Title
Collecting Activity Data from GPS Readings

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

Publication Date
2002-07-01

UCI-ITS-AS-WP-02-3

Collecting Activity Data from GPS Readings

James E. Marca [1]
Craig R. Rindt [2]
Michael G. McNally [3]

UCI-ITS-AS-WP-02-3

[1] Department of Civil and Environmental Engineering and Institute of Transportation Studies
University of California, Irvine; Irvine, CA 92697-3600, U.S.A., jmarca@translab.its.uci.edu

[2] Department of Civil and Environmental Engineering and Institute of Transportation Studies
University of California, Irvine; Irvine, CA 92697-3600, U.S.A., crindt@translab.its.uci.edu

[3] Department of Civil and Environmental Engineering and Institute of Transportation Studies
University of California, Irvine; Irvine, CA 92697-3600, U.S.A., mmcnally@uci.edu

July 2002

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.

Collecting activity data from GPS readings

James E. Marca, Craig R. Rindt, and Michael G. M^{c}Nally

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697
tel: (949) 824-6571
fax: (949) 824-8385
jmarca@translab.its.uci.edu

August 16, 2002

Abstract. GPS recording devices offer a painless way to collect travel data, but are not directly useful to a standard activity survey. This paper documents one method for linking activities with location data. Based on a small but extended pilot survey, a technique has been developed to estimate the most likely activity at a destination, based on the respondents' past responses. If destinations and activities were randomly paired, this information would be irrelevant. But the pilot survey also demonstrated that activities are tightly clustered in space.

Wordcount: 4160 + 4 figures

INTRODUCTION

A GPS-enabled data collection system, such as the Tracer system *(1)*, can collect approximate activity destinations by recording the stops and starts of vehicle trips. The collection of this kind of data is relatively painless to the survey participant, requiring little more than a willingness to be traced by a GPS device. This is in contrast to the labor intensive task of entering activity destinations in an activity diary. However these GPS-tagged locations do not in themselves reveal activity data. A web-based activity survey has been developed *(2)* which is tightly integrated with the wireless Tracer GPS data collection system. The GPS device broadcasts continuously via a wireless data modem to a base station database. The on-line survey uses the database to generate dynamically survey questions. The website collects activity data related to each destination recorded by the GPS units. Unlike the GPS data, this information is time consuming for the respondent, and therefore results in the usual decline in response levels over time.

This paper documents a method for using any and all activity data a person is willing to provide to attempt to tag all incoming activity destinations with an activity name. The technique relies heavily upon the fact that activities and locations, as reported, appear to be highly correlated. That is, people repeat what they do and where they do it.

An example application is as follows. Suppose a respondent volunteers to participate in a travel and activity survey for one month, but stops filling in the activity destinations after a week or so. Rather than discarding the extra three weeks of location data, using the methods described in this paper it is possible to estimate with varying degrees of certainty the most likely activities at each of the locations, based on the information the respondent was willing to provide. This data can be used to further streamline the activity survey, and to generate a list of "most-wanted" answers to ask the respondent after a reminder call. In addition, after the data has been collected, this activity prediction technique can be used to generate "augmented" activity sequences, ones which have less certainty than using just the actual reported activities, but which may allow conclusions to be drawn about the larger patterns of behavior.

ASSOCIATION OF ACTIVITIES WITH LOCATIONS

The Tracer data collection system *(1)* and the ANNE web-based activity survey *(2)* were tested with a small pilot study consisting of four volunteers. The purpose of the pilot study was to test the survey hardware, software, and web-based interface. A side effect of the pilot study was to collect several months of GPS data for each volunteer, along with a much shorter period of activity survey completion. A companion paper *(3)* makes the case that the destinations recorded in this pilot survey are clustered in space. That is, any randomly chosen destination is likely to have many more close neighbors than would be expected if the process were truly spatially random. The implication is that destinations are revisited; this paper looks at associating a repeated activity with the revisited destinations.

Further evidence of the coincidence of activities and destinations was provided informally as the respondents completed the web-based survey. As documented in Marca *(2)*, the on-line survey rendered each page of questions dynamically, including ordering past responses to activity options other input checkboxes according to a distance and time based metric. The pilot survey volunteers commented anecdotally that one of the first options was often the one chosen.

The sort algorithm worked well to the extent that time and space were the most important factors in determining the relevance of an option. But this was not always the case. First the frequency of a particular response was ignored. The ANNE methodology allowed free-text and even multiple responses to all questions. Suppose a respondent decided to change the name of an activity over time, say from a long entry such as "Eating dinner at home and watching the kids" to multiple short entries, such as "Eating dinner" and "Watching the kids." The sorting algorithm would not know about the shift in data entry styles, and would continue to suggest the initial response. In a more likely scenario in a more formal survey, a respondent might perform one activity much more frequently than another at the same location, such as working at work and occasionally eating lunch at the same place. Another situation is one in which the EDCU may have missed a stop, and the respondent felt compelled to document the missed activity at an incorrect location. Again, the "wrong" activity would bubble up to the top of the list anytime that destination was revisited. A more complete and robust solution is to develop a probability distribution for the likelihood of engaging in different activities at different locations. That is the topic of this paper.

There do not seem to be any similar efforts in the activity analysis literature. This is probably due to the short durations of most GPS-related surveys, and the focus on using GPS devices to verify the accuracy of larger activity survey samples. Wolf et al. *(4)* discuss eliminating the travel diary by carefully merging GPS records with a detailed geographic information system (GIS), but they do not discuss searching for or relying upon repeated spatial or temporal patterns within the GPS data itself, nor do they appeal to the responses of the survey participants. Again, this may be due to the shorter durations of GPS data collected, where repetition is unlikely to be a dominant feature.

ESTIMATING A SEMI-VARIOGRAM

A naive approach to associating observed activities with locations is to first bin the data by imposing a grid or tessellation on the observation area, and then use the frequency of observations in each bin to compute an empirical distribution. This is easy to do, but misses relationships between neighboring bins. Especially given the clustered nature of destinations, the likelihood of engaging in an activity at any given longitude and latitude is likely to be highly correlated with the same likelihood at all nearby points. For example, parking in the same spot would create a highly localized cluster of observations over time, all relating to the same activity; another activity location might not provide the respondent with a dedicated parking spot, and so would have a different distribution of destinations. These local features of the destination process should be included in the prediction effort. A technique called kriging has been developed in mining exploration, meteorology, and other disciplines where spatial processes are natural and unavoidable *(5)*. Kriging incorporates point estimates of spatial correlation, as well as global and local trends, when estimating a prediction surface.

In order to estimate a kriging prediction surface, one must first estimate what is known as the semi-variogram *(5, 6)*. The semi-variogram approximates the spatial variance of the observations. In a mining application, one would measure variables relating to the porosity of the rock, and so on. In a forestry application, one might measure the diameter of trees or the types of plants. In this application, the goal is to associate activities with destinations, and so the best variable to use is the frequency of observations—the numbers of unique activities (including repeats) at a

particular longitude–latitude pair.

Within the observed data from the pilot survey, the frequency of visits to an exact latitude and longitude ranged from 1 to 4. It is to be expected that the frequency of any given location is likely to be low, since GPS readings of longitude and latitude are precise to 6 decimal places. However, the general vicinity of a destination will have a relatively high concentration of visits if a repeat activity occurs there. An example is shown graphically for one of the respondents in figure 1; the other three respondents had essentially identical plots. As more data is collected, one would expect the observation of unique points to steadily increase in the neighborhood of all frequently visited points, since there are an infinite number of points available as measurements become more precise. If the precision of observation is reduced, the effect of the random noise diminishes and the underlying spatial clustering can be observed.

[FIGURE 1 about here.]

The semi-variogram is designed to characterize autocorrelation in spatial data. This includes both gross clustering of points, as well as finer details of clustering such as local directional patterns or trends. As was mentioned above, for small distances data tend to be autocorrelated. If the exact activity destination could be measured each time, then one would observe a uniform plane with sharp spikes of high activity at a handful of points on a map. Instead, the spikes are flattened and spread out by a number of dispersing factors. These include the inherent error of GPS measurements, the GPS recording parking spots rather than actual activity destinations, the variable polling rate of the EDCUs, the aliasing of the last reported time versus the actual time the vehicle was stopped, and so on. In addition to these data collection sources of variability, the exact destination itself—in this case, where the vehicle is parked—is subject to a natural spreading in space. Factors such as the availability and location of a dedicated parking lot, versus the potentially more spatially distributing effect of first-come, first-served on-street parking will cause the final, measured destination to vary even though the "actual" destination of the activity is the same.

At larger distances, this effect declines and the variance stabilizes, as can be seen from the low precision curve figure 1. At the coarsest granularity, the numbers of new destinations increase quickly at first, and then grow at a slower rate. To give an idea of the scales involved, in Orange County, California 0.01 degrees of longitude by 0.01 degrees of latitude is roughly 0.9 km by 1.1 km, or 1.44 km^2. This area defines a rectangle that approximately bounds South Coast Plaza, a large regional shopping mall. One might say that the travel patterns observed are clustered at the scale of a shopping mall (which may be peculiar to Southern California). Of course a much larger study is needed to verify this claim.

In terms of the mechanics of estimating a semi-variogram, there are many factors that one should consider. For example, there may be significant directional or systematic variations. This is a natural result of the fact that vehicles tend to drive on linear roads, although it is softened somewhat by the existence of parking lots at destinations. Examining the scatter plots of recorded destinations show that the EDCU device would often record its last point on streets leading up to the final destination. This was common for all four vehicles. This will result in a large variation in one or two directions, and very little variation in other directions. But it is difficult to specify these local spatial and directional effects in general terms. Given the informal nature of this particular data set, the goal of this analysis is only to be able to estimate a kriging surface so as to examine the *potential* for this technique. For a general method that is applicable to any individual traveling

in any area, it is safer to presume that the autocorrelation is uniform in all directions. Bear in mind that a large scale study would do well to consider the local effects of the transportation network on the expected variance of destination measurements.

It is also theoretically possible to estimate separate variograms for each activity type. However, this was not a practical option due to the small numbers observed of each kind of activity. In a larger study, it is likely that the numbers of activities will be the same or fewer per person, given realistic expectations of the maximum duration of the survey. On the other hand, it is reasonable to assume that the spatial autocorrelation is largely independent of the activity at the destination. Effects such as small GPS errors, the scattering of parking locations around the destination, and the slight aliasing introduced by the periodic polling and updating scheme of the GPS data collection unit occur independent of the destination or the activity. These effects might be balanced or even swamped by factors unique to each destination, such as the availability of a dedicated parking space. Again, the estimation of the semi-variogram is an area which would benefit greatly from a larger survey and further analysis.

For the above reasons, the semi-variogram was estimated for each respondent in the pilot survey using all of the respondents' labeled locations, regardless the name of the activity assigned. Unlabeled activities were not used, as the intent is to try to use only the labeled activities to infer information concerning the unlabeled destinations. For three of the four sets of points, the semi-variogram was estimated to a maximum distance of 5 km, ensuring that local effects are captured, but not extending the influence of a point too far. For user id 5, whose trips covered very long distances and who labeled only 12 destinations, the estimation of a semi-variogram had to be performed with a distance limit of 15 km. A robust estimator from Cressie *(5)* implemented in the `geoR` library package of the R statistical language and environment *(7)* was used to estimate the semi-variogram. These estimates curve over a largely horizontal line through the observed data, with a short initial curve capturing the local spatial auto-correlation. The semi-variograms were then used to estimate kriging surfaces for each of the labeled activities. Again the library `geoR` was used, with the estimate computed using a weighted least squares (WLS) estimator, with the so-called nugget effect (spikes at data points; see *(6, p. 441)*) allowed to be non-zero and free to vary in the estimation. The resulting surfaces are discussed in the next section. Once again, note that a much more extensive treatment of the estimation of the semi-variogram is warranted, and should be performed when a larger survey sample is collected. As this is only a pilot study, the most important contribution is to demonstrate the potential applications of this data.

TAGGING UNKNOWN ACTIVITIES USING ESTIMATED KRIGING SURFACES

Only some of the estimated surfaces are shown, due to space constraints. For uid 4, only a small number of recorded events were labeled with activity names. However, only a few activities took place outside of a small geographic area. This contrasts directly with the observations of another volunteer, in which there were many "false" stops generated spread over a very large geographic area. The other two volunteers has a more balanced mix of activity entries and spatial spread. Figure 2 depicts the total likelihood for engaging in any named activity based on the responses of user id 6, with dark areas representing higher likelihood. The starred points represent named activity destinations, while the small dots represent the much more numerous unlabeled destinations that fall within a polygon containing the named points. Outside of this polygon the kriging surface is

largely invalid, as there is no data to support such extrapolation.

[FIGURE 2 about here.]

Similar kriging surfaces were generated individually for each of the activities entered. An example of an activity-specific kriging surface is shown in figure 3. Unlike figure 2, only those points corresponding to the activity in question were used to estimate the kriging surface. Therefore the black areas reveal locations where *that* activity is most likely to occur, rather than where *any* activity might occur. The figure shown is for an activity with several observations. Activities with only one or two observations describe less interesting surfaces, although the existence of *other* activities enables a rigorous definition of "not" observations.

[FIGURE 3 about here.]

All of the activity-specific kriging surfaces were used to predict the expected number of events of each type at all of the unlabeled locations. When these values are summed and normalized, the result approximates the probability that a certain kind of act might have occurred at a particular location, given the past survey entries. Figure 4 shows the results of these predictions for all of the unlabeled destinations for all of the respondents. For example, for user id 4, this procedure assigns labels with a 90% or greater probability to 79 unlabeled activities, which is only a small number of the 296 unlabeled activity destinations, but it is somewhat remarkable in that only 20 activities were provided labels by user id 4 in the on-line survey.

[FIGURE 4 about here.]

As the numbers of repeated observations goes up for a particular activity in an area, the kriging surface becomes that much more complex and that much better able to predict whether an unlabeled observation might correspond to a particular activity. With very little information, it is hard to get much specificity out of the predictions. The pilot survey participants ran the gamut, with a very low response rate and a correspondingly bad predicting surface for user id 5, to a high response rate, complex kriging surfaces, and a pretty good prediction rate for user id 2. Finally note that while the number of activities "labeled" by the kriging surfaces are quite high, this should probably be interpreted as a rate, related to the response level. Further work needs to be performed on a wider range of respondents, but it should be possible to specify some minimum response rate to guarantee that a percentage of unlabeled destinations can be labeled with a belief greater than 0.9.

CONCLUSION

This paper presents the examination of the results of a small pilot study exploring the use of a GPS device used in an activity survey. While the respondents quickly tired of entering data in the activity survey, they continued using the GPS data collection devices for much longer periods of time. One can imagine similar scenarios in real surveys, perhaps where respondents are instructed to respond to an on-line survey for only a subset of days, or only at the beginning and end of a survey period.

The GPS measurements provide accurate time and position data for the parking spot of the vehicle at the start of an activity, which approximates the actual time and place of the activity. This paper demonstrates that it is possible to estimate activity-specific kriging surfaces for these data, using a semi-variogram estimated from all of the observed and labeled destinations. The kriging surface gives the expected frequency of events for each named activity at any given longitude and latitude. The combination of the frequencies from all kriging surfaces (one for each reported activity) can be used to generate a probability for the activity names at each unlabeled point.

It was shown that this method can generate most likely activities that have a high probability, relative to other entries. The net conclusion is that activity destinations, at least for the four individuals observed, are *not* random spatial processes, which means that a small set of labeled destinations can be used with some confidence to estimate the activities at unlabeled points. Further refinement of the technique is also possible by including the time of day of the activity, which was not considered here.

The proposed initial application of this technique is to further streamline the dynamically generated activity survey. Instead of presenting the respondent with a list of *all* past activities, appealing to the kriging surfaces as a synthesis of all past observations allows the survey to generate a list of only the most likely activities. This will reduce the respondent burden slightly by eliminating the visual clutter caused by irrelevant activity choices.

Another application that could result from further development of this technique is to augment a short activity survey with an approximate activity survey, based only on GPS data. This application would pair a GPS data collection device and an on-line survey, as in this pilot survey example. However, the activity survey would be designed to end prior to the GPS data collection period. The extended time period of GPS data collection (which is relatively painless to the respondent) can then be post-processed using the activity-specific kriging surfaces to determine the most likely activity at each destination. Those activities which are assigned with some high degree of certainty can be used to approximate an extended duration activity survey.

Alternately, the areas of low or medium predicted frequency can be the subject of focused questions. For example, suppose an on-line survey is designed to bracket a period of GPS data collection. The initial survey can collect activity data for all destinations visited, as with this survey. Then the second survey can be designed to prompt the respondent to fill in activity data for days which have high numbers of unknown destinations. Alternately it might be more important to further refine some area of space which has multiple kinds of activities situated next to each other. In any case, the goal is to use the results of the kriging analysis to get the most useful information possible from a respondent, rather than wasting the respondent's good will answering questions about destinations which are already well known.

An interesting extension of this work that would have important ramifications for survey design would be to establish a relationship between the point pattern of destinations, the numbers of activities labeled via the on-line survey, and the fraction of unknown points that can be assigned an activity label with a high degree of confidence. As the point process becomes more random, the power of the kriging surface will diminish as the peaks become hills or slight ripples. A very small number of responses has a much greater impact when those responses are near all or most of the respondent's unlabeled activities. But just gathering more activity data is not likely to guarantee a good match to every unknown point, as there are always new destinations or activities. Determining a good balance is the key to a cost effective survey design which gets the most mileage out of a respondent's limited patience.

REFERENCES

[1] James E. Marca, Craig R. Rindt, and Michael G. M^cNally. The tracer data collection system: Implementation and operational experience. CASA working paper AS-WP-02-2, August 2002.

[2] James E. Marca. The design and implementation of an on-line activity survey. CASA working paper AS-WP-02-1, August 2002.

[3] James E. Marca. A preliminary model of activity destinations as random point processes. Working paper, May 2002.

[4] Jean Wolf, Randall Guensler, and William Bachman. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record*, (1768):125–134, 2001.

[5] Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., New York, revised edition, 1993.

[6] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Statistics and Computing. Springer-Verlag, New York, third edition, 1999.

[7] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

LIST OF FIGURES

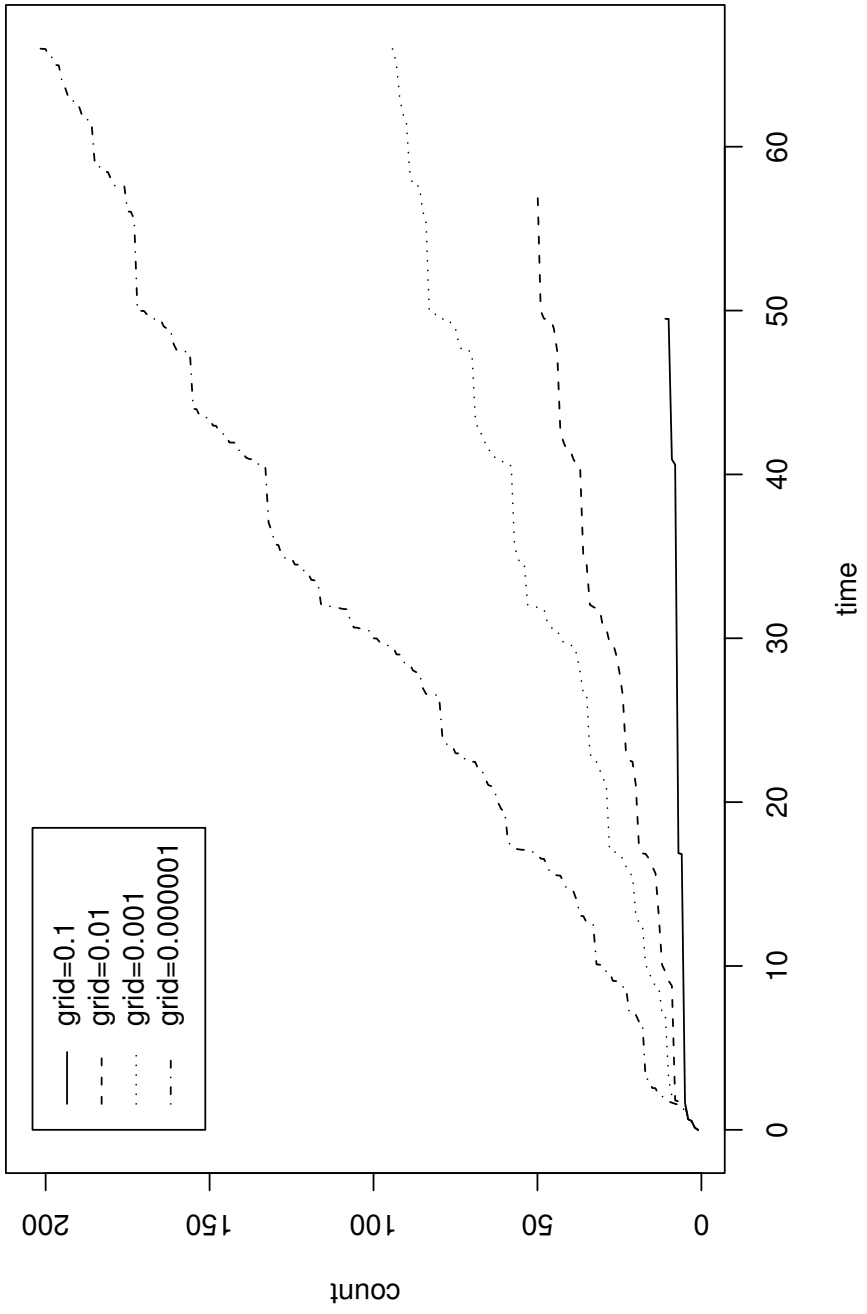

FIGURE 1: Growth of new destinations versus elapsed survey days, for user id 6. The tapering off of low precision measurements shows that destinations are repeated over time. The constant rise of finer precision measurements demonstrates the natural variability of GPS measurements.

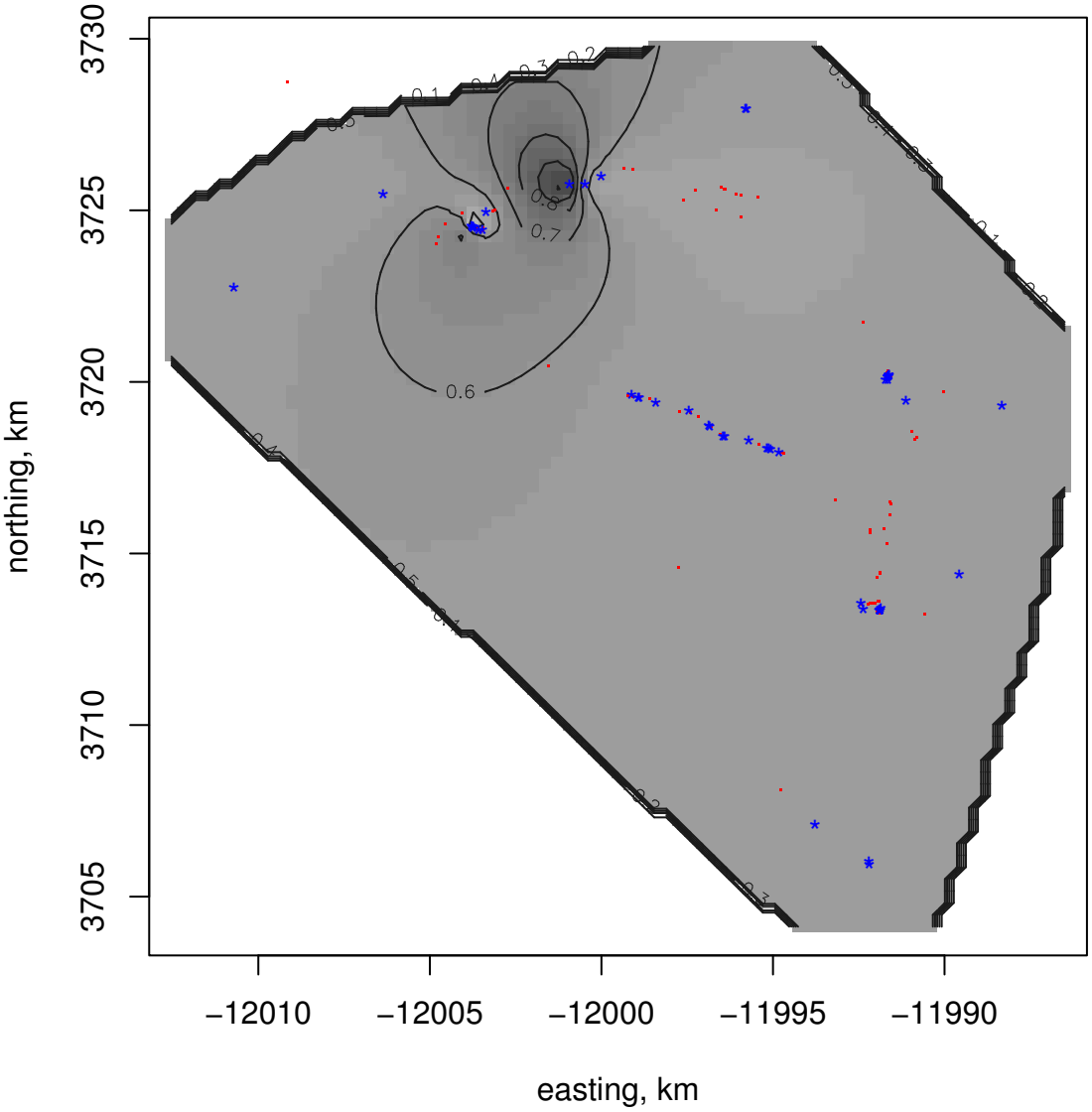

FIGURE 2: Kriging surface for user id 6. The surface was computed using the named (starred) activities. The small dots represent unlabeled points. The dark areas represent higher expected frequencies, and the white areas lower frequencies.

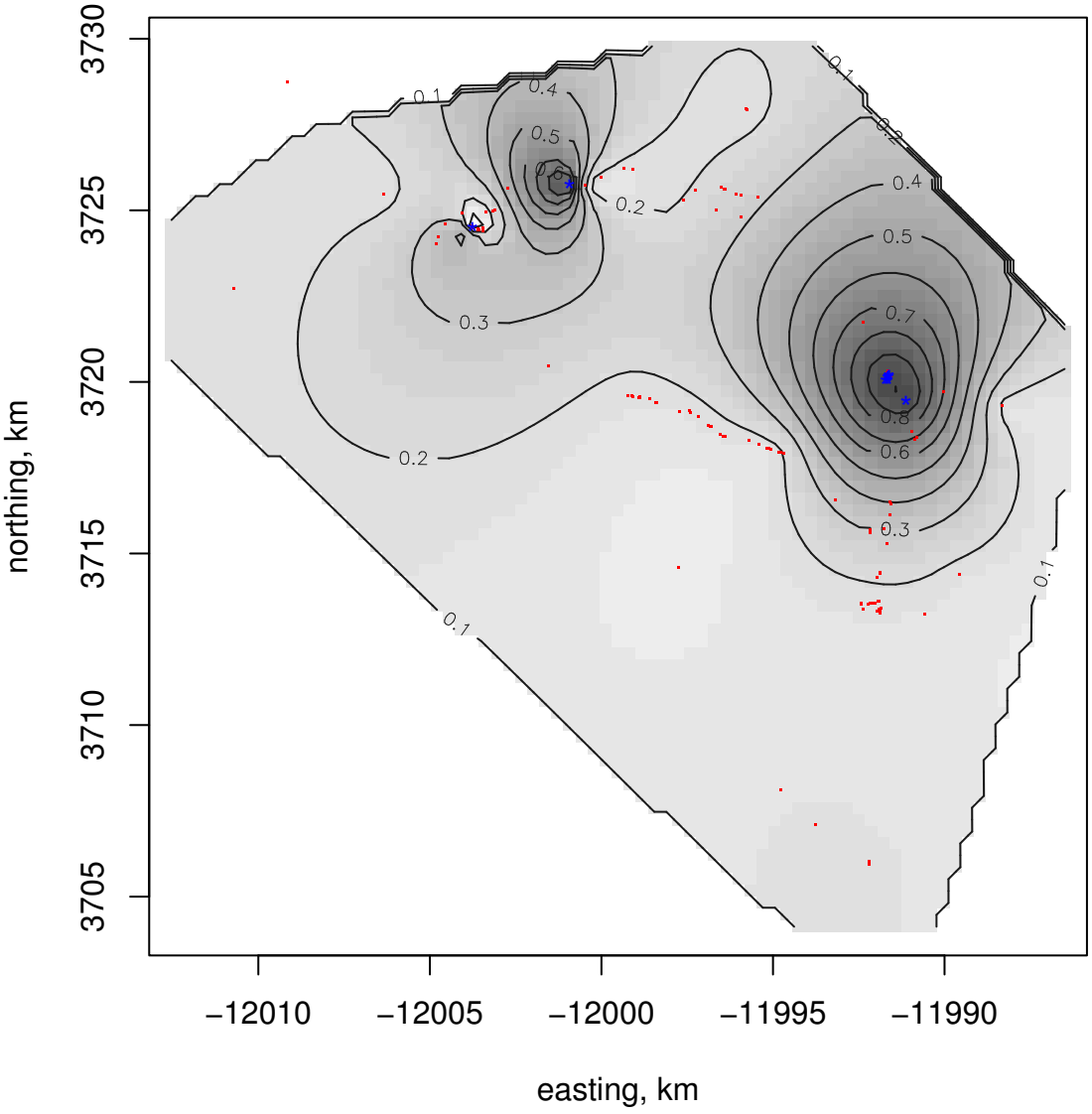

FIGURE 3: Kriging surface for user id 6 activity 1. The stars are activity 1 observations, while the points are both labeled and unlabeled destinations. Compare areas of high and low expected frequency to the "any activity" surface of figure 2.

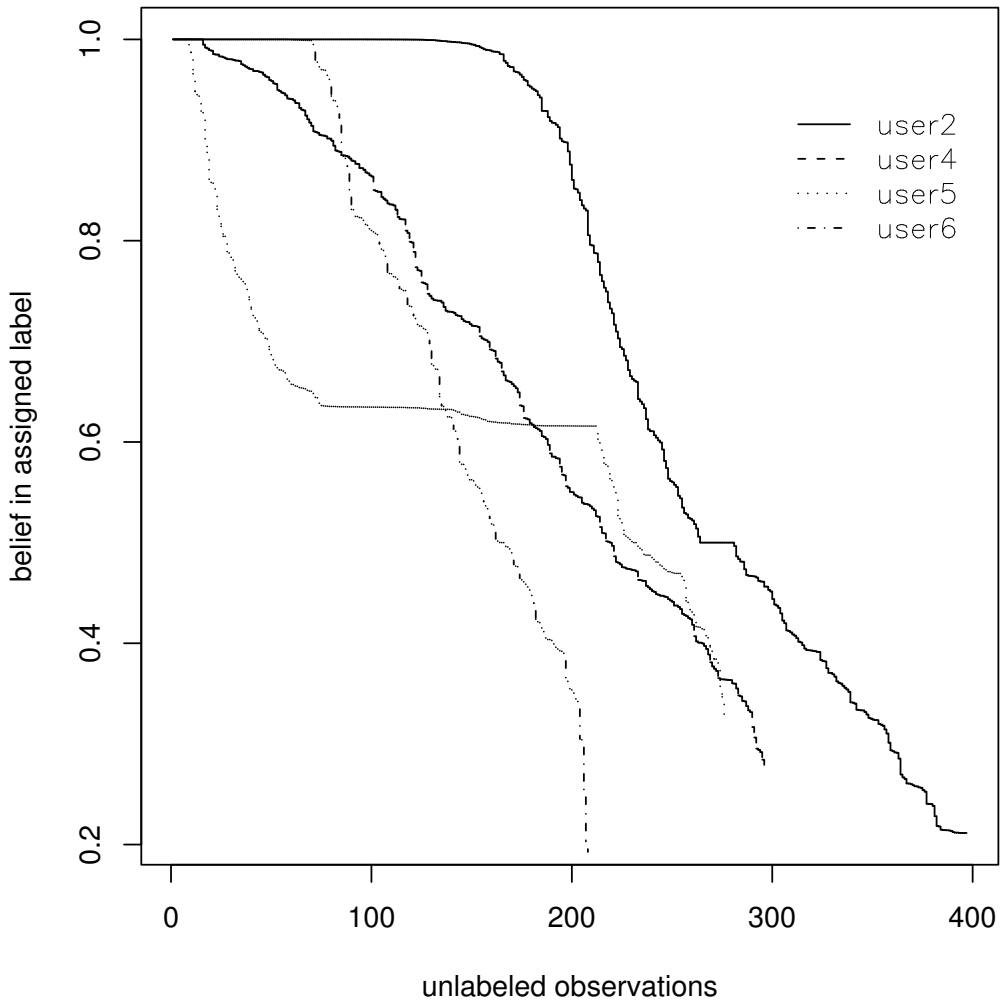

FIGURE 4: Probabilities of the most likely activity labels assigned to each unknown activity for each of four pilot survey volunteers, given the responses to the on-line survey. The more activities are entered, the more refined the activity-specific kriging surfaces, and the higher the corresponding belief associated with unlabeled activities. (Note that activities have been resorted from best prediction to worst prediction for each respondent.)