Notes on GPS Data Quality

select excerpts from: Characterization of urban commuter driving profiles to optimize battery size in light-duty Plug-in Electric Vehicles. R. Smith, S. Shahidinejad, D. Blair and E.L. Bibeau, 2010 [under review for publication].

The GPS-based data logger used in this study was the Otto Driving Companion model PM2626 manufactured by Persen Technologies Inc., hereafter referred to as ‘Otto’. Powered by the 12V cigarette lighter socket in a vehicle, it is a GPS receiver that records latitude, longitude, speed, date and time at one second intervals (Persen Technologies, 2010).

Seventy-six volunteers were selected on a first-come-first-served basis using a ‘snowball method’ of recruitment. Volunteers were asked to complete a survey to identify their gender, level of education, employment status, household income, number of household members typically using the vehicle in which the GPS would be used, primary uses for the vehicle, and the type, make, model and age of the vehicle.

Positional data quality suffered due to signal interference and multi-bounce caused by natural and artificial barriers such as tall buildings and tree canopies. In contrast, the quality of the speed data was much higher than the positional data, since the Otto measures speed independently of location. The Otto automatically numbers vehicle trips sequentially, with a new trip beginning each time the unit is powered off and then on.

More than 44 million data points were recorded by the 76 Ottos over the one year study. Not unexpectedly, several data quality issues were found to be prevalent in the raw GPS data. Location errors due to signal impedance or multi-bounce frequently placed vehicles outside of roadways or even outside of the study area. Unrealistically high speeds were occasionally recorded. The length and duration of vehicle trips were often underreported, sometimes severely so, due to the lag time involved in acquiring satellite signals after initial power-up. Problems with the 12V DC power supply cable frequently interrupted data capture, splitting what should have been one trip into two or more parts, resulting in an over-reporting of the total number of trips and the under-reporting of trip lengths. Finally, inconsistent use of the GPS devices by project volunteers resulted in a general under-reporting of data; for example, it was apparent that some volunteers often forgot to keep the Otto plugged in for every trip or did not always place the device in a location where it could maintain a good fix on the satellites.

Several procedures were involved in the removal of erroneous data and the reconstruction of missing values. The first was to paste together two or more trips which had been recorded as separate trips due to power interruptions. Following the method of Schönfelder et al. (2002), adjacent trips were pasted together if the gap between ending and starting times was less than or equal to 120 seconds. The assumption was that legitimate breaks between trips would almost always be longer than 120 seconds. The second procedure was based on the observation that the lag time in acquiring a satellite fix usually placed the start of a trip far from its true starting location, but the end of a trip was recorded much more accurately and reliably because there was an established link between the GPS and the satellites by the end of trips. Therefore, the true location of the start of a new trip was forced to be equal to the location of the end of the previous trip, but only if the difference in straight-line distance between these locations was less than 1 km. Also, if the difference in distance between the previous end of trip and the subsequent start of trip was less than 100 m, no adjustment to the start-up position was made.

In those instances when the adjustments were made to the start-up locations, it was necessary to model the missing speed data for the gap in the trip. By determining the straight-line distance between the location of the recorded trip start and the location of the adjusted trip start, the distance missing from the record was calculated. This gap was then filled by randomly selecting micro-trips (bounded by speeds of zero) from the previous trip and pasting them into the gap until it was filled; this procedure was ended when the distance remaining in the gap was less than 100 m. In other words, the "signature" from the previous trip is used to interpolate the missing events in the new trip, the assumption being that the path followed in the previous trip would be an appropriate surrogate for the beginning of the subsequent trip. In the end, 17 percent of the published GPS data was comprised of data reconstructed to fill start-up gaps. Furthermore, for the infrequent occasions when the filling of a gap between two trips with micro-trips from the first trip resulted in a time gap between the trips of no more than 120 seconds, the two trips were joined, as per the first procedure outlined above.

With trip start-up locations and ‘trip stitching’ completed, the next task was to record the locations of the origins and destinations of all trips. These were used to categorize the purpose of trips and the type of parking at that destination. Using the home addresses supplied by the project volunteers, homes were geo-referenced and these locations were compared to mapped trip destinations overlain on a high-resolution digital aerial ortho-photograph of Winnipeg (ATLIS Geomatics Inc., 2005) to verify the location of each volunteer’s home location. For those volunteers who reported that they were employed, the clustering of the trip destinations mapped using ArcGIS was used to identify the most likely location of the place of employment. The five densest clusters, as defined by ArcGIS Spatial Analysis tools, were visually assessed for this process. For all volunteers, all trip destinations in the five densest clusters of destinations not already identified as home or work were categorized into a city zone type as defined by the DMTI Spatial Inc. (2003), shape-file of Winnipeg land-use categories, by identifying the shortest distance between the trip destination and the centroid of the city zone; a destination was required to be within at least 1 km of a centroid to be associated with the zone in question. Examples of zone types include commercial, industrial, residential, governmental and institutional, and recreational. Similarly, with an additional shape-file it was possible to associated trip destinations with building footprints, including schools, shopping centers, churches and hospitals; a destination was associated with a building footprint if it was within 100 m of that building.

The database has been posted online so that other researchers may evaluate our results and use it in their studies (Smith and Blair, 2010). The original raw GPS data has been transformed in several ways. As previously discussed, recording errors as well as gaps created by power losses have been filled and stitched. Data records with speeds greater than 150 km/h were removed because these higher speeds most often appeared to be the result of recording error and were always associated with highway driving not germane to this study. All latitude and longitudes were transformed to protect the confidentiality of the volunteer drivers. More specifically, the position of the starting point of each trip was mapped to (0,0) and subsequent points for that trip were mapped as latitude and longitude deviations from this point. Consequently, it is impossible to identify the location of the home or any other location visited by any volunteer. Data points located outside of a box bounding the City of Winnipeg were retained in the database and accounted for a relatively small proportion (12.4 percent) of the total database.

References:

ATLIS Geomatics Inc., 2005. Digital Aerial Images Ortho-rectified by ATLIS Geomatics Inc.

DMTI Spatial Inc., 2003. CanMap Streetfiles.

Schönfelder, S., Axhausen, K., Antille, N., Bierlaire, M., 2002. Exploring the potentials of automatically collected GPS data for travel behavior analysis – a Swedish data source. GI-Technologien für Verkehr und Logistik 13, Institut für Geoinformatik, Universität Münster, Münster, 155-179.

Smith, R., Blair, D. 2010. Raw data from AUTO21 study of driver behavior, available online, http://auto21.uwinnipeg.ca.