Using GPS Data to Understand Driving Behavior

Joe Grengs, Xiaoguang Wang, and Lidia Kostyniuk

TRAVEL-DEMAND modeling is evolving away from relying on crude data aggregated at large urban zones toward using highly disaggregated approaches where individual travelers are modeled by interacting with fine-grained spatial settings represented by parcel data. And the conceptual framework that supports travel-demand modeling is moving away from understanding travel as a series of trips toward a view of people interacting with their surroundings as a series of activities. This evolution in travel-demand modeling has given rise to analysis that combines highly detailed travel data collected through global positioning systems (GPS) with techniques in geographic information systems (GIS). Collecting travel-behavior data by GPS offers several important advantages over conventional trip diary surveys: GPS data can be collected over much longer periods of time than the typical two-day diary; they do not rely on the memory and estimates of a survey respondent; and they provide linkages among complex trips, tours, and daily travel patterns. The most important advantage of GPS data is that they allow us to address the dynamic properties of travel behavior by capturing detailed spatial, temporal, and attribute conditions throughout the full length of the traveling experience. Unlike conventional travel diaries that provide no information between origins and destinations, GPS data offer insights into the traveler’s choices and decisions while en route. However, despite these advantages,
GPS data present significant challenges that hinder their widespread adoption for travel behavior studies. The volume of data is massive, and converting points of data into a meaningful model of highly complex travel—with trip-chains of multiple activities and purposes—makes for a cumbersome database design.

This paper investigates driving behavior based on GPS data collected by the University of Michigan Transportation Research Institute (UMTRI). The database contains driving data for 78 drivers living in the Detroit metropolitan region in 2004, with automobile use tracked on a day-to-day basis for four weeks, with geographic positions captured every second by GPS. We combine the GPS data with geocoded street addresses of business establishments, land-use polygons, aerial photographs, census data, and road attributes. The paper has two main objectives. The first is to explain methodological challenges of converting an enormous set of geocoded data points into a meaningful database that describes the complexity of trips and tours. The second objective is to describe in a detailed manner the driving characteristics of a single driver over the course of a month of driving, to illustrate the kinds of valuable lessons that transportation analysts can learn from GPS data. We find that common travel patterns are more complex than generally understood from traditional travel surveys and that transportation engineers and planners can benefit from GPS data used as a new technology for travel study.

Why Use GPS Data for Travel Behavior Research?

Transportation engineers and planners in the United States generally collect data on daily travel patterns using self-reported written diaries and telephone surveys. These conventional travel surveys have the advantage of being fairly straightforward to administer and the data collected are easy to manage. Self-reported surveys are particularly useful for travel behavior studies because a person can describe the exact nature of the purpose for taking a trip, such as to go shopping, to visit friends, or to eat a meal at a restaurant.

But conventional travel surveys have several serious limitations for travel behavior research. First, the self-reporting of data is known to be unreliable. People typically underreport short trips, and underestimate trip durations and misrepresent the time that a trip starts and ends. Trip destination locations are
reported inconsistently, such as listing the nearest main intersection when a street address is unknown. A second disadvantage is that self-reported surveys are collected over very short time periods, typically over two days. Third, these self-reported surveys fail to capture important spatial information about trips because they collect data on individual trips by aggregating them to traffic analysis zones (TAZs) for analysis and modeling. Points of origin and destination for each trip are coded to a TAZ, so that travel behavior at small scales within a TAZ is lost. Furthermore, this method of limiting data collection to trip end locations leaves no information about travel behavior between origin and destination, so that the actual route traveled between TAZs is unknown. Transportation planners must resort to such methods as route choice modeling or shortest path networks to ascertain the route between TAZs.

Collecting data repeatedly with GPS as a person travels offers the possibility of overcoming the shortcomings of conventional travel surveys. Because GPS data do not rely on the memory, estimations, and diligence of a person’s self-reporting, they provide precise locations and times. GPS data can readily be collected over long durations such as 30 days. And because data points are collected throughout an entire trip, exact route locations are known.

Despite these many advantages, GPS data have several disadvantages. How to maintain the privacy of subjects and protect the confidentiality of the data must surely be a top priority if GPS travel data is ever to become widely adopted. A second limitation is that the collected data are massive and difficult to work with. To take full advantage of the data requires advanced database techniques, computer programming, and substantial skills with GIS software. But the single most limiting factor in working with GPS data is the difficulty in ascertaining the purpose of a trip. In the sections that follow, we explain some of the challenges in using GPS data and illustrate how GPS data can help urban planners solve some of the most urgent questions about travel behavior.

Data

The naturalistic driving data (NDD) were collected by UMTRI’s Engineering Research Division as part of the Intelligent Vehicle Initiative Road Departure Crash Warning Field Operational Test.
The project was conducted under a cooperative agreement with the U.S. Department of Transportation along with Visteon Corporation and AssistWare Technologies. The project developed, validated, and field-tested a set of technologies for two main safety purposes: 1) to find a way to warn drivers when they were drifting from their lane, and 2) to alert drivers to slow down if they were approaching a curve too fast and to negotiate it more safely. The study used 11 instrumented vehicles given to 78 subjects to use for their normal driving for one month each, spread over ten months between May 2004 and February 2005. The drivers were randomly selected from licensed drivers from Southeast Michigan, from three age groups (20–30, 40–50, 60–70) equally divided by sex. The vehicles were equipped with an array of sensor, communication, and GPS technology to gather information on vehicle position, speed, position, yaw rate, heading, and time. Other sensors acquired information on various vehicle systems including turn signal status, headlight status, windshield wiper status, and steering wheel angle. The vehicles were also equipped with front-facing and side-facing radar, yielding radar measurements of nearby vehicles. Video data collected both the forward scene and the driver’s activities. The resulting data captured a total of 9,582 trips over 83,000 miles of driving.

**Challenges in Using GPS Data**

A GPS device, by collecting a continuous stream of data throughout travel, is capable of producing extremely large quantities of data. A transportation analyst faces two main challenges in converting the massive quantity of data into useful information for decision making. First, a simple table of GPS points—including such data items as latitude, longitude, speed, time of day, and heading—must be processed and translated into a trip-log format that describes travel behavior in terms of a related set of origins, destinations, trips, routes, and tours. Second, in order to construct relationships between travel behavior and the surrounding built environment, the trip-log items must be associated with spatial features in GIS layers. We illustrate these challenges with several examples.

The first illustration of converting tabular GPS data into useful travel behavior information is the task of detecting trips ends. The GPS data in this study contains indicators of engine starts and shut-downs, which can be used to define points as the
start and end of a trip. However, using engine shut-downs as an indicator of a trip end misses several situations where multiple trips are chained together without an engine shut-down. Drivers routinely let their engines run when dropping off or picking up a passenger; when visiting a drive-through restaurant or bank; when stopping to drop a letter in a post office box; and even when refilling the car’s tank with gasoline.

We refer to such cases of stops that occur while the engine is running as intermediate stops. Under normal driving conditions, a GPS device collects location data at regular time intervals. In our case, we collect latitude and longitude positions for every one second of elapsed time. Therefore, as a vehicle moves through space we have both distance traveled and the time it takes to cover the distance. When an intermediate stop occurs, the vehicle covers no distance even while time elapses. This stationary time gap is the main variable used for identifying intermediate stops. A common approach is to define a trip end if the stationary dwell time exceeds some minimum time period. However, relying on dwell time alone runs the risk of mistakenly identifying a vehicle stuck in traffic as a trip end. Some traffic stops can exceed several minutes, such as a left-turn at an unsignalized intersection in busy traffic. Following Stopher et al., who conducted experiments of actual traffic conditions to settle on a dwell time of two minutes, we define an intermediate stop as any location where the following criteria are met: the difference in successive latitude and longitude values is less than seven meters; the heading is unchanged or zero; speed is zero; and the elapsed time during which these conditions hold is equal to or greater than 120 seconds. A computer algorithm flags all points that meet these conditions as intermediate stops.

Another challenge in processing raw GPS data is to combine multiple trip ends into a single destination location. The raw GPS data provide multiple points clustered in close proximity where a transportation analyst would prefer to know a single destination. Figure 1 below illustrates this problem. The figure shows four separate trips ends as circles, all clustered in the parking lot of a regional shopping center. To be useful in travel behavior studies, an analyst would want to know that all four points are associated with the single shopping destination. Working only with a table of data, an analyst has no way of knowing that these four seemingly distinct points are related. Only by laying these trip ends over a map for visual context do we see that the four points are all sitting in one parking lot, presumably visiting a single destination.
With over 12,000 trip ends in our data base, visually inspecting such points on a map is not practical. Instead, we devised a method for aggregating multiple trip ends into a single destination by writing a computer program that tests for clustering of points within a threshold distance.

We associated multiple trip ends with a single destination location if the distance between trip ends is within 100 feet. After carrying out a sensitivity analysis by testing various threshold distances ranging from 50 feet to 1,000 feet, and after confirming, by visual inspection of aerial photographs of all trip ends from one driver, we settled on a threshold distance of 100 feet. The algorithm used is what computer scientists refer to as a union-find algorithm that seeks to identify a grouping of points by solving a maximal clique problem. A clique is a set of vertices in a graph in which there is an edge between every pair of vertices. To solve the maximal clique problem is to find the largest clique in the graph. In our case, a graph is constructed containing nodes representing each latitude and longitude of a trip end for one driver. Edges are formed connecting nodes if and only if the
edge length is less than the threshold distance (100 feet). This graph is then solved to find all maximal cliques, which represents a cluster of trip ends that are located within 100 feet of each other. The geographic center is then calculated for each maximal clique and the trip end closest to the geographic center is assigned to represent the unique destination point.

A third illustration is in ascertaining the purpose of a trip. Unlike with conventional travel surveys where the traveler reports what his or her purpose was in carrying out a trip, an analyst using GPS data must deduce the purpose of a trip by observing the types of activities near the destination point. We found that estimating the purpose of a trip was the single most challenging task in converting GPS points to useful travel behavior information. We could successfully identify home, work, and school destinations—the places that tend to be visited regularly—but we found that other types of trip destinations that are more discretionary could not be reliably identified. Any subsequent analysis would, therefore, require either collecting information on trip purposes directly from the drivers themselves, or would require categorizing trip purposes in a highly aggregated manner.

Several data items are useful for identifying home, work, and school destinations, and we designed an algorithm to query the sets of data. We obtained a GIS land-use layer from a regional planning agency so that we could determine whether the trip ends sat within such categories as residential, commercial, or industrial land developments. We purchased business establishment data and geocoded some 76,000 business firms in GIS for the Detroit metropolitan region so that we could identify types of nearby locations with Standard Industrial Classification (SIC) codes. By combining these land-use data and business establishments with time of day, activity duration, and visit frequency, we could make fairly reasonable estimates of the trip purpose.

Figure 2 below provides an example of the difficulty in identifying trip purpose, and uses the same location as the previous map of Figure 1. As illustrated in Figure 1, four trip ends (shown as circles) were aggregated into a single clique (shown as a triangle). The task in Figure 2 was to associate the clique with a destination. The clique sits within the land-use category of Commercial and Office. Note that the business establishments shown as dark circles are geocoded with a uniform offset distance from the street and do not provide a precise location for the actual position of the business relative to the street. Several nearby
business establishments are candidates for the destination, but how might we choose just one?

A destination is defined as a driver’s home if all of the following criteria are met: 1) the destination is visited at least 15 times in four weeks; 2) the average activity duration at the destination exceeds eight hours; and 3) the destination occurs within a land-use category related to residential development. For nearly all drivers, these criteria produced a single location, and we then defined this location as the home. For several drivers where these criteria produced more than one location, we defined the home location as the one with the highest visit frequency. For a few drivers for whom these criteria yielded no locations, we selectively relaxed the criteria until we could reliably assign a single home location.

The data included information on a driver’s occupation and employment status, allowing us to determine full-time workers and students, which we then used to identify work and school locations. The criteria to identify work locations for full-time workers working outside the home included: 1) destinations are not home locations; 2) destinations are visited at least ten times in four weeks; and 3) the average activity duration exceeds four hours. We did not consider land use as one of the criteria.
because we assumed that jobs could be located in any type of land use category.

We used similar criteria to identify school location for full-time students except we assumed that visit frequency of school locations was more than eight times in four weeks, average activity duration was longer than 1.5 hours, and the location occurred within an institutional land-use category.

Although this procedure allowed us to identify locations of home, work, and school, we could not reliably identify other trip purposes such as shopping, eating out, or conducting personal businesses. Several factors made these other trip purposes difficult to identify. First, dissimilar trip purposes may share similar data characteristics. Consider, for example, attempting to distinguish between eating out at a fast-food restaurant and taking a trip to a convenience store. Both are located on similar land-use sites, require a duration of about 20–30 minutes, can occur within a wide range of starting times throughout the day, and may have similar visit frequencies. A second difficulty in identifying trip purposes other than home, work, and school results from the way business establishment data are geocoded onto a map. As shown in Figure 2, business establishments contain latitude and longitude coordinates that place them in close proximity to roads, even though the actual physical location of the establishment may be set back considerably from the road. GPS trip end locations may occur at considerable distances from the establishment locations making it difficult to estimate the match between trip end and establishment.

The fourth and final illustration of processing GPS data is the problem of linking the set of GPS points to spatial features in a GIS. Each of the GPS points in our database is associated with values of latitude and longitude, so importing them into GIS is a simple matter. But linking each of the over one million points to the appropriate segment of a road network is not nearly as simple as it might seem. To link a point to the nearest road segment produces serious errors, as shown in Figure 3 below. The panel on the left of the figure shows a series of seven GPS points that were collected as a vehicle traveled along the road running east to west. But by zooming into the middle point, shown in the second panel, we can see that the straightforward technique of linking points to the nearest road segment would mistakenly assign this point to a segment that sits perpendicular to the actual road traveled. To solve this problem that occurs commonly at road intersections, we developed a computer program that
accounts for the sequencing of nearby GPS and combines linear referencing with shortest-path network tools in GIS.

Applications to the Study of Driving Behavior

Despite the challenges in processing the enormous quantity of data produced, GPS data offer substantial advantages over conventional travel survey data. This section demonstrates several types of problems that can be addressed with GPS data.

Longitudinal Analysis of Trip-Making Behavior

First, unlike conventional travel surveys, GPS data can readily be collected over long periods without burdening the traveler. The data, therefore, allow for longitudinal analysis that provides insights into spatial and temporal patterns in driving that the standard two-day survey cannot cover. Here we provide the results of an in-depth longitudinal analysis of one driver’s travel behavior, including variations in daily trip purpose, destination visit frequency, and route selection.

Figure 4 shows the variation in trip frequency over four weeks for a single driver. The figure confirms that drivers tend to exhibit weekly patterns in their trip making behavior. This driver usually makes fewer trips during the weekends, while from Monday to Friday the trip frequency tends to increase. Yet, important variations appear from week to week. For instance,
Sundays show a great deal of variation, ranging from the lowest to nearly the highest number of trips per day over the month.

Figure 5 shows patterns of daily time devoted to various activities for this driver over four weeks. Work time is fairly constant at about eight hours each weekday, although this driver does occasionally work on Sundays and occasionally does not work on weekdays. Somewhat surprising is the degree of variation in drive time from day to day.

Figure 6 shows the variation of trip generation by different times of day for each day of the week. The y-axis is the number of trip starts. Although trips are generally clustered around the morning and afternoon peak periods, this driver tends to make trips throughout the day, often late into the evening and early morning hours. Figure 7 shows data by week rather than by weekday, to illustrate how trip making varies from one week to the next. Although the
same general pattern appears from one week to the next suggesting that this driver engages in a regular commute pattern, important variations are revealed by the figure. For example, despite a regular work commute pattern, this driver took far fewer trips in Week 4 than in Week 1, presumably by cutting down on discretionary trip making.

Next, we compare how patterns of this single driver’s trip making behavior compares to two other drivers living in distinctly different parts of the metropolitan region. The driver of this study (Driver 1) lives in the outer suburbs. Driver 2 lives in an inner-ring suburb, and Driver 3 lives in the central city of Detroit. Figure 8 compares these drivers by showing trip generation patterns by time of day. Drivers 1 and 2 show similar trip-making patterns, with trip starts typically clustered around morning and afternoon peak periods. Driver 1 also exhibits a third peak period over the lunch hour. By inspecting maps, we found that the land-use patterns at Driver 1’s work place are considerably more mixed and compact than Driver 2’s work place, perhaps offering more opportunity for leaving work to eat lunch or run errands. Driver
3, by contrast to the other two, generates substantially more trips, and the trips are spread throughout the day, a pattern that might be partly explained by this driver’s flexible work schedule.

We can gain further insight into the trip-making behavior of these three drivers by separating the data by trip purpose, as shown in Figure 9 below. As suggested in the previous figure, Driver 1 generates a large share of nonwork trips during the lunch hour, while Driver 3 participates in nonwork activities throughout the day.

**Data Visualization and GIS Mapping**

A challenge with analyzing disaggregate GPS travel data is the multiple interaction of several dimensions, including spatial location, time of day, duration of activity, and sequencing of trips and tours. Visual displays of the data with GIS maps is an effective way of communicating the interaction of these many dimensions, and a useful technique for the analyst to gain insight into the data. In this section we provide a few examples of visualizing data with GIS mapping, by investigating the spatial trip-making patterns of one driver over a month.
FIGURE 8
Trip Generation by Time of Day, Over Four Weeks, Three Drivers

FIGURE 9
Trip Generation by Time of Day and Trip Purpose, Over Four Weeks, Three Drivers
The first illustration deals with repetitive travel patterns. By collecting data over 30 days of driving, we were able to discover temporal and spatial patterns not possible in two-day travel surveys. Do drivers typically cling to familiar routes, or do they vary their travel depending on surrounding conditions? How much do fixed activities (such as work, home, and childcare) constrain the options for participating in more discretionary activities (such as recreation, shopping, or visiting friends)? These are questions that are difficult to address without longitudinal data like those collected by GPS.

The map of Figure 10 shows the visit frequency for both destinations and routes for one driver over four weeks. We observe that, as expected, the activity space is highly bounded by the key locations of

![Figure 10: Destination and Route Frequency, One Month, One Driver](image_url)
home and work. Destinations that are visited frequently (shown by the proportionally-sized circles) tend to be near either home or work, and the routes traveled most frequently (proportionally-sized lines) occur largely between work and home. However, it appears that this driver does not adhere to a single route between work and home, in order to visit destinations that lie off the shortest-path route.

Figure 11 provides a zoomed-in view of a portion of Figure 10, showing more detail of the area south of the home location. Here we find that this driver has a very regular route to

FIGURE 11
Closer View of Destination and Route Frequency, One Month, One Driver
engage in frequent social visits—over ten in one month—at a single point to the south, an example of the kind of pattern that cannot easily be detected by the typical two-day travel survey.

Another approach to understanding this driver’s patterns is to observe changes from one week to the next. Figure 12 shows the evolution of the activity space for one driver over four weeks, with destinations and routes cumulatively added to each previous map panel. Although the activity space is generally bounded by
work and home locations, the map also makes clear that this driver continuously visits new places, with the geographic scale of activity growing over time. Figure 13 extends this point by showing that this driver continuously visits new places throughout the entire month.

A second illustration of data visualization through mapping is in comparing actual routes to shortest-path routes. Conventional travel surveys collect no information on the route traveled between TAZs and many analysts are left to rely on assuming that drivers take the path that minimizes distance or time. We find that actual routes rarely come close to approximating the shortest path. Figure 14 shows such an example, a case where the driver has chosen a route that is 37 percent longer than what would be predicted by a shortest-path algorithm. Why do drivers regularly choose inefficient routes, and what role does the built environment play in such choices, such as in facilitating the chaining of multiple trips? Questions like these are difficult to address without the detail provided by GPS data.

The third and final example of data visualization is in learning about trip-chaining behavior, a highly complex phenomenon.

**FIGURE 13**
Cumulative Frequency of Visits to Previously Unvisited Locations
that requires a careful examination of both spatial and temporal variation. Figure 15 illustrates the power of visualizing multiple variables in a map by simultaneously showing the sequencing of driving behavior.
multiple activities, the speed of travel between activities, and the duration of each activity. The map shows all activities of one driver over the course of a single day. This driver begins his day at home denoted by point 7 in the lower right corner of the map. He travels to his first activity of the day, denoted by point 1, where he spends 518 minutes (a little more than 8-1/2 hours). His return trip home is far more circuitous than the trip he took to work, making five stops ranging in durations from 4 to 24 minutes, before returning home at point 7 to spend the night. Visualizing trip-chaining behavior in a detailed manner such as this provides insights into the choices that drivers face. Furthermore, by combining trip-chaining data with other spatial data in a GIS, transportation analysts can begin to assess drivers’ choices by comparing sites visited to the range of site options that were not visited.

Conclusion

This paper is an exploration of how an enormous and richly detailed set of GPS-based data might be transformed into a database to help transportation engineers and planners better understand driving behavior. It explains the challenges in converting geocoded data points into a meaningful database that describes the complexity of trips, routes, and tours of chained trips. And it investigates in detail the driving characteristics of a single driver over the course of a month to test how GPS data can be applied to better understand trip-making behavior.

The main lesson from the study is that travel patterns are more complex than traditional travel surveys reveal. Although GPS data require substantial investments in data handling techniques, the database we developed offers several advantages over conventional travel survey data: travel routes between destinations are known; patterns can be observed in a single driver over a full 30 days; and events are captured as a driver is in motion, including speed, acceleration, turns, and sudden stops. The explorations carried out in this paper demonstrate the advantages of GPS data and suggest that the emerging technology will allow transportation engineers and planners to measure more precisely than in previous studies the effect of driving on such urgent issues as pollution emissions and energy consumption.
Bibliography


H. Li, Investigating Morning Commute Route Choice Behavior Using Global Positioning Systems and Multi-day Travel Data, unpublished dissertation (Atlanta, Georgia: Georgia Institute of Technology, 2004).


