Mathematical techniques to ensure privacy in large-scale travel behavior datasets

Jeffrey Dumont, Senior Data Scientist
Joann Lynch, Senior Consultant
Rachel Schmidt, Senior Analyst
Ben Stabler, GISP, Director

RSG, Inc.
White River Junction, Vermont

Introduction

As part of a typical household travel study, researchers collect a vast amount of personal data on survey participants including the socio-demographics of all household members including children (e.g., gender, age, relationship, ethnicity, income) in addition to the household travel behavior (the number and types of trips, exact locations visited in terms of latitude and longitude coordinates, time spent at destinations, travel party, etc.). It has been established that survey participants can reveal a lot about themselves by the way they travel. At the most basic level, even single trips can reveal sensitive information about a traveler. For example, knowing that a person travels to and spends significant time at a house of worship can reveal religious affiliation, or spending time at a rehabilitation facility can reveal addictions. These are examples of information data subjects may prefer to keep private.

Over the past 5 years, the travel behavior survey industry has shifted to using smartphone-based applications to collect household travel data. The data collected from a smartphone-based travel survey is a unique combination of passive data (point traces of exact locations and exact times) and active survey data (self-reported modes, purposes, etc.). This combination allows for more accurate data recording, better trip recall, and more detailed trip traces potentially revealing additional information about the participant. Prior to smartphone-based data collection, respondents self-reported travel behavior through an online or paper-based instrument. From a privacy perspective, this allowed respondents to self-select which trips they reported in order to potentially withhold any sensitive travel. However, with the smartphone approach, all trips are recorded passively (even if the trip survey is not completed). This combination of passive trip data with highly detailed self-reported data results in more concern over data privacy with smartphone-based data collection.

The end users of these datasets are primarily transportation planners and travel forecasters who use these data to inform urban planning decisions. Since these datasets are usually commissioned by public agencies, there can sometimes be a requirement to deliver a privacy-protected public version of the dataset.

Goal

The objective for this work is to quantify the trade-off between different privacy protecting techniques for improving participants' privacy while maintaining the usability, accuracy, and precision of data products for end users according to the metrics specified below. The techniques of interest include data

perturbation (e.g., moving a household 500m in random direction), data switching, the addition of synthetic data, and spatial aggregation (e.g., analysis at the Census block group level as opposed to the GPS point level), all evaluated under differential privacy definitions. This could also potentially include the creation of new differentially private mechanisms that are needed for transportation-specific data.

Travel behavior metrics:

- Socio-demographic comparisons to Census
- Trip rates (overall, by trip purpose, by mode)
- Activity distributions by time of day
- Travel mode shares
- Trip destination distribution
- Vehicles miles traveled (VMT)
- Shared mobility usage

All metrics for various levels of geography (ranging from the region as a whole down to the Census blockgroup level) and time periods.

Related publications by RSG, Inc.: https://meetrmove.rsginc.com/publications/