Longitudinal GPS Travel Data and Breach of Privacy via Enhanced Spatial and Demographic Analysis

Vetri Venthan Elango, Sara Khoeini, Yanzhi Xu and Randall Guensler

Overview
- Longitudinal GPS travel data contain a wealth of useful information for model development and policy analysis.
- High resolution GPS data provide accurate spatial and temporal detail for use in analyzing:
  - Household travel and driver behavior
  - Safety
  - Emissions
  - Electric vehicle range requirements
  - Etc.
- Public access to data will benefit other researchers.
- However, participant privacy is a concern.
- Aggressive driving behavior could be subpoenaed.
- Home security can be compromised.
- Human subject's requirements:
  - To protect the privacy of participants, no data may be distributed in any form that can in any way be linked back to an individual, a specific household, a specific vehicle, or a specific work location.
  - The methodological goals of this research effort are:
    - Post-processing high resolution GPS data
    - Retain enough detailed data to be useful in various research communities
    - Ensure privacy of participants.
    - Ensure that separate data sources cannot be “stitched together” to compromise privacy.

Data
- Commute Atlanta Data:
  - Instrumental vehicle research that collected high-resolution GPS (2004-2006)
  - Assess the effects of converting operating costs into variable per-mile driving costs
- Approximately 300 vehicles in 270 households
- 1.2 million vehicle trips
- Approximately 100 households in the pricing study
- It is possible to identify household and work locations when accurate daily trip-level data and detailed second-by-second data are made available

Proposed Commute Atlanta Dataset for Public Access
- Travel diary data (trip-level travel data)
- On-road vehicle activity data (second-by-second)
- Heuristic analysis of each household or vehicle usage data over long periods, behavioral habits can be identified.
- Longitudinal nature of the study – presence of repeated travel
- It is possible to identify household and work locations when accurate daily trip-level data and detailed second-by-second data are made available

Travel Diary Data
- Attributes:
  - Trip Summary Data
  - Primary driver identification, vehicle identification, and household identification
  - Trip details such as distance, start date-time, and end date-time, vehicle non-activity time, origin zone, and destination zone
  - Socio-economic variables
  - Number of vehicles in household, number of vehicles monitored, household size, income, ethnicity, number of workers and students
  - Vehicle Characteristics:
    - Model Year Group, body type, fuel type, and model code
  - Privacy Concerns:
    - Attributes do not directly identify individuals
    - Longitudinal nature of the study – presence of repeated travel
    - Heuristic analysis of each household or vehicle usage data over long periods, behavioral habits can be identified
    - Behavioral habits + socio-economic data can yield the identity of the households and individuals

Figure 1 shows the proposed methodology to identify Home Location using spatial analysis.

On-Road Vehicle Activity Data
- Attributes:
  - GPS data
  - Latitude and Longitude, speed, heading, date and time
  - Number of satellites, position quality information, etc.
  - Background Characteristics
  - HEPMs attributes
  - Georgia Tech Household Classification Group (income, vehicle ownership, and household size)
  - Vehicle Characteristics:
    - Fuel Type, Engine Type, Body Type and Model Year Group
  - Driver Characteristics:
    - Age group and Gender
  - Privacy Concerns:
    - High resolution GPS data can identify participant’s home, work locations,
    - Shopping, recreational and social preferences
    - Driver risk parameters
    - High resolution GPS data + Vehicle Characteristics + Driver Characteristics, yield the identity of individual participants

Figure 4 shows the proposed methodology to identify Home location using spatial analysis.

Summary
- High resolution GPS data can compromise privacy.
- Splitting Commute Atlanta data into two data sets is insufficient because advanced spatial analysis can identify household within a small subset.
- Using demographic data and other data sources such as vehicle registration data without names makes it easier to identify households
- Liability associated with participant privacy protection lies with data collector.
- Future Work:
  - Refine the methodology and use household demographic data purchased from private companies.
  - Develop new techniques and identify variables that need to be withheld or purposefully modified to protect participant privacy.
  - Conduct detailed analysis and make recommendations to reduce liabilities for data collection agencies.

This research was sponsored by the Georgia Department of Transportation under contracts 10-03. Opinions expressed here are those of the authors and not necessarily those of the Georgia Department of Transportation.