When is Big Data Big Enough? Implications of Using GPS-Based Surveys for Travel Demand Analysis

Akshay Vij
Kalyanaraman Shankari

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-141
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-141.html

August 1, 2014

Acknowledgement

This research was funded in part by the Jim Gray Fellowship and in part by the NSF ActionWebs CPS-0931843. We also wish to thank Mogeng Yin, Shanthi Shanmugam and Ryan Lei for their help in developing E-Mission, the smartphone app used by this study for data collection.

When is Big Data Big Enough?
Implications of Using GPS-Based Surveys for Travel Demand Analysis

July 30, 2014

Akshay Vij
University of California at Berkeley
214 McLaughlin Hall
Berkeley, CA 94720-1720
a.vij@berkeley.edu

K. Shankari
University of California at Berkeley
465 Soda Hall
Berkeley, CA 94720-1776
shankari@eecs.berkeley.edu

Abstract

A number of studies in the last decade have argued that GPS-based surveys offer the potential to replace traditional travel diary surveys. GPS-based surveys impose lower respondent burden, offer greater spatiotemporal precision and incur fewer monetary costs. However, GPS-based surveys do not collect certain key inputs required for the estimation of travel demand models, such as the travel mode(s) taken or the trip purpose, relying instead on data-processing procedures to infer this information. This study assesses the impact that errors in inference can have on travel demand models estimated using data from GPS-based surveys. We use simulated datasets to compare performance across different sample sizes, inference accuracies and estimation methods. Findings from the simulated datasets are corroborated with real data collected from individuals living in the San Francisco Bay Area, United States. Results indicate that the benefits of using GPS-based surveys will vary significantly, depending upon the sample size of the data, the accuracy of the inference algorithm and the desired complexity of the travel demand model specification. In many cases, gains in the volume of data that can potentially be retrieved using GPS devices may be offset by the loss in quality caused by inaccuracies in inference. For example, a Monte Carlo experiment finds that a relatively parsimonious model of travel mode choice behavior that could reliably be estimated using 100 high-quality observations could need 10,000 observations and more, depending upon the accuracy of the inference algorithm. This study argues that GPS-based surveys may never entirely replace traditional travel diary surveys. For data from GPS-based surveys to be useful for existing modes of travel demand analysis, it needs either to be supplemented with data collected from traditional surveys or GPS-based surveys need to allow for direct interaction with the study participant. Alternatively, newer modes of analysis need to be developed that can compensate for inaccuracies in data from existing GPS-based surveys.

1. Introduction

Traditional models of individual and household travel and activity behavior are estimated using travel diary datasets that ask a small subset of the population of interest to record over a period of one or two days which activities were conducted where, when, for how long, with whom and using what mode of travel. For example, SF-CHAMP (San Francisco Chained Activity Modeling Process), the state-of-the-art activity-based model of travel demand developed by the San Francisco County Transportation Authority, was estimated using travel diary data from about 30,000 individuals collected over a period of two days (Cambridge Systematics, 2002). The size of that and other similar travel diary datasets pales in comparison to the volume of information that can potentially be retrieved from new technologies, such as Global Positioning System (GPS) sensors and smartphones, and social media platforms, such as Twitter and Facebook, now and in the future. Advances in GPS technologies in particular have received substantive attention in the recent past. Multiple studies have sought to develop GPS-based surveys to collect all the information that is usually collected by extant mail-back, phone-based or door-to-door travel diary surveys, but with very little input from survey participants (for a recent review of the literature, the reader is referred to Shen and Stopher, 2014). Though the use of these surveys thus far has been limited to pilot projects (e.g. Pereira et al., 2013), they are expected eventually to replace traditional travel diary surveys (Wolf et al., 2001).

The benefits of using GPS-based surveys are manifold. They impose fewer requirements on survey respondents, offer greater spatiotemporal precision and are cheaper to implement. Nevertheless, as pointed out by Shen and Stopher (2014), GPS-based surveys "cannot record travel mode, trip purpose or the number of occupants in a private vehicle — all important attributes in a traditional travel survey. Therefore, data-processing procedures become critical to the usefulness of GPS surveys, because there would be insufficient information for travel modelling purposes without the results of the processing." Numerous algorithms have been proposed for inferring one or more of these missing pieces of information from the GPS data, augmented in many cases with additional sources, such as accelerometer readings from smartphones (e.g. Reddy et al., 2008) or land use characteristics from Geographic Information Systems (GIS) databases (e.g. Bohte and Maat, 2009). However, even the most successful inference algorithm will have some error associated with it. For example, most published studies in the literature, including those cited here, report average accuracies of 60-90%. Errors in inference could potentially compromise the quality of data collected through GPS-based surveys and the validity of travel demand models developed using this data. And yet, to the best of our knowledge, no study has systematically examined the implications of using low-quality big data for traditional modes of analyses.

The objective of this study is to evaluate the impact of errors in inference on travel demand models estimated with GPS data. The paper is structured as follows: Section 2 describes a Monte Carlo experiment that compares model performance across different sample sizes, inference accuracies and estimation methods; Section 3 uses validated GPS data collected from individuals residing in the San Francisco Bay Area, United States to corroborate findings from the Monte Carlo experiment; and Section 4 concludes the paper with a summary of findings and implications.

2. Monte Carlo Experiment

In this section, we simulate a Monte Carlo experiment to assess the impact of inference errors on estimation results. A Monte Carlo experiment is especially useful because the true parameters underlying the data generating process are known, and the impact of inference errors can be evaluated under a wide variety of conditions, leading to more generalizable results that aren't specific to any one dataset. Section 2.1 describes how the data is generated. Sections 2.2 and 2.3 compare the parameter values recovered from two different estimation methods.

2.1 Data Generation

To measure the impact of errors in inference on estimation results, we create a two-step Monte Carlo experiment. First, we simulate datasets for a hypothetical Random Utility Maximization (RUM) model of travel mode choice behavior. The RUM model is by far the most popular model among studies on individual and household travel and activity behavior, and travel mode choice behavior perhaps the most widely studied problem (see, for example, Ben-Akiva and Lerman, 1985). Second, we simulate the probability that the chosen travel mode is correctly identified using a hypothetical inference algorithm modeled along the lines of a decision tree. Decision trees are classifiers

developed in machine learning and data mining that have proven to be popular with studies on travel and activity inference using location traces and other data (see, for example, Zheng et al., 2010).

We begin by describing the RUM model of travel mode choice behavior. We assume that for a given trip, a decision-maker can choose between four travel modes: walk, bike, car and transit. The utility of each travel mode is defined as a linear function of the travel time, cost and greenhouse gas emissions incurred by that mode:

$$\mathrm{U_{nj}} = \mathrm{V_{nj}} + \varepsilon_{\mathrm{nj}} = \mathrm{ASC_j} + \beta_{\mathrm{tt}}\mathrm{tt_{nj}} + \beta_{\mathrm{cost}}\mathrm{cost_{nj}} + \beta_{\mathrm{ghg}}\mathrm{ghg_{nj}} + \varepsilon_{\mathrm{nj}}, \quad \varepsilon_{\mathrm{nj}} \sim \mathrm{Gumbel}(0, \pi^2/6) \tag{1}$$

, where $\mathrm{U_{nj}}$ is the utility of travel mode j as perceived by decision-maker n; and $\mathrm{V_{nj}}$ is the systematic component of the utility and $\varepsilon_{\mathrm{nj}}$ is the stochastic component, assumed to be i.i.d. Gumbel with location zero and scale parameter $\pi^2/6$. The systematic component is some function of the variables, $\mathrm{tt_{nj}}$, $\mathrm{cost_{nj}}$ and $\mathrm{ghg_{nj}}$, denoting respectively the travel time, cost and greenhouse gas emissions incurred by travel mode j on decision-maker n, and the model parameters $\mathrm{ASC_j}$, β_{tt}, β_{cost} and β_{ghg}, denoting respectively the alternative-specific constant and the mean sensitivities to travel time, cost and greenhouse gas emissions. Decision-makers are assumed to be utility-maximizing in that they choose that travel mode that offers them the greatest utility:

$$\mathrm{y_{nj}} = \begin{cases} 1; \text{ if } \mathrm{U_{nj}} \geq \mathrm{U_{nj'}} \text{ for } \mathrm{j'} = 1, \dots, \mathrm{J} \\ 0; \text{ otherwise} \end{cases} \tag{2}$$

, where $\mathrm{y_{nj}}$ is an indicator of the actual choice. The assumption that $\varepsilon_{\mathrm{nj}}$ is i.i.d. Gumbel with location zero and scale parameter $\pi^2/6$ results in the familiar multinomial logit expression for the choice model:

$$\mathrm{P}\left(\mathrm{y_{nj}} = 1 | \mathbf{ASC}, \beta_{\mathrm{tt}}, \beta_{\mathrm{cost}}, \beta_{\mathrm{ghg}}; \mathbf{tt_n}, \mathbf{cost_n}, \mathbf{ghg_n}\right)$$

$$= \frac{\exp\left(\mathrm{ASC_j} + \beta_{\mathrm{tt}}\mathrm{tt_{nj}} + \beta_{\mathrm{cost}}\mathrm{cost_{nj}} + \beta_{\mathrm{ghg}}\mathrm{ghg_{nj}}\right)}{\sum_{\mathrm{j'}} \exp\left(\mathrm{ASC_{j'}} + \beta_{\mathrm{tt}}\mathrm{tt_{nj'}} + \beta_{\mathrm{cost}}\mathrm{cost_{nj'}} + \beta_{\mathrm{ghg}}\mathrm{ghg_{nj'}}\right)} \tag{3}$$

We describe the process of generating synthetic datasets for the model framework over the subsequent paragraphs. In terms of the variables, we employ the following distributions:

$$\mathrm{tt_{walk}} \sim \mathcal{U}(1.5\mathrm{tt_{car}}, 2.5\mathrm{tt_{car}}), \quad \mathrm{cost_{walk}} = 0, \quad \mathrm{ghg_{walk}} = 0 \tag{4}$$

$$\mathrm{tt_{bike}} \sim \mathcal{U}(\mathrm{tt_{car}}, 1.5\mathrm{tt_{car}}), \quad \mathrm{cost_{bike}} = 0, \quad \mathrm{ghg_{bike}} = 0 \tag{5}$$

$$\mathrm{tt_{car}} \sim \mathcal{U}(10, 50), \quad \mathrm{cost_{car}} \sim \mathcal{U}(0, 20), \quad \mathrm{ghg_{car}} \sim 0.89\left(\frac{\mathrm{tt_{car}}}{60}\right)\mathcal{U}(40, 60) \tag{6}$$

$$\mathrm{tt_{transit}} \sim \mathcal{U}(\mathrm{tt_{car}}, 2.5\mathrm{tt_{car}}), \quad \mathrm{cost_{transit}} \sim \mathcal{U}(0, 4), \quad \mathrm{ghg_{transit}} \sim 0.28\left(\frac{\mathrm{tt_{transit}}}{60}\right)\mathcal{U}(20, 30) \tag{7}$$

, where travel time is measured in minutes, cost in dollars and greenhouse gas emissions in pounds of CO_2 equivalent; and $\mathcal{U}(a, b)$ denotes a continuous uniform distribution over the range (a, b). The parameter values are enumerated below:

$$\mathrm{ASC_{walk}} = 0, \quad \mathrm{ASC_{bike}} = -4.5, \quad \mathrm{ASC_{car}} = 0.5, \quad \mathrm{ASC_{transit}} = -1 \tag{8}$$

$$\beta_{\mathrm{tt}} = -0.30, \quad \beta_{\mathrm{cost}} = -0.45, \quad \beta_{\mathrm{ghg}} = -0.10 \tag{9}$$

Following the methodology proposed by Williams and Ortúzar (1982) and the approach outlined by Raveau et al. (2010), values for each of the model parameters are chosen such that: (1) the marginal rates of substitution between travel time, cost and greenhouse gas emissions are consistent with values observed by studies in the literature; (2) the part-worth utilities of each of the explanatory variables are comparable in terms of magnitude; (3) the model is identifiable; and (4) the error in the data is roughly 25%, i.e. one in four simulated decision-makers change their

choice because of the stochastic component, thereby ensuring that the decision-making process is neither deterministic nor completely stochastic.

Equations (1)-(9) provide a blueprint for generating any number of observations. However, for any given observation, the travel mode *actually* chosen, as identified by equation (2), will not always be the same as the travel mode *inferred* to have been chosen. Over the course of the subsequent paragraphs, we describe how we simulate errors in travel mode inference. Let y'_{nj} be an indicator of the inferred choice, i.e. y'_{nj} equals one if decision-maker n is inferred to have chosen travel mode j, and zero otherwise. We assume that the classifier used for inference employs a decision tree learning algorithm trained on a single feature - the average speed across the trip, denoted r_n for the trip made by decision-maker n. The average speed is assumed to have the following lognormal distributions conditional on the chosen travel mode (the parameters for the distributions were estimated using real data), and draws are taken for each simulated decision-maker:

$$r_n \sim \begin{cases} \ln \mathcal{N}(0.28, 0.43)\,; \text{ if } y_{n,walk} = 1 \\ \ln \mathcal{N}(1.38, 0.38)\,; \text{ if } y_{n,bike} = 1 \\ \ln \mathcal{N}(2.05, 0.63)\,; \text{ if } y_{n,car} = 1 \\ \ln \mathcal{N}(1.89, 0.79)\,; \text{ if } y_{n,transit} = 1 \end{cases} \tag{10}$$

For some desired value for the average accuracy of the decision tree, we tune two parameters: the maximum depth of the tree and the minimum number of samples required at any leaf, until the accuracy for the sample is within 1% of the desired value. Once we have an acceptable classifier, we use it to infer the travel mode most likely to have been used for each trip in the sample. In all, 100 datasets each are generated for 100, 1000 and 10000 pseudo-observed decision-makers hypothesized to behave according to the decision-making process described above, and average accuracies of the inference algorithm between 60% and 100%, implemented in 5% increments, resulting in a total of 2700 datasets. Over the following subsections, we describe our attempts to recover estimates for the model parameters for each of these datasets, and how the estimates compare with the true values.

2.2 Maximum Likelihood Estimation

For each of the 2700 datasets, estimates for the model parameters are recovered by maximizing the following likelihood function for the multinomial logit model:

$$\begin{aligned} &L\left(\mathbf{ASC}, \beta_{tt}, \beta_{cost}, \beta_{ghg} | \mathbf{y'}, \mathbf{tt}, \mathbf{cost}, \mathbf{ghg}\right) \\ &= \prod_n P\left(\mathbf{y'_n} | \mathbf{ASC}, \beta_{tt}, \beta_{cost}, \beta_{ghg}; \mathbf{tt_n}, \mathbf{cost_n}, \mathbf{ghg_n}\right) \\ &= \prod_n \prod_j \left[P\left(y_{nj} = 1 | \mathbf{ASC}, \beta_{tt}, \beta_{cost}, \beta_{ghg}; \mathbf{tt_n}, \mathbf{cost_n}, \mathbf{ghg_n}\right)\right]^{y'_{nj}} \\ &= \prod_n \prod_j \left[\frac{\exp\left(ASC_j + \beta_{tt} tt_{nj} + \beta_{cost} cost_{nj} + \beta_{ghg} ghg_{nj}\right)}{\sum_{j'} \exp\left(ASC_{j'} + \beta_{tt} tt_{nj'} + \beta_{cost} cost_{nj'} + \beta_{ghg} ghg_{nj'}\right)}\right]^{y'_{nj}} \end{aligned} \tag{11}$$

, where $\mathbf{y'_n}$ is a $(J \times 1)$ vector whose j^{th} element is y'_{nj}. All models are estimated in Python using an implementation of the BFGS algorithm contained in the SciPy library (Jones et al., 2001). We assess the impact of errors in inference by comparing estimates for the value of time, defined as the ratio of the estimates for β_{tt} and β_{cost}, and the value of green, defined as the ratio of the estimates for β_{ghg} and β_{cost}, with the true values for these willingness-to-pay measures, assumed implicitly by equation (9) to be 40$/hr and 22¢/lb, respectively. For each of the 100 datasets belonging to a particular combination of the number of observations and the average accuracy of the inference algorithm, Figure 1 plots the mean and standard error in the estimates for these measures. For example, the left most point in the top-left plot denotes the mean estimate of value of time across the 100 datasets with 100 observations each and an average inference accuracy of 60%. For the model specification at hand, the bias in estimates

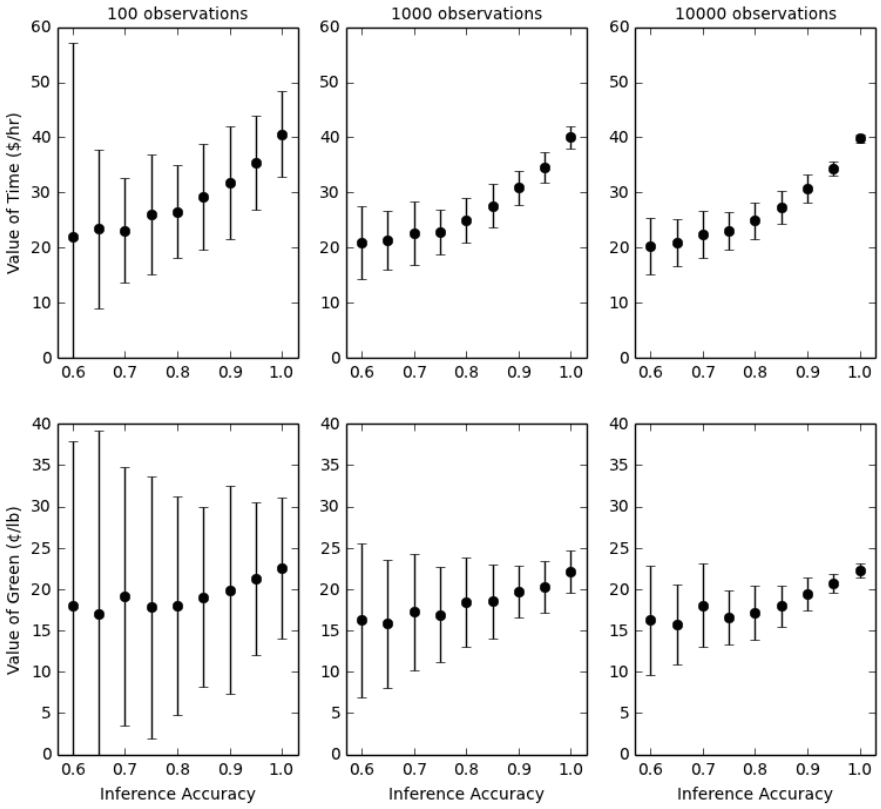

Figure 1: A plot of the mean and standard error in the value of time ($/hr) and value of green (¢/lb) estimated by maximizing the likelihood function given by equation (11) for each of the 100 datasets belonging to a particular combination of the number of observations and the average accuracy of the inference algorithm

expectedly decreases as the average accuracy of the inference algorithm increases, eventually converging to the true values at an accuracy of 100%. However, the bias appears to be independent of the number of observations, though this could be an anomaly arising out of the particulars of the RUM model used to generate the data. What's even more interesting though is the magnitude of the bias. The mean estimate for value of time for 10000 observations and an inference accuracy of 80% is 24.8$/hr, off by nearly 40% from the true value of 40$/hr. Even for higher accuracies, such as 95%, the mean estimate for 10000 observations is 34.3$/hr, off by 14%. Estimates for the value of green appear to be somewhat more robust. The mean estimates for value of green for 10000 observations and inference accuracies of 80% and 95% are 17.2¢/lb and 20.7¢/lb, off by 22% and 6%, respectively, from the true value of 22¢/lb. For lower inference accuracies, such as 60%, the mean estimates are off by as much as 49% in the case of value of time and 26% in the case of value of green.

As has been shown by numerous studies across statistics and econometrics, a purely stochastic measurement error in the dependent variable does not bias estimation results for probabilistic models, though there is a loss in statistical efficiency (see, for example, Greene, 2002). This is because the relationship between the dependent variable and the independent variables has a stochastic component built into it, as represented by ε in equation (1). However, errors in inference are usually systematic and likely some function of the dependent variable itself. For example, in the context of travel mode inference, trips made on foot are easier to identify than trips made by bike, car or public transit (see, for example, Zheng et al., 2010). As a consequence, parameter estimates are likely to be biased, even when working with large datasets and high inference accuracies. In general, the magnitude of bias will vary from dataset to dataset. The purpose of the Monte Carlo experiment is not to determine an absolute range, but to demonstrate the variability in estimates that can be expected from inaccuracies in inference.

2.3 Maximum Weighted Likelihood Estimation

As mentioned in Section 2.1, classifiers such as decision trees predict for a given observation the probability of occurrence associated with every possible outcome. The outcome that has the greatest probability is assumed to have taken place and the probability distribution across outcomes is subsequently discarded. In doing so, the analyst is introducing measurement error into the choice model, resulting in biased parameter estimates, as evidenced by Section 2.2. Readers familiar with discrete choice models will recognize the analogy with sequential estimation in the context of Integrated Choice and Latent Variable (ICLV) models or Latent Class Choice Models (LCCMs). In the case of these models, mean estimates for latent variables obtained using standard estimators in the first step, such as principal components analysis or cluster analysis, are treated as observable explanatory variables in the discrete choice model in the second step. Ignoring the measurement error associated with the predicted estimates from the first step when estimating parameters in the second step can bias these estimates (see Ben-Akiva et al., 2002 or Walker and Li, 2007 for a discussion on the subject). For these reasons, ICLV models and LCCMs are usually estimated simultaneously. In our case, the solution may be translated as maximizing the following weighted likelihood function, assuming that the features used to train the classifier are either independent of the explanatory variables used in the choice model they or do not affect the choice outcome:

$$
\begin{aligned}
&\mathrm{L}\left(\mathbf{ASC}, \beta_{\mathrm{tt}}, \beta_{\mathrm{cost}}, \beta_{\mathrm{ghg}} | \mathbf{tt}, \mathbf{cost}, \mathbf{ghg}, \mathbf{r}\right) \\
&= \prod_{\mathrm{n}} \sum_{\mathrm{j}} \mathrm{P}\left(\mathrm{y}'_{\mathrm{nj}} = 1 | \mathrm{r}_{\mathrm{n}}\right) \mathrm{P}\left(\mathbf{y_n} | \mathbf{ASC}, \beta_{\mathrm{tt}}, \beta_{\mathrm{cost}}, \beta_{\mathrm{ghg}}; \mathbf{tt_n}, \mathbf{cost_n}, \mathbf{ghg_n}, \mathrm{y}'_{\mathrm{nj}} = 1\right) \\
&= \prod_{\mathrm{n}} \sum_{\mathrm{j}} \mathrm{P}\left(\mathrm{y}'_{\mathrm{nj}} = 1 | \mathrm{r}_{\mathrm{n}}\right) \mathrm{P}\left(\mathrm{y}_{\mathrm{nj}} = 1 | \mathbf{ASC}, \beta_{\mathrm{tt}}, \beta_{\mathrm{cost}}, \beta_{\mathrm{ghg}}; \mathbf{tt_n}, \mathbf{cost_n}, \mathbf{ghg_n}\right)
\end{aligned}
\tag{12}
$$

, where $\mathrm{P}\left(\mathrm{y}'_{\mathrm{nj}} = 1 | \mathrm{r}_{\mathrm{n}}\right)$ is the probability that decision-maker n chose travel mode j, as predicted by the inference algorithm. Note that as the average inference accuracy approaches 100%, equation (12) converges to equation (11). Ideally, the analyst should jointly estimate the inference and choice model but for practical reasons that isn't always possible. It is usually not straightforward to recast classifiers employed for inference as probabilistic models that can subsequently be estimated using maximum likelihood estimation. Most classifiers continue to be estimated using metrics derived from information theory that differ from the probabilistic framework employed by models of travel and activity behavior. But more importantly, the inference algorithm is often trained on a dataset where the explanatory variables used in the choice model are not available and/or collected. The alternative, as represented by equation (12), is to estimate the inference model independently, and to treat the outcomes predicted by the inference model as stochastic variables in the choice model, marginalizing over the estimated distribution. Such an approach leads to consistent but inefficient estimates (Ben-Akiva et al., 2002) in cases where it can be assumed that the features used to train the inference model are either independent of the explanatory variables used in the choice model or they don't affect the choice outcome, as in the case of the Monte Carlo experiment described here. If the two sets of variables are correlated and the features do influence observed choice, as will likely be the case in reality, then estimates from the choice model may still suffer from omitted variable bias.

Estimates for the unknown parameters were recovered by maximizing the weighted likelihood function using the same optimization routines as in Section 2.2. Similar to the analysis in that sub-section, Figure 2 plots the mean and standard error of the estimates for value of time and value of green recovered from each of the 100 datasets belonging to a particular combination of the number of observations and the average accuracy of the inference algorithm. The vertical lines running through the plots indicate large standard errors in the parameter estimates, indicating a failure in the optimization routine to consistently recover unbiased estimates for either measure. However, the frequency with which these vertical lines appear decreases as both the number of observations and the accuracy of inference increase, and for large datasets with high inference accuracies the optimization routine is able to recover unbiased estimates for both the value of time and the value of green. For example, with 100 observations anything less than complete accuracy is unable to recover parameter estimates consistently, but with 10000 observations the parameter estimates can be recovered consistently with average accuracies of 85% and above.

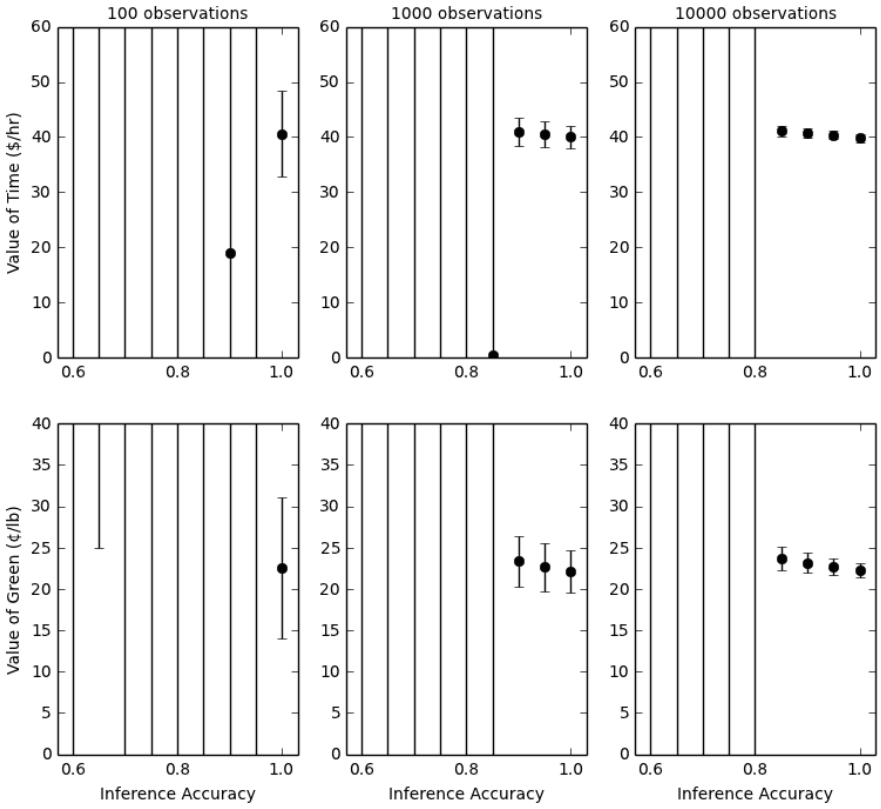

Figure 2: A plot of the mean and standard error in the value of time ($/hr) and value of green (¢/lb) estimated by maximizing the weighted likelihood function given by equation (12) for each of the 100 datasets belonging to a particular combination of the number of observations and the average accuracy of the inference algorithm

These results indicate that increases in the amount of data that can potentially be retrieved using newer technologies are often offset by the loss in quality incurred by inaccuracies in inference. For example, Figure 2 suggests that the same information that could reliably be retrieved from 100 high-quality observations could potentially need 10,000 observations and more, depending upon the accuracy of inference and the consequent quality of data. Conversely, one could argue that with sufficiently large datasets, e.g. location data retrieved from sparse sources such as cellphone towers that cannot provide the same accuracy in inference as denser sources such as GPS sensors, the parameters of interest could still be recovered with a high degree of precision. As before, the purpose of this analysis is not to develop normative guidelines on when to use low quality big data and when to forego it in favor of high quality 'small' data, but to illustrate the kinds of issues that might be encountered when working with these larger datasets, and the ways in which they can be controlled for, if only partially.

3. Case Study: GPS-based Survey in the San Francisco Bay Area, United States

In this section, we corroborate findings from Section 2 using real data collected from 45 smartphone users living in the San Francisco Bay Area, United States through the means of an app called E-Mission. The app is being developed by a team of researchers at the University of California, Berkeley. One of the objectives of E-Mission is to collect all the information that is usually collected by travel diary surveys, but with minimal input from the smartphone user. For more details about the app, the reader is referred to Shankari et al. (2014). For the purpose of our analysis, we will be limiting our attention to trip data. For a given trip, E-Mission records two pieces of information: the series of raw location traces that constitute the trip, used for inferring the travel mode(s) taken; and the travel mode chain that was actually used by the smartphone user to make the trip, used to validate the inference. In all, data from 3381 trips collected over a three-month period in 2014 is used for our analysis. First, for each trip in

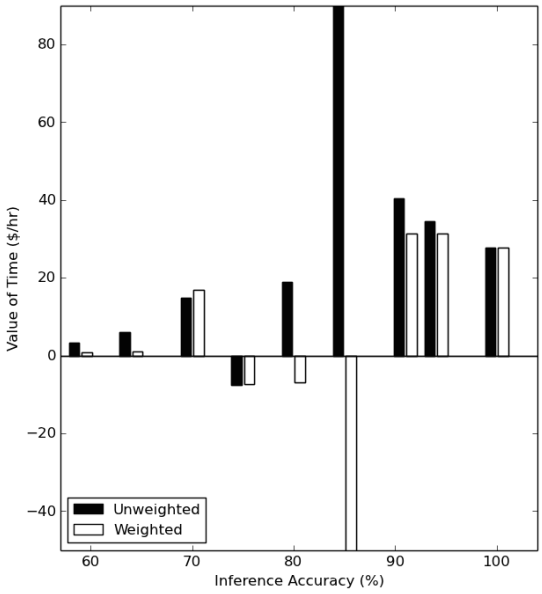

Figure 3: A plot comparing the value of time ($/hr) as estimated by maximizing the unweighted and weighted likelihood functions given by equations (11) and (12), respectively, as a function of the average accuracy of the inference algorithm

the dataset, we use the location traces to construct a vector of features comprising trip attributes such as speed, acceleration, heading change, etc. Second, we employ different subsets of the full set of features and the ground truth collected by E-Mission to train decision tree learning algorithms for travel mode inference with average accuracies between 60% and 100%. Third, for each trip in the sample, we generate the predicted outcome and the predicted probability distribution across all outcomes, as predicted by each of the decision trees trained in the previous step (we do not split the data into training and test sets, using the classifier trained on trips where we know the outcomes to predict outcomes for those same trips as if we didn't know what the outcomes were[1]). Fourth, we derive the travel times and costs incurred by different travel modes for all trips in the sample using skims from the San Francisco Metropolitan Transportation Commission (SF MTC). And finally, we use the predictions from the inference algorithms and the level-of-service attributes derived from the skims to estimate multinomial logit models of travel mode choice, using both maximum likelihood estimation and maximum weighted likelihood estimation. For each trip, the decision-maker is hypothesized to have at most four travel modes to choose from: walk, bike, car and public transit, and the systematic component of the utility of each travel mode is defined as a linear function of the travel time and cost incurred by that travel mode.

Figure 3 plots estimates for the value of time, as recovered by the multinomial logit model from both maximum likelihood estimation and maximum weighted likelihood estimation, as a function of the average accuracy of the decision tree used for travel mode inference. The value of time estimated when then inference algorithm has 100% accuracy is 27.6$/hr. For our analysis, we will treat this as the true value, using it as a baseline when calculating bias. Figure 3 reveals a number of key trends. As the average accuracy of the inference algorithm increases, the magnitude of bias tends to decrease for both estimation methods. At lower average accuracies, the value of time recovered from maximum likelihood estimation is closer to the true value than that recovered from maximum weighted likelihood estimation, consistent with the high standard errors observed at lower average accuracies for the latter in the Monte Carlo experiment. At higher average accuracies, maximum weighted likelihood estimation performs better than maximum likelihood estimation, consistent once again with findings from the Monte Carlo experiment. However, both estimation procedures fall apart when the average accuracy is 85%, with the value of

[1] We do not concern ourselves with overfitting because our objective here is not to train a classifier that can subsequently be used for prediction.

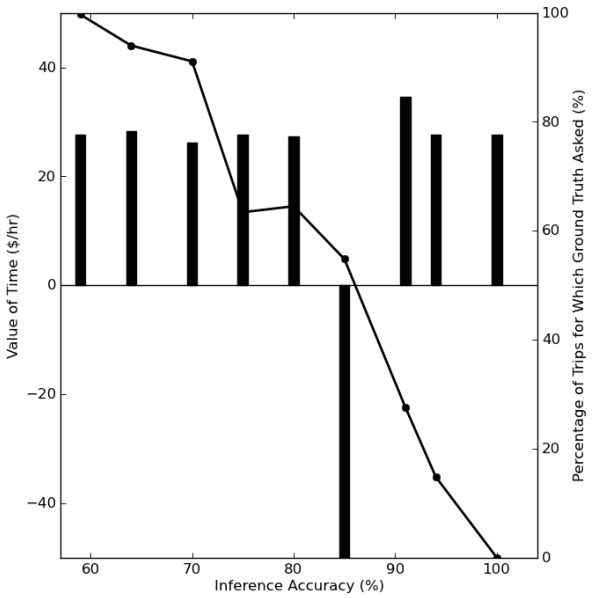

Figure 4: A bar and line plot where the bar plot represents the value of time ($/hr) estimated by maximizing the weighted likelihood function given by equation (12) and the inferred data is supplemented with ground truth, and the line plot represents the percentage of trips for which the accuracy of the inference algorithm is below 90% and ground truth is used, plotted as a function of the average accuracy of the inference algorithm

time tending towards positive infinity for maximum likelihood estimation and negative infinity for maximum weighted likelihood estimation. This is likely due to incorrect inferences for a handful of trips that appear to exert a disproportionate influence on the estimation results. We could have removed these trips from our analysis, as would be the course to take in reality, but we decided to include them in our analysis to illustrate potential problems that the analyst may face when working with low-quality data. In general, the choice between the two estimation methods comes down to a tradeoff between bias and variance: maximum likelihood estimation yields efficient but inconsistent estimates, whereas maximum weighted likelihood estimation provides consistent but inefficient estimates. For a given dataset, the appropriate method will depend upon the sample size, the accuracy of the inference algorithm and the desired complexity of the travel demand model specification. In our case, for a sample size of 3381 trips and a relatively simple travel mode choice model specification, our results suggest that maximum likelihood estimation ought to be preferred when the average accuracy of the inference algorithm is below 85%, and maximum weighted likelihood estimation ought to be preferred when the average accuracy is above 85%.

Regardless of the chosen estimation method, even at high average accuracies the bias in estimates is sizeable. For example, at an average accuracy of 95%, the values of time recovered by maximum likelihood estimation and maximum weighted likelihood estimation are 34.5$/hr and 31.3$/hr, respectively, off by 25% and 13% from the true value, respectively. These results serve as a cautionary warning against the use of low-quality big data for travel demand analysis. Unless the average accuracy of the inference algorithm is close to 100%, the magnitude of bias in parameter estimates may render the use of such data for model development infeasible. In practice, no inference algorithm will ever be 100% accurate, and data collected passively through mobile sensors or social media platforms may always need to be augmented with ground truth for it to be usable. However, as the average accuracy of the inference algorithm deteriorates, the number of observations for which ground truth is needed will increase, as will the consequent burden on study participants. To get an estimate of the trade-off between estimation accuracy and participant burden, we perform the following experiment with each of the inference algorithms trained previously. For trips where the outcome predicted by the inference algorithm has a probability of occurrence abo ve 90%, we use the predicted probability distribution in calculating the weighted likelihood function. For trips where the outcome predicted by the inference algorithm has a probability of occurrence under 90%, we use the ground truth in calculating the weighted likelihood function, assuming that in these cases the study participant can be asked what they did, as they would be in a traditional travel diary survey. Figure 4 plots both the value of time recovered

from maximizing the weighted likelihood function thus constructed, and the percentage of observations for which study participants would potentially be required to provide ground truth under the scheme described above, as a function of the average accuracy of the inference algorithm. As is apparent from the plot, even at low accuracies, estimates are within 2-3% of the true value (though again, at an average accuracy of 85%, the estimation routine breaks down). However, as the average accuracy decreases, the burden on study participants increases. For example, at low average accuracies between 60% and 70%, using 90% as the threshold to determine when to ask for ground truth and when to rely on the inference algorithm, we would need to ask for ground truth for more than 80% of the trips. However, at higher accuracies, the burden is more acceptable. For example, when the average accuracy of the inference algorithm is 94%, and the threshold is still 90%, ground truth is needed for only 15% of the trips, but the bias in our estimate for value of time is less than 0.1%. In general, the analyst can decide upon an appropriate threshold for the partial collection of ground truth data based on the levels of participant burden and errors in estimation that are acceptable given the objectives of the study.

4. Conclusions

The last few years have been witness to great excitement over big data and its potential to address a multitude of societal problems, within transportation engineering and without, on an unprecedented scale and level of detail. The National Science Foundation's recent call for research proposals on "Critical Techniques and Technologies for Advancing Big Data Science and Engineering", the Transportation Research Board's call for papers last year on "Big Data, ICTs, and Travel Demand Models" and Transportation Research Part C's recent call for papers on "Big Data in Transportation and Traffic Engineering" reflect some of the ongoing interest. With regards to travel demand analysis, attention has centered on the development of fully automated GPS-based surveys that can allow for the collection of travel diary data from a greater subset of the population over a longer period of time at a fraction of the cost incurred by more traditional survey methods. The collection of richer travel diary datasets can lead to significant advances in our understanding of travel behavior and consequently, our ability to design transportation systems that serve the immediate needs of the population and satisfy long-term societal objectives.

GPS-based surveys record an individual's location over time, augmented in some cases by information from additional sensors, such as accelerometers and Wi-Fi devices. However, certain vital inputs to the travel demand modeling process, such as the travel mode(s) taken by the individual to make a trip or the purpose of the trip, must necessarily be inferred from this data. Errors in inference can compromise the quality of the data thus collected, raising questions about the validity of travel demand models estimated using this data. In an attempt to address these questions, this study examined the impact that errors in inference can have on estimation results. We used simulated datasets to compare performance across different sample sizes, inference accuracies and estimation methods. Findings were corroborated using real data collected from smartphone users living in the San Francisco Bay Area, United States. Results indicate that the benefits of using GPS-based surveys will vary significantly, depending upon the sample size of the data, the accuracy of the inference algorithm and the desired complexity of the travel demand model specification. If the data is truly big enough, the quality of inference may not matter. But in many cases, gains in volume could potentially be neutralized by losses in quality. For example, a Monte Carlo experiment finds that a relatively parsimonious model of travel mode choice behavior that could reliably be estimated using 100 high-quality observations could need 10,000 observations and more, depending upon the accuracy of the inference algorithm. In practice, no algorithm will ever guarantee complete accuracy. For data from GPS-based surveys to still be useful for travel demand analysis, it will need either to be incredibly big, or to be supplemented with data collected from traditional survey methods that require direct interaction with the study participant.

Acknowledgements

This research was funded in part by the Jim Gray Fellowship and in part by the NSF ActionWebs CPS-0931843. Jim Gray, who did pioneering work on the management of large amounts of data, disappeared while sailing in the San Francisco Bay in 2007. We hope that he would have found this exploration into data sizes and accuracies interesting. We also wish to thank Mogeng Yin, Shanthi Shanmugam and Ryan Lei for their help in developing E-Mission, the smartphone app used by this study for data collection.

References

Ben-Akiva, M., Walker, J. L., Bernardino, A. T., Gopinath, D. A., Morikawa, T., and Polydoropoulou, A. (2002), "**Integration of choice and latent variable models**," in: Hani S. Mahmassani (ed.): *In perpetual motion: Travel behavior research opportunities and application challenges*, Elsevier, Amsterdam, pp. 431–470.

Bohte, W., and Maat, K. (2009), "**Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands**," *Transportation Research Part C: Emerging Technologies*, Vol. 17, No. 3, pp. 285-297.

Cambridge Systematics (2002), "**San Francisco Travel Demand Forecasting Model Development: Executive Summary**," prepared for San Francisco County Transportation Authority.

Greene, W. H. (2003), "**Econometric analysis**," Pearson Education India.

Jones, E., Oliphant, T., Peterson, P., and others (2001), "**SciPy: Open source scientific tools for Python**," http://www.scipy.org/.

McFadden, D. (1986), "**The choice theory approach to marketing research**," *Marketing Science*, Vol. 5, No. 4, pp. 275-297.

Pereira, F., Carrion, C., Zhao, F., Cottrill, C. D., Zegras, C., & Ben-Akiva, M. (2013), "**The Future Mobility Survey: Overview and preliminary evaluation**," In *Proceedings of the Eastern Asia Society for Transportation Studies*, Vol. 9.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010), "**Using mobile phones to determine transportation modes**," *ACM Transactions on Sensor Networks (TOSN)*, Vol. 6, No. 2, pp. 13:1-13:27.

Shankari, K., Yin, M., Shanmugam, S., Culler, D., and Katz, R. (2014), "**E-Mission: Automated transportation emission calculation using smart phones**," *Tech. rep. UCB/EECS-2014-140*. EECS Department, University of California, Berkeley. url: http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-140.html.

Shen, L., and Stopher, P. R. (2014), "**Review of GPS travel survey and GPS data-processing methods**," *Transport Reviews*, Vol. 34, No. 3, pp. 316-334.

Walker, J. L., and Li, J. (2007), "**Latent lifestyle preferences and household location decisions**," *Journal of Geographical Systems*, Vol. 9, No. 1, pp. 77-101.

Wolf, J., Guensler, R., and Bachman, W. (2001), "**Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data**," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1768, No. 1, pp. 125-134.

Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W. Y. (2010), "**Understanding transportation modes based on GPS data for web applications**," *ACM Transactions on the Web (TWEB)*, Vol. 4, No. 1, pp. 1:1-1:36.