaluation methodologies, ubiquitous computing, mobile computing.

ACM Classification Keywords
H.1.m. Models and Principles: Miscellaneous.

INTRODUCTION
Ubiquitous computing has directed much research attention lately toward inference of everyday human activities [2,11,13,20,1,10,19]. Activity inference is a subproblem common to many applications in areas like health monitoring [20], information delivery [2], and transportation prediction [9]. It also shows promise for many more applications that benefit from accurate user models, such as helping people understand how they spend their time, providing ethnographers with more data to help them better understand human behaviors, and supplying epidemiologists with information that helps them understand the relationship between behavior and health.

UbiComp'08, September 21-24, 2008, Seoul, Korea.

A common way that researchers build an activity inference system is to collect sensor readings and ground truth activity data in a user experiment, and then use that data to build an activity classifier. One challenge with this approach is the effort and expense of collecting enough realistic data. Laboratory experiments can produce activity-specific data, but the data may be biased by the artificial setting. Deployed systems are more realistic, but require more robust engineering and longer experiment running times to observe infrequent activities. A recently deployed indoor system [12] reported that in 104 observed hours of a single participant, a couple activities were observed for less than a minute each. Much more observation time would be necessary to draw statistically significant conclusions.

A second challenge of data collection is the difficulty of comprehensive coverage. Some activity inference applications—such as health monitoring and time accounting—require that everyday life be monitored continuously and ubiquitously. Mobile devices would seem to address this issue, but they are often carried in ways that limit their sensing capabilities, and often not even carried at all [17].

Fortunately, governments and other large institutions have been collecting large amounts of coded activity data for decades. These time-use studies list all activities performed by each participant over a 24 hour period (or more). Among other uses, the data is collected to inform significant commercial, political, and economic decisions. Large studies contain tens or hundreds of thousands of participants, and cost millions of dollars. Some of these data sets are available to the public for free.

Time-use studies hold considerable value for ubicomp systems and applications. Among their uses:

1. **Construct Activity Classifiers.** Activity data in time-use studies are linked to other variables such time of day, day of week, participant demographics, copresent individuals, and, in some cases, location and emotion. These variables can be treated as features for a classifier that is cheap to build and covers a broad class of activities. Although time-use data differs in resolution and coverage from sensor data, these two data types

can be combined, for example, in a system that uses Bayesian methods.

2. **Estimate Which Features Predict Best.** Time use data can help determine the relative value of demographics, location, time, and previous activity in making activity predictions either in general or for specific activities of interest. Knowing this helps system designers determine the features they need to achieve a minimum prediction accuracy.
3. **Inform Understanding about Simultaneous Activities.** Time use studies can confirm recent observations from ubicomp studies such as [12] about which activities happen simultaneously.
4. **Identify Circumstances for Rare Activities.** If a new study is designed to collect more information about rare activities, time-use data can identify the situations in which such activities are most likely to happen, thereby minimizing data collection expense.
5. **Validate Study Sites.** For a detailed study at a single location, time-use data can validate that the general activities at the site approximate population norms. This validation supports the generalization of the study's findings to other sites.
6. **Provide Field-Tested Activity and Location Taxonomies.** Longer-running studies have refined their activity and location taxonomies in reaction to the millions of activity records they have collected. Such classifications are more likely to be complete and unambiguous than taxonomies created for a new ubicomp activity study.

This paper contributes to the first three uses of time-use data for ubiquitous computing. After presenting an introduction to time-use studies, it analyzes how well a recent time-use study predicts activity using time, location, demographics, and previous activity. The paper then examines how activity predictability varies at different locations, and for different activities. Finally, it gives an example of how time-use study analysis can easily calculate statistics about simultaneous activities that are much more expensive to collect in an instrumented environment.

RELATED WORK

Other ubicomp research has utilized large data sources. Closest to our approach is Predestination [9], which uses land-use data from the United States Geological Survey and the National Household Transportation Survey to better predict destination places. Time-use data covers the entire day, not just transportation episodes, so it can benefit a broader class of applications. Also, better modeling of activities can lead to better prediction of trips, as suggested by a popular research direction in the transportation modeling field [14].

Much recent research has studied the effectiveness of particular sensor types for determining activities [12,20]. Our approach instead starts from available large-scale authoritative data and the contextual variables they contain. Time-use studies do not have the same detail as sensor-based studies, and are biased by participants' self-reported interpretations, but because these studies include far more participants, they are less biased by individual differences and more likely to cover rare activities well. Logan specifically cites small numbers of episodes for certain activities as a significant problem for (home) activity recognition systems [12].

LifeNet [16,24] and Pentney et al. [19] also use a large, general data set to study how contextual variables might influence activity. But rather than using diary data, these projects start from tens of thousands of interrelated common-sense logical statements, and derive conclusions from reasoning over these statements. We view this approach as complementary to ours. Conclusions from common sense databases may be affected by biases in the database statements, and conclusions from time-use studies may be biased by the way activity is coded.

TIME-USE STUDIES

Table 1 shows an excerpt of time-use data from the American Time-Use Survey (ATUS). ATUS is the largest, most recent time-use study in the United States, and is run by the Bureau of Labor Statistics. Its purpose is to estimate work not included in economics measures (e.g., home childcare). ATUS codes activities hierarchically into three tiers of differing granularity that contain 18, 110, and 462 activity codes [22]. Location is coded more simply, as a 27-valued symbolic variable (see Figure 4 for a subset of the

RESPID	TIME	ACTIVITY (TIER 3)	LOCATION
20060101060033	07:00 - 07:20	Physical care for hh children	Respondents home or yard
20060101060033	07:20 - 09:20	Playing with hh children, not sports	Respondents home or yard
20060101060033	09:20 - 10:20	Physical care for hh children	Respondents home or yard
20060101060033	10:20 - 10:30	Travel related to grocery shopping	Car, truck, or motorcycle (driver)
20060101060033	10:30 - 11:30	Grocery shopping	Grocery store
20060101060033	12:40 - 12:50	Travel related to grocery shopping	Car, truck, or motorcycle (driver)
20060101060033	12:50 - 13:10	Physical care for hh children	Respondents home or yard

Table 1: Seven of the 263,286 activity episodes collected from 12,943 households in the 2006 American Time-Use Study (ATUS). "hh" abbreviates "household." Other variables (not shown) include demographics, family demographics, employment, simultaneous child care, and copresent individuals.

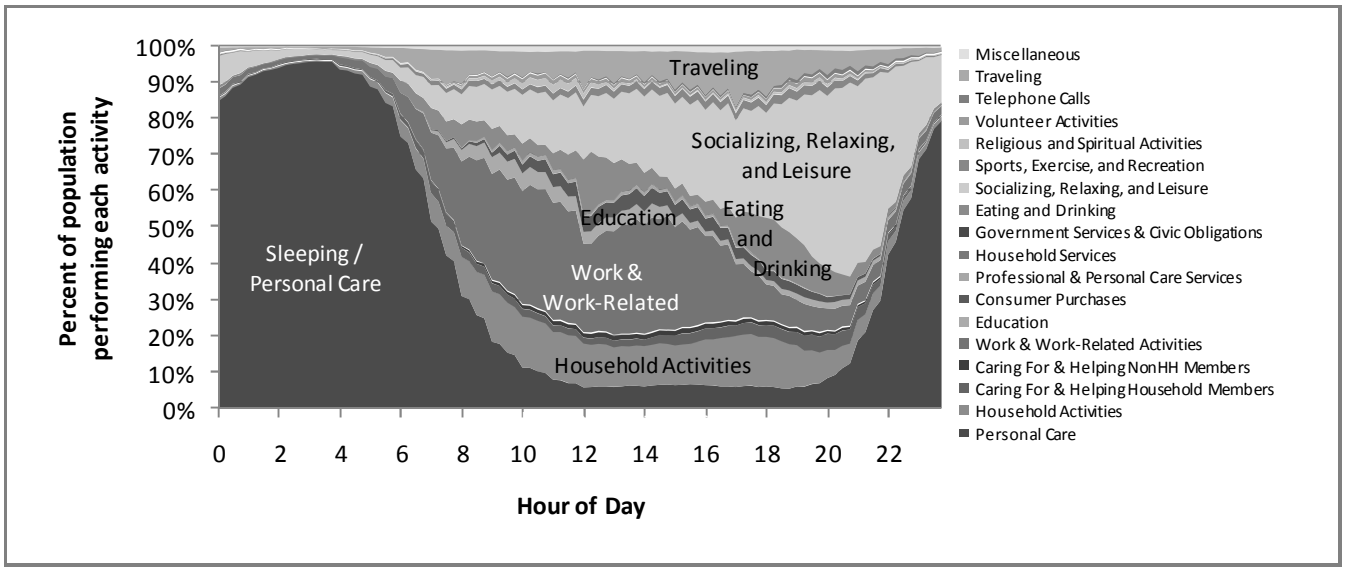

Figure 1: Estimate of US population performing each ATUS Tier 1 activity by time of day (from ATUS 2006). Finer coding is available in other Tiers. Jagged edges reflect biases toward reporting activities on hour and half-hour boundaries.

location codes). Figure 1 shows the overall proportions of time spent in each of the 18 Tier 1 activity codes, and Figure 2 gives examples of codes in the different Tiers.

In addition to ATUS, many other time-use studies have been conducted for many different purposes. Some of the more common motives are to quantify unpaid work, study how behaviors vary by demographic, measure exposure to environmental pollutants, report on the activities parents do with children, and investigate how people spend leisure time. Studies also vary by their duration (24 hours or more), season (year round or all on one day), data collection method (interview or questionnaire), response rates (e.g., 55.1% for ATUS vs. 94.7% for a recent Korean study [23]), number of participants (e.g., over 200,000 in a Japanese study), activity coding (categorical or free text [8], single activity or multiple), additional questions (e.g., pollutant exposure), reporting method (activity episodes or time per activity per participant), and data availability (public or restricted). The Centre for Time Use Research maintains a long list of such studies from over eighty countries at www.timeuse.org [4].

There have been several efforts to unify different time-use study data sets under a common format that simplifies data access and analysis. For the United States, the American Heritage Time Use Study (AHTUS) combines data from ATUS and four older American studies. The Harmonized European Time Use Study (HETUS) currently unifies data from fifteen European countries. AHTUS data is publicly available. HETUS data is restricted, but does publish several charts similar to Figure 1.

The largest effort to unify time-use data is the Multinational Time Use Study (MTUS). Started in the early 1980s, MTUS now combines data from over 50 datasets from 19 countries. It includes records from over 300,000 participants. Access to summary data (time per participant per activity) is available to anyone, but access to episode

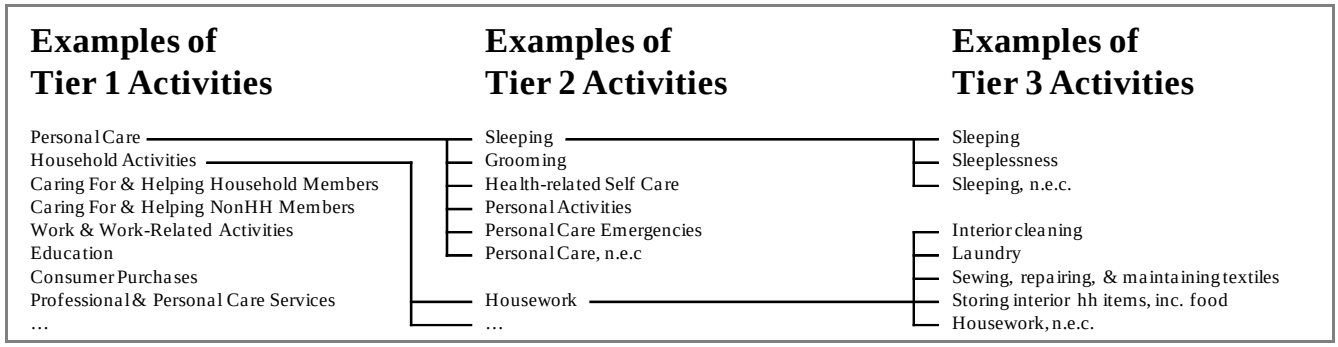

Figure 2: Examples from ATUS's activity classification. There are 18 Tier 1 activities (see Figure 1 for the complete list), 110 Tier 2 activities, and 462 Tier 3 activities. Tier 2 and Tier 3 activities are hierarchically grouped under activities in the preceding Tier. "n.e.c." stands for "not elsewhere classified."

data requires separate permission, in some cases from the supplier of the data.

The time-use research field is older than ubiquitous computing. The International Association of Time-Use Research (IATUR) will hold its 30th conference in 2008. Publications span a variety of topics, including analysis techniques, trends, and subgroup comparisons.

While time-use studies can provide a lot of useful data, they are not problem-free. One of the most critical questions concerns data quality. A participant must consciously recall every activity episode and its details, and communicate the activity accurately to the interviewer. Depending on his or her impressions of the interviewer and beliefs about how the data may be used, the participant may describe activities differently from how they happened, choose to report false activities, or omit some activities. For activities that are reported, the interviewer must judge which code most closely matches the participant's description. In some cases, a description might have more than one possible coding. Or there may be multiple activities, and the interviewer must choose the most important [21]. Some studies use multiple coders to reduce individual coding biases.

Smaller studies have explored alternative methods, such as the Experience Sampling Method (ESM) [3], telephone-based sampling at random times, and direct observation. Efforts to measure validity of the different approaches have shown that correlations between diaries and ESM, and random hour questioning vary between about 70% and 80% [21] (p. 82).

This paper only examines results from American studies because these data are easily downloadable and contain full activity episodes. To our knowledge, most other studies either restrict access or publish only summary statistics, not individual activity episodes.

For more information about time-use studies in general, see Pentland, et al. [18] and Michelson [15].

SUITABILITY OF TIME-USE DATA FOR UBICOMP

Like activity data collected from ubicomp studies, time-use study data records activities as a function of time, place, and demographic data. But there are differences. Activities in a time-use study may last a couple hours, whereas an activity in a sensor-driven study often lasts between an instant to tens of minutes. Also, time-use studies usually cover all activities in an entire day, whereas ubicomp studies may focus on a limited domain, such as physical motion, in-home activities of daily living, or mechanical repair. Finally, all data in a time-use study is cognitively processed by the study participant, and, in many studies, also by an interviewer. Data in ubicomp studies comes from sensors.

These three differences, the "duration difference," the "domain specificity difference," and the "cognitive interpretation difference" may cast doubt on the appropriateness of time-use study data for ubicomp systems. We now argue that despite these issues, time-use data offer important benefits.

The duration difference arises because participants have only so much patience for reporting their activities. When asked to give a detailed account of what they did, participants may combine many short activities into a single, longer, more abstract one. "Scooping granola, pouring milk, lifting spoon, …" becomes "eating breakfast." While the former activities may be easier for sensors to detect, the latter expression more closely matches how a person would communicate the morning's events, unless something unusual happened during the "eating breakfast" routine. Because of their coarser granularity, time-use studies cannot predict activities at the detail they are sensed, but they can bias predictions toward more likely activities.

The domain specificity difference results from the effort required to bring together data from various sensors and sensor types in different domains, and to associate all the collected data with the same individual. Some application-driven activity-inference projects avoid facing these problems by working on a domain-specific application (e.g., bicycle repair). Today, general time-use data is of limited use for these applications. However, we believe that systems can benefit from cross-domain inferences (e.g., purchasing a tire at a bicycle shop may suggest a particular bicycle repair later), and that the cross-domain platforms that make such systems possible will benefit from broad time-use data.

Finally, the cognitive interpretation difference causes inaccuracies through misunderstandings. However, cognitive interpretation can also enable data collection of privacy-sensitive activities, such as bathroom use. Participants may feel more comfortable describing these activities than having them sensed and recorded. Time-use studies may therefore contain more accurate data about these activities than can be collected from sensors.

INFERRING ACTIVITY FROM CONTEXT

How accurately can activity be inferred by time-use study data? We investigate this question by using the activity distributions in the ATUS data to construct maximum-likelihood classifiers. That is, given input variables $v_{1..n}$, the classifier infers the activity a that maximizes the conditional probability $\Pr(a \mid v_1, v_2, \ldots v_n)$. We compute the full joint conditional probabilities for our analysis.

The input variables we consider are hour of day, day of week, sex, age group, previous activity, and location. To simplify analysis, where necessary we limit the range of input variable values to a small discrete set. That is, we use hour instead of time to the minute, and age in groups of five years instead of by year. Because each input variable can only take one of a few values, and because of the large number of activity episodes, it is possible to calculate

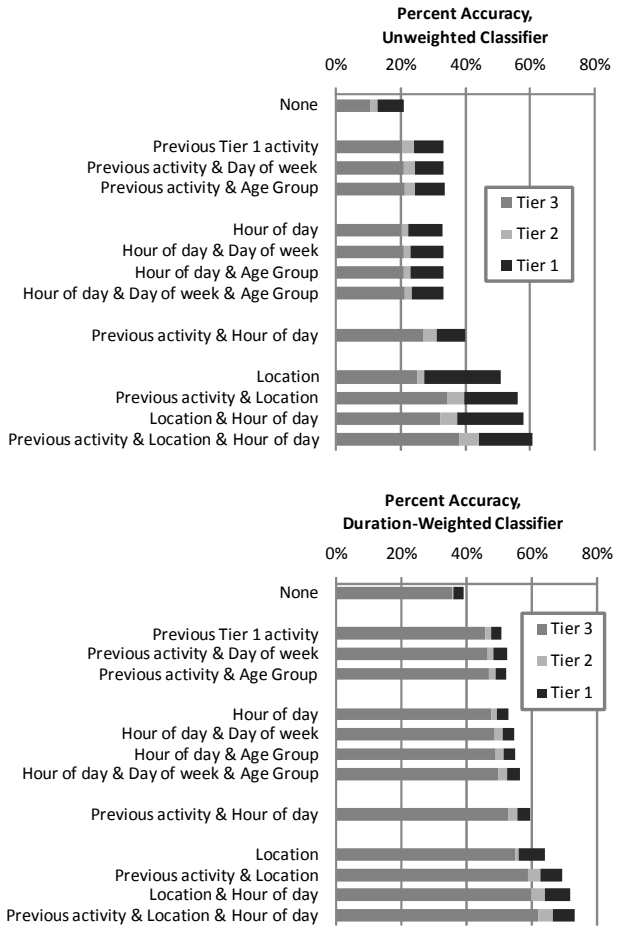

Figure 3: Percentage of activities inferred correctly in the ATUS dataset according to various contextual variables and Tiers. Location is the most informative variable, but higher accuracies are possible by using hour of day and the previous activity.

maximum-likelihood estimates over joint distributions that have hundreds to tens of thousands of activity episodes for the most likely activity.

We evaluate two kinds of classifiers, an "Unweighted Classifier" and a "Duration-Weighted Classifier". In the Unweighted Classifier, each activity episode has equal influence regardless of its duration. It favors shorter activities that happen more often (such as "Telephone Calls") over longer activities that happen less frequently (such as "Sleeping"). An Unweighted Classifier is appropriate if classifications are made a fixed number of times per activity. For example, a classifier that ran with every location change would perform approximately like an Unweighted Classifier.

The Duration-Weighted Classifier multiplies each activity episode by its duration before computing the activity distribution. The Duration-Weighted Classifier is more appropriate if classifications are made independently of activity duration, such as on an hourly basis.

ATUS codes the location of three activities (sleeping, grooming, and sexual activities) as "Not Specified" to protect respondent privacy. Since these activities account for 99% of activities at the "Not Specified" location, a classifier using the ATUS location code shows an unrealistic ability to distinguish them from other in-home activities. We therefore recode these activities as though they happened in the respondent's home. Although some of these activities likely happen elsewhere, we believe that the classifier accuracy figures are more accurate with this recode than without. Also, ATUS codes each type of transportation as a separate location (e.g., "Bus," "Bicycle," "Boat," etc.). For simplicity, we have combined these locations together into a single "Transportation" location.

Overall Accuracy

Figure 3 summarizes the overall results for both the Unweighted and Duration-Weighted Classifiers. The x-axis measures the percentage of time that activities are correctly inferred (the true positive rate). All figures are calculated using tenfold cross-validation.

The y-axis splits the classifiers into five groups. Each group contains a different combination of input variables. Each bar shows the accuracy for a classifier for Tier 1, Tier 2, and Tier 3 activities. Because a lower-numbered Tier need distinguish among fewer activities, it always classifies more accurately than a higher Tier.

The first classifier group, "None," shows the base accuracy using no contextual variables. In the absence of context, the maximum-likelihood activity in the ATUS dataset is "Personal care" (Tier 1) and "Sleeping" (Tiers 2 and 3). Tiers 2 and 3 predict less accurately because "Personal care" covers several other categories that take much less time overall (see Figure 2). Note how the difference is much smaller in the Duration-Weighted Classifier because "Sleeping" has a proportionally larger effect than the other activities.

The next group of three classifiers uses knowledge of the immediately preceding Tier 1 activity to compute the next activity. Because any real system cannot know the previous activity with certainty, these results give an upper bound on the expected true results. We also assume that the first activity of the day is preceded by itself (because ATUS contains no data for these cases). Unsurprisingly, these classifiers perform noticeably better than the uninformed "None" classifier.

"Hour of day" gives a small improvement in performance over "Previous activity." "Hour of day" is more reliably sensed then "Previous activity," so these figures should be obtainable in practice. This classifier is most accurate during the night when "Sleeping" is by far the most common activity.

Location clearly adds the most predictive power, especially for the Unweighted Classifier. However, it is still possible to do better by adding "Hour of day" and "Previous

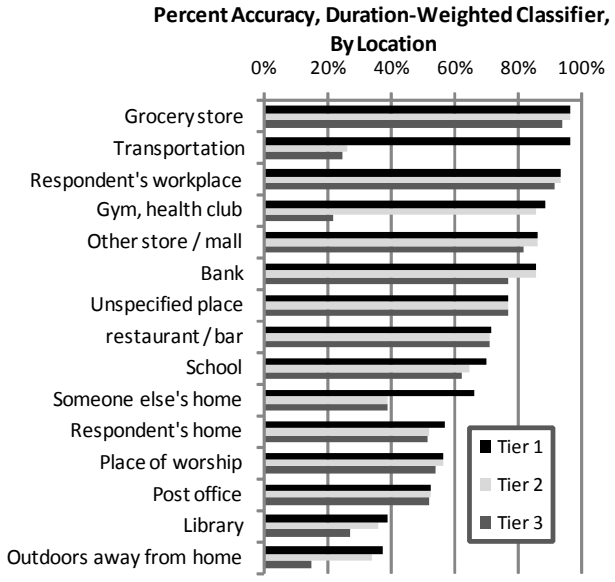

Figure 4: Percentage of ATUS activities inferred correctly, based only on location, for different locations and Tiers.

Activity." Adding "Day of Week" and "Age Group" (not shown in the figure) has minimal effects, less than 0.5%.

Also not shown in the figure are our experiments with these classifiers augmented with the Sex of the participant. Improvements here also made less than 0.5% difference. Note also that although these low-effect variables ("Sex," "Day of week," and "Age Group") do not affect the maximum-likelihood results much, they do alter the activity distribution, so a Bayesian estimator may benefit from their inclusion.

The Unweighted Classifier performs worse than the Duration-Weighted Classifier. We believe this result arises because in the Unweighted Classifier, "Sleeping" is treated as one of many activities. But in the Duration-Weighted classifier, correctly predicting it (which, as Table 3 shows, all classifiers generally do) accounts for several hours of the day and therefore a large fraction of the performance.

Note that the Unweighted Classifiers using Location have wider "Tier 1" bars than the other Unweighted Classifiers. This indicates that a classifier using Location gets many Tier 1 activities correct but fails to distinguish the proper Tier 2 or Tier 3 activity. This effect arises mainly because the "Transportation" location predicts the Tier 1 "Traveling" code well, but does not distinguish among the reasons for traveling, which are included in the Tier 2 codes (e.g., "Travel Related to Personal Care," "Travel Related to Household Activities," etc.).

In summary, these results indicate that hour of day, previous activity, and location all predict activity better than a fixed-activity classifier. Time is easy to sense, which makes it particularly useful. Combining previous activities

Activity (Respondent's Home)	Percent
Sleeping	58.5%
Television and movies (not religious)	12.8%
Washing, dressing and grooming oneself	3.4%
Eating and drinking	3.4%
Food and drink preparation	2.0%
Interior cleaning	1.9%
Reading for personal interest	1.8%
Socializing and communicating with others	1.4%
Relaxing, thinking	1.3%
Work, main job	1.1%
Other	12.4%

Activity (Someone else's home)	Percent
Socializing and communicating with others	38.8%
Television and movies (not religious)	12.7%
Eating and drinking	7.4%
Playing games	5.5%
Attending or hosting parties/receptions/ceremonies	5.2%
House & lawn maintenance & repair assistance for non-household adults	2.8%
Work, main job	1.7%
Housework, cooking, & shopping assistance for non-household adults	1.5%
Relaxing, thinking	1.5%
Food and drink preparation	1.3%
Other	21.6%

Table 2: Breakdown of the time spent in the top ten Tier 3 activities at "Respondent's Home" and "Someone Else's Home." The set of activities suggests that location within the home will partially, but not completely, predict activity.

with time is even more accurate. However if location is available, then it provides the best clues about activity. Note that this is true only when location is accurate and is properly interpreted for a particular individual. What is one person's coffee shop is someone else's workplace.

Activity Inference Accuracy, by Location

Although a classifier that uses location performs better than classifiers using other variables, it is far from perfect. It never does better than 80%. Figure 4 shows that in fact accuracy ranges from above 95% down to around 15%, depending on location and Tier.

Unsurprisingly, prediction is best in cases where location suggests a particular activity. The most common activity at a grocery store is "consumer purchase." At "transportation," it's "traveling." Workplace is "work;" Gym is "sports;" other store, "consumer purchase;" and bank, "professional services." "Unspecified place" is mostly "sports" (coded as "Walking" below Tier 1), and "restaurant/bar" is mostly "eating." These places are either designed for commercial transactions, employment, or transportation.

Interestingly, however, there is a class of locations that do not predict activity well, and at which multiple activities often occur. Generally speaking, these locations are public facilities (school, library, post offices, and outdoors), churches, and homes. An activity-inference system should

not expect to be able to accurately infer activity in these places using building-level location alone. Figure 4 shows that this is true even for the coarse, 18-activity Tier 1 classification.

This result raises a question: Is poor prediction of these activities an artifact of the coarse location taxonomy? Would a finer location classification—say, that included indoors—predict activity better? Or is location not enough to predict activity in these cases? ATUS does not contain the data to make a quantitative conclusion, but it does offer some insights through the distribution of activities at these locations.

For example, Table 2 shows the most likely Tier 3 activities when the respondent was at their home and when at someone else's home. Although the locations of the listed activities are not known, it is apparent that some of them happen in different locations (e.g., "Television and movies" and "Food and drink preparation") whereas others are more likely to overlap in even the most precisely measured location (e.g., "Reading for personal interest" and "Relaxing, thinking"). We can state that for homes, better location fidelity will help improve activity predictability, but not for all activities. A similar analysis could be performed for other locations.

As an aside, Table 2 also shows how activities are affected not only by the type of building, but also by a person's role within it. We can see, for example, that sleeping is much more common in one's own home than in another's, where socializing is the most common activity. Again, the activities on this list are not surprising, but the list is useful for a system designer who might otherwise forget an important activity.

This kind of data is especially time-consuming to collect through instrumentation, because social events may be less frequent than other activities. Furthermore, getting a representative sample of activities may depend on the individuals present, their social relationships, and the occasion. It is much faster to reuse a large, already-collected data set.

Finally, observe that Figure 4 also shows that at some locations there are great differences in predictability between Tiers. This provides another perspective on the effect seen above in Figure 3, namely that some locations strongly predict a single, popular activity at a lower-numbered Tier, but weakly select among a higher-entropy set of activities at a higher-numbered Tier. These breakdowns are artifacts of the activity taxonomy; a different taxonomy, for example, might break down workplace activities, thereby reducing predictability at a higher Tier for "Respondent's workplace."

Activity Inference Accuracy for Different Activities

We study next the accuracy with which specific activities can be inferred. We use the same maximum-likelihood classifiers based on contextual variables as described above. For a given activity (say, personal care), we measure the fraction of instances of that activity that are correctly classified (the Recall), and the fraction of activities given that label that are labeled correctly (the Precision). Naturally, high Recall and Precision are desirable. Our measures again use tenfold cross-validation.

Other researchers measure the area under an ROC curve to compare activities [5,12]. We only report the Recall and Precision for the single point on the ROC curve determined by the maximum-likelihood classifier. Since this classifier outputs only the most likely activity, there is no meaningful parameterization for producing a ROC curve. The results for Tier 1 are given in Table 3. To facilitate calibration and interpretation of our results, the table also shows the average time spent per day on each activity.

The results show large differences in predictability among the activities. Consider the classifier based on location alone. Table 3 reveals that while some activities (e.g., "Personal care") are predicted very accurately from location, the classifier fails to predict other activities (e.g. "Household activities"). This may seem counterintuitive since most household activities happen in the home. But the classifier does not detect those activities because, knowing only location, it must guess "Sleeping," as sleeping is a safer guess than household activities. A classifier that uses both Hour of day and Location can better identify household activities, but only unreliably, as there are other activities during the day that are often more likely.

Different activities are improved by different variables. For example, educational activities cannot be predicted well based only on hour of day. But if we add the age group of the respondent as a contextual variable, "education" can be predicted with 39.0% Recall and 32.9% Precision.

Note that adding a feature can reduce the Recall and/or Precision for some activities. For example, the recall for "Personal care" is worse using Location & Hour of day than when using just Location. This happens because the overall predictability of other activities improves (such as "Socializing, relaxing, and leisure" in this case.) In other words, the time of an activity allows the classifier to correctly classify activities such as "Socializing" in situations that it previously misclassified as "Personal care." But these activities are not carved out perfectly: some "Personal care" activities are erroneously categorized (thus lowering Recall). The precision of "Personal care" is better with Location & Hour of day than with Location alone. In other cases, adding a feature may reduce precision.

Some activities are always predicted poorly. Telephone calls are impossible to predict from any of the features in the time-use study. Fortunately, they are easy to detect in an activity-inference system that has access to the user's cellphone. Other hard-to-predict activities, such as caring for others, volunteer activities, household services (supervising others' work at home), and government and civic services may not be so easy to predict.

Activity	Avg hh:mm per day	Accuracy of Tier 1 Activity Classification (in percent)											
		Hour of day		Hour of day & Age Group		Prev Acty & Hour of day		Location		Location & Hour of day		Location & Prev Acty & Hour of day	
		Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre
Personal care (inc. sleep)	9:23	87.2	74.5	85.8	77.5	88.2	84.0	100	56.8	89.0	82.3	89.1	86.6
Socializing, relaxing, and leisure	4:31	52.1	39.3	57.5	39.8	61.0	41.6	14.3	47.8	71.2	47.1	73.8	49.0
Work & work-related activities	3:27	61.3	30.2	68.5	36.7	73.5	44.8	87.9	93.7	87.9	93.6	87.6	95.0
Household activities	1:49	-	-	1.0	21.8	11.6	28.6	0.1	52.3	14.2	31.5	22.2	34.0
Telephone calls	1:14	-	-	-	-	-	-	-	-	-	-	0.0	0.0
Eating and drinking	1:06	-	-	-	-	8.6	51.9	20.2	71.6	19.5	73.2	29.1	65.2
Education	0:27	-	-	39.0	32.9	4.4	44.1	64.7	69.9	64.3	69.6	59.8	72.2
Caring for household members	0:26	-	-	-	-	1.4	27.7	-	-	0.0	1.5	2.4	35.7
Consumer purchases	0:24	-	-	-	-	-	-	89.2	88.9	89.2	88.9	88.2	89.9
Sports, exercise, and recreation	0:18	-	-	-	-	0.4	16.8	36.4	48.5	35.5	47.2	31.3	48.8
Traveling	0:11	-	-	-	-	31.8	49.4	96.0	97.0	95.9	96.8	94.8	96.1
Caring for non household members	0:08	-	-	-	-	0.1	11.9	-	-	0.0	24.6	3.8	25.4
Religious / spiritual activities	0:07	-	-	-	-	2.6	17.7	80.7	56.4	80.0	56.4	75.8	60.1
Volunteer activities	0:07	-	-	-	-	0.8	15.7	-	-	0.0	2.9	2.9	24.7
Prof & personal care services	0:05	-	-	-	-	3.5	34.6	5.1	85.7	24.5	22.0	40.0	32.5
Household services	0:01	-	-	-	-	0.0	0.0	-	-	-	-	0.0	0.0
Government and civic services	0:00	-	-	-	-	3.0	22.3	-	-	-	-	4.0	67.9

Table 3: Precision and Recall, by Tier 1 activity, for classifiers based on various contextual variables. In cells containing a hyphen, the activity is never predicted because all combinations of input variables favor other activities. Dark shading indicates a classifier with an F-measure in the top 25% percentile of all non-degenerate classifiers. Light shading indicates a classifier in the top half.

Finally, we would like to emphasize that all these activity inference figures are calculated from general population statistics. There is no learning of any particular user's patterns. When such mechanisms are combined with techniques that incorporate time-use data, overall accuracies should be better.

SIMULTANEOUS ACTIVITIES

Some studies report "secondary activities" that happen in parallel with the primary activity. ATUS does not, because of the difficulty of collecting and coding these data. Interviewers must ask many more questions and often code different stop and start times for primary and secondary activities [21].

To investigate secondary activities, we studied the 1985 American's Use of Time study (AUT) (2923 participants). AUT codes activities differently from ATUS, using a flat variable with 92 codes. The main code is supplemented by a secondary code if a secondary activity was performed at the same time. 45% of all activities were accompanied by a secondary activity. On a time-weighted basis, 31% of the time there was a secondary activity.

The most common activities that were either accompanied by a secondary activity or were themselves secondary activities to a primary activity were 1) "conversation, phone, texting," 2) "watch television, video", 3) "wash, dress, personal care", 4) "other meals & snacks", and 5) "listen to radio." Note that "other meals & snacks" includes all eating not at work or in a restaurant. Very often, these activities were multitasked with each other.

This result confirms observations from other ubicomp studies. For example, Logan, et al. [12] noted that their own intensive study uncovered the tendency of participants to overlap eating with other activities, and to perform eating in a variety of places. AUT also shows that 51.0% of all "other meals & snacks" activities either occurred with a secondary activity or were themselves secondary activities. AUT cannot, however, show how eating is spread out over places because it does not have fine-grain location information. However, another time-use study, the 1992

National Human Activity Pattern Survey (NHAPS), does. For the NHAPS activity "Eat", the top five locations are "Home, Kitchen" (47%), "Home, Living Room, Family Room, Den" (14%), "Home, dining room" (12%), "Indoors, Restaurant" (11%), and "Home, bedroom" (2%).

Even though we must draw from several time-use studies to perform this analysis, the results confirm Logan et al.'s observations, and even go further, ranking the frequency of eating out within the frequency of eating in different rooms. The original observation of the nature of eating emerged from weeks of data collected about a single individual. Although our analysis is biased by self-reporting issues and draws on old data from multiple studies, different years, and different activity codes, it took only a couple hours to perform, and it does aggregate the activity patterns of thousands of people (2,923 in AUT, 7,513 in NHAPS). We are not arguing that analyses of time-use studies will replace original research, but rather that they offer a different perspective that is inexpensive and often worthwhile to explore.

RESEARCH QUESTIONS

Although the data from time-use surveys can be immediately useful for activity-inference systems, we see several interesting research questions that, if answered, could make significant new contributions to ubicomp activity-inference.

How much do time-use activity and location taxonomies vary? There are differences among the classifications used by time-use studies; to what extent are these differences subjective? What aspects of activity and location are universally or near-universally agreed on? How much do classification differences contribute to inaccuracies in activity prediction?

What issues arise when adopting an activity taxonomy for a ubicomp application? A few ubicomp systems [6,20] have already adopted classifications used in time-use studies such as healthcare's Activities of Daily Living [7], or the Multinational Time-Use Study activity classification [25]. Using a standard classification is beneficial because it is less likely to omit important activities, and more likely to interoperate with other systems if they adopt the same standard. However, our initial efforts in using time-use data uncovered significant activity mismatch challenges in adapting a time-use taxonomy to a mobile recommender system (codenamed Magitti [2]). How serious are these issues, and how can they be addressed?

What methodologies used by time-use studies can be adopted in ubicomp systems? Because of the granularity gap and domain specific difference described earlier in this paper, time-use data may not always be adequate for certain kinds of activity inference. For example, a study might compare the differences in activity patterns after the introduction of a new technology. In this case, although time-use study data itself may not be useful, the practices adopted by time-use studies (such as recruitment, collecting and coding data, and treatment of simultaneous activities) may help researchers avoid mistakes that would reduce the quality of their results.

How can ubicomp contribute to time-use study research? Ultimately, ubiquitous computing may benefit time-use studies more than time-use studies may benefit ubicomp. Because time-use data is so critical for sociology, public health, economics, and media assessment, automated techniques of collecting the kind of data that time-use studies have traditionally provided would give these researchers tools to make more accurate and precise conclusions, to answer different questions, and to reduce their costs.

CONCLUSION

This paper has studied the applicability of time-use study data for ubicomp activity-inference systems. We argue that these data are useful because they enable cheap and comprehensive activity classifiers, and we analyze the accuracy of these activity classifiers. We find that location is the most useful classifier feature, and that when combined with time of day, activity can be predicted with up to about 70% accuracy, depending on the activity taxonomy's granularity. We further show how time-use studies provide a less expensive path for answering activity-related research questions, such as the amount and nature of simultaneously-performed activities. We also describe several other uses for time-use data, and several research questions that would make this data even more valuable for the ubicomp community.

The ubicomp and time-use research communities have barely interacted until now, yet it seems inevitable that they will influence each other more strongly in the near future. Already we have seen the application of survey data to transportation. Health-care applications are increasingly prominent in ubicomp, and have a long history in time-use research. Finally, ubicomp is becoming increasingly data-driven and activity-oriented, whereas time-use research is becoming increasingly interested in how new technology might assist data collection [21,26].

Unfortunately, at the time of this paper's writing, funding for ATUS, the American Time Use Survey, had been eliminated from the proposed US 2009 federal budget. Given the value to ubicomp applications that we have found for time-use data, and the rarity of unrestricted sources of recent episode data, it is our hope that this decision will be reversed, and that this study will continue to supply researchers with fresh data for many years to come.

ACKNOWLEDGEMENTS

We thank Dai Nippon Printing Co., Ltd. "Media Technology Research Center" and "Corporate R&D Division" for their sponsorship during the early stages of this research, Bo Begole and Norman Su for feedback on drafts of this paper, and the anonymous reviewers and our

shepherd John Krumm for their constructive comments and suggestions.

REFERENCES

[1] Bao and Intille, "Activity Recognition from User-Annotated Acceleration Data," *Pervasive Computing*, 2004.

[2] V. Bellotti, J.B. Begole, E.H. Chi, N. Ducheneaut, J. Fang, E. Isaacs, et al., "Activity-Based Serendipitous Recommendations with the Magitti Mobile Leisure Guide," *Proc. of SIGCHI conference on Human Factors in Computing Systems*, pp. 1157-1166, 2008.

[3] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*, Harper & Row USA, 1990.

[4] K. Fisher, J. Tucker, and A. Jahandar, *Technical Details of Time Use Studies*, Institute for Social and Economic Research, University of Essex, ; http://www.timeuse.org/mtus/.

[5] J. Fogarty, R.S. Baker, and S.E. Hudson, "Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction," *Proceedings of Graphics Interface 2005*, 2005, pp. 129-136.

[6] S. Intille, E.M. Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, et al., "Tools for Studying Behavior and Technology in Natural Settings," *Ubiquitous Computing*, 2003.

[7] S. Katz, A.B. Ford, R.W. Moskowitz, B.A. Jackson, M.W. Jaffe, and K.L. White, "Studies of illness in the aged--The index of ADL: A standardized measure of biological and psychosocial functions," *JAMA 185:914-919 (1963).*

[8] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, et al., "The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants," *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 11, 2001.

[9] J. Krumm and E. Horvitz, "Predestination: Inferring Destinations from Partial Trajectories," *UbiComp*, 2006, pp. 243–260.

[10] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, "A Hybrid Discriminative/Generative Approach for Modeling Human Activities," *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2005)*, 2005.

[11] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition using relational Markov networks," *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[12] B. Logan, J. Healey, M. Philipose, E.M. Tapia, and S. Intille, "A Long-Term Evaluation of Sensing Modalities for Activity Recognition." *Proc. of Ubicomp 2007.*

[13] Lukowicz, Ward, Junker, Stäger, Tröster, Atrash, et al., "Recognizing Workshop Activity Using Body Worn Microphones and Accelerometers," *Pervasive Computing*, 2004.

[14] M.G. McNally, "The activity-based approach," *Handbook of Transport Modelling*, 2000, pp. 53-69.

[15] W.M. Michelson, *Time Use: Expanding Explanation in the Social Sciences*, Paradigm Publishers, 2005.

[16] B. Morgan, "Learning Commonsense Human-language Descriptions from Temporal and Spatial Sensor-network Data," Massachusetts Institute of Technology, 2006.

[17] S.N. Patel, J.A. Kientz, G.R. Hayes, S. Bhat, and G.D. Abowd, "Farther than you may think: An empirical investigation of the proximity of users to their mobile phones," *Proceedings of Ubicomp*, 2006.

[18] W.E. Pentland, *Time Use Research in the Social Sciences*, Kluwer Academic Publishers, 1999.

[19] W. Pentney, A. Popescu, S. Wang, and H. Kautz, "Sensor-Based Understanding of Daily Life via Large-Scale Use of Common Sense," *Proceedings of AAAI*, 2006.

[20] M. Philipose, K.P. Fishkin, M. Perkowitz, D.J. Patterson, D. Fox, H. Kautz, et al., "Inferring Activities from Interactions with Objects," *IEEE Pervasive Computing*, vol. 3, 2004, pp. 50-57.

[21] J.P. Robinson, "The Time-Diary Method: Structure and Uses," *Time Use Research in the Social Sciences*, 1999.

[22] K.J. Shelley, "Developing the American Time Use Survey Activity Classification System," *Monthly Labor Review*, vol. 128, 2005, pp. 3–15.

[23] A. Shon, "Methodological and Operational Dimensions on Time-Use Survey in the Republic of Korea," *International Seminar on Time Use Studies*, 1999.

[24] P. Singh and W. Williams, "LifeNet: a propositional model of ordinary human activity," *Proceedings of the Workshop on Distributed and Collaborative Knowledge Capture (DC-KCAP)*, 2003.

[25] A. Szalai, *The use of time: daily activities of urban and suburban populations in twelve countries*, Mouton, 1972.

[26] J. Wolf, R. Guensler, and W. Bachman, "Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data," *Transportation Research Record*, vol. 1768, 2001, pp. 125-134.