



**Using GPS Data to Investigate and Adjust for
Household Diary Data Non-Response**

**PRESENTED AT THE
10th TRB Planning Applications Conference
Portland, Oregon**

April 2005

Dr. Stacey Bricka, NuStats
Austin, Texas

Dr. Jean Wolf, GeoStats
Atlanta, Georgia

Mr. Mark Bradley, Bradley Research and
Consulting
Santa Barbara, California

PTV NuStats LLC
206 Wild Basin Road
Building A, Suite 300
Austin, TX 78746

Phone: 512.306.9065 x 2261

Fax: 512.306.9077

E-mail: mmcaffrey@nustats.com

**Paper for the 10th TRB Planning Applications Conference
Portland, Oregon
April 2005**

Using GPS Data to Investigate and Adjust for Household Diary Data Non-Response

Mark Bradley, Bradley Research and Consulting, Santa Barbara, CA
Jean Wolf, GeoStats, Atlanta, GA
Stacey Bricka, NuStats, Austin, TX

1. Introduction

A number of papers have been written in recent years looking at the differences between travel patterns reported in household travel/activity diary surveys and corresponding travel patterns captured for those same persons using vehicle-based GPS devices (Bachu, et al. 2001; Casas and Arce 1999; Pearson 2001; Pierce, et al. 2003; Wolf, et al. 2003b, Zmud and Wolf 2003; NuStats and Battelle 2004). This paper follows a similar approach using data from a very recent household survey in the Kansas City region. In this paper, however, we dig deeper into the data using various modeling methods, with the intention of being able to say more about two issues in particular:

- Reasons why diary-based data and GPS-based data may show different numbers of trips for the same types of individuals and households; and
- What those differences reflect about the strengths and weaknesses of using vehicle-based GPS data, either as a way of validating/adjusting diary-based data, or as a substitute for diary-based data.

2. Discussion of the issues

Suppose that the members of a given household report all of the trips that they make during a specific day in a travel diary format. This data is collected post-hoc via a CATI telephone survey. On that same travel day, GPS devices connected to each of that household's vehicles also collect data on the location and movement of those vehicles at each moment of the day. This GPS data is later downloaded and analyzed to identify trip start and end times and start and end locations. In a perfect world, all of the self-reported vehicle trip start and end times and locations reported in the travel/activity diaries would also be identified (matched) in the GPS data. Furthermore, in a perfect world, validation measures based on the match between diary data trips and GPS data trips for those in the GPS sample would be transferable to those diary respondents who are not in the GPS sample—if the diary data appears accurate for the GPS respondents, it can be assumed to be accurate for all diary respondents.

In the real world, however, there are various possible reasons why the two types of data may differ within the GPS sample, and why the similarities/differences found for those within the

GPS sample may not hold for other respondents outside the GPS sample. There are three types of bias that are most likely to influence these issues, as outlined below.

Non-response bias in completing the diary

It is generally recognized that people who complete travel/activity diary surveys may not report all of their trips, and that certain types of trips and individuals are more prone to this type of non-response bias than others. This is the type of bias that GPS data collection has been aimed at investigating and correcting. Differences between GPS and diary data that are related to this type of bias can take several forms and have various causes, as the following example illustrates.

Suppose that the GPS record for a certain vehicle includes the following trips/stops, where a trip end is specified as any place that the vehicle stops for more than two minutes. The table and Figure 1 below show a schematic diagram of these trips (though GPS data is capable of reproducing more detail, including a mapping of the specific route chosen, as well as the travel time and speed on each link).

From place	Leave at	Dir. of travel	To Place	Arrive at
Home	8:02	N	B	8:10
B	8:13	N	C	8:15
C	9:12	E	D	9:22
D	9:38	SW	Home	...

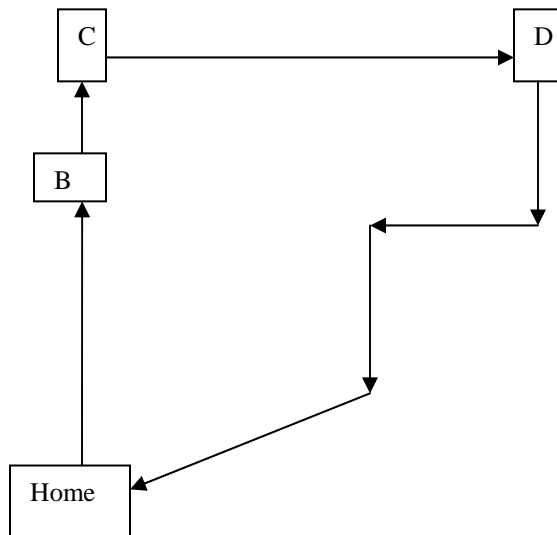


Figure 1: Example Travel Tour

Case A below shows what might be called a “perfect” match, of the diary data to the GPS data from above, allowing for certain rounding tolerances on the self-reported departure and arrival times. (There are also tolerances for locations, as people may park the vehicle and walk to their final reported destination location.). The diary data contains additional information such as activity purpose, as shown below, as well as the number of passengers in the vehicle.

CASE A: “Perfect Match”

GPS Data					Diary Data				
From place	Leave at	Dir. of travel	To Place	Arrive at	From place	Leave at	To Place	Purpose	Arrive at
Home	8:02	N	B	8:10	Home	8:00	B	Gas	8:10
B	8:13	N	C	8:15	B	8:15	C	Groceries	8:20
C	9:12	E	D	9:22	C	9:15	D	Bank	9:25
D	9:38	SW	Home	9:57	D	9:40	Home	At home	10:00

Case B below shows a typical case of survey diary non-response. The 3-minute stop at location B was not reported in the diary data. Because in this case we do not have the diary information to tell us why the person stopped at location B, the issue is whether the stop at B is for a real activity (e.g. for gas or shopping), or just a long stop at a traffic light, in a traffic jam, to make a left turn, etc. According to the GPS data, the vehicle continued in the same direction (north) after making the “stop” at B. If the person instead had turned around at B and gone south to location C, then the likelihood that B is a real destination would be greater. To properly identify real missing trips in such cases is a challenge in designing the software used to process the GPS data and match it to the diary data. Of course, this challenge has been taken on by various firms and research teams – and often the exact characteristics of the GPS data at these potential trip ends confirms or refutes that a stop was made, and either the GPS data or the diary data is flagged accordingly. (These characteristics include the departure of the vehicle from the primary road, the heading/orientation of the vehicle while stopped, and the direction of travel once the vehicle begins moving again.)

Case B: An Omitted Trip

GPS Data					Diary Data				
From place	Leave at	Dir. of travel	To Place	Arrive at	From place	Leave at	To Place	Purpose	Arrive at
Home	8:02	N	B	8:10					
B	8:13	N	C	8:15	Home	8:00	C	Groceries	8:20
C	9:12	E	D	9:22	C	9:15	D	Bank	9:25
D	9:38	SW	Home	9:57	D	9:40	Home	At home	10:00

To illustrate such factors, Case C below is also missing a stop, but this time it is the stop at D rather than B that is not reported in the diary data. In this case, it seems clearer that it was a real missed trip/activity because (a) the GPS record shows that the person changed directions twice in going from C to D to Home, so D was not on the path from C to Home, (b) the GPS dwell time at D was 16 minutes, so very unlikely to be a traffic-related delay, and (c) the travel time of 45

minutes inferred from the diary data to get from C to Home is unrealistically long for a non-stop trip between those two points.

Case C: An Omitted Trip (2)

GPS Data					Diary Data				
From place	Leave at	Dir. of travel	To Place	Arrive at	From place	Leave at	To Place	Purpose	Arrive at
Home	8:02	N	B	8:10	Home	8:00	B	Gas	8:10
B	8:13	N	C	8:15	B	8:15	C	Groceries	8:20
C	9:12	E	D	9:22	C	9:15	Home	At home	10:00
D	9:38	SW	Home	9:57					

In Case D below, the diary data contains an extra trip, picking up a child at location E on the way home from the bank. The GPS data analysis may not identify this as a separate stop/trip if (a) the dwell time at the stop is very short – less than 2 minutes, and (b) the stop does not require the vehicle to change direction radically. (In such cases, it is straightforward to adjust the GPS data to include the stop once it is identified as an actual valid stop in the diary data.)

Case D: An Extra Trip

GPS Data					Diary Data				
From place	Leave at	Dir. of travel	To Place	Arrive at	From place	Leave at	To Place	Purpose	Arrive at
Home	8:02	N	B	8:10	Home	8:00	B	Gas	8:10
B	8:13	N	C	8:15	B	8:15	C	Groceries	8:20
C	9:12	E	D	9:22	C	9:15	D	Bank	9:25
D	9:38	SW	Home	9:57	D	9:40	E	Pick up child	9:50
				...	E	9:51	Home	At home	10:00

Finally, we have Case E below, where, the diary household neglected to report the entire chain of trips (home-based tour), reporting that the vehicle was at home the entire time. If no one in the diary household reports a similar chain of trips in a different vehicle—a probable case of misreporting the vehicle ID—then it is most likely a case of missing diary trips (although it is remotely possible that a non-household member used the vehicle for the particular tour).

Case E: A Missing Tour

GPS Data					Diary Data				
From place	Leave at	Dir. of travel	To Place	Arrive at	From place	Leave at	To Place	Purpose	Arrive at
Home	8:02	N	B	8:10	Home	At home	
B	8:13	N	C	8:15					
C	9:12	E	D	9:22					
D	9:38	SW	Home	9:57					...

Once this type of data matching and processing is done, an analysis can be done to determine which types of vehicle trips are most likely to be present in the GPS data but missing in the diary data, giving some insight into non-response bias in the diary data. Such an analysis is presented in the following section. First, however, it is useful to consider whether other types of survey bias might influence the GPS data as compared to diary data.

Self-selection bias

It may be the case that people who are willing to have GPS devices connected to their vehicles and used to track their movements may be systematically different from those people who are not willing. If such a self-selection bias is not related to travel patterns, then it is not a problem. In reality, however, it may be the case that those willing to be in the GPS sample tend to be those that use their vehicles more (or less) than other people, even after all other variables are taken into account (age, income, gender, location type, household composition, vehicles per driver). An analysis of this possibility is reported in Section 3.

Survey instrument bias

It may also be the case that people who have GPS devices in their vehicles collecting details about their trips will tend to complete the travel/activity diaries differently than those without GPS-equipped vehicles. In particular, the respondents in the GPS sample may be more conscientious in filling out their diaries (at least for vehicle trips) because they know that their diary data can be “tested” against the GPS data. Conversely, it is also possible that GPS-instrumented households may be less likely to spend time accurately recording trip details if they suspect that the GPS devices are already collecting the same details. An analysis of this issue is also reported in Section 3.

3. Data Collection and Analysis

The Survey

The data used for this analysis is from the 2004 Kansas City Regional Household Travel Survey, sponsored by the Mid-America Regional Council (MARC), and the Kansas and Missouri Departments of Transportation. The primary objective of the survey was to document travel behavior data characteristics of regional households in order to update the regional transportation model.

As documented in the study's final report (NuStats 2004), the household travel survey was conducted using standard travel survey methods and computer-aided telephone interviewing (CATI) technology. It entailed the collection of activity and travel information for all household members regardless of age during a specific 24-hour period. The survey relied on the willingness of regional households to (1) provide demographic information about the household, its members and its vehicles and (2) have all household members record all travel and activity for a specific 24-hour period, including address information for all locations visited, trip purpose, mode, and travel times. No incentives were provided to respondents, although an extensive public information campaign was used at the start of the project to emphasize the importance of and benefits from participating.

This study included a technology supplement that involved equipping a subsample of household vehicles with global positioning system (GPS) data loggers (NuStats and GeoStats 2004). The objectives of the GPS component were twofold: (1) to provide an independent data stream of vehicular travel in order to measure the level of accuracy of the travel data reported over the telephone and (2) to obtain details about those trips that were captured by GPS but not reported over the telephone, to be used later to derive trip correction factors.

The survey ran from January through May 2004. In total, 4,001 households were recruited to participate in the study and 3,052 provided travel data. The overall response rate, calculated according to standards established by the Council of American Survey Research Organizations, was 35% (this included a 46% recruitment rate and a 76% retrieval rate).

GPS data collection was done using the GeoStats "GeoLogger". This is a rugged yet simple GPS data-logging device developed specifically for use in household travel surveys and travel time studies throughout the world. As shown in Figure 2, the GeoLogger consists of three components: the data collection device, a GPS receiver/antenna that mounts to the windshield using a suction cup, and the power cord, which plugs into the cigarette lighter or an auxiliary 12-volt power outlet in the vehicle. Installation of the unit is very simple, and requires only plugging the unit into the power source and affixing the suction-cup mount to the windshield. This device is totally passive; once it is installed, no further action (or interaction) with the unit is required.

FIGURE 2: THE GEOSTATS GEOLOGGER



Source: GeoStats

The device is designed to log the vehicle's position at either one-second or five-second intervals. It can be programmed to log all valid GPS points or only those for which the speed is greater than 1 MPH (to screen out non-movement events), and has 4 MB of data storage, which was ample for the needs of this project. For this study, data was logged at 1-second intervals and the unit was programmed to log all valid GPS points where the speed was 1 MPH or greater. The standard GPS data stream elements recorded by the GeoLogger included date, time, latitude, longitude, speed, heading, altitude, number of satellites, and horizontal dilution of precision (a measure of positional accuracy). These elements were stored in the logger in standard National Marine Electronics Association (NMEA) formats and converted to user-specified formats upon downloading.

GPS Data Processing

Each GeoLogger was programmed to store date, time, position, and speed information for each second that the vehicle was in motion. As the GPS data were received, GeoStats analysts converted the second-by-second GPS data streams into a GIS-compatible format (using software developed in-house) then reviewed it for completeness of data. Next, each file was processed to identify potential trip ends based on time intervals between consecutively logged points. (For this study, all initial dwell times of 120 seconds or more were flagged as potential trip stops.) The data were then loaded into a GIS-based application and manually reviewed by analysts to screen out traffic delays and other falsely identified stops with dwell times of 120 seconds or more, as well as to add in stops that had dwell times of less than 120 seconds but had stop characteristics. Once this step was completed, the updated GPS-based trip file for a given household vehicle was ready to be compared with the CATI data for that same vehicle.

Initial Matching of Diary and GPS Data

Prior to the start of the trip comparison process, GeoStats analysts identified 2,309 vehicle driver trips in the GPS data stream for the 228 households where complete GPS and CATI data were available. This represented vehicular travel from 426 vehicles. The corresponding CATI data indicated that members of the participating households reported 2,083 vehicle-driver trips using those same vehicles.

The trip comparison process was done with customized software developed by GeoStats (the Trip Identification and Analysis System or TIAS). This program was designed to compare individual vehicle driver trip records in each vehicle file using location and time as the significant variables for matching. The location and times in the GPS data represented where the vehicle stopped or was parked and were logged automatically by the unit. Respondents reported the CATI location and time data during the retrieval interview. The different sources of information introduce some error into the automatic matching process, as respondents who parked near to their final destination then walked reported their arrival time at the destination, while the GPS equipment recorded the time the vehicle was turned off. In recognition of this, analysts reviewed the times and locations to confirm accurate matches using time and distance thresholds as guidelines to reflect the imperfect reporting mechanisms. Matched results and discrepancies fell into the following categories:

Matched Trips. GPS and CATI trips that matched according to location and time comprise the first category. To account for measurement differences (i.e., where vehicle was parked vs. where the respondent went), time thresholds were 12.5 minutes and thresholds for locations were 100 meters. This means that times were considered a match if within 12.5 minutes of each other and/or locations were considered a match if found to be within 100 meters of each other.

Of all trips made by the 426 vehicles, 209 vehicles had initial perfect matches (within the set thresholds) between the CATI and GPS data. The manual review process identified an additional 38 vehicles with perfectly matching trips. To this were added the 58 vehicles that had no GPS data for the travel day and were reported as not used on the travel day in the CATI data, which is also a “match.” Thus, the total number of vehicles with fully matched trips was 305 (209 + 38 + 58) or 72%. This equates to 1,840 trips of the 2,083 CATI reported vehicle driver trips (88%).

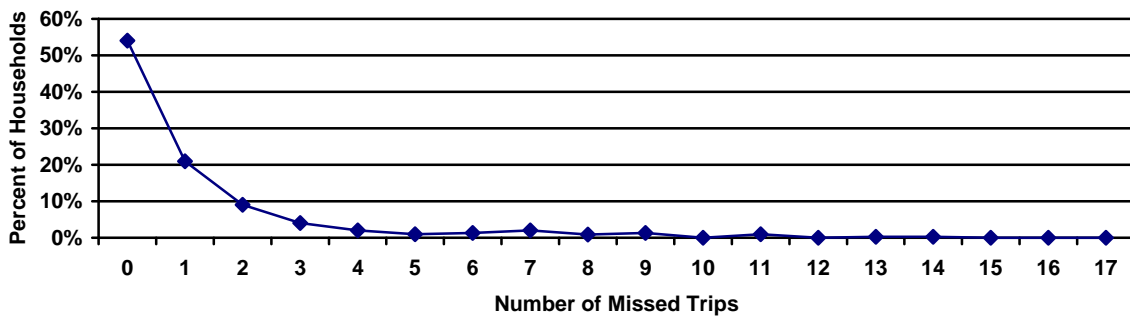
GPS Trip Detected but No CATI Trip Reported. The second category contained those cases where trips were identified within the GPS data stream but not within the CATI data. These “missed” trips were tagged as single links within a trip chain, multiple links within a trip chain, or as complete round-trips missing all links in a tour based on where the missed trip fell within the vehicle’s travel for the 24-hour period. (These cover Cases B, C and E in the earlier example.)

GPS Trip Not Detected but CATI Trip Reported. Finally, the third comparison was regarding CATI vehicle driver trips that had no corresponding GPS trips. During the matching process, 77 CATI trips were identified as having no corresponding GPS trip. This often happens when respondents forget to plug in the GPS unit at the start of the day, unplug it too early at the end of

the travel day, or unplug it during the middle of the day to use the power source for other equipment (such as charging the cell phone). There is also the potential for GPS trips to appear to be missing because they were not actually made on the travel day but reported for that day anyway by the respondent. The missing GPS trips captured through the CATI retrieval process were tagged and are included in the 2,083 CATI trips mentioned above.

Figure 3 summarizes the incidence of missed trips across the 228 households. As shown in that figure, more than half of the households (123 or 59%) had no missed trips (the CATI and GPS data matched exactly). Of the remaining 105 households, 21% had only 1 missed trip, 13% had 2-3 missed trips, 4% had 4 or 5 missed trips, and 3% had more than 5 missed trips. The highest number of missed trips was 17 for any given household, although the average was 1.4 missed trips per household. These data suggest that most households are fairly accurate reporters of their travel, with the majority of missed trips coming from a very low number of households.

Figure 3: Incidence of Missed Vehicle Driver Trips in CATI Data
(source NuStats and GeoStats, 2004)



The initial comparisons suggested a very low rate of vehicle driver trip underreporting in the CATI data file, meaning very accurate reported diary data. Further analysis below helps to explain why this is the case. First, however, it is useful to take a more qualitative look at some of the reasons for the underreporting of diary trips that was discovered.

Follow-up Survey on Missing Trips

A follow-up survey was conducted with a small sample of GPS households for whom the GPS data identified trips that were not found in the diary data. A short questionnaire was sent to those households that listed the reported travel in that particular household vehicle, along with a map of the unmatched, GPS acquired trips. The participants were then asked to identify the unreported stops, the driver of the vehicle, how many household members were with them at the time, the trip purpose, and the reason for not reporting the missing stops/trips in the travel diary. Such follow-up surveys were mailed to 32 households, with 27 ultimately completing the survey, providing details about 47 “missed stops.”

For each missed stop/trip, the respondent was asked to indicate why that stop/trip was not reported during the diary retrieval interview. Of the 47 trips identified in the GPS data but not found in the diary data, participant-provided data showed that 9 (19%) were not true stops/trips.

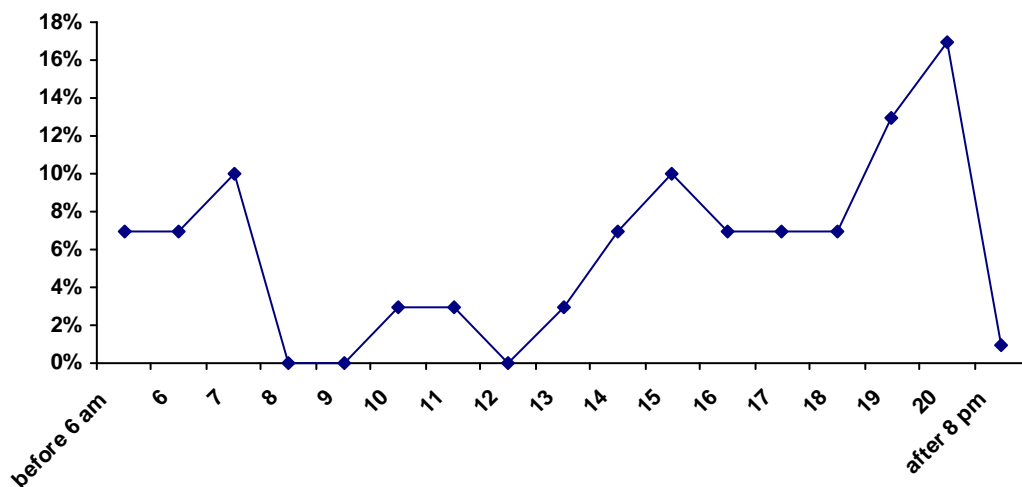
They were either traffic delays, where a respondent made a wrong turn and had to turn around, or the respondent moved their car within the same parking lot (not recorded in log as there was no change in address). An additional 21% of the stops/trips were work-related by someone who drives for a living (these are typically not collected in household travel surveys, but instead are captured through workplace or commercial vehicle surveys).

For those 28 trips that were true trips that should have been reported, the explanations were mainly that the respondent didn't think the stop was important enough to record in the travel diary or report over the telephone (14 cases), or that the respondent simply forgot (10 cases). These cases included 7 stops to drop off or pick up someone, 7 stops to "grab some food," 4 stops to refuel, 4 shopping stops (ran in and bought one item), and 3 stops to mail a letter. The average duration for these forgotten/"not important" trips was only 5.7 minutes. Due to the small sample sizes, this information is for qualitative consideration only.

Each missed trip was categorized in relation to the other travel reported by the respondent. Three types of trips were identified: single stop in a chain, multiple stops in a chain, or all links in a round trip tour. Most of the missed trips (76%) were a single stop among a series of trips in a chain (with all other trip chain stops reported). 14% of the missed trips included multiple stops in a trip chain not reported, while 10% were entire round trips not reported.

Some earlier GPS studies had found that most missed trips took place later in the day, when the respondent had returned home from normal travel and ran out unexpectedly at the end of the travel day. A summary of the times these missed trips took place, shown in Figure 4, confirms this finding, as most missed trips were reported after 6 pm.

Figure 5: Time Missed Trips
(source NuStats and GeoStats, 2004)



Analysis of willingness to participate in the GPS survey and actual participation

We now turn to a more quantitative analysis of the potential biases in both the diary data and the GPS data. All households who agreed to participate in the household travel survey were eligible for inclusion in the GPS study, provided three criteria were present. First, they had to own at least one household vehicle. Second, all household vehicles had to have a functioning cigarette lighter or 12-volt power adapter (to the best of the owner’s knowledge). Finally, they had to indicate interest in participating. Indication of interest was accomplished through a question that was asked of all households who owned at least one vehicle with functioning power outlets.

As summarized in Table 1, of the 3,052 households who provided complete diary data, all but 717 (23%) were eligible to be in the GPS sample and were asked if they were willing to participate. Of the 2,335 eligible households, 669 (29%) were not willing to participate. Of the 1,666 households who were willing to participate, 294 were selected to receive GPS devices, and 228 of those actually participated in the GPS survey and used the devices to report their travel.

Table 1: Distribution of households by willingness to participate and participation

Category	Number of households	Used to test for self-selection bias (Model A below)	Used to test for instrument bias (Model B below)
Not eligible for GPS sample	717		
Not willing to participate	669	669	
Willing, but did not participate	1,438	1,438	1,438
Participated in the GPS survey	228		228
TOTAL	3,052	2,107	1,666

In our analysis of potential biases, we use two other interesting pieces of information captured during the CATI retrieval of the diary data. The first data item is whether or not the travel and activities for each person in the household were recorded on the paper diary. If not, then those trips were presumably reported over the telephone completely from memory. The second data item is whether each person reported their own travel and activities during the CATI retrieval interview. If not, then someone else in the household (usually the main respondent) reported that person’s trips by proxy.

Because our first analysis is at the household level and concerns only vehicle driver trips, it is useful to look at both of these variables in terms of all licensed drivers in the household. Table 2 shows that of the 5,372 licensed drivers in the sample, about 21% did not use the paper diaries, and about 39% had their trips reported by proxy. When these numbers are broken down by whether or not the drivers’ households were willing to participate and actually participated in the GPS survey (the same categories as in Table 1), there is one striking result. Of the 430 licensed drivers taking part in the GPS survey, only 5% did not use the paper diary instrument, compared to over 21% in all other categories. It appears that participating in the GPS survey made drivers much more likely to use the paper diary to record their trips as well. For the proxy variable, there is no notable difference between categories. Reporting by proxy is mainly a function of household size—the main respondent often reports the trips for all household members during

the retrieval call, so drivers in larger households are more likely to have their trips reported by proxy.

Table 2: Use of paper diary and proxy by willingness to participate and participation

Category	Number of licensed drivers	Percent of drivers who did not use paper diary	Percent of drivers whose trips were reported by proxy
TOTAL	5,372	21.2%	38.9%
Not eligible for GPS sample	1,022	24.9%	39.4%
Not willing to participate	1,191	23.9%	36.0%
Willing, but did not participate	2,729	21.1%	40.1%
Participated in the GPS survey	430	4.9%	38.4%

These diary use and proxy variables were included in two separate models, reported in Table 3. Both the willingness to participate in the GPS survey and the actual participation were analyzed separately using binary logistic models as a function of various household characteristics. Model A is for the willingness to participate in the GPS survey, estimated *only* among households who did not actually end up participating, so that any self-selection effect could be estimated separately from the effect of actual participation (instrument effect). Households with more workers per vehicle and higher incomes tended to be more willing to participate, while older respondents and households with fewer vehicles were less willing to participate. The strongest variable in the model, however, is related to the number of diary vehicles driver trips per adult in the household—the more driver trips per adult a household reported in the survey diary, the more likely they were to be willing to participate in the GPS survey. This can be seen as evidence of self-selection bias, where households that make more car trips are more likely to participate. The fraction of drivers in the household who did not use the paper diary is not significant, nor is the fraction of drivers whose trips were reported by proxy.

Table 3: Binary Logistic Model of GPS Willingness to Participate/Participation

Dependent Variable	Model A: Willing to participate?		Model B: Actually participated?	
	Coefficient	T-stat.	Coefficient	T-stat.
Sample size	2107		1666	
Rho-squared	.071		.035	
Independent Variables	Coefficient	T-stat.	Coefficient	T-stat.
Workers per vehicle in the HH	0.150	1.6	-0.126	-0.4
HH Income is \$15-35K	0.400	5.2	-0.100	-0.1
HH Income is \$35-75K	0.491	9.1	0.250	0.7
HH Income is over \$75K	0.565	9.9	-0.011	-0.0
Head of HH is age 60-69	-0.361	-6.3	-0.144	-0.4
Head of HH is age 70+	-0.800	-28.6	-0.317	-1.2
HH owns only one vehicle	-0.310	-5.4	0.080	0.1
Diary vehicle trips per adult in the HH	0.135	34.8	0.065	5.3
Fraction HH drivers not use paper diary	0.084	0.4	-1.765	-25.5
Fraction HH drivers reported by proxy	0.064	0.1	0.357	1.1
Constant	-0.052	-0.6	-2.039	-30.0

Model B was estimated only among those who were willing to participate in the GPS survey, and attempts to distinguish those who actually participated from those who did not. Because this was not a choice made by the household, but was rather a selection made by the survey administrators (using stratified random sampling), one would expect to find very little difference between those who did and did not participate. The results in Table 3 confirm this—none of the variables related to age, income, etc. that were significant in the model for willingness to participate were significant in the model for actual participation. The only exception is the variable for the number of driver trips per adult recorded in the diary survey. Those adults who participated in the GPS survey reported significantly more driver trips than those who were willing to participate but did not. This is strong evidence of an instrument bias, above and beyond the self-selection bias identified in Model A. In the next section, we simultaneously analyze the relative influence of these two types of bias, along with non-response bias, on diary vehicle driver trip rates.

The strongest effect found in model B is that those households who participated in the GPS survey were much less likely to have drivers who did not use the paper diary, confirming the pattern seen in Table 2. Put the other way around, participating in the GPS survey appears to have the beneficial side effect of getting respondents to use the paper diary instrument to record their trips more diligently. It does not appear to have any effect on whether trips are reported in person or by proxy.

A simultaneous analysis of sources of bias

A typical type of analysis that has been done to compare GPS data to diary data is regression analysis on vehicle trips per household. Because different households have different numbers of drivers and vehicles, a more suitable dependent variable to use is vehicle trips per driver or vehicle trips per vehicle. In the GPS data, the vehicle is known, but the driver is not, so, for an analysis that uses both GPS data and diary data, the most informative unit of analysis is the number of trips per specific vehicle either reported (diary data) or recorded (GPS data) during the diary day. The results of such a regression analysis are reported in Table 5 below. Because the number of trips per vehicle is typically not a normally distributed variable, normal linear regression, as used in Model C, is not the best method for analysis. Table 5 also shows the results of Model D using a logarithmic transformation of the vehicle trip rate, used because the log-normal distribution is closer to the distribution of trips/day across the sample. (The constant 1.0 is added before taking the log, to accommodate cases with 0 trips.)

First, Table 4 shows the distribution of the 6,044 vehicle-days used for regression analysis, broken down by whether or not they were willing or eligible to participate in the GPS survey, whether or not they actually did participate, and, for those who did participate, whether the vehicle trip rate is from the diary data or the GPS data. As indicated in the table, by estimating effects of each of these different sub-samples compared to the base group, over and above any other effects related to household type, vehicle type, etc., we can isolate the effects of three different types of bias:

1. *Self-selection bias*: Those unwilling to participate in the GPS survey vs. those who are.
2. *Instrument bias*: Those actually participating in the GPS survey versus those who do not.
3. *Non-response bias*: Trips reported in the travel diaries of the GPS sample respondents versus trips captured in the GPS data.

Table 4: Distribution of the sample for regression

Data/sample type	Observed vehicle-days	Mean trips/day
Willing to participate in GPS survey but did not (base group)	2,859	3.89
Not eligible to participate in GPS survey	1,129	3.28
Not willing to participate in GPS survey (provide evidence of self-selection bias)	1,176	3.39
Participated in GPS survey- Diary data trip rate (provide evidence of instrument bias)	440	4.73
Participated in GPS survey-GPS data trip rate (provide evidence of non-response bias)	440	5.20
Total	6,044	3.84

Both models C and D in Table 5 give similar findings regarding the characteristics of the households and vehicles in the sample:

- The more vehicles a household owns, the fewer trips that are made in each vehicle, indicating a substitution of trips between vehicles. (The “base” level of 2 vehicles is constrained to have a coefficient of 0.)
- The more workers per vehicle in a household, the greater the number of trips made in that vehicle. Smaller positive effects are found for the number of non-working adults and the number of children per vehicle in the household.
- The higher the household income, the higher the number of trips in the vehicle. (The “base” level of income is below \$15,000.)
- As the age of the head of household increases above 40, the number of trips in the vehicle increases to a maximum in the age group 50-59, and then starts to decrease with age. (The fact that the trip rate is still higher for those in their 70’s than for those in their 30’s is surprising, and may be due to age-specific non-response bias; an issue we return to later.)
- The number of trips tends to be higher for vans and SUV’s than for regular cars, but lower for pickups.
- The number of trips made decreases with the age of the vehicle.

Table 5: Linear Regression Analysis of Trips per Vehicle

Model	C-Linear trip rate		D-Log trip rate	
Adjusted R-squared	0.178		0.221	
Dependent variable	Vehicle trips reported/recorded		LN(1+vehicle trips reported/recorded)	
Independent variables	Coefficient	T-stat.	Coefficient	T-stat.
Constant	2.605	9.4	1.049	16.0
One vehicle in the HH	0.619	4.2	0.220	6.3
Three vehicles in the HH	-0.435	-3.9	-0.171	-6.5
Four vehicles in the HH	-0.808	-5.4	-0.350	-9.9
Workers per vehicle in HH	1.219	7.3	0.360	9.1
Non-workers per vehicle in HH	0.831	5.3	0.184	4.9
Children per vehicle in HH	0.863	10.7	0.138	7.3
HH Income is \$15-35K	0.567	2.5	0.116	2.2
HH Income is \$35-75K	0.669	3.0	0.147	2.8
HH Income is over \$75K	0.583	2.5	0.154	2.8
Head of HH is age 40-49	0.339	3.2	0.068	2.7
Head of HH is age 50-59	0.476	4.1	0.086	3.1
Head of HH is age 60-69	0.336	2.4	0.053	1.6
Head of HH is age 70+	0.238	1.5	0.016	0.4
Vehicle type is a van	0.959	7.4	0.184	6.0
Vehicle type is an SUV	0.309	2.7	0.085	3.1
Vehicle type is a pickup truck	-0.349	-3.1	-0.103	-3.9
Age of the vehicle	-0.110	-12.0	-0.032	-14.9
Frac.drivers who use diary / by proxy	0.213	1.0	0.154	3.2
Frac.drivers not use diary / no proxy	-1.196	-6.8	-0.270	-4.3
Frac.drivers not use diary / by proxy	-1.495	-5.7	-0.305	-7.4
Not willing to participate in GPS	-0.483	-4.6	-0.085	-3.4
Not eligible to participate in GPS	-0.131	-1.2	-0.023	-0.9
Participated in the GPS survey	0.495	3.2	0.078	2.2
Trip rate from the GPS survey rather than the diary survey	0.470	2.3	0.056	1.2

The most interesting results for this study, however, are found at the bottom of the table. These reflect estimates of various types of bias, after all other household and vehicle characteristics have been considered.

- As the fraction of household drivers who used the paper diary but had their trips reported by proxy increases, trip rates increase slightly, but not significantly in the linear model.
- As the fraction of drivers in the household who did not use the paper diary increases, the trip rate per vehicle becomes much lower. This variable is also interacted with the proxy variable, and the negative effect is even stronger if the trips by those drivers who do not use the diaries are reported by proxy. .
- Those not willing to participate in the GPS survey have significantly fewer trips per vehicle in the diary data than those who are willing to participate. This indicates a strong *self-selection bias* in trip rates.

- Those not eligible to participate in the GPS survey (mainly due to non-functioning power outlets in their vehicles to plug in the GPS) also reported fewer diary vehicle trips than those willing to participate, but the effect is smaller than that found for those not willing to participate. This result makes sense, since the ineligible sample contains a mix of people who would and would not have been willing to participate if they had been asked.
- Those who actually did participate in the GPS survey reported a significantly greater number of trips per vehicle in the diary data than those who were willing to participate but were not selected to (the base group with coefficient 0). This result indicates that the *survey instrument bias* is separate from the self-selection bias mentioned above.
- Finally, since the models are based on both GPS data and diary data, we can test whether or not the GPS data has more trips per vehicle after the other biases are accounted for. (Remember that some non-response bias is already captured in the model by the variables for not using the paper diaries and reporting by proxy.) The models show that, beyond those effects, there is also *non-response bias*, with the trip rates in the GPS data higher than in the diary data for the same sample. This bias appears to be of a magnitude similar to both the self-selection and survey instrument biases.

Additional analysis of non-response bias

It is also instructive to focus only on those vehicles for which both GPS and diary data are available, and to analyze what variables are related to the ratio of the trip rates for the two data types. The analysis is based on 366 vehicle days (the 440 vehicle days in the GPS sample, minus 74 cases for vehicles that made no trips on the diary days in either data set). A ratio of 1.0 indicates no non-response bias in the diary data compared to the GPS. The regression analysis results in Table 6 indicate the following systematic variations in non-response bias:

- Vehicles in very low income households have more missing trips, while those in households with more workers per vehicle have fewer missing trips.
- Vehicles in households where the head is under age 30 have more missing trips, while those in households where the head is over 60 have fewer missing trips. (This result helps to explain the age-related results reported above in Table 5.)
- Vehicles in households with 3 or 4 vehicles have more missing trips. Older vehicles also tend to have more missing trips. Both of these effects may be related to less diligent trip reporting by teenagers and young adults, who tend to drive the older vehicles in multi-vehicle households.
- Pickup trucks have more missing trips. This result could be related to commercial trips that were not intended to be recorded in the diary survey, as discussed earlier.
- The fraction of household drivers who used the paper diary but had their trips reported by proxy has no significant effect. This result indicates that proxy reporting in itself is not a major contributor to non-response bias, as long as the paper diaries for those individuals are available.
- The fraction of drivers who did not use the paper diary is related to more missing trips, but *particularly* in the case where those drivers' trips were reported by proxy. This effect is not most significant single variable in the model. Although these cases (no diary and proxy reporting) are only 10% of the drivers in the survey sample, they account for much more than 10% of the non-reported trips when compared to the GPS data.

Table 6: Linear regression of the ratio of diary trips to GPS trips for the same vehicle

Dependent variable	GPS trips/diary trips	
Adjusted R-squared	0.117	
Independent variables	Coefficient	T-stat.
Constant	0.901	64.1
Workers per vehicle in HH	0.043	3.8
HH income under \$15K	-0.141	-4.4
Head of HH is under age 30	-0.036	-1.9
Head of HH is over age 60	0.016	1.3
Three vehicles in the HH	-0.049	-4.1
Four vehicles in the HH	-0.105	-5.4
Vehicle is a pickup truck	-0.047	-4.0
Age of the vehicle	-0.0058	-5.3
Frac.drivers who use diary / by proxy	0.018	1.0
Frac.drivers not use diary / no proxy	-0.084	-2.2
Frac.drivers not use diary / by proxy	-0.346	-9.1

Adding trip-specific variables would increase the explanatory power of the type of model described above, but there are problems with doing so. First, we do not have any information for the missing diary trips, other than the trip distance, duration and departure time from the nearest corresponding GPS trips. Other reported analyses of GPS versus diary non-response have tended to find such variables to be insignificant. Similarly, the GPS data is missing information on trip purpose, trip chaining and vehicle occupancy that could be related to similar information in the diary data. As a result, it is not possible to use those variables in the analysis, because they would be confounded with the dependent variable of whether or not a trip is in both data sets. However, there may be ways that the location information in the GPS data can be used to infer activity purpose/type. This applies especially to frequently-visited locations such as home, work and school, but may also apply to other locations that have a non-ambiguous land use in a GIS database, such as shopping malls, restaurants, and parks. For more information on this type of data inferences, see Wolf, et al. (2001).

4. Conclusions and Recommendations

With many interesting findings above that deserve further investigation, these analyses need to be repeated on similar diary and GPS data sets from other regions in order to verify that the findings are transferable. Based on the analysis thus far, however, we can offer a number of preliminary conclusions and recommendations.

The use of GPS data to correct non-reporting bias in diary surveys: The qualitative discussion and follow-up survey described earlier indicate that one needs to be very careful in using GPS data to adjust diary-based trip rates. About 40% of the “missing” stops/trips that were identified in the GPS data turned out not to be stops/trips that were meant to be in the diary data—either they were stops at traffic lights or other types of delays, or they were trips in commercial vehicles outside the scope of the study.

The quantitative analysis shows that instrument bias due to the use of the GPS method is liable to be even larger than the non-reporting bias found in the diary data. In particular, this is because the GPS respondents are more likely than other respondents to make use of the paper diaries that are sent them, and they are less likely to omit trips. This means that adjustment/correction factors that are calculated only on the GPS vs. diary trip rates *within* the GPS sample may not be applicable to the travel diary sample as a whole because the diary vehicle trips for the GPS sample are already biased upwards relative to those who are not in the GPS sample. A more correct (and more complicated) correction scheme would also need to measure and adjust for this type of instrument bias, for self-selection bias, for the use/non-use of the paper diary forms, and for proxy reporting, before the adjustment factors can be applied to the entire household survey sample.

We are intending to carry out further research to find out if these same conclusions hold true when analyzing corresponding diary and GPS datasets from other regions. The extent of non-reporting in diary data compared to GPS data has typically been higher in other surveys than in the Kansas City data, and it is not clear why that is the case. (It may be partially due to steadily improving techniques for GPS data processing and trip matching.) In the meantime, we urge caution in using GPS data to adjust trip rates or expansion factors in diary surveys. Also, our findings point to other uses for GPS data collection that may be more fruitful in the long run:

The use of GPS data in travel model calibration

During the process of calibrating travel model forecasts against count data, it is often found that the model does not predict enough trips, presumably because the survey data used to estimate the model was missing trips to begin with. Because many travel models are segmented by trip purpose and OD type (home-based versus non-home-based), using GPS-based correction factors that adjust the total number of trips is not appropriate if the non-response bias tends to affect one type of trip more than another. The follow-up survey reported above suggests that most non-reported diary stops/trips tend to be intermediate stops on tours, meaning that the diary data does not contain enough non-home-based trips. Even if the GPS data does not contain enough quantitative information to allow a full re-weighting or replacement of the missing diary trips, by providing such qualitative information through the use of follow-up surveys, it can give important clues to indicate which specific types of vehicle trips tend to be missing in the data and need to be adjusted in model calibration. This issue is also very important for missing walk and transit trips that can be identified through the use of person-based, “wearable” GPS devices.

Using the lessons learned from GPS surveys to improve the collection of diary-based survey data: The evidence reported above can provide some hints as to how CATI-based travel/activity diary surveys could be enhanced. For one thing, much of the diary non-response appears to be related to those people who do not make use of the paper diaries, particularly in combination with proxy reporting. It may ultimately be more cost-effective to not include such people/households in sample in the first place (or else assign them a new travel day on which to use the diaries and then call them back again later).

It was also found that many of the “true” missing diary trips take place at the end of the travel day. This suggests that it is important to ask all respondents specifically whether or not they ran out to pick-up someone or something at the store on their way home from work or after they

were settled in for the evening, similar to probing for a lunch trip when a worker reports being at the office all day. Similarly, most of the “true” missing diary trips are associated with short intermediate stops on the way to another destination. Interviewers typically probe for these during retrieval, asking “did you make any stops along the way?” However, as shown in the small follow-up survey reported above, 76% of the true missed trips were a short stop on the way to another location. This suggests the importance of that probe question, as well as the need to fine-tune the wording to be more specific – any stops to get fuel or food, or drop off or pick up someone or something.

The follow-up survey of missed trips reported in this paper was quite small in scope, but the usefulness of the findings suggest that more extensive follow-up surveys of this sort should be carried out in the future. Such additional details would be crucial in extending these analyses to more accurately measure the propensity and characteristics of “missed trips” in the diary survey data, particularly in dealing with some of the more difficult issues such as commercial trips and other types of trips that GPS units pick up but are legitimately excluded from the diary data.

The use of GPS data collection as an integral part of household travel diary surveys: We see the most promising future of GPS data not as a means of validating or adjusting diary-based data, but as a major integral component of the data collected for most or all households. Although the presence of self-selection bias and instrument bias tend to complicate the use of GPS data as validation/adjustment data, each is actually quite good news in its own way. First, the types of self-selection biases found in the willingness to participate in the GPS survey—with household making more trips more likely to participate and older households less likely to participate—are the direct opposite of those biases that seem to most effect travel diary survey samples, which are typically skewed toward older households that travel less than average. So, the use of GPS as an optional, supplementary mode of data collection could serve to help obtain a more representative overall survey sample (although some other self-selection biases, such as the difficulty of recruiting low income, non-car owning households, would still persist).

Second, the fact that the presence of the GPS device makes people more conscientious in using and completing the diary surveys suggests that supplementary paper-based surveys given to GPS respondents would be very effective in capturing any details that the GPS device cannot (trip purpose, driver of vehicle, number of passengers, etc.). It is an interesting research question as to whether this effect would also cause vehicle-based GPS respondents to also diligently report their non-vehicle (walk and transit) trips. Another interesting research question is whether or not this type of instrument effect also extends to respondents equipped with person-based “wearable” GPS units to capture all trips. A pilot study now taking place in the Portland, OR region is testing person-based GPS data collection with supplementary diaries as a full replacement for paper-based diary data collection. The Portland study is also testing the efficacy of doing continuous surveying of small samples rather than a single large-scale survey, in order to save costs by using each unit of GPS equipment for as many different households as possible.

In the future, as personal, handheld computers become more powerful and ubiquitous (probably in the form of multi-functional cell phones), then it is may be possible to include both the GPS receiver/transmitter and CAPI software for supplemental data collection (using voice prompts and/or voice recognition) into a single device that respondents are already used to carrying with

them. At that point, the use of paper-based diary surveys may truly become obsolete. Until then, however, it is likely that we continue to see a mix of paper-based, internet-based and GPS data collection methods, and we will need to keep a close watch on the types of biases associated with each of these methods

Acknowledgments

The authors are grateful to Todd Ashby of Mid-America Regional Council (MARC) in Kansas City for permission to use the data analyzed in this paper.

References

Bachu, P., R. Dudala, and S. Kothuri (2001), "Prompted Recall in Global Positioning Survey: Proof of Concept Study," *Transportation Research Record*, No. 1768, pp. 106-113.

Casas, J. and C. Arce (1999), "Trip Reporting in Household Travel Diaries: A Comparison to GPS-Collected Data," Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C., January.

Chung, E. and A. Shalaby (2004), "Development of a Trip Reconstruction Tool to Identify Traveled Links and Used Modes for GPS-based Personal Travel Surveys," Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., January.

Draijer, G., N. Kalfs, and J. Perdok (2000), "GPS as Data Collection Method for Travel Research," *Transportation Research Record*, No. 1719, pp. 147-153.

Kreitz, M., and Doherty, S. T. 2002. Collection of Spatial Behavioral Data and their Use in Activity Scheduling Models. *Transportation Research Record: Journal of the Transportation Research Board* 1804: 126-133.

NuStats (2004). Kansas City Regional Travel Survey Final Report. (Kansas City, MO: Mid-America Regional Council).

NuStats and GeoStats (2004). Kansas City Regional Travel Survey GPS Study Report. (Kansas City, MO: Mid-America Regional Council), 2004.

NuStats and Battelle (2004), "Year 2000 Post-Census Regional Travel Survey: GPS Study Final Report," Submitted to the Southern California Association of Governments.

Pearson, D. (2001), "Global Positioning System (GPS) and Travel Surveys: Results from the 1997 Austin Household Survey," Presented at the Eighth Conference on the Application of Transportation Planning Methods, Corpus Christi, Texas, April.

Pierce, B., J. Casas, and G. Giamo (2003), "Estimating Trip Rate Under-Reporting: Preliminary Results from the Ohio Household Travel Survey," Presented at the 82nd Annual Meeting of the Transportation Research Board. Washington D.C., January.

Wolf, J., R. Guensler, and W. Bachman (2001), "Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data," *Transportation Research Record*, Number 1768, pp.125-134.

Wolf, J., M. Loechl, M. Thompson and C. Arce (2003), "Trip Rate Analysis in GPS-Enhanced Personal Travel Surveys," in P. Stopher and P.M. Jones (eds.) *Transport Survey Quality and Innovation*, Pergamon, Oxford, chapter 28, pp. 483-498.

Wolf, J., S. Schönfelder, U. Samaga and K.W. Axhausen (2004b), "80 weeks of GPS-traces: Approaches to Enriching Trip Information," Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., (in press).

Wolf, J., M. Oliveira, and M. Thompson (2003b). The Impact of Trip Underreporting on VMT and Travel Time Estimates: Preliminary Findings from the California Statewide Household Travel Survey GPS Study. *Transportation Research Record* No. 1854, pp. 189-198.

Zmud, J. and J. Wolf (2003), "Identifying the Correlates of Trip Misreporting - Results from the California Statewide Household Travel Survey GPS Study" Presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August.