The Cleveland GPS Household Travel Survey: Survey Design, Imputation of Trip Characteristics, and Secondary Uses of the Data

8.1.2013

Jeremy Wilhelm
Westat | Geostats Services
530 Means St NW Suite 310
Atlanta, Georgia 30318
Phone: (404) 588-1004
Email: jeremywilhelm@westat.com

Jean Wolf, PhD
Westat | Geostats Services
530 Means St NW Suite 310
Atlanta, Georgia 30318
Phone: (404) 588-1004
Email: jeanwolf@westat.com

Eunah Kang, AICP
Northeast Ohio Areawide Coordinating Agency
1299 Superior Avenue
Cleveland, Ohio 44114-3204
Phone: (216) 241-2414
Email: EKang@mpo.noaca.org

Derek Taylor
Northeast Ohio Areawide Coordinating Agency
1299 Superior Avenue
Cleveland, Ohio 44114-3204
Phone: (216) 241-2414
Email: DTaylor@mpo.noaca.org

Word Count: 7,321 words (including 11 tables/figures)
The Cleveland GPS Household Travel Survey: Survey Design and Secondary Uses of the Data

Jeremy Wilhelm, Jean Wolf, Eunah Kang, and Derek Taylor

Abstract
In 2011, the Ohio Department of Transportation (ODOT) and the Northeast Ohio Areawide Coordinating Agency (NOACA) spearheaded the second large-scale GPS-based household travel survey conducted in the United States. The survey covered the five counties in and around Cleveland, OH that comprise the planning area of NOACA. Of the 6,542 households recruited into the Cleveland GPS-based travel survey, 4,001 households were asked only to return their GPS devices after use (with no additional participation requirement). Of the remaining 2,541 recruited households, 1,940 were asked to complete a prompted recall (PR) survey based on recorded GPS data while the other 601 households (which were comprised entirely of household members over the age of 75) were provided with simple travel logs to use for recording and reporting travel. All GPS households were asked to use the GPS devices for a minimum of three consecutive days (with at least two weekdays), resulting in the collection of 108,441 GPS trips covering 708,557 miles. Regularly visited locations were collected by the survey research firm during the recruitment interview, and were leveraged along with recorded GPS traces and land use data to assign trip purpose, mode, and travel costs to GPS-Only households based on a probabilistic ‘best fit’. The algorithms were calibrated by PR participant responses. This paper starts with an in-depth overview of the survey design and concludes with a review of several spatial analyses conducted by NOACA transportation planners using the collected GPS dataset.
1 Survey Background

The Cleveland GPS household travel survey (HTS) covered five counties in the region that make up the NOACA planning area; these counties include Cuyahoga, Geauga, Lorain, Lake, and Medina. This geographic area contains roughly 2.1 million residents (Per the US Census 2010). The goal of the survey was to collect socio-demographic and travel behavior data from a minimum of 4,250 households residing across the region. Demographic information, obtained during the recruitment interview, and detailed travel information, obtained from wearable GPS data loggers and/or prompted recall interviews, was collected for all household members. Households in which there was no household member under the age of 75 were asked to complete the study using traditional travel logs only, in order to decrease burden and to ensure that these households were represented in the final data file.

The overall project goal was exceeded, and the final data set contains demographic and trip information for 4,545 households.

2 Survey Design

Recent innovations in travel surveys include a focus on multi-modal data collection and the use of Global Positioning System (GPS) data as a replacement for travel reporting. These innovative approaches are appealing because of their potential to both lower respondent burden and to collect high quality data from a demographically diverse and representative sample of the population. The Cleveland survey was the second large-scale multi-day GPS-based household travel survey conducted in North America, and the design utilized the aforementioned innovations as well as others to be discussed in subsequent sections of this paper. The first GPS-based travel survey was a research project sponsored by the Ohio DOT in Cincinnati and was conducted in 2009-2010 (Stopher, Wargelin, et al. 2012), (Giaimo, et al. 2011).

Data collection for the main survey occurred for one full year beginning in February 2012 and ending in March 2013. The goal was to collect complete travel survey information from at least 4,250 households. This sample was to be divided so that 60% of participating households would be part of the GPS sub-sample with no requirement for a final interview to collect travel details (GPS-Only), 30% would participate in a GPS-based prompted recall (GPS+PR) survey to confirm travel details, and the final 10% would participate with travel logs as the basis of travel reporting (Log-Only). Once the GPS data were collected and processed, the GPS+PR household results were used to calibrate imputation processes used for assigning trip characteristics to the GPS-Only household results.

2.2 Public Outreach

An extensive effort was made by both the consulting team and NOACA staff to proactively inform the public about the existence, value, and importance of the regional travel survey effort. A survey brand was developed, with logos and color schemes that were used in all communications about the survey, including advance invitations, press releases, and the public facing website and online surveys.

Regional stakeholders were contacted and asked to inform their jurisdictions about the survey. Press releases were sent periodically throughout the one-year data collection period, and a public relations representative from NOACA appeared on morning radio programs to explain the importance of the
survey. Press coverage was posted in a ‘News about the survey’ section of the survey public website. Citizens of the area who had not been invited to join the survey were provided an opportunity to share their thoughts via an online opinion survey also available on the survey public website.

2.3 Multiple Survey Modes

It is increasingly evident that the provision of multiple modes for participants to complete a travel survey can lead to responses coming from a more representative mix of the population than what a single mode approach would provide (Wolf, Wilhelm, et al. 2013), (Millar and Dillman 2011) (de Leeuw 2005). This recognition led survey designers to include the following modes for participation:

- Computer-assisted telephone interviews (CATI) with GPS-assisted prompted recall (PR)
- Web-based computer-assisted self-interviews (CASI) with GPS-assisted prompted recall (PR)
- GPS Only travel detail reporting
- Mail-back pen and paper survey (PAPI)

The inclusion of a large subsample of participants who did not complete a final interview to retrieve trip details is a major element in this survey; 60% of the sample was asked to use and return the GPS devices assigned to each member of the household between 14 and 75 years of age without any further reporting requirement.
Table 2-1 shows the differences in burden based on survey modes offered in typical surveys as well as in the Cleveland design. The ratings provided for burden are simple qualitative assignments based on research experience and assessments of burden; e.g., it is more difficult to carry a GPS device and also complete travel logs than it is to do only one of these tasks. The largest proportion of the Cleveland sample participated in the least burdensome mode available (GPS Only).

<table>
<thead>
<tr>
<th>Mode</th>
<th>Recruit Interview</th>
<th>Recruit Burden*</th>
<th>Travel Recording</th>
<th>Travel Burden</th>
<th>Retrieval Interview</th>
<th>Retrieval Burden</th>
<th>Total Burden</th>
<th>Cleveland Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical PAPI</td>
<td>Demographic Questionnaire via CATI/CASI</td>
<td>2</td>
<td>Completion of multiple page travel diary</td>
<td>3</td>
<td>Report of all travel details via CATI/CASI</td>
<td>3</td>
<td>8</td>
<td>10% of sample</td>
</tr>
<tr>
<td>Dual Method</td>
<td>Demographic Questionnaire via CATI/CASI</td>
<td>2</td>
<td>GPS for all HH persons or vehicles plus multiple page travel diary</td>
<td>4</td>
<td>Report of all travel details via CATI/CASI</td>
<td>3</td>
<td>9</td>
<td>Not applicable</td>
</tr>
<tr>
<td>GPS+PR</td>
<td>Demographic Questionnaire via CATI/CASI</td>
<td>2</td>
<td>GPS &amp; travel log/memory jogger for all persons</td>
<td>3</td>
<td>Return GPS devices and logs, Confirmation of GPS Trips via CATI/CASI</td>
<td>2</td>
<td>7</td>
<td>30% of Sample</td>
</tr>
<tr>
<td>GPS Only</td>
<td>Demographic Questionnaire via CATI/CASI</td>
<td>2</td>
<td>GPS; Travel log for non-GPS HH Members</td>
<td>2</td>
<td>Return GPS devices and logs</td>
<td>1</td>
<td>5</td>
<td>60% of Sample</td>
</tr>
</tbody>
</table>

### 2.3.1 Telephone and Web Recruitment

The need to provide a complete, professional, and well-designed web survey has become essential with recent research showing that 78% of Americans have internet access (Zickhur and Smith 2012), and additional research pointing to an increasing decline in the ownership of landline telephones (making previously utilized Random Digit Dialing methods for recruitment dated, if not obsolete).

The survey public website was a resource for participants to learn about the survey and see the progress updated throughout the study period. The site also hosted links to media coverage (web articles, radio programs, and when applicable, videos) giving a sense of the legitimacy of the effort. Cleveland area residents invited to participate in the survey were directed to the public website. Households with a stated preference of completing the retrieval survey online were given a pre-determined time window within which they could complete the web survey before the first CATI call was made.
2.3.2 GPS-Only, GPS and Prompted Recall, and Log-Only Sub-Samples

As previously stated, the sample for the Cleveland HTS was divided so that 60% of households would be part of the GPS sub-sample with no travel reporting requirement other than GPS device usage and return, 30% would be part of a GPS-based prompted recall (GPS+PR) survey, and 10% would participate with travel logs only. One of the major motivations for this split design was to allow for the largest possible proportion of the participating households to take part in the least burdensome mode possible. However, the GPS+PR component was necessary to provide a sample that could be used to calibrate the processes used to impute trip characteristics for the larger GPS only sub-sample.

2.3.3 Trip Builder and Web Based Responses

The use of TripBuilder™ for the collection of travel details allowed for a consistent portal for data collection given that the web-based system supports diary data collection as well as GPS-based prompted recall and can be used for telephone interviews, web self-complete surveys, or mail-back travel log data entry. The integrated online map (provided via the Google Maps API) enables real-time geocoding and point-of-interest lookup for all location data collected or confirmed.

TripBuilder™ collects travel information in two steps. The first step focuses on generating a sequential list of visited places along with their basic attributes, including arrival and departure times, mode of travel, vehicle used, place type, location information, and travel companions. For GPS-based studies, this initial list of trips is preloaded from the processed GPS data, along with a map display of the actual routes taken. The second step then collects additional place details, such as activity, transit fare, and parking details, if such level of detail is required. Figure 2-1: Example TripBuilder Screen shows a typical TripBuilder screen.
Using a single system for the collection of all travel survey data ensured that all collected data undergoes the same underlying questions and corresponding response options, the same branching, and the same logic checks. In addition, it guarantees that all collected data is stored in identical formats and tables with the reporting mode – web, telephone, GPS equipment return or mail-back – captured as an additional data element.

### 2.4 Results of Data Collection

The results shown in Table 2-2 indicate that the expected lower burden of the GPS-Only households did, in fact, translate to a slightly better retrieval rate than the performance of those households in the GPS+PR sub-sample (69.48% versus 67.66% respectively).

<table>
<thead>
<tr>
<th>Survey Type</th>
<th>Mailings</th>
<th>Recruits</th>
<th>Retrievals</th>
<th>Retrieval Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPS PR</td>
<td>62,102</td>
<td>1,938</td>
<td>1,312</td>
<td>67.69%</td>
</tr>
<tr>
<td>GPS Only</td>
<td>128,420</td>
<td>4,002</td>
<td>2,780</td>
<td>69.46%</td>
</tr>
<tr>
<td>Log Only (Over 75)</td>
<td>NA</td>
<td>602</td>
<td>453</td>
<td>75.25%</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>190,522</strong></td>
<td><strong>6,542</strong></td>
<td><strong>4,545</strong></td>
<td><strong>69.47%</strong></td>
</tr>
</tbody>
</table>

The high retrieval rate for the Log Only households is attributable to the nature of that sub-sample being an elderly population with lower mobility and higher survey accessibility via landline telephone. Overall,
the survey results exceeded the goals of NOACA and ODOT with respect to sample sizes as well as demographic distributions.

3 Uses of the Data

The data collected using GPS data loggers in a household travel survey provide rich, detailed, and precise information that can be used for a variety of additional analyses not available from a traditional dataset collected via a survey or interview alone. The following sections summarize efforts by NOACA to understand the potential ways in which the GPS data collected as part of their GPS-based HTS can be leveraged to gain insights into travel behavior in the Cleveland region.

3.1 Dataset Background

The GPS dataset provided to NOACA and used for the following analyses consisted of MS Access databases with tables for GPS Points and GPS Trips, as well as household and person level socioeconomic data. The point and trip data provide the date, time, speed, location, route and origin/destination for all recorded trips as seen in Figure 3-1.

![Figure 3-1: GPS Trip Example](image)

While GPS datasets have traditionally been used for trip rate correction factor calculations, there is additional value in the data including being able to analyze spatial and temporal aspects of travel which allow for the visualization of travel patterns. Also of interest is the association of this rich dataset with household socioeconomic characteristics, as well as information regarding non-motorized forms of transportation.

The data utilized in this analysis, while only representing 75% of the final dataset, included over 15 million GPS points. The GPS coverage was extensive, spanning the entire modeling network and included over 1,000 freeway trips as shown in Figure 3-2. Travel days were evenly distributed across a
one-year time period via a rigorous sampling plan, thus resulting in a broad geographic coverage with highway link trip frequencies that appropriately represent regional travel behaviors.

Figure 3-2: Coverage and Volume of GPS Trips for the Modeling Network

GPS data have also been visually checked against aerial photos to assess whether the accuracy of GPS points allows various scales of spatial analyses at regional, municipal, and parcel levels. The majority of GPS points fell within the Right of Way as seen in Figure 3-3. There were a few locations where the devices lost GPS signal, such as the underground rail tracks at Terminal Tower or at Cleveland Hopkins Airport. GPS points located beyond 60 feet of the street centerline were filtered out to avoid potential GPS data noise that could adversely affect the analysis results.
3.2 Method and Data Processing

Data processing and spatial analysis of the GPS datasets required advanced tools and computing capacity due to their large size and complex data characteristics. Traditional GIS tools and applications could not efficiently support the data processing in a reasonable amount of time. This section describes NOACA’s in-house data processing effort, the advantages of the new tools utilized, and effective ways to analyze and visualize the information derived from GPS Points.

3.2.1 Data Processing using PostGIS/PostgreSQL

The GPS Points table in the MS Access database was converted to ESRI’s file geodatabase to perform various visual checks to ensure the data was representative and of high quality. The analysis started by using the file geodatabase to assign GPS Points to highway network links and applying the following data processing steps:

- Imported a feature class of a file geodatabase to PostGIS/PostgreSQL DB
- In parallel, created a clean highway links by removing duplicate links
- Projected GPS Points to State Plane Ohio North in Feet to align with highway network
- Spatially joined the clean highway link table with GPS Point tables using 60ft buffer
- Selected the joined tables by time of day, modes, and speeds
- Created summary statistics by highway link IDs

PostGIS and PostgreSQL, which are both open source spatial database software, were used to process the data due to their ability to handle massive datasets in a reasonable amount of time. Among the many advantages in managing databases, PostGIS and PostgreSQL provided not only efficient data
computations, but also the ability to analyze GIS geometry such as data projections and spatial joins. In addition, these spatial database software allowed for the easy creation of summary statistics that would traditionally have been limited using standard GIS databases.

### 3.2.2 Visualization using ArcGIS Spatial Analyst

Another software product that was extensively used in this analysis was ESRI’s ArcGIS Spatial Analyst. It provided effective summarization and visualization of GPS Point data. Primary tools used in this visualization process include Kernel density, line density, and neighborhood statistics, which aggregate vector data into grid cells.

For the highway network linked with GPS Points, band width plots, desire lines, and realized paths were employed to visualize and analyze the GPS dataset. Using collected pairs of X and Y coordinates of each trip route, Euclidean distance lines, also known as desire lines, were drawn and calculated. In addition, by connecting each GPS point by following the trip sequence, realized trip paths were drawn and calculated. An example of each visualization method used is shown in the following sections.

### 3.3 GPS Data for Exploratory Spatial Analysis

GPS Points data collected as part of the regional household travel survey have great analytical potential in explanatory spatial analysis. Not only do the GPS points contain location, speed, and time data, but they also contain information regarding direction, origin and destination, frequency, and route of travel by mode (the latter of which can be derived using speed profiles combined with GIS transportation layers). When applied with visual data analysis techniques and descriptive statistics methods, the GPS data become an invaluable source to perform pattern analyses at any given scale of geography and provide insights to better understand travel patterns.

NOACA’s researchers paid special attention to the advantages of GPS data collected from the region-wide travel survey. The data have associated household and socioeconomic characteristics, such as household size, household income, and vehicle ownership, which are distinct from various other analyses using GPS probe data. This spatial analysis was intended to expand on the survey’s traditional uses by exploring and displaying examples of the analytical capabilities that GPS data offer using various GIS analysis tools and methods.

#### 3.3.1 Major Destinations using Kernel Density

Major regional destinations can be identified using trip destination points, an analysis approach which is distinguishable from conventional household or firm location-based analyses. This analysis used a Kernel density function embedded in the ArcGIS Spatial Analyst extension to visualize the magnitude of major destinations. Since this function generalizes a large number of point values into small raster cells, the advantage is that it produces a point density layer in a relatively small file size. For the example depicted in Figure 4-4, the cell size was set to ¼ square mile.

Since the destinations were drawn from trip end points that included intermediate stops along the entire trip, Figure 3-4 identifies not only the major employment centers and high density residential areas, but also major rail stations and regional shopping centers that often are visited mid-trip. The resulting analysis identifies downtown Cleveland as a dominant trip destination followed by the
University Circle area. It also verifies major rail and transit hubs as frequently visited locations, such as Windermere station, Shaker Square station, and University Circle station, and thus could be used to provide analytical support for transit-oriented development opportunities.

3.3.2 Origin-Destination Analysis Using Desire Lines

Each set of origin and destination pairs for trips arriving at the University Circle area were visualized in Figure 4-5. This analysis, known as select zone analysis in transportation modeling applications, identifies where the trips originate. Powerful features of this type of origin-destination analysis include two different aspects. One is that GPS trip points present analytical capabilities in various spatial resolutions, such as at the Traffic Analysis Zone, or at Census Tracts or Blocks, and even at parcel levels, and allow flexibility either in aggregating or disaggregating based on the purpose of the analysis. The second advantage is that GPS trip point data provide a detailed understanding on trip characteristics and associated socioeconomic characteristics of travelers. It helps to answer the questions of who travels, why they travel, and where they travel to and from, with many customizable summary details.

Figure 3-5 shows that the University Circle area draws trips not only from areas in its immediate proximity, but also from all over the region including Lake County, western Cuyahoga County, and Lorain County. Also the relative thickness of the desire lines between the northeast area of the region and University Circle could provide useful insights on existing trip patterns to be used in future studies. The analysis was performed using the “XY to Line” ArcGIS tool that creates straight lines between trip start and end points displayed in X and Y coordinates.
3.3.3 Route and Mode Choices

This analysis is similar to the previous analysis in that it uses origin and destination information, but different in terms of data processing methods. Realized path is considered the network distance between trip origin and destination points. However, this analysis utilized observed GPS traces which identified that one’s route choice might be different from the calculated shortest path. Travel routes can be delineated by travel mode, and thus this analysis is useful in displaying major travel corridors by different modes of transportation.

Figure 3-6 shows trips coming to the University Circle area by chosen routes and transportation modes. Routes highlighted in red present the major routes taken by trip makers. Martin Luther King Boulevard connecting from I-90 East highlighted in red has the most travelers, followed by the Carnegie Ave corridor highlighted in yellow connected from I-90 West.

The analysis first employed the ArcGIS “Points to Line” tool to convert the GPS point sequence of each trip to line segments. Then the Line Density function, which is embedded in ArcGIS Spatial Analyst, was utilized to create the color scheme that effectively displays the magnitude of trip frequencies.
3.4 GPS Data as Alternative Data Source

GPS-based household travel survey data also provides valuable opportunities as an alternative data source when a more commonly used data source is either unavailable in a timely manner or expensive to obtain. Over the years, fast development of GPS technologies along with enhanced big data processing capabilities have enabled a wide variety of highway-related data products to be available for use in transportation planning. Transit-travel related data, however, remained relatively unchanged. For example, acquiring a transit speed table by area type and facility type still requires field data collection efforts to be made. GPS data can work as a supplemental source. This section focuses on the areas where GPS survey data can fill the void when the availability of observed data is uncertain or limited. In this section, a couple of examples in exploiting transit speeds from GPS data are introduced and examined.

3.4.1 Transit Travel Time Delay

Transit travel time in transportation modeling is often reflected in a function of highway speeds and expressed in units of minutes per mile delays to auto travel time. A frequently cited source to meet the data need for transit travel time is transit vehicle speeds recorded using the GPS component of the Automatic Vehicle Locator (AVL). As a means to automatically determine geographic locations of transit vehicles, this AVL data provides rich information on transit travel times. Despite these advantages, however, the biggest barrier for use in other purposes, such as travel demand modeling, often comes from non-technical issues. Since most transit operators do not have experience and dedicated...
manpower to extract and process the data from the system, it often takes long periods of time to obtain raw AVL data and to process these data before they can be used in transportation models.

GPS-based household travel survey data was examined by NOACA as an alternative data source for transit travel time. After assigning GPS Points to nearest highway network links, auto and transit travel times were calculated and aggregated by four different area types and facility types. The result of the analysis can be seen in Figure 3-7. Auto and transit travel times along freeways were similar to one another, and transit travel time delays along major arterials are higher than minor arterials. However, the central business district (CBD) area shows a flipped pattern, which is most likely due to the provision of bus-only lanes in the CBD, such as those along Superior Avenue.

Figure 3-7: Delay by Area Type and Facility Type

3.4.2 Rail Speed Variations

Not only can GPS data derive average transit speeds, they can also provide detailed speed information along entire fixed transit routes, which have been passively collected by transit passengers who carried GPS devices during their transit trips. This section presents detailed speed variations along a fixed route rail transit line.

An example of rail speed variations is displayed in Figure 3-8. GPS Points of rail passengers were separated from the database, and then each point containing speed information was averaged into
neighboring cells using ArcGIS Spatial Analyst. As seen in gradual color patterns, rail speed ranged from 2 MPH to approximately 60 MPH. When the analysis result was overlaid with a street shapefile and a rail station shapefile, other factors affecting speed variations were identified, such as dwell time at station locations, slower operations due to railroad geometries, and intersection delays along light rail.

Figure 3-8: Averaging Point Speeds into Neighborhood Cells

3.5 Lessons Learned

The motivation for these analysis efforts was to expand the use of GPS data collected from the household travel survey to other planning and data collection purposes. Over the course of developing analysis ideas and conducting analyses, there were several noteworthy lessons learned.

First, traditional data processing and visualization techniques were not sufficient enough to handle GPS points within a reasonable amount of time. The massive size of the spatial data required the use of advanced tools and enhanced computing capacity. Also, a Graphic User Interface should be avoided in order to expedite data processing.

Secondly, the importance of accurate base street files, highway network link files, and transit network line files cannot be emphasized enough. When accurate GPS points are assigned to poor quality highway network link files, the results are poor or erroneous spatial joins. True shaped highway and transit network are an absolute necessity for more accurate results.

Thirdly, valuable analysis prototypes need to be identified from various research methods and case studies. When detailed mode or trip purpose information was used in finer scale of geography, the data may not provide statistically significant results. Accumulated knowledge on appropriate analysis types will allow analysts and data users to understand limitations of the data and the analysis being conducted.
Last but not least, NOACA researchers realized that there were many concerns and issues involved with the sharing of GPS data with other potential users. GPS data not only allows access to personal information, but also to specific locations visited by all household members including children. Extra careful attention needs to be made in storing, using, and sharing GPS data. NOACA plans to prepare guidelines related to this issue. In addition, it is vital that GPS data never be available to the public in its raw data format. Creating public use files or processing the data using a GPS data anonymizer are two current options that should be considered at the MPO level.

4 Acknowledgments

The authors would also like to acknowledge Greg Giaimo and Rebekha Anderson from the Ohio Department of Transportation for their guidance and support on the project.
5 Bibliography


