

# Real Time Destination Prediction Based On Efficient Routes

John Krumm  
Microsoft Research

Copyright © 2006 SAE International

## ABSTRACT

This paper presents a novel method for predicting the location of a driver's destination during the drive. Such a prediction can be used to help decide which information to automatically present to the driver, depending on where the driver is going. The prediction is based on the common intuition that drivers tend to choose efficient routes. We quantify this preference for efficiency probabilistically based on a database of driving trips we gathered with GPS receivers. We show how to use this probability along with a map of driving times to compute the probability of any candidate destination. Our tests show that halfway through the drive, we can predict the destination to within about 10 km, and at three quarters of the way, the error drops to about 3 km.

## INTRODUCTION

Knowledge of a driver's destination is an important parameter for delivering useful information during the drive. For instance, an in-car navigation system could automatically show traffic jams, gas stations, restaurants, and other points of interest that the driver is expected to encounter along the way. If the navigation system can make an accurate guess about the general

region to which the driver is heading, then it can intelligently filter the information it displays, reducing the cognitive load. While it would be possible to explicitly ask the driver about his or her destination, drivers would likely not bother to provide this information at the beginning of every trip.

It would be much more convenient to automatically predict the destination. Toward this end, we have developed an algorithm to predict driving destinations based on the intuition that the driver will take a moderately efficient route to the destination. Using GPS data we gathered from 118 driving volunteers who made about 4300 different local trips in the Seattle area, we derived a probability distribution for the amount of driving time drivers normally waste by taking less-than-optimal routes. We used this probability distribution to repeatedly compute the probability of a grid of candidate destinations as the drive progresses. The probability computation is set up to reduce the probability of destinations for which the driver has passed up efficient routes. We tested the algorithm on our GPS data and found that we can predict the destination to within about 10 km at the trip's halfway point, and at three quarters of the way, the error drops to about 3 km.

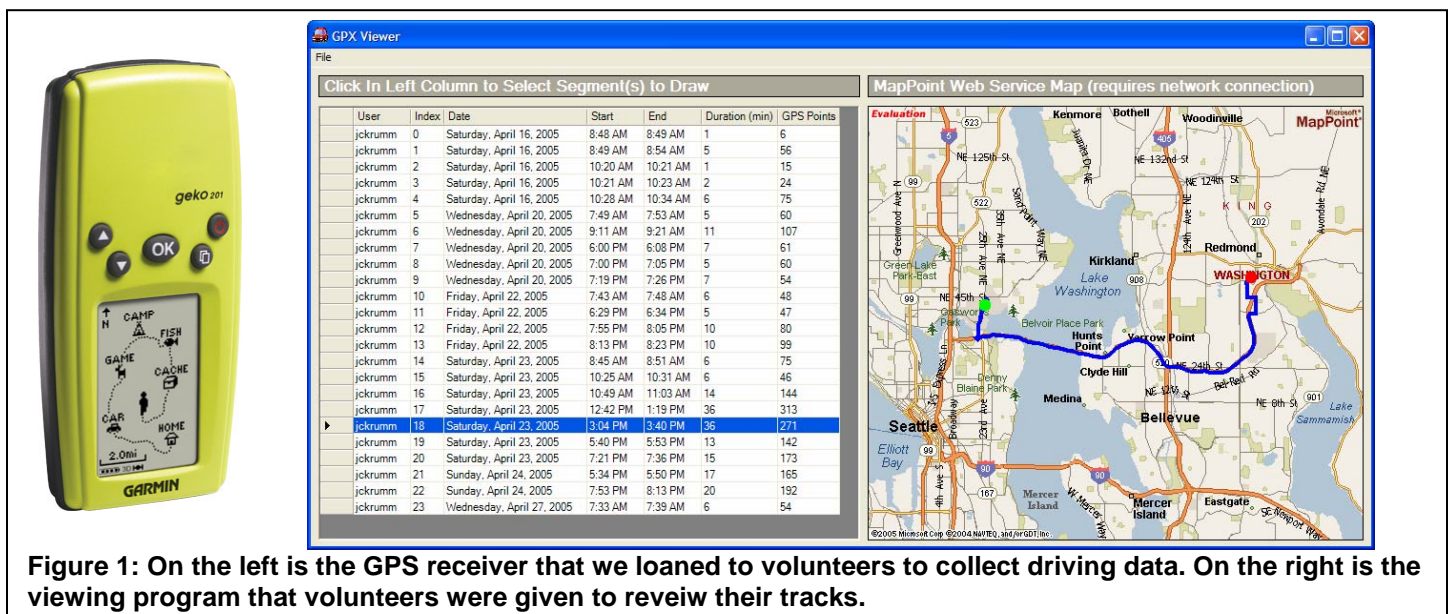


Figure 1: On the left is the GPS receiver that we loaned to volunteers to collect driving data. On the right is the viewing program that volunteers were given to review their tracks.

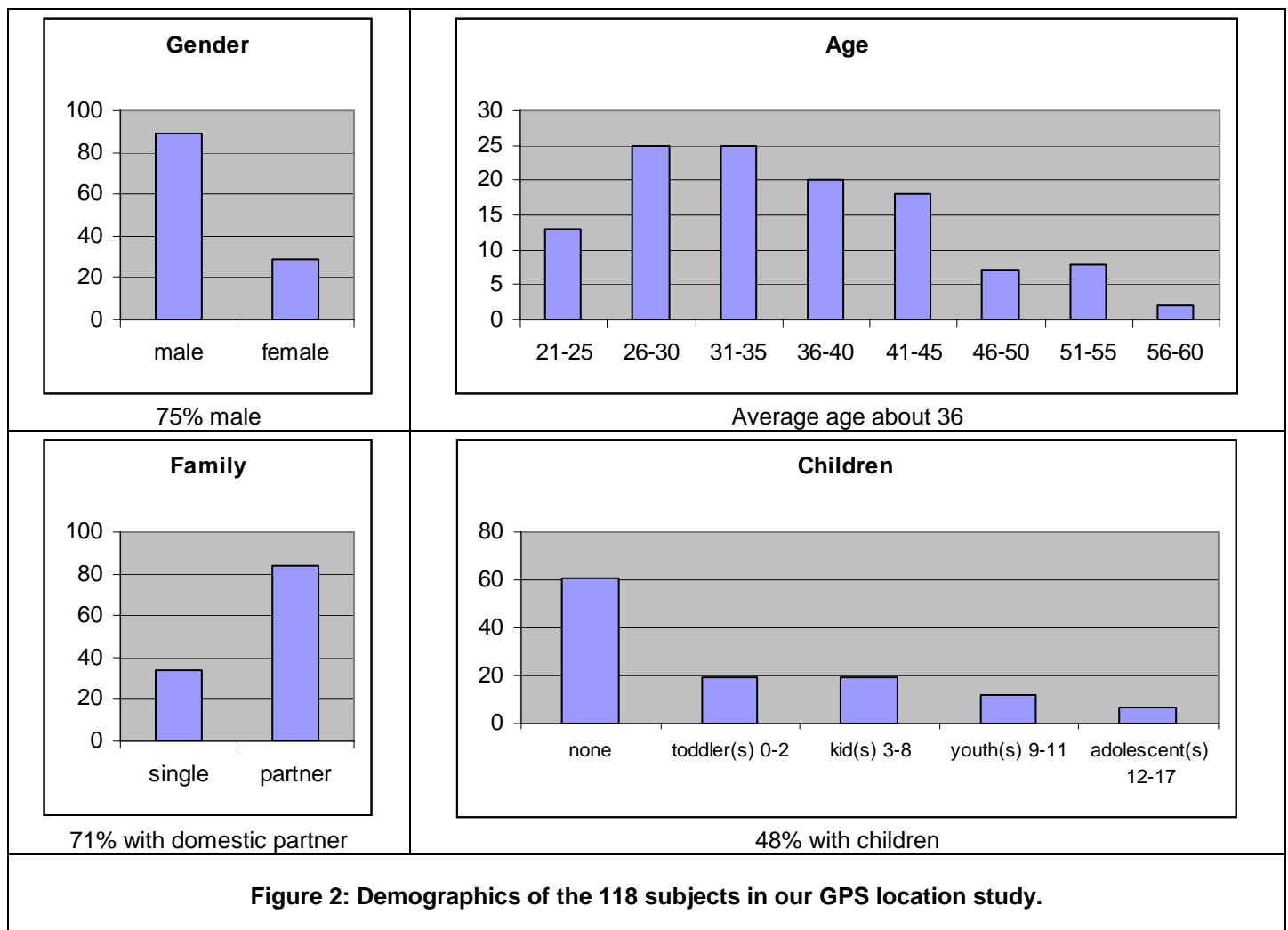
Our work differs from previous work in destination prediction in that we do not use a model of an individual's travel behavior. For instance, Ashbrook and Starner[1] use GPS traces to find a user's meaningful locations and then apply a second-order Markov model to predict which of these locations a user will go to next. In a body of work represented by [2], Patterson, Liao *et al.* present dynamic Bayes networks that learn about a user's travel behavior to predict where the user will go among a set of previously learned destinations. Our work is different in that we assume no prior knowledge of a driver's usual destinations (*e.g.* work, home, school). This means that our system can work "out of the box" in a new car, a rental car, or in a city the driver has not visited before.

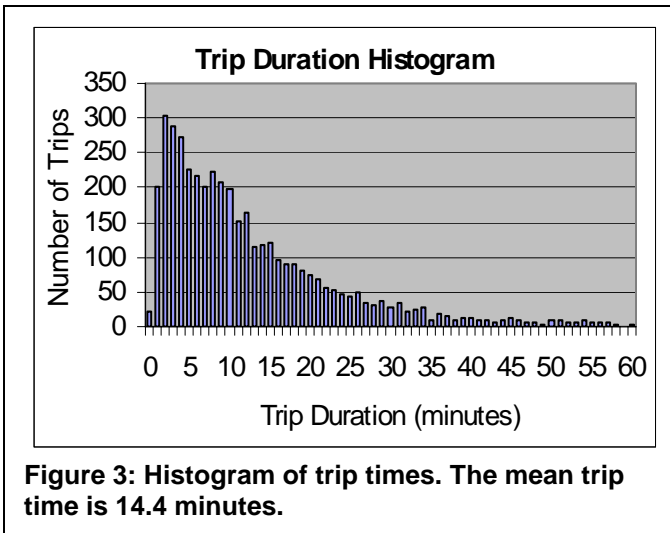
Another relevant area of research is predicting locations for users of mobile wireless devices, like cell phones and Wi-Fi[3]. These algorithms are designed to predict where a wireless user will go to facilitate efficient handoffs between antennas. In contrast, our work is designed to predict the ultimate destination of a driver, not the next few locations. Also, our locations are defined geographically, not in terms of antenna locations, which means we can be sensitive to road networks and driving behavior.

## MULTIPERSON LOCATION SURVEY

We developed our model of efficient driving and took our test data from the Microsoft Multiperson Location Survey (MSMLS)[4]. The MSMLS is an ongoing project aimed at gathering data about where people go in their daily lives. Volunteer subjects for our survey are loaned one of 40 Garmin Geko 201 GPS receivers (Figure 1) for nominally two weeks. The default use of the receiver is to power it from the cigarette lighter in the subject's vehicle, with the receiver resting on the dashboard or some place with a clear view of the sky for GPS satellite reception.

Our GPS receivers are capable of recording up to 10,000 time-stamped (latitude, longitude, altitude) triples. The Geko 201 can be programmed to record at regular intervals in time or distance. We used a third mode that adaptively records more points when the receiver is accelerating or turning, presumably by thresholding on the deviation between the measured point and the receiver's internally extrapolated estimate. This mode offers five resolution settings varying from "highest" to "lowest". We chose the "highest" setting. For the approximately 480,000 points we recorded, the median distance between the points was 62.0 meters, and the median time between points was 6.0 seconds. Our GPS receiver uses the wide area augmentation system





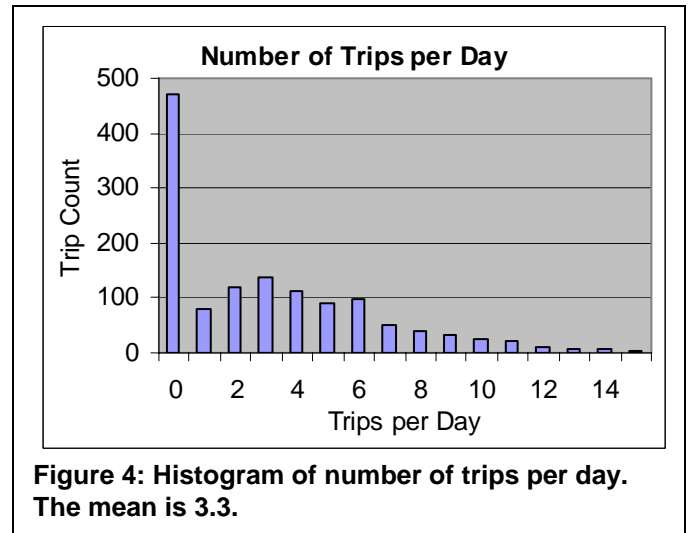
(WAAS), whose RMS error has been measured as 1.13 meters[5].

We solicited volunteer subjects primarily from Microsoft Research, but some volunteers were other Microsoft employees and spouses. The demographics of the subjects are shown in Figure 2. Overall, of the 118 subjects, 75% were male, 71% had a domestic partner, 48% had children, and the average age was about 36.

For the purpose of determining trip destinations, we segmented each subject's data into discrete trips. After downloading from the GPS into our database, each subject's raw data consisted of a sequence of time-stamped (latitude, longitude) coordinates. (We ignored altitude.) We split these sequences into discrete trips by looking for places in the sequence that met either of the following criteria:

- Gap of at least five minutes – This indicates that the GPS was not moving and, because of its adaptive recording mode, not recording new points. Such a gap can also come from vehicles whose cigarette lighter turns off with the car, which would turn off the cigarette lighter-powered GPS.
- At least five minutes of speeds below two miles per hour – This accounts for the fact that, even when parked, GPS noise can make it appear that the vehicle is moving slightly. Five minutes or more of this extremely slow apparent movement is considered a split between trips.

Segmentation resulted in a total of 4300 discrete trips. To check the plausibility of our segmentation scheme, we computed statistics about the trips to see if the results were reasonable. One statistic is the average temporal length of each trip, which was 14.4 minutes. A histogram of trip times is shown in Figure 3. Lacking any other source of trip length statistics, this result appears reasonable. Another statistic is the average number of trips per day, which we computed as 3.3, also a

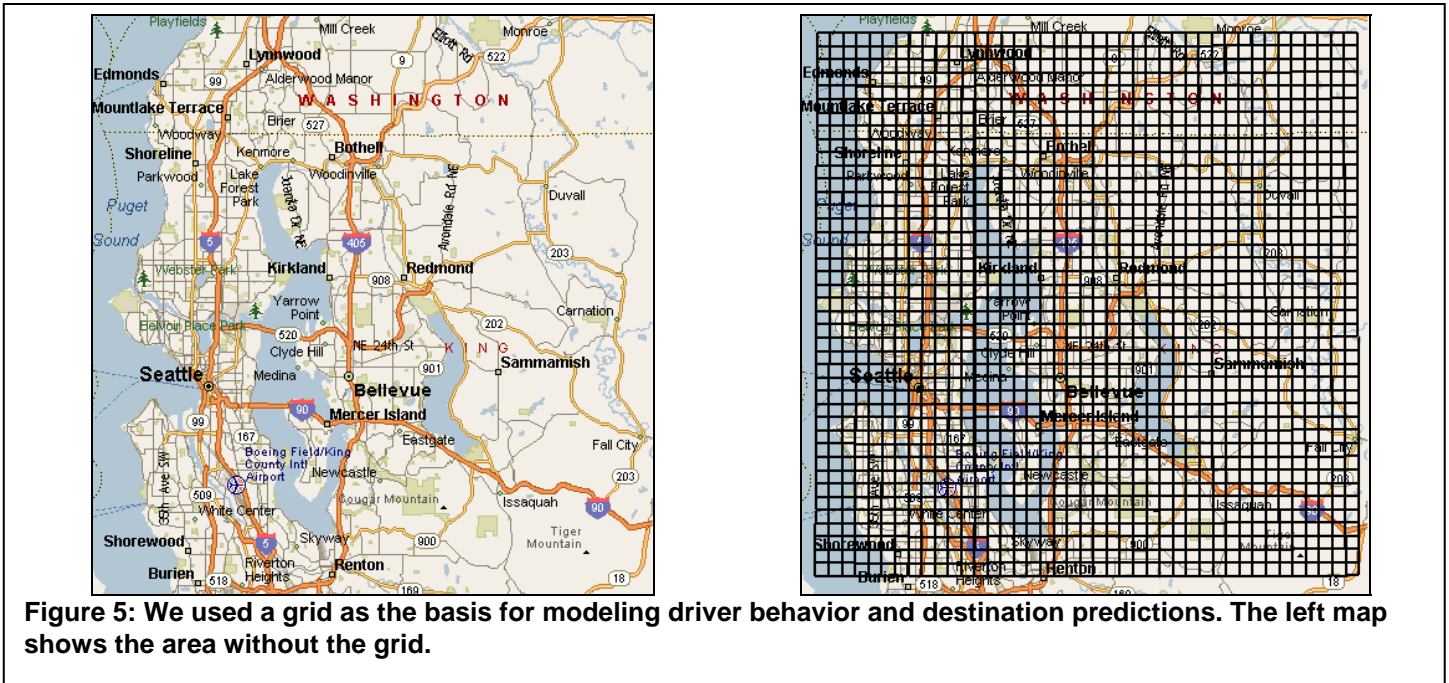


reasonable number. A histogram of the number of trips per day is shown in Figure 4.

## GRID REPRESENTATION

The computational substrate for our assessment of driving behavior and destination predictions is a grid placed over the Seattle area, as shown in Figure 5. This grid is 41 km X 41 km, with each square cell being 1 km on a side. Each cell is represented as an integer index,  $i = 1, 2, 3, \dots, N_c$ , with  $N_c = 1681$  being the total number of cells in our case. We represent each trip through the grid as a sequence of traversed cells, as shown in Figure 6. We convert from a sequence of (latitude, longitude) coordinates to a sequence of cell indices by making a time-ordered list of all the traversed cells. Then we replace all subsequences of repeated cells with a single instance of the repeated cell, giving a list of traversed cells with no adjacent repeats.

Our destination prediction is based on the assumption that drivers chose efficient routes. We quantify efficiency using the driving time between points on the driver's path and candidate destinations. Thus, for each pair of cells  $(i, j)$  in our grid, we estimate the driving time  $T_{i,j}$  between them. A first approximation to the driving time could come from a simple Euclidian distance and speed approximation between each pair of cells. Instead, we used the Microsoft MapPoint desktop mapping software to plan a driving route between the center (latitude, longitude) points of pairs of cells. MapPoint provides a programmatic interface which returns the estimated driving time of planned routes. Using a driving route planner takes into account the road network and speed limits between cells, giving a more accurate driving time estimate. For  $N_c$  cells, there are  $N_c(N_c - 1)$  different ordered pairs, not including pairs of identical cells. Our route planning software plans routes at the rate of about four per second on a 2.8 GHz PC, meaning it would take about 196 hours to plan routes for all  $N_c(N_c - 1) \approx 2.82 \times 10^6$  pairs. We cut this time in half by assuming that the travel time from cell  $i$  to  $j$  is the



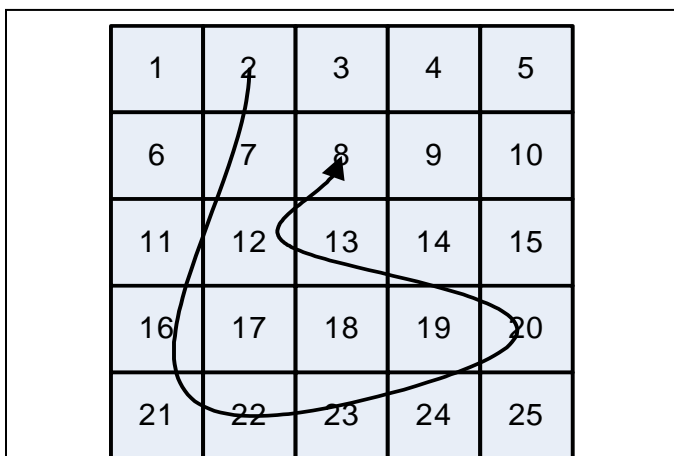
same as from cell  $j$  to  $i$ , i.e.  $T_{i,j} = T_{j,i}$ . The computation time for route planning was the main barrier to increasing the resolution of our grid. Fortunately, this computation must be done only once for the grid. The results can be stored on vehicle's computer.

### DRIVERS' ABILITY FOR EFFICIENT ROUTING

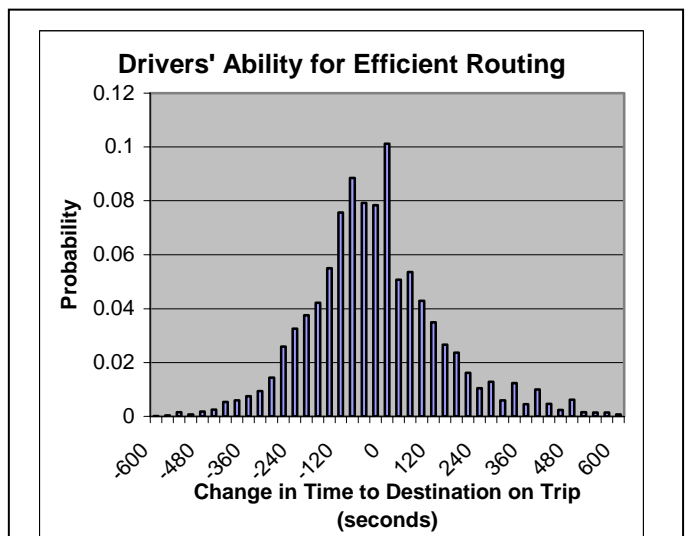
Our intuition says that drivers will not pass up an opportunity to get to their destination quickly. For instance, if a driver comes close to his or her destination at one point during the trip, he or she is unlikely to subsequently drive farther from the destination. In other words, as a trip progresses, we expect the time to the destination to decrease monotonically. We tested this assumption using our trip data. We first converted each trip into a sequence of cells (as explained above) and examined each sequence one cell at a time. As we went

through each sequence, we kept track of the minimum time to the sequence's last cell (the destination cell) encountered so far. An efficient route would reduce this minimum time as the sequence progresses. For each cell transition in the sequence, we computed  $\Delta t$ , the change in estimated driving time achieved by transitioning to the new cell over the minimum time to the destination encountered so far. We would expect this time to be usually negative, meaning that the cell transition reduced the time to the destination. We tested this by computing all the  $\Delta t$ 's for all the cell transitions from our GPS data. The normalized histogram of these  $\Delta t$ 's is shown in Figure 7.

The normalized histogram of  $\Delta t$ 's is an estimate for  $p_{\Delta}(\Delta t)$ , which gives the probability of the change in trip



**Figure 6: Trips are represented as a sequence of cells. On this grid, this trip would be represented as  $S = \{2, 7, 12, 11, 16, 21, 22, 23, 24, 19, 20, 19, 14, 13, 12, 13, 8\}$ . There are no adjacent repeats, although cells 12, 13, and 19 appear twice.**



**Figure 7: As a trip progresses, drivers sometimes reduce the time to their destination (negative time changes) and sometimes increase the time to their destination (positive time changes).**

time that a driver's transition to the next cell will cause, with reference to the closest the driver has been to the destination so far. The probability that the driver will reduce the minimum time to the destination is  $p = \int_{\Delta t < 0} p_{\Delta t}(\Delta t) d\Delta t = 0.625$  based on our data. Surprisingly,

this means that  $1 - p = 0.375$ , or 37.5% of the time, the driver's move to a new cell actually increased the time to the destination. Most of our volunteer drivers live and work within the test grid, so we expected their familiarity with the area would lead to more efficient driving. However, this number may be artificially high due to drivers' specialized knowledge that our route planner didn't have, such as shortcuts, changes in the road network, and traffic conditions. Also, the mean and median of  $p_{\Delta t}(\Delta t)$  are -22.2 seconds and -39.0 seconds, respectively, so on average the data shows that drivers do proceed toward their destinations with each transition to a new cell in the grid.

### DESTINATION PREDICTION

Our goal is to assign a destination probability to each cell in the grid, with higher probabilities meaning more likely destinations. In particular, we want to estimate  $p(c_i|S)$  for  $i=1,2,3,K,N_c$ , where  $c_i$  represents cell  $i$  and  $S = \{s_1, s_2, s_3, K, s_{N_c}\}$  represents the sequence of cells traversed so far. Thus, for any sequence of traversed cells  $S$ ,  $p(c_i|S)$  gives the probability of the destination being cell  $c_i$ .

Applying Bayes rule, we get

$$p(c_i|S) = \frac{p(S|c_i)p(c_i)}{\sum_{j=1}^{N_c} p(S|c_j)p(c_j)}$$

The terms of the right side are:

- $p(S|c_i)$  – the likelihood of the sequence  $S$  given the destination  $c_i$ .
- $p(c_i)$  – the prior probability of cell  $c_i$  being the destination. Since we have no prior bias about the destination, we use  $c_i = 1/N_c$ , giving all the cells the same prior probability.
- $\sum_{j=1}^{N_c} p(S|c_j)p(c_j)$  – a normalization factor to make  $\sum_{i=1}^{N_c} p(c_i|S) = 1$ .

To compute  $p(S|c_i)$ , we compute the probability of the sequence of traversed cells, assuming the driver's ability to move toward the destination is probabilistically independent at each step. This leads to the following simple formula:

$$p(S|c_i) = \prod_{j=2}^N \begin{cases} p & \text{if } s_j \text{ is closer to } c_i \text{ than any previous cell in } S \\ 1-p & \text{otherwise} \end{cases}$$

For each cell transition in the sequence of cells traversed so far, this formula multiplies the probability  $p$  if the transition moved closer to the candidate destination  $c_i$  than the closest point so far in the sequence, otherwise it multiplies  $1-p$  if the transition moved farther away from the candidate destination. As long as  $p > 0.5$  this formula will favor candidate destinations that the driver is more often driving toward.

An example of the results of these computations is shown in the three maps in Figure 8. In the first map, the trip starts at the highlighted square near the middle of the

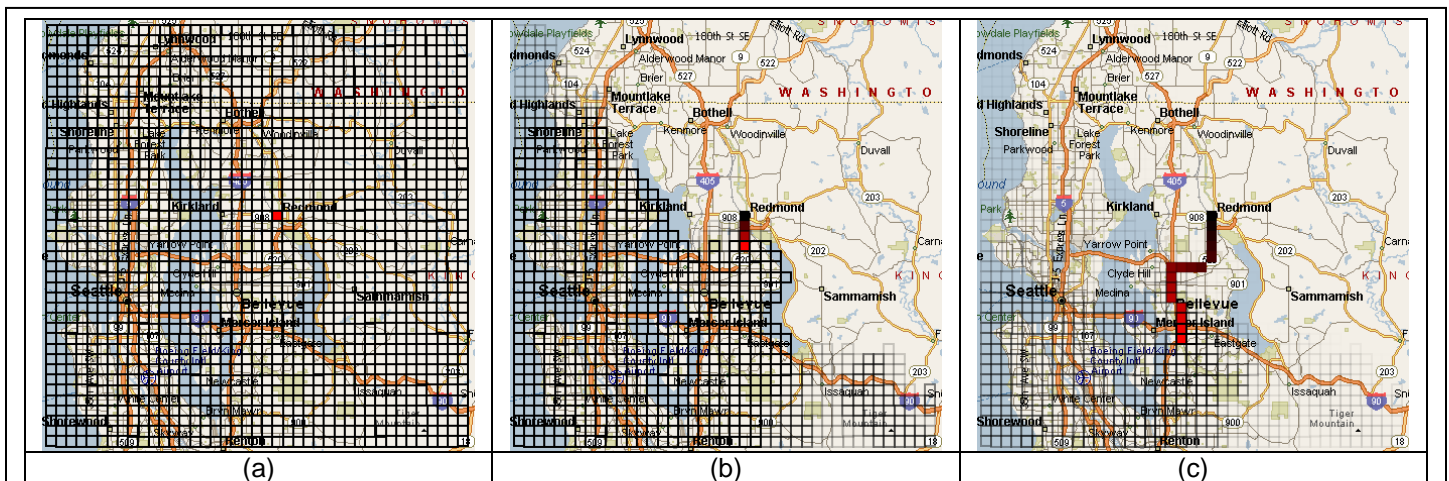


Figure 8: Probabilities of some cells are reduced significantly as a trip progresses. Darker cell outlines represent higher probabilities. In (a), the trip has started at the highlighted cell near the middle of the grid. The probability distribution is uniform. After traveling four cells south in (b), most of the northeast upper triangle is eliminated. After going further south, all but the southwest quadrant is eliminated in (c).

Percentage of Trip Completed	Number of Test Trips	Median Error (km)
25%	756	21.00
50%	1428	10.05
75%	576	3.00
90%	278	3.16

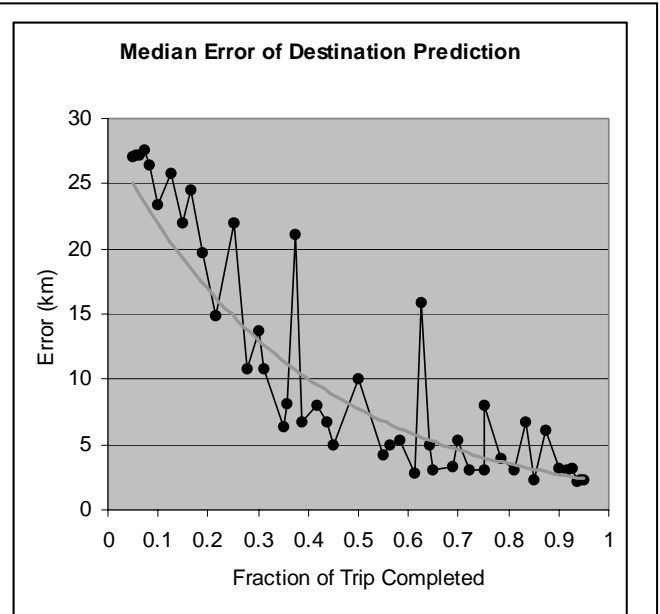
**Table 1: Median error of destination prediction as a function of the percentage of the driving trip completed.**

map. The computed destination probability of all the cells is equal. After traversing four squares south in the second map, nearly the entire northeast upper triangle of the grid is eliminated as a destination. The probabilities computed for these candidate locations are low because the driver has consistently transitioned to cells that would increase the travel time to these cells. After continuing south in the third map, all but the southwest quadrant of the grid is eliminated, because the driver would have likely taken different routes to get to destinations in the other three quadrants.

We tested the accuracy our destination predictions using a randomly selected 20% of our GPS trip data to compute  $p$  (the probability of an efficient transition) and the other 80% to test. The training data led to a value of  $p = 0.684$ , close to  $p = 0.625$  that we computed earlier from all the data. For testing, we computed the destination probability of each cell in the grid for each cell of each of the test trips. To quantify the performance, we computed the median error between the actual destination and the mode (maximum peak) of  $p(c_i|S)$  on the grid of cells. For each test trip, as the trip progresses, we can compute the fraction of the trip that has been completed for each traversed cell. For instance, a 4-cell trip will have fractions  $\{0.00, 0.33, 0.67, 1.00\}$ , and a 5-cell trip will have fractions  $\{0.00, 0.25, 0.50, 0.75, 1.00\}$ . We aggregated destination errors for all the observed trip fractions. For instance, there were 756 test trips where one traversed cell represented exactly 25% of the trip. The median error of the destination prediction for these 756 trips was 21 km. For the 1428 trips at 50%, the median error dropped to 10 km. Table 1 gives some of the results. Results from all the trip fractions with 100 or more representative trips are plotted in Figure 9. As expected, the prediction error drops as the trip progresses, because the driver is passing up more and more destinations.

## CONCLUSION

A simple model of driving efficiency over a known road network is an effective technique for predicting a driver's destination. Our prediction is based on the intuition that drivers will usually follow an efficient route, and we quantified this behavior using actual driving data and computed route times. The ability to predict destinations could be used to filter out irrelevant data presented to drivers, concentrating instead on information concerning places and situations the driver is most likely to encounter during the trip.



**Figure 9: The median error of the predicted destination drops as the trip progresses. The gray curve is a smoothed version of the results.**

There are other sources of information that can be used to make destination predictions, such as a distribution of likely trip times, likely destination types (e.g. drivers rarely end up in lakes), and a history of past destinations. We plan to investigate these other sources and combine them with the driving efficiency method used in this paper to increase the accuracy of our predictions.

## REFERENCES

1. Ashbrook, D. and T. Starner, *Using GPS To Learn Significant Locations and Predict Movement Across Multiple Users*. Personal and Ubiquitous Computing, 2003. 7(5): p. 275-286.
2. Patterson, D.J., et al. *Opportunity Knocks: A System to Provide Cognitive Assistance with Transportation Services*. in *UbiComp 2004: Ubiquitous Computing*. 2004. Nottingham, UK: Springer.
3. Cheng, C., R. Jain, and E.v.d. Berg, *Location Prediction Algorithms for Mobile Wireless Systems*, in *Wireless Internet Handbook: Technologies, Standards, and Applications*. 2003, CRC Press: Boca Raton, FL, USA. p. 245-263.
4. Krumm, J. and E. Horvitz, *The Microsoft Multiperson Location Survey (MSR-TR-2005-103)*. 2005, Microsoft Research.
5. Coyne, P.I., S.J. Casey, and G.A. Milliken, *Comparison of Differentially Corrected GPS Sources for Support of Site-Specific Management in Agriculture*. 2003, Kansas State University Agricultural Experiment Station and Cooperative Extension Service.