Plan Overview

A Data Management Plan created using DMP Tool

Title: Reducing Rural Driving with Store Siting

Creator: John Krumm - **ORCID:** 0000-0003-4394-6704

Affiliation: University of Southern California (usc.edu)

Principal Investigator: John Krumm, Cyrus Shahabi

Funder: United States Department of Transportation (DOT) (transportation.gov)

Template: Digital Curation Centre

Project abstract:

People who live in rural areas have to travel farther, because their destinations are farther away. It may be that the addition of a single grocery store, hardware store, movie theater, or some other business could significantly reduce rural travel demand in a region. For instance, it may be that adding a grocery store at a certain location would eliminate thousands of miles of rural vehicle travel per month. While we cannot mandate the creation of a new store, governments can offer a variety of incentives to entice new stores to a specific region.

We propose to study actual movement data from rural residents to discover how adding which types of stores, and where, would be most effective for reducing rural travel. Commercial companies have GPS trajectory data from phones of regular people that reveal the driving habits of hundreds of thousands of people in California. With our access to this data, and our long expertise in geospatial processing geospatial, we can compute where rural residents drive, including the types of stores at their destinations. From this, we can show good places to site new stores for the maximal reduction in driving miles.

Our results will show suggested sites for new stores along with the associated reduction in driving miles. Following our suggestions should significantly reduce miles driven by rural residents.

Start date: 01-01-2024

End date: 12-31-2024

Last modified: 07-08-2024

Reducing Rural Driving with Store Siting

Data Collection

What data will you collect or create?

We will create three types of data:

1. Home locations - These will be computed from existing GPS mobility data that is tagged with an anonymous ID for each person. The home locations will be stored in a table, where each row contains a person's anonymous ID and the latitude/longitude of their home.
2. Trips - These will be stored as a table with one row for each trip. Each row will contain the person's anonymous ID, a trip ID, the start and end locations of the tip as latitude/longitude pairs, the start and end date/times of the trip, the type of start and end points (home, store, etc.), the driving distance of the trip, and the driving time of the trip.
3. Trip statistics - This will include frequent clusters of end points (e.g. hardware store, pet food store, grocery store) and amount of travel to clusters in terms of time and distance.

We expect to create this data for several thousand people in the Los Angeles County area covering a period of 3-12 months. We understand the sensitivity of this data, which we address later in our data management plan.

How will the data be collected or created?

The data will be created from

1. GPS mobility traces that our lab already has. One major source is Veraset, who aggregates GPS data from smart phones.
2. Store and house location data from various public sources

We will develop algorithms to use the above data to compute home locations and trip data.

Documentation and Metadata

What documentation and metadata will accompany the data?

The documentation of the data we create will include:

1. Data access - Where to find the data, which will be in Dryad.
2. Source of data - This will describe the raw GPS mobility traces that we used to compute the data as well as the names of the contributors who computed the data and a pointer to the project's final report that describes the overall goals, methods, and results.
3. Inference algorithms - We will describe the computer algorithms that we used to compute the home and trip data.
4. Schemas - We will describe the schemas of our data. The data will be stored as tables, which means the schemas will describe the data in each column, including its meaning, units, and data type.

Ethics and Legal Compliance

How will you manage any ethical issues?

Source data - The source data consists of GPS mobility traces. We have obtained an IRB exemption from USC to use the GPS mobility data for our research. We understand its sensitivity, and we do not share it beyond USC researchers in our lab who need it for their work. The data is anonymized and secured.

Computed data - The computed data consists of the home locations and trips of people represented in the GPS mobility data. We understand the sensitivity of data, and we will not share it beyond the researchers working on this project. We will publicly share statistics computed about the trips, such as frequent clusters of visited stores and travel distances and travel times.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

We are restricted from sharing the GPS mobility data, and thus we will not share it outside the USC lab members who need it for their research. There are not copyright nor IP/IPR issues with the GPS mobility data, nor with the data we will derive from it.

Storage and Backup

How will the data be stored and backed up during the research?

We will store and back up the GPS data and computed data on our lab's internal servers. The backups will run automatically every week. Existing, experienced graduate students in our lab will be responsible for backups, recovery, and granting access.

How will you manage access and security?

We will grant password-protected access to the computd data to internal USC lab members who need the data for their research on this project. Passwords will be rotated on a regular basis. No one other than current researchers with a need to use the data have access to the data. The data will be properly labeled and segregated as “sensitive” or “linkable to personal information.” We perform regular security audits to confirm these standards.

We have experience storing GPS mobility traces, and the data we derive from these traces for this project will be treated with the same procedures.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

All the data we compute from the GPS mobility traces will have long term value, and thus it should be retained for future validation of our results, for testing against new inference algorithms, and for new research.

What is the long-term preservation plan for the dataset?

There are two types of data to preserve.

1. Sensitive data - This consists of peoples' home locations and trips. We will store this data for three years on our internal servers, with access permissions described above. We will not share it due to its sensitivity.
2. Aggregate trip data - This consists of statistics describing peoples' trips, which is not traceable back to the people making the trips. We will also preserve this on our internal servers for three years as well as share it on Dryad.

Data Sharing

How will you share the data?

We will share the non-sensitive data via Dryad. We will give the link to the Dryad data in our research publications, and there will be no restrictions on who can use it. The Dryad repository will be set up as part of Task 6 of our proposal in October 2024.

Are any restrictions on data sharing required?

There will be no restrictions on sharing the non-sensitive data.

Responsibilities and Resources

Who will be responsible for data management?

John Krumm (jkrumm@usc.edu), the project's Co-PI, will be responsible for all data management. He will supervise the graduate student(s) on this project as they produce, review, store, back up, secure, and share the data.

What resources will you require to deliver your plan?

We will use our lab's existing servers for storing the data.

Planned Research Outputs

Text - "Research Paper"

Our research paper will describe the methods we used to recommend the placement of new stores to reduce the travel burden of rural residents.

Text - "Store Siting Recommendations"

We will create a list of recommended store sitings that will reduce the travel burden of rural residents.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Research Paper	Text	2024-06-08	Open	None specified	1 MB	Custom Data Use Agreements/Terms of Use	None specified	No	No
Store Siting Recommendations	Text	2024-12-30	Open	None specified	1 MB	Custom Data Use Agreements/Terms of Use	None specified	No	No