UCI-ITS-AS-WP-00-2

A Latent Factor Model of Observed Activities

UCI-ITS-WP-00-2

James E. Marca
Michael G. McNally
Craig R. Rindt

Department of Civil Engineering and
Institute of Transportation Studies
University of California, Irvine
jmarca@uci.edu, mmcnally@uci.edu, crindt@uci.edu

December 2000

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

A latent factor model of observed activities

James E. Marca

Dr. Michael G. McNally

Craig R. Rindt

Author address:

The Institute of Transportation Studies, University of California, Irvine, CA 92697-3600, USA

E-mail address: jmarca@translab.its.uci.edu

ABSTRACT. This paper examines the problem of describing an activity in a concise, usable way. An activity is defined by a vector of observed attributes. Including more observed attributes improves the explanatory power and theoretical completeness of any model of activities, but simultaneously leads to a combinatorial explosion when considering questions about choosing between activities, or sequences of activities—questions which arise in simulation applications. This paper first builds a description of individual activities using a vector of observed attributes. Then latent variable analysis is used to reduce this vector to just two latent variables, which together explain most of the variation in the original variables.

1. Introduction

Activity analysis is rooted in the idea that people travel in order to get from activity to activity. The locations at either end of a trip are now seen as less important to the understanding of travel than the activities being performed at those locations. However, describing activities is not as easy as describing trip ends. Activities are described by a large number of characteristics. Activities have names, occur over some length of time, fulfill obligations, reinforce social constructs, and so on. Depending upon one's theoretical point of view, any or all of these attributes could be an important contributor to the definition of an activity, which in turn leads to the trip.

For example, say there is a trip from one zone to another. A survey might reveal that the trip in question occurred because the person wanted to eat a meal. The destination becomes a place to perform the meal activity. But there are different meal activities, ranging from a quick bite to eat to a relaxed four-course meal. One might inquire about such details, but the specific details of a meal pertain only to meals, not to other categories of activities, such as work or entertainment. A more general approach is to inquire about generic attributes of the particular meal activity in question, such as the length of time, the amount of money spent, and the participation of others. These facts apply to any named activity, and so they can be collected in a survey without too much trouble.

The difficulty comes in trying to explain the observations, and then make use of them. With a detailed survey, each activity is ultimately defined by a vector of observations. Even with a small number of variables taking on a small number of discrete values, the combinations of those values produce a very large potential activity space. The exploding number of alternatives is compounded even more by any analysis of sequences of activities.

This paper presents a way to describe activities in a concise, usable way, while still preserving most of the explanatory power of the complicated descriptors of an activity. We use latent variables to capture the variations in the observed features of activities. Section 2 introduces the data we used for this project, and discusses some of the issues in applying latent variable analysis to activity data. Section 3 presents the mechanics of latent variable models, and the results of the estimation procedure. Section 4 interprets the estimated model, and section 5 discusses future extensions and applications.

2. A latent variable model of an activity

A latent variable is defined as a variable that cannot be directly measured. For example, it is difficult to measure whether someone is "in a hurry." However, the latent variable "in a hurry" is directly responsible for observable attributes such as shorter travel times, higher peak driving speeds, and shorter activity durations. This paper's goal is to estimate latent factors that are responsible for the observed variations in activities.

Much work in activity analysis has focused on the constraints placed on an activity. Even at the beginning of activity analysis, Hägerstrand (1970) focused his comments on the description of constraints that limited what was possible for a person to do. More recently, Recker (1995) formulates an optimization framework which attempts to solve the best path through time and space, given an input of activities that must be performed.

In addition to constraints, there must be some positive, motivational force that causes activities to happen. The motivation is usually considered to be utility maximization. Examples of this approach are numerous, since the point of view is similar to the mode choice literature. For example, the STARCHILD model (Recker, M^{c}Nally and Root, 1986) and its recent extensions (M^{c}Nally, 1997; M^{c}Nally, 1998) simulate the generation of activity sequences by assuming that each individual's actions belong to one or more classes of activity patterns. Similarly, Pas (1988) develops a typology of multi-day activity patterns, and then associates different types of weekly patterns with different socio-demographic variables. Ben-Akiva and Bowman (1995) develop a model of activity choice based on the hypothesis that travelers optimize the simultaneous choice of the features a linked chain of activities. Vaughn, Speckman and Pas (1997) take a slightly different approach, matching up simulated individuals with actual, observed activities, leaving the motivating factors to the real people.

The process of engaging in an activity is the result of a dynamic balancing between the internal motivations of the person, and the external possibilities of the environment. An activity that has been observed exists. The reasons for its existence are a combination of the individual's motivation to perform the act, and the environment's intrinsic capacity to allow the act to happen. High motivation can overcome a certain amount of environmental impediment, while low motivation is typically only paired with activities that are very easily done in an environment. This paper recognizes the balance between internal motivation and external potential by proposing a two-factor latent variable model of activities. An observed activity is the result of some level of motivation on the one hand, and some degree of environmental potential on the other. Again, it is important to stress that only observed activities are modeled, and so none of the constraints are insurmountable, and none of the lower levels of apathy result in inactivity.

With these two ideas for latent factors in mind, the next step is to examine observed activities. The data used for this research are the responses to the Portland Oregon two-day activity diary survey (Portland METRO, 1994). This survey contains 129,188 separate activities, performed by 10,048 people belonging to 4,451 households. Each activity has been measured on a number of different dimensions in this survey. In this research, our goal is to focus on each activity in isolation, as the product of the motivation and constraint factors. Obviously, there are many repeated measures of activities over people and households, and so a thorough treatment would control for

these effects. We will leave these more complicated treatments for future research once we have proven the basic concept.

Since we are isolating individual activities, we chose to model only those features which pertain to a single activity. In other words, we did not model the sequencing of activities, or the number of activities. We also eliminated time of day as a descriptive feature. If we had included time of day, then future sequencing and ordering analyses would be complicated by the existence of a time variable in the description of a particular activity. We did explore keeping time of day in as a general category (such as morning, evening, and night), but the final results were not different enough to justify keeping the variable in the analysis. After weeding out any observed feature relating to sequencing and timing, the remaining variables are:

- the duration of an activity,
- the duration of the trip to get to the activity,
- the name of the activity, and
- the location of the activity.

2.1. Activity duration. Activity duration is calculated by subtracting the reported activity starting time from the ending time. This is ostensibly a continuous variable, and one would expect it to be correlated with the concepts of motivation and limitation. However, examining the reported durations reveals a surprising degree of choppiness. Figure 1 shows the spikes in frequency, despite the overall decaying shape of the activity duration. These spikes occur primarily every half-hour, indicating that respondents were rounding reported start and end times to half-hour intervals.

The spiky nature of the actual duration value makes it quite difficult to use as is in an estimation procedure. Therefore we decided to discretize the duration into 5 different intervals, defined by:

$$\text{act duration category} = \begin{cases} 1 & \text{act duration} \leq 0:30; \\ 2 & 0:30 < \text{act duration} \leq 1:15; \\ 3 & 1:15 < \text{act duration} \leq 2:30; \\ 4 & 2:30 < \text{act duration} \leq 4:00; \\ 5 & 4:00 < \text{act duration}. \end{cases} \tag{1}$$

The selection of the intervals was somewhat arbitrary, although they were chosen to result in reasonably sized groupings. Future activity surveys, with advances in data collection devices, will be able to use more exact estimates of duration. This will remove the need to categorize the data.

2.2. Rounded versus exact activity duration. As durations get shorter, the incidence of rounding off reported durations picks up for 15, 10 and even 5 minute intervals. A histogram of the reported minutes value of the duration sheds more light on the rounding tendencies. Figure 2 shows the steep peaks at zero and thirty, indicating that most durations were rounded off to these values. Next in importance are the 15 and 45 minutes peaks, followed by the tens and the fives. This result is in accordance with the findings of Murakami and Wagner (1999). They reported that unbiased global positioning system measurements showed that trip departure time and trip duration were evenly distributed over the 60 minutes, whereas respondents routinely rounded off

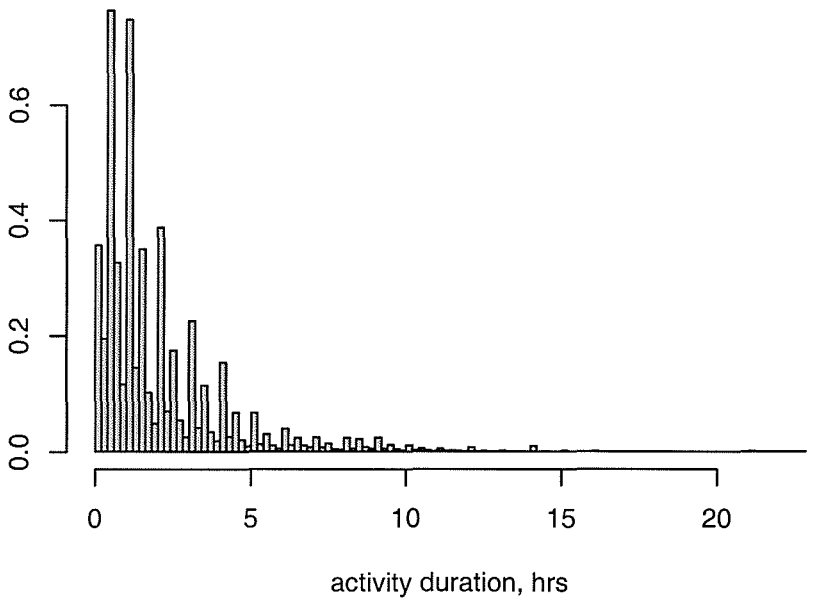


FIGURE 1. Relative frequency (estimate of the pdf) of activity duration (area of all bins sums to one). The spikes occur on the hour, half-hour, and to a lesser extent on the quarter-hour, indicating rounding of reported times.

these values. Although not shown, the start and end times show very similar peaking behavior, with most activities starting and/or ending on the hour or half hour.

The presence of a dust of unrounded values at the bottom of figure 2 indicates that for certain kinds of activities, people do not round off their reported activity durations. We suppose that the that activities reported with more *exact* durations are for some reason memorable to the respondent. Further, we suppose that the quality that makes the activity memorable is not measured by the other activity features. In other words, these exact duration activities, while small in number, represent a unique type of activity simply due to the fact that the person decided to report the exact duration. In all other ways they may look identical to other acts, but since we have this extra information, we will incorporate it into our analysis. The rounding attribute is defined as:

$$\text{(2)}\quad \text{round} = \begin{cases} 1 & \text{if duration} \leq 1 \text{ hr \& reported minutes} = \text{multiple of 5 minutes;} \\ 1 & \text{if duration} > 1 \text{ hr \& reported minutes} = \text{multiple of 15 minutes;} \\ 2 & \text{otherwise.} \end{cases}$$

A counter hypothesis is that the unrounded dust at the bottom of figure 2 is due to just a few survey respondents who were very conscientious about their timing of activities. To check this, figure 3 was created using the definition of equation 2. It shows a histogram (an estimate of the pdf) of the fraction of each person's reported activities that were reported in exact terms. The spike at 0 captures the large group

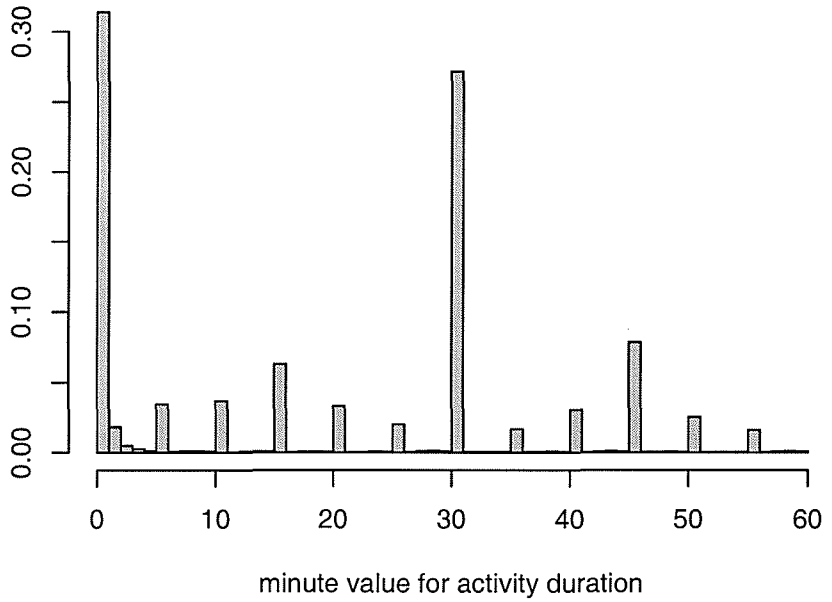


FIGURE 2. Relative frequency (estimate of the pdf) of reported duration minutes. Most reported durations were multiples of either an hour or half-hour, as is shown by the peaks at 0 minutes and 30 minutes.

of 3,010 people (30%) who only report rounded activity durations. At the other end of the scale, the low value at 1 captures the 11 people (0.1%) who reported *all* of their activities with exact durations. In between are those people who reported some of their activities with exact durations. Clearly for most persons reporting one or more exact durations, only a small fraction of a person's activities were reported in unrounded minutes. This indicates that the exact durations are not due to a few people who kept really thorough activity diaries.

2.3. Trip duration. Similar considerations apply to the trip duration. We experimented with building a rounding flag for trip duration as well, but since trips were generally much shorter than activity durations, the definition of what "rounding" meant was not as clearly defined. In addition, the results showed that the two factors captured much of the same tendencies. However, trip duration was also choppy, and so was broken down into 4 categories as follows:

$$\text{trip duration category} = \begin{cases} \text{no trip made;} \\ 0:00 < \text{trip duration} \leq 0:10; \\ 0:10 < \text{trip duration} \leq 0:20; \\ 0:20 < \text{trip duration.} \end{cases} \tag{3}$$

2.4. Activity name. Another important dimension of an activity is its name. We followed the general combinations used by Golob and M^{c}Nally (1997), with two additional

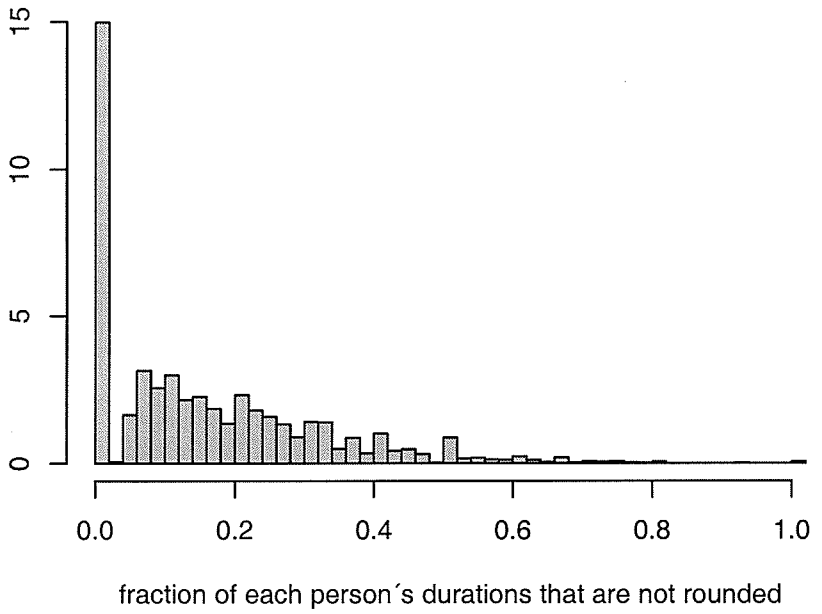


FIGURE 3. Relative frequency (estimate of the pdf) of the fraction of a person's activity durations that were *not* rounded (as defined by equation 2). The unrounded durations are typically just a small fraction of the total number of activities a person reports.

categories produced by separating out *meals* from *maintenance*, and *amusements—at-home* from *discretionary*. This was done due to the large size of these two categories relative to others. Our five activity name categories are:

discretionary: combining *visiting, casual entertaining, formal entertaining, culture, civic, volunteer work, amusements—out-of-home, hobbies, exercise/athletics, rest and relaxation, spectator athletic events, incidental trip*, and *tag-along trip*;
meals
work: combining *work, work related,* and *volunteer work*;
maintenance: combining *shopping—general, shopping—major, personal services, medical care, professional services, household or personal business, household maintenance, household obligations, pick-up/drop-off passengers, school*, and *religious/civil services*; and
amusements—at-home

2.5. Activity location. The final feature of an activity to consider is its location. The raw description of a location was discarded as a variable because of its *sparse* categorical nature. While the Portland survey contains quite a large number of responses, the positional spread of those responses barely leave a mark on the city of Portland. There are statistical techniques for exploring point data spread over a plane, such as kriging and spatial interpolation (Ripley, 1981; Venables and Ripley, 1999), but the sparseness of the data, combined with a lack of knowledge about the exact locations of activities and the paths between them (as one would get from a global positioning

system), led to the decision to drop the raw positional information from the analysis. Instead, we used a binary variable indicating whether the activity occurred in or out of the home to represent the location features of an activity. This variable captures activities that are performed exclusively in or out of the home, as well as those that can be performed in either place. Future surveys which include GIS path data may be able to treat location in a more complete manner.

3. Estimating a two-factor latent model

This section will describe the estimation of latent variable models. The interested reader is referred to Bartholomew and Knott (1999), for more detail on the methods used. Structural equations modeling (Bollen, 1989) is a closely related field to latent variable modeling, the difference being that the structural effects are *not* specified in a latent variable model.

Latent variables are defined as variables that cannot be observed, but which govern the attributes of variables that can be observed. The estimation of a latent variable assumes that the latent variable explains *all* of the variation in the observed variables. Following Bartholomew and Knott (1999), the vector $\mathbf{x}$ of randomly distributed, observed variables are dependent upon a vector $\mathbf{y}$ of randomly distributed latent variables. The probability density function of $\mathbf{x}$ is given by

$$f(\mathbf{x}) = \int h(\mathbf{y}) g(\mathbf{x}|\mathbf{y}) d\mathbf{y}, \tag{4}$$

where the integral is over the full range of $\mathbf{y}$, $h(\mathbf{y})$ is the prior distribution of $\mathbf{y}$, and $g(\mathbf{x}|\mathbf{y})$ is the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$.

Since $\mathbf{y}$ cannot be known or observed, in practice one tries to specify a small number (q) of independent latent variables, such that all of the $\mathbf{x}$s are uncorrelated for a given $\mathbf{y}$. This implies that the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is composed of the product of the independent conditional distributions of the components of $\mathbf{x}$, or, for p different observable variables,

$$g(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{p} g_i(x_i|\mathbf{y}). \tag{5}$$

Substituting into equation 4 gives

$$f(\mathbf{x}) = \int h(\mathbf{y}) \prod_{i=1}^{p} g_i(x_i|\mathbf{y}) d\mathbf{y}. \tag{6}$$

Bartholomew and Knott (1999) show that the choice of the prior distribution of the latent variables, $h(\mathbf{y})$, is more or less arbitrary, and that one can safely assume a normal distribution with zero mean and standard deviation of one (or the identity matrix). Bartholomew and Knott (1999) propose that the $g_i(x_i|\mathbf{y})$ distributions fall into the one-parameter exponential family, or

$$g_i(x_i|\theta_i) = F_i(x_i) G_i(\theta_i) \exp(\theta_i u_i(x_i)), \qquad (i = 1, 2, \ldots, p). \tag{7}$$

If one assumes that θ_i for each of the p observed variables is a linear function of the q latent variables, then one can form the so-called General Linear Latent Variable Model (GLLVM).

$$\theta_i = a_{i0} + a_{i1}y_1 + a_{i2}y_2 + \ldots + a_{iq}y_q, \qquad (i = 1, 2, \ldots, p). \tag{8}$$

As was discussed above, the data used for this analysis are categorical, rather than continuous variables. For p observed categorical variables, in which variable i can take on c_i categories with each category being indexed by s, one can define $X_1, \ldots, X_p$ as p polytomous variables having c_i categories, $i = 1 \ldots p$, such that

$$X_{i(s)} = \begin{cases} 1 & \text{if the response falls in category } s, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The conditional distribution $g_i(x_i|\theta_i)$ from equation 7 becomes a response function conditional on the latent variable vector $\mathbf{y}$. This response function is defined as $\pi_{i(s)}(\mathbf{y})$, or

$$\Pr[X_{i(s)} = 1|\mathbf{y}] = \pi_{i(s)}(\mathbf{y}) \tag{10}$$

In the case of a binary response function such as this, one can assume the convenient logit-type form. For a two-factor model there are two latent variables, y_1 and y_2. Assuming that the GLLVM of equation 8 holds in this case, then the probability is:

$$\pi_{i(s)}(\mathbf{y}) = \frac{\exp(a_{0i(s)} + a_{1i(s)}y_1 + a_{2i(s)}y_2)}{\sum_{r=1}^{c_i} \exp(a_{0i(r)} + a_{1i(r)}y_1 + a_{2i(r)}y_2)} \tag{11}$$

Put in words, given the $\mathbf{y}$ vector and having estimated the coefficients A which relate the observed values to the latent variables, each category s of variable i will have a non-zero probability $\pi_{i(s)}(\mathbf{y})$ of being 1—meaning the observation falls into that category. The categories for a variable are mutually exclusive, and so it is sufficient to estimate the model for all but one set of a_is, setting the coefficients corresponding to the first category of each variable to zero arbitrarily (see table 1).

From the Portland data set, 99,999 activities were used to estimate the model, and the remaining 29,189 were used as a holdout set for validation. The model estimation was performed using the program `latvpoly.exe`, available from the online notes to Bartholomew and Knott (1999). The estimation output is shown in table 1. The right hand column represents the probability of belonging to a particular category given that the two latent variables are zero. Since the latent variables are normally distributed with a mean of zero, this column represents the first order marginal probability of membership in each category based on the estimated A matrix.

Table 2 presents the first order marginal totals for each of the variables. The predicted values are the result of generating 29,189 values of Latent Factor 1 and Latent Factor 2, assuming the two factors are distributed $N(0, I)$. The estimated A matrix was used to transform these latent factors into probabilities, and then an actual category value was chosen for each variable via a random drawing. The holdout column consists of the 29,189 observations that were held out of the original estimation process. A χ^2 test was performed for each marginal, testing the hypothesis that the distributions were

Category	A(0,I,J)	A(1,I,J)	A(2,I,J)	median prob
act dur $\leq 0:30$	0	0	0	0.25
$0:30 <$ act dur $\leq 1:15$	0.25	-0.19	-0.55	0.33
$1:15 <$ act dur $\leq 2:30$	0.39	1.66	-2.48	0.38
$2:30 <$ act dur $\leq 4:00$	-2.09	4.03	-4.22	0.03
$4:00 <$ act dur	-3.27	5.14	-4.12	0.01
no trip	0	0	0	0.40
$0:00 <$ trip dur $\leq 0:10$	-0.50	1.44	1.66	0.25
$0:10 <$ trip dur $\leq 0:20$	-0.76	1.47	1.38	0.19
$0:20 <$ trip dur	-0.92	1.49	1.22	0.16
act duration not rounded	0	0	0	0.12
act duration rounded	1.99	-0.83	-0.26	0.88
Discretionary	0	0	0	0.34
Meal	-0.53	-2.04	0.39	0.20
Work	-3.13	2.82	-0.10	0.01
Maintenance	0.13	0.23	0.54	0.38
Amusements—at home	-1.54	-1.53	-2.49	0.07
At home	0	0	0	0.97
Not at home	-3.40	10.02	8.69	0.03
		% of G^2 explained		80.8854
		Loglikelihood value		-493631.42
		Likelihood ratio stat.		26088.303
		Degrees of freedom		299

TABLE 1. Estimated A matrix for a two-factor latent variable model

different. This hypothesis was rejected in all cases. The results of analyzing the higher order marginal totals are not shown, due to space limitations. In all cases, the χ^2 test showed that the observed and predicted distributions were not significantly different from each other.

4. Interpretation of latent factors

Rather than struggle with the 5 dimensional import of the different A matrix values of table 1, it is somewhat easier to examine the probability surfaces generated for each of the dimensions over the latent factor plane. These surfaces are generated by solving equation 11, and then plotting the probability of the most likely category for each latent factor pair. The results are plotted in figures 4 through 8. By examining these plots closely, one can build a conception of the impact of the latent factors on the different observed variables. The following subsections discuss each plot in turn.

4.1. Activity duration. Looking at figure 4, it is clear that activity duration generally increases as Factor 1 increases and as Factor 2 decreases. The exceptions to this are at the positive limits of the latent factor plane. On the right, large Factor 1 results in a high probability of a very long activity, no matter the value of Factor 2. Along the top, large values of Factor 2 result in a high probability of a very short activity, regardless of the value of Factor 1.

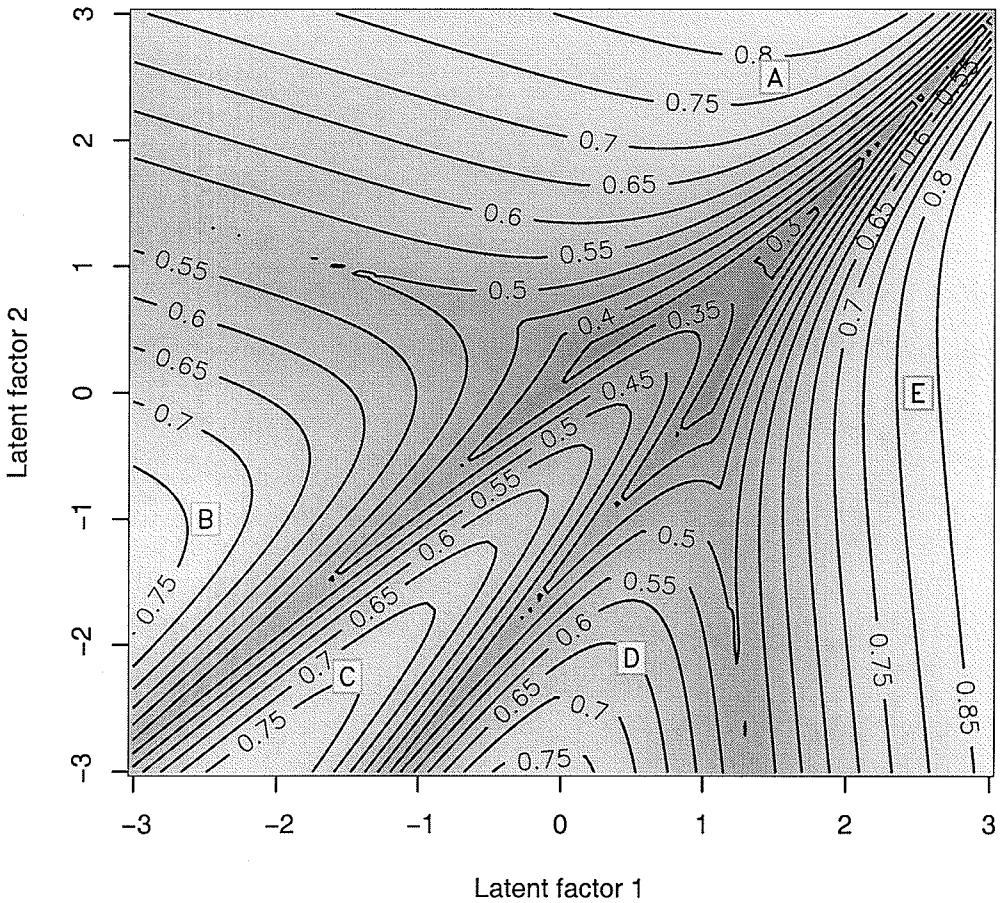
In general, as the latent factors move farther away from zero, the probability of being in one particular activity duration category tends to dominate the other four, with

Category	Holdout	Predicted
act dur $\leq 0:30$	7155	7307
$0:30 <$ act dur $\leq 1:15$	7762	7994
$1:15 <$ act dur $\leq 2:30$	6720	6886
$2:30 <$ act dur $\leq 4:00$	3864	3733
$4:00 <$ act dur	3688	3269
$\chi^2 = 72.1986$	$df = 4$	$p = 7.772 \times 10^{-15}$
no trip	13032	12961
$0:00 <$ trip dur $\leq 0:10$	6763	7228
$0:10 <$ trip dur $\leq 0:20$	5188	5015
$0:20 <$ trip dur	4206	3985
$\chi^2 = 48.528$	$df = 3$	$p = 1.644 \times 10^{-10}$
act duration not rounded	5984	4272
act duration rounded	23205	24917
$\chi^2 = 803.7107$	$df = 1$	$p \leq 2.2 \times 10^{-16}$
Discretionary	5107	5743
Meal	6923	6921
Work	3115	2911
Maintenance	7754	7943
Amusements—at home	6290	5671
$\chi^2 = 156.7917$	$df = 3$	$p \leq 2.2 \times 10^{-16}$
At home	17937	17639
Not at home	11252	11550
$\chi^2 = 12.7232$	$df = 1$	$p = 0.0003612$

TABLE 2. First order marginal totals, holdout versus predicted

the probability of the most likely category quickly rising above 50%. The exceptions to this are the valleys between the peaks, which consist of the five troughs between pairs of categories, and the general depression in the middle of the plot. The valleys actually take up very little area of the latent factor plane, due to the steep changes in probability. Therefore, the net effect of changes in the latent factors is similar to a membership function. Finally, it is worth noting that for the most likely values of the two latent variables—pairs of values falling in the circle between -1 and 1—no single duration category dominates.

As was noted earlier, activity duration is a continuous variable that has been divided into categories somewhat arbitrarily in this analysis. It would be better to leave duration a continuous variable, but this was made quite difficult given the severe rounding characteristics of the data. As one inspects figure 4, there is a rather smooth progression from short activities along the top, through longer categories of activities as one moves counter-clockwise, until one reaches the four or more hours category on the right. At the same time, there is an especially sharp division between the shortest category and the longest category. This characteristic indicates that while the model captures some of the continuous progression, the specification of the model would probably improve with accurate measurements of (continuous) activity duration. As GPS-based data collection tools become more common, the improved data will become available.

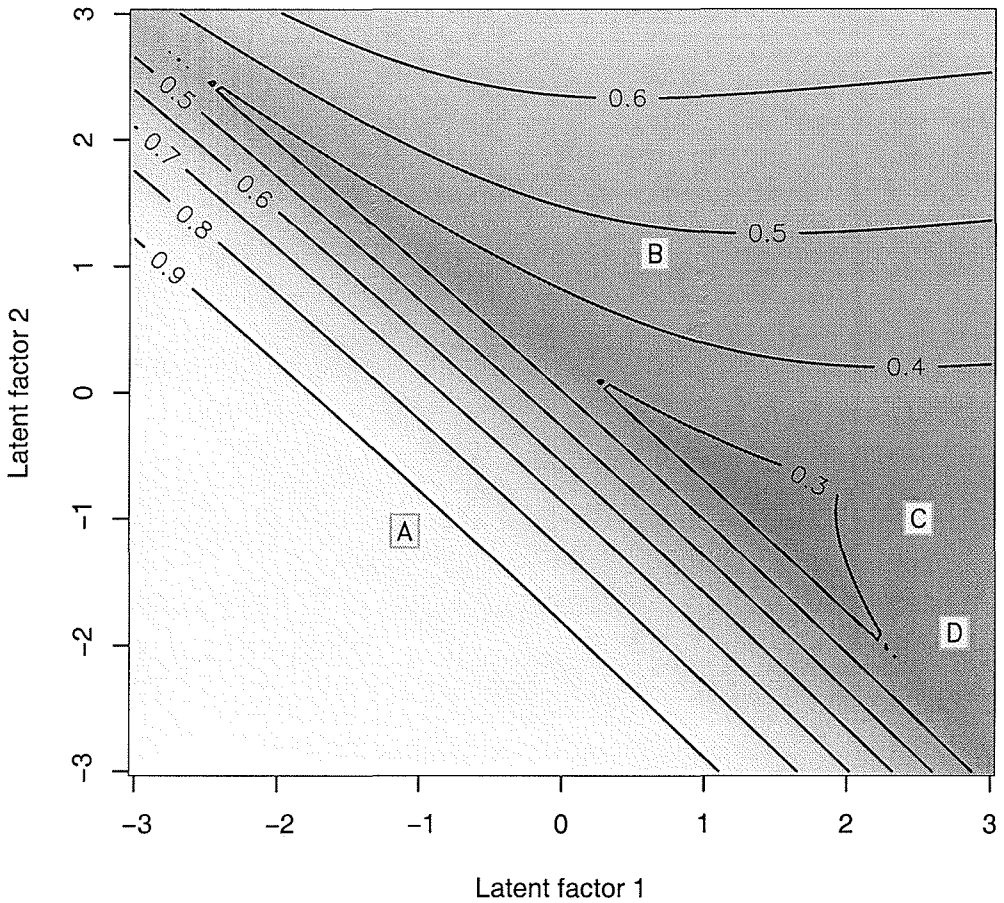


Region	Category
A	act duration $\leq 0:30$
B	$0:30 <$ act duration $\leq 1:15$
C	$1:15 <$ act duration $\leq 2:30$
D	$2:30 <$ act duration $\leq 4:00$
E	$4:00 <$ act duration

FIGURE 4. Maximal probability surface for activity duration categories. Latent factors are assumed normally distributed $N(\mathbf{0}, \mathbf{I})$.

4.2. Trip durations. The effect of the latent factors upon the trip duration is shown in figure 5. The impact of the numerous activities without preceding trips is quite strong, as would be expected from a category that contains more than 40% of the observations. Reflecting that fact, the lower left of the latent factor plane is given over primarily to activities that do not require a trip. In general, negative values of Factor 1 and Factor 2 will result in a trip not being taken to the activity.

Looking at the portion of the plane where trips are taken, increasing Factor 1 doesn't have much of an effect on moving from category B to a higher duration category. In contrast, increasing Factor 2 will tend to move towards a higher likelihood of engaging in a shorter duration trip to the activity. Unlike figure 4, the latent factor plane is not divided into distinct regions of dominance for each trip duration category. Positive Factor 2 will typically result in a short trip being taken. But zero or negative values of



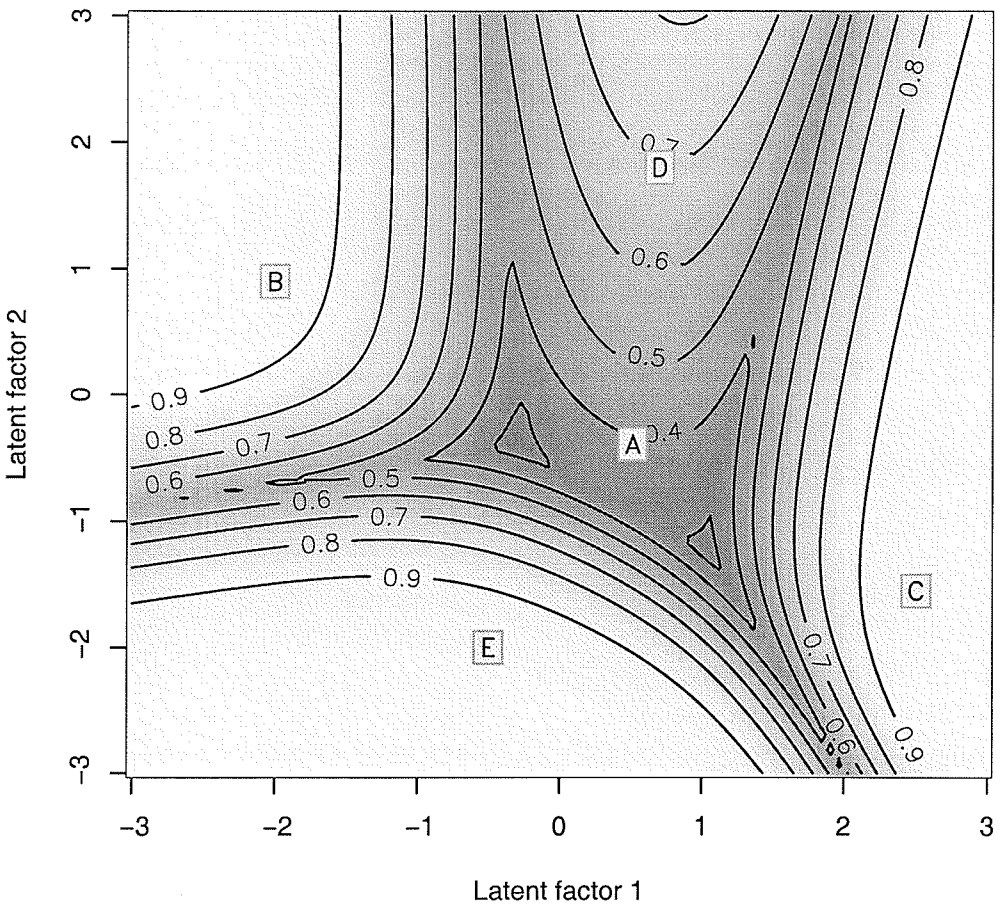
Region	Category
A	no trip made
B	$0:00 <$ trip duration $\leq 0:10$
C	$0:10 <$ trip duration $\leq 0:20$
D	$0:20 <$ trip duration

FIGURE 5. Maximal probability surface for trip duration categories. Latent factors are assumed normally distributed $N(\mathbf{0}, \mathbf{I})$.

Factor 2, combined with positive values of Factor 1, will result in nearly equal likelihood of belonging to any one of the three trip duration categories.

The lack of differentiation between categories suggests that trip duration is not handled very well by the latent variable model. This might be caused by two separate effects. First, one problem is probably related to combining trips and non-trips into a single variable. The latent effects that cause an individual to travel longer for an activity probably influence membership in all three of these categories in a consistent, linear fashion (as is assumed by the GLLVM of equation 8). But it is unlikely that the same latent effects contribute linearly to the transition between traveling and not traveling. It appears that the estimation process focused on discriminating between traveling and not traveling, rather than separating out the different categories of trip durations.

This leads to the second explanation for the vagueness of figure 5. The need to travel to an activity, as well as the extent of travel needed, is as much a result of the prior activity as it is the result of the current activity. The supposition being applied in



Region	Category
A	discretionary
B	meal
C	work
D	maintenance
E	amusements—at home

FIGURE 6. Maximal probability surface for activity names categories. Latent factors are assumed normally distributed $N(\mathbf{0}, \mathbf{I})$.

this paper is that two latent factors explain the activities alone—not the sequencing of activities, and not the characteristics of the person. But by including travel duration as a descriptive dimension of an activity, we have indirectly included some information about the preceding activity locations.

The intent behind including travel was to capture the idea that for some activities, people are willing to travel longer distances to get to a particular location at which to perform the activity. By superimposing figure 5 and figure 6, one can observe some of this effect. *Meals* activities are preceded by short trips, or no trip at all. Meals are a common part of life, and there are plenty of opportunities to eat all around us. *Work* activities, on the other hand, tend to require more travel time, as would be expected by the fact that people generally have exactly one place where work may be performed.

However, note that work also extends up to the short duration trip range. One can imagine leaving home on a long trip to *work* (positive Factor 1, negative Factor 2), then

taking a short trip from work to lunch (negative Factor 1, positive Factor 2), followed by a short return trip to work (both latent factors positive). Finally a day would end with a long trip back home to relax in front of the television before eating dinner (long trip to *recreation—in home*, then no trip to *meal*). Obviously, the sequencing of activities in this example has as much influence over the latent factor space as does the features of the activity. The same *work* activity can require a short or a long trip, depending upon where a person is relative to the work site.

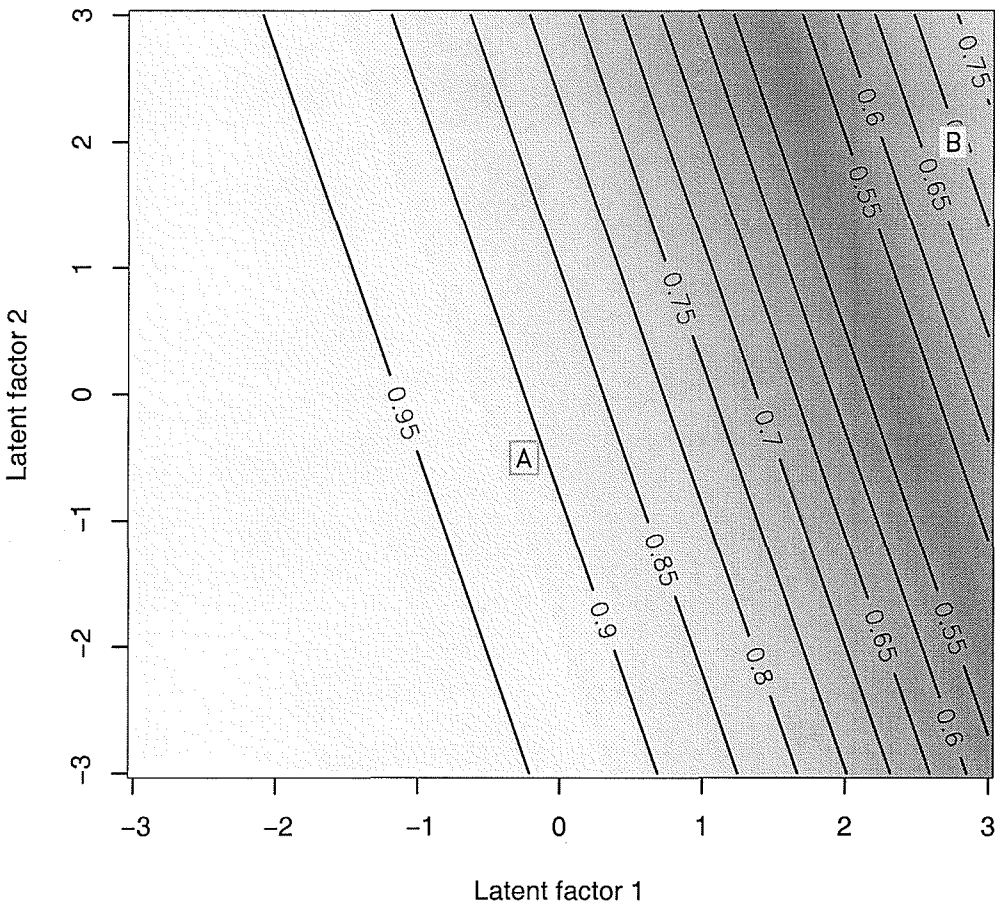
4.3. What's in a name? The next plot to analyze is figure 6, which shows the impact of different factor values on the name of the activity. There are very steep, sharp differences for four of the five categories of activities. The exception is for *discretionary* activities, which never form a dominant region in the latent factor plane. The label A in figure 6 has been placed at the maximum value of the probability of belonging to the *discretionary* category. This value is 37%, which is just less than the 39% probability of belonging to the *maintenance* category at the same point. The probability of engaging in a *discretionary* act decreases in all directions from point labeled A.

The relative positions of the distinct probability regions of *meals, work, maintenance,* and *amusements—in home* activities are likely due to their association with distinct features of other explanatory dimensions. For example, *meals* may be in the home or out, require a short trip or none at all, and rarely last longer than 2 hours, which places them in the upper left corner of the latent factor plane. In contrast, *discretionary* activities evidently do not have distinctive features in the other descriptive dimensions. Furthermore, identifying them by name also does not serve to differentiate them sufficiently, given the two latent factors. Therefore, they are placed towards the middle of the latent factor plane by the estimation procedure. The implication for future analyses is that the definition of discretionary trips should be reexamined, and perhaps combined with maintenance trips.

In general, as with activity duration, different values of the latent factors tend to locate in regions where a single activity name is dominant. Once again, the exception to this is towards the center of the latent factor plane, where 2 or 3 different activity names coexist in roughly equal proportions. On a final note, small positive values of Factor 1, combined with positive values of Factor 2 make one more likely to engage in either *discretionary* or *maintenance* activities. This is also the region where reporting exact values begins to increase, as is discussed in the next section.

4.4. Rounding off reported activity duration. An easy figure to interpret is figure 7. As Factor 1 goes from negative to positive, the chance that the respondent will report a more exact activity duration increases. Superimposing the previous plot of activity names (figure 6) on figure 7 shows that the likelihood of reporting exact values increases with the increasing likelihood of engaging in *maintenance* and work *activities*, and to a lesser extent, *discretionary* activities. In contrast *meals* and *amusements—in home* are generally reported as rounded durations.

This coincides well with the supposition that reporting exact values is done for activities that have some importance to the person reporting the duration. One would expect that relaxing in front of the television or eating dinner would be hard to recall in great detail, and so they will be rounded off in a survey response. On the other hand, running household errands, or getting to work just a few minutes early or late are situations that are probably easier to recall due to their unique and important nature in a day's



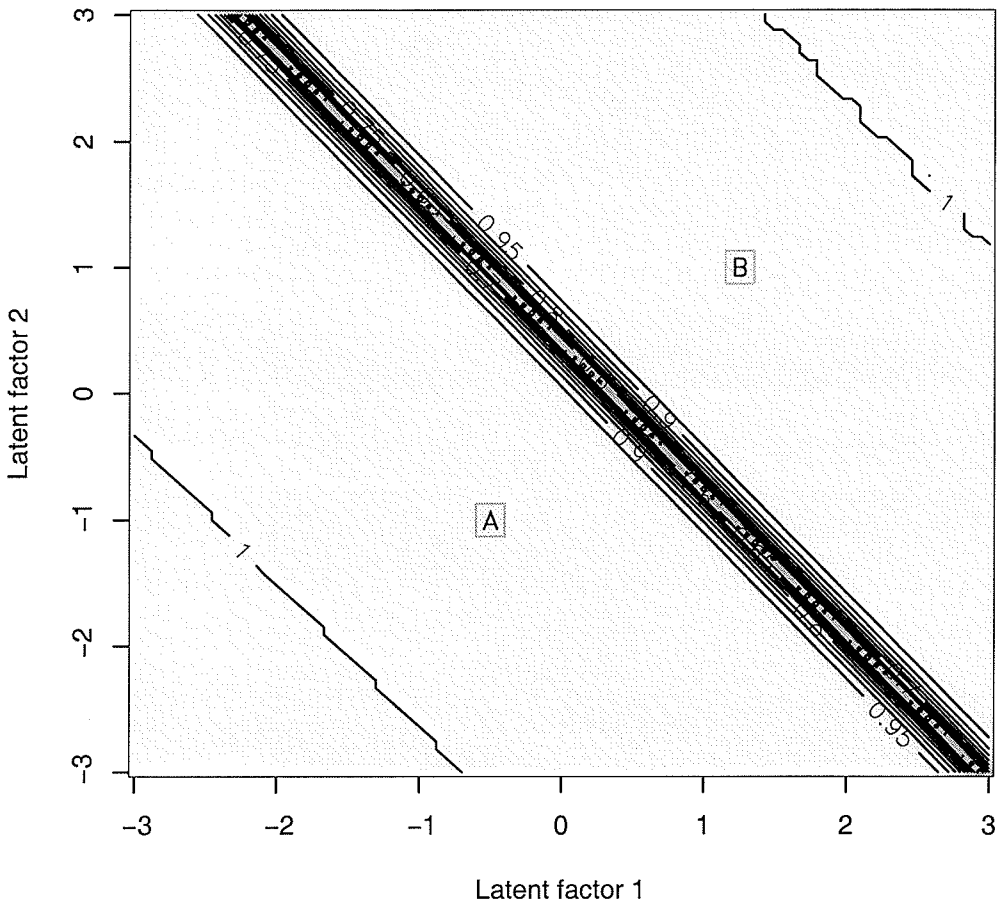
Region	Category
A	reported activity duration rounded off
B	activity duration was *not* rounded off

FIGURE 7. Maximal probability surface for whether or not the reported activity duration was rounded off. Latent factors are assumed normally distributed $N(\mathbf{0}, \mathbf{I})$.

events. Also note that reporting exact durations never becomes the dominant category, which means that two latent factors cannot define exactly the conditions that lead to reporting exact durations.

As noted earlier, Murakami and Wagner (1999) showed that people tend to report trip start times that are rounded, despite the fact that the GPS monitors show that the distribution of trip starting times is roughly uniform over all minutes of the hour. The analysis in this paper begins to offer an explanation for why people might *not* report a rounded off value for time. More research is necessary to determine if the effect is simply that some activities and situations are more easily recalled in a survey situation, or whether there is some deeper relationship with the particular situation in which the respondent found himself or herself that day that led to reporting exact values for activity times.

4.5. Getting out of the house. The final plot in figure 8 shows the influence of the two latent variables on whether or not an activity is conducted in or out of the home.



Region	Category
A	in home
B	out of home

FIGURE 8. Maximal probability surface for in home flag. Latent factors are assumed normally distributed $N(\mathbf{0}, \mathbf{I})$.

This plot shows a rather surprising, sharp delineation of the latent factor plane. There is very little area of the plane where in-home and out-of-home activities coexist. The sharply defined boundary makes interpretation rather easy. Values of the two latent factors located in the upper right half of the plane correspond to activities that occur out of the home, while values in the lower left half of the plane correspond to activities in the home.

Due to the sharp division of the factor plane, we considered building separate models for in-home and out-of-home activities. It is possible that the estimation procedure is devoting too much effort to explaining the small area that falls within the transition band between the two categories. Building two separate models could produce a better model in each case. However, we decided not to do this for three reasons. First, the use of a holdout set to test our model showed that even with this variable included, the model was doing an excellent job of reproducing the observed activities. Second, by including this variable in the model, we are able to describe all of the features of an activity by generating just a single pair of latent variables. If we had two separate models, first we would have to draw a random number to choose whether the activity

was in or out of the home, and then we would have to load up a different matrix of coefficients corresponding to the selected model, and then draw the corresponding latent variables.

The third reason for keeping the in-home variable in the analysis is that the consistent set of coefficients (the A matrix) allows the inspection and evaluation of all activities at once. For example, by superimposing figure 6 over figure 8, it is clear that breaking out two separate models would bifurcate most of the named activities. Thus the analysis of travel time versus activity name performed earlier would be that much more difficult, requiring 4 reference graphs instead of just two. The middle of the in-home/out-of-home valley also cuts right across the maximum likelihood point for engaging in a *discretionary* activity in figure 6, as well as across the transition areas (in figure 4) between duration categories B, C, and D (30 minutes to 4 hours) with categories A (less than 30 minutes) and E (more than 4 hours). For these reasons we decided to keep the in-home flag in this model.

4.6. General interpretation of the two latent variables. It is difficult to draw general conclusions about the *meaning* of the two latent factors described by this research. To do this properly, one would rotate the latent factor plane, so as to align the two factors with the observed characteristics in such a way as to simplify the description of each factor. For example, one might rotate the plane such that Factor 1 was orthogonal to the contour lines of the rounding flag in figure 7. After careful consideration, we have decided that it would be misleading to do perform this analysis. While we had hoped to ascribe the characteristics of environmental opportunity and individual motivation to the two factor dimensions, it is not appropriate to do so at this stage of the research.

The reasons for this are as follows. First, as was noted earlier, the effect of travel duration seems to have confounded both the characteristics of an activity and the characteristics of a sequence of activities. Therefore the latent factors as estimated are describing both the activities, and some portion of the sequencing of activities. Second, the next step of this research is to examine the nature of sequences of activities. Rather than redo the current analysis without the trip duration variable, it better to move on an include a full consideration of activity sequencing. The impact of latent factors capturing motivation and environmental opportunity should also apply to sequences of activities, as people strive to a greater or lesser degree to organize and optimize their behavior over time and space. Finally, given that we are analyzing activities that were not measured with the express purpose of testing our hypotheses, it is premature to make significant claims about the meaning of the latent variables.

5. Conclusions and directions of future research

This paper set out to describe a very large activity space in a simple and concise way. This result has been achieved. The description of an activity proposed in this paper required five different categorical variables, with between two and five categories each, for a total of 400 different permutations of descriptive categories of activities. Using a large set of observed activities, a latent variable model was estimated, with two normally distributed latent variables that explained most of the observed variation in the five observed variables. When the estimated model was used to simulate another, holdout set of observed activities, the hypothesis that the simulated and the

observed distributions of activity characteristics were different was rejected. Thus the 400 categories could be reduced to just two bivariate-normal random variables.

Despite this success, there are many areas where the current work can be expanded and improved. First, we decided against ascribing meaning to the two latent factors as estimated, since such conclusions are premature. Future iterations of this work, with more complete models, should have more interpretive power assigned to the latent factors.

Second, we will need to analyze sequences of activities in order to apply the model to simulating activities. This will definitely require isolating the effects of repeated measurements of individuals, something that was postponed in the current research. The research presented in this paper focused on latent variables that described the characteristics of individual activities—not of sequences of activities, nor of individuals, nor of households. Since each individual in the survey performed about 13 activities, and each household accounted for roughly 29 activities, controlling for person and household effects is important.

Third, the effect of location should be expanded beyond the in-home binary flag and some small consideration of trip duration. While applying a full analysis of all spatial characteristics is probably not warranted, as we expand our analysis to include sequences of activities we can also include other measures of space, such as distance from home and distance from the Portland center. Including the change in these variables associated with any movement prior to engaging in an activity would also address some of the problems with the reported trip duration variable. Future data collection efforts, with detailed GPS trip data, may provide opportunities for a full analysis of spatial effects.

As we expand the analysis methods begun with this paper, we will also apply the latent variables that are estimated. Our initial goal is to generate sequences of activities for a simulation. Obviously this cannot be done with the current results, since sequencing effects in the observed data were explicitly excluded. Another application that has promise is to relate the real-valued latent variables to behavioral traits of the population. Future data collection efforts using GPS devices could be linked with questions that ask about the respondent's knowledge and opinions of their environment. The latent variable techniques could be applied to sort out the differences between habitual activity behavior, and exploratory activity behavior. Another application of this research is to produce pretty, three-dimensional animations of the activity probability surface generated by changing the latent variables over time, or by varying some *a priori* features of activities. This kind of tool would improve the qualitative and quantitative understanding of the range of realistically possible behavior as a person moves through time and space. Such a visualization tool would allow decisionmakers to see directly the results of different transportation policy options.

This paper has demonstrated that applying latent variable modeling techniques to the analysis of activities is highly effective. In purely practical terms, complicated descriptions of activities can be reduced to just a few random variables. In theoretical terms, the latent factors help visualize the linkages between different aspects of activities. While some technical aspects must be ironed out in future research, the approach has the potential to produce a number of useful applications, as well as focus theoretical developments in activity based transportation research.

REFERENCES

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis*, Vol. 7 of *Kendall's Library of Statistics*, second edn, Arnold, 338 Euston Road, London NW1 3BH.

Ben-Akiva, M. E. and Bowman, J. L. (1995). Activity-based disaggregate travel demand model system with daily activity schedules, *EIRASS Conference on activity-based approaches: Activity scheduling and the analysis of activity patterns*, Eindhoven University of Technology, Eindhoven, The Netherlands.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.

Golob, T. F. and McNally, M. G. (1997). A model of activity participation and travel interactions between household heads, *Transportation Research B* **31B**(3): 177–194.

Hägerstrand, T. (1970). What about people in regional science?, *Papers of the Regional Science Association* **24**: 7–21.

McNally, M. G. (1997). An activity-based microsimulation model for travel demand forecasting, *in* D. F. Ettema and H. J. P. Timmermans (eds), *Activity-based approaches to travel analysis*, Pergamon, Elsevier Science, Oxford, U.K., chapter 2.

McNally, M. G. (1998). Activity-based forecasting models integrating GIS, *Geographical Systems* **5**: 163–187.

Murakami, E. and Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting?, *Transportation Research Part C* **7**(2/3): 149–165.

Pas, E. I. (1988). Weekly travel-activity behavior, *Transportation* **15**: 89–109.

Portland METRO (1994). Oregon and Southwest Washington 1994 Activity and Travel Behavior Survey, 600 Northeast Grand Avenue, Portland, OR 97232-2736.

Recker, W. W. (1995). The household activity pattern problem: General formulation and solution, *Transportation Research B* **29B**(1): 61–77.

Recker, W. W., McNally, M. G. and Root, G. S. (1986). A model of complex travel behavior: Part I—Theoretical development, *Transportation Research A* **20A**(4): 307–318.

Ripley, B. D. (1981). *Spatial Statistics*, John Wiley and Sons, New York.

Vaughn, K. M., Speckman, P. and Pas, E. (1997). Generating household activity-travel patterns (HATPs) for synthetic populations, 76^{th} *annual meeting of the Transportation Research Board*, TRB, Washington, D. C.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, Statistics and Computing, third edn, Springer-Verlag, New York.