



RSG

the science of **insight**

Mathematical Techniques to Ensure Privacy in Large-Scale Travel Behavior Datasets

Mathematical Problems in Industry Workshop

14 June 2021

Outline

- A brief introduction to RSG (more later during the panel)
- Background on the problem
- A brief introduction to the data we are sharing





Introduction to RSG

Who we are



Jeffrey Dumont

Senior Data Scientist

15 years at RSG

Based in our White River Junction VT office

Professional interests include:

- Survey sample design
- Advanced statistical analysis including discrete choice analysis
- The application of data privacy



Who we are



Rachel Schmidt

Senior Analyst

4 years at RSG

Based in our Burlington VT office

Professional interests include:

- Survey sample design
- Tool development for processing travel data
- Data QA/QC



We've been in business for 35 years and have expert staff across the country



Founded by Dartmouth College professors



35 years in business



85+ professional staff across six offices



Core market is transportation-related consulting



100% Employee Owned!



Areas of expertise - Research & Analytics

- Survey-based Research
- Passive mobility data analytics
- Discrete choice modeling
- Travel model development
- Freight modeling
- Noise control engineering
- Software development



Areas of expertise - Strategic Insights

- Transportation forecasting
- Strategic transportation planning
- Pricing strategies
- Product & service development



Academic roots

Collaborations with:

- Dartmouth College/Tuck School of Business
- University of Leeds
- MIT
- UT Austin
- University of Washington
- Harvard
- Arizona State University
- To name just a handful ...





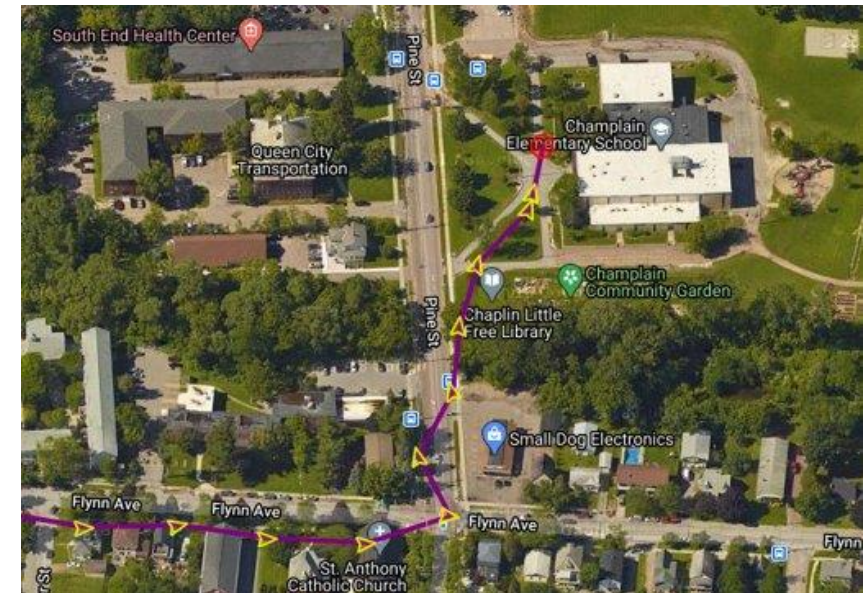
Introduction to our problem

What is a household travel study (HTS)?

Household travel studies are the standard data collection process to gather information to inform investment and urban planning decisions around transportation infrastructure.

A typical household travel study collects details on:

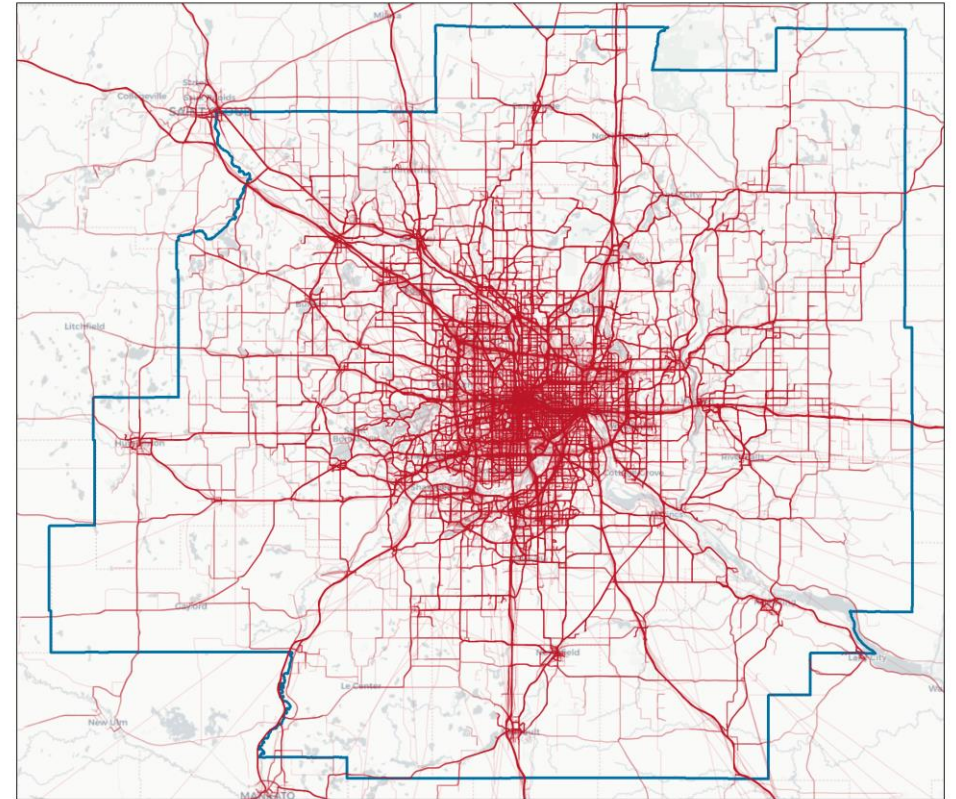
- The household – relationships, income
- Individuals in the household – age, gender, race/ethnicity
- Habitual locations – home, work, and school
- Trips made during travel assigned travel period – the reason for the trip, who you traveled with
- Trip traces – breadcrumbs along the way



How are the data used?

Important metrics include:

- Trip rates (overall, by trip purpose, and by mode)
- Activity distribution by time of day
- Travel mode shares
- Trip destination distributions
- Vehicle miles traveled (VMT)
- Shared mobility usage
- Socio-demographic comparisons to census data
- Data also support the estimation of travel related models
 - Choice of route, destination, mode



Who commissions them?

These studies are often commissioned by public agencies who often have requirements on providing collected data to the public.

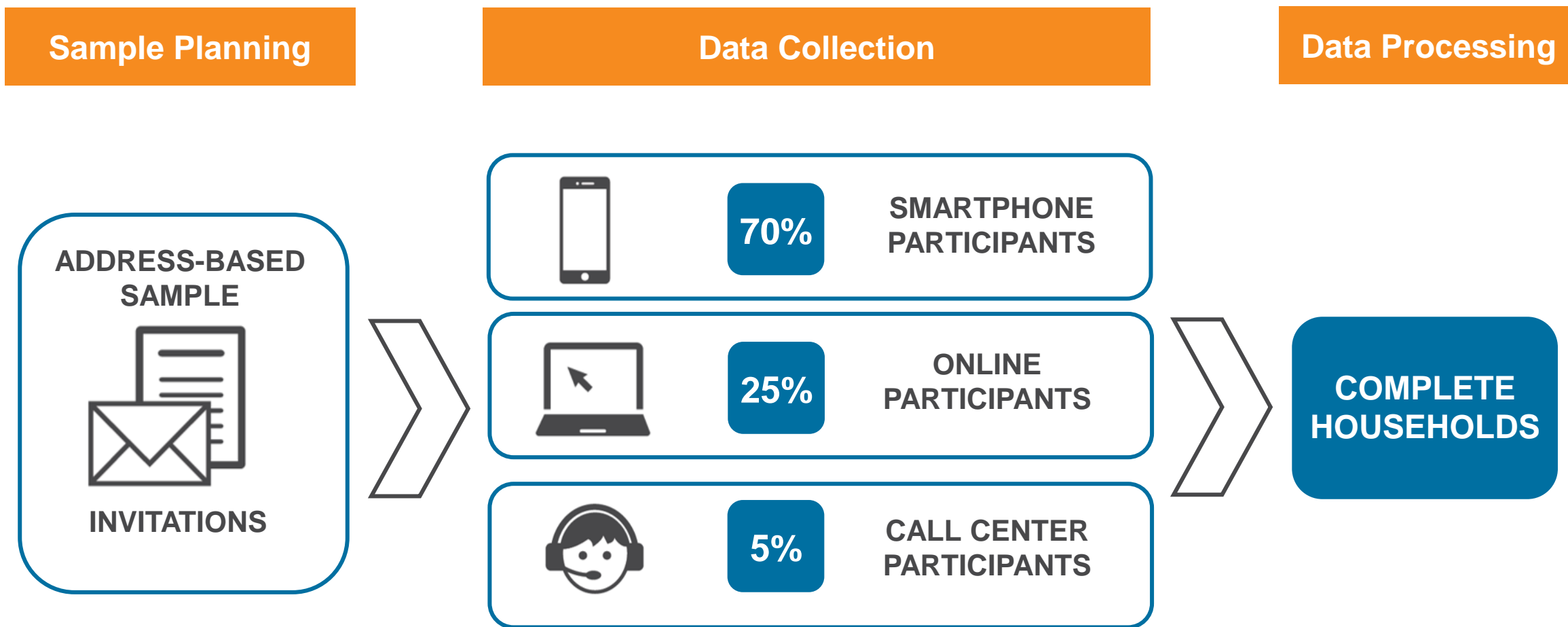
The current privacy protection solution is spatial aggregation to the census block group level.



Institute for Transportation Research and Education



Our approach to a traditional HTS

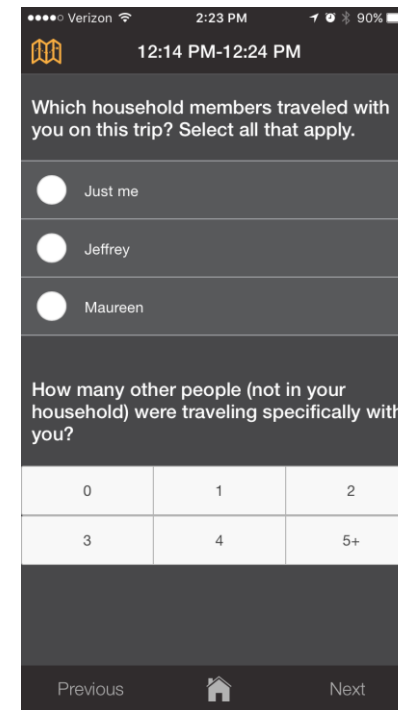
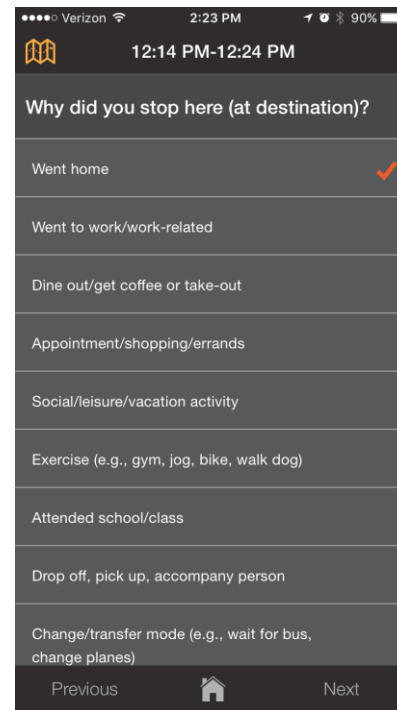
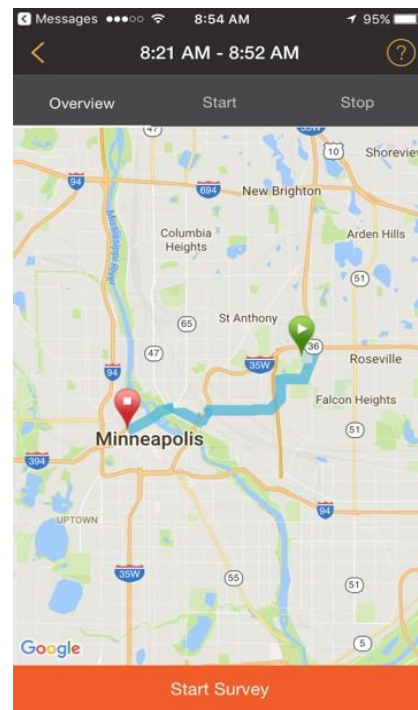
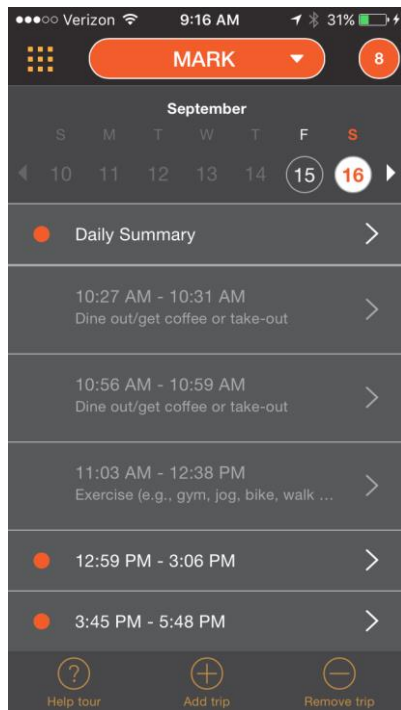


Smartphone app participation

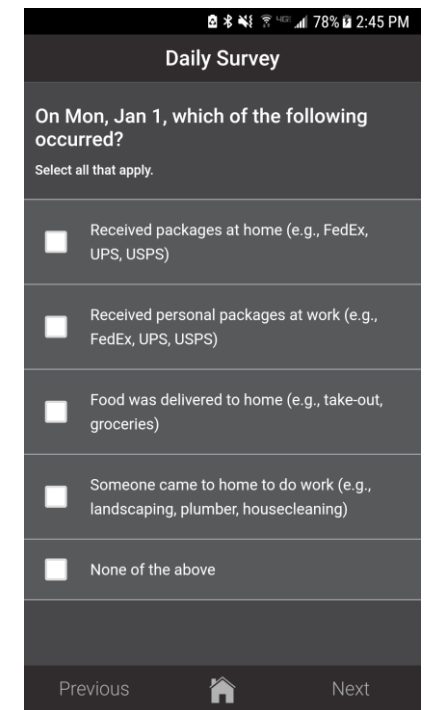


Passive/automatic collection of spatial data for seven days **AND** prompted in-app surveys

TRIP SURVEY



DAILY SURVEY



Changing data privacy landscape

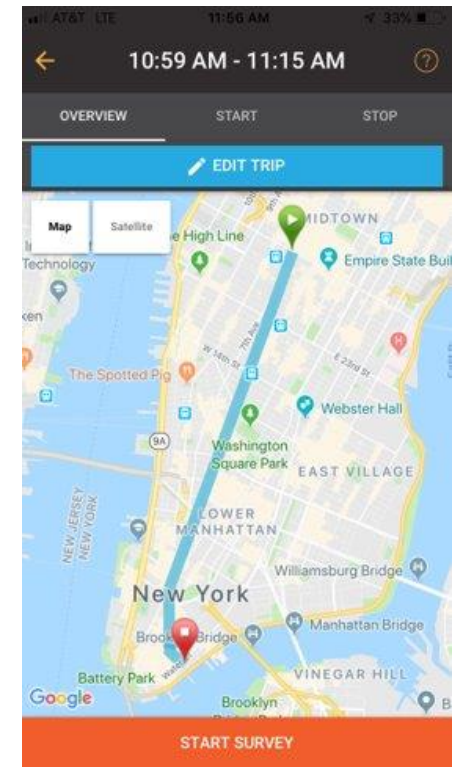
On top of RSG's desire to be good data stewards, there is an ever-developing data privacy legal landscape.

In the last 3 years, we have had:

- General Data Protection Regulations (GDPR)
- California Consumer Privacy Act/California Privacy Rights Act
- Numerous other state and countries passing privacy regulations

These new laws all have in common:

- Provide for data subject rights
 - To view, correct, or delete any data collected on an individual
- Need a legal justification for collecting and processing data
 - Consent is the most robust and universal method of justification



Privacy

Participants are presented with detailed privacy documentation and opt-in to our studies.

Project-Specific Privacy Documentation

Who is conducting the survey?

The *Travel Behavior Inventory* is being conducted on behalf of the Metropolitan Council by Resource Systems Group, Inc. (RSG). RSG conducts market research on behalf of both public and private sector clients using software applications, smartphones, websites, surveys, computers, tablets, and other means of collecting data (collectively, "Research Tools"). The data RSG collects is governed by this [privacy policy](#).

In the event of a conflict between the terms of the Project-Specific Documentation below and the terms of RSG's [privacy policy](#), the terms in the Project-Specific Documentation below will govern.

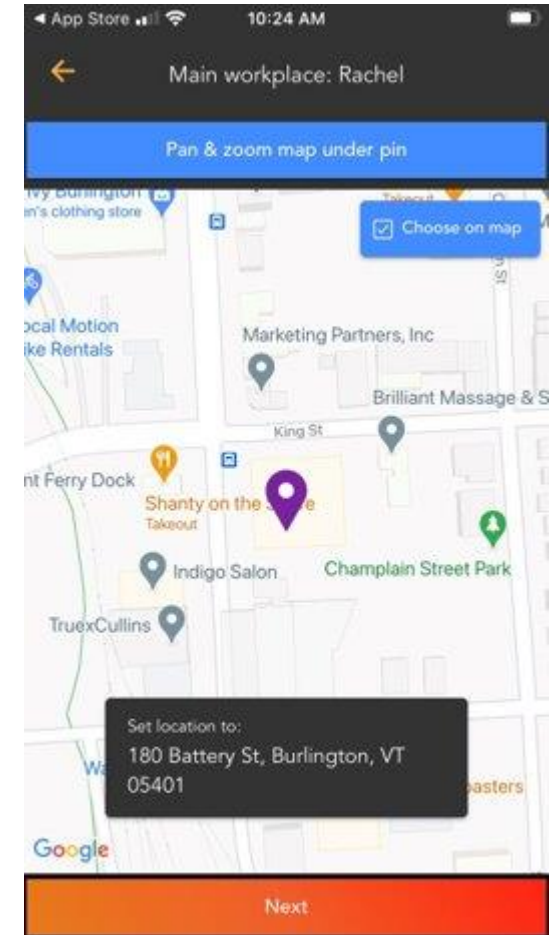
What are we collecting?

RSG will need to collect and use certain personal information from you for this survey, including:

- Travel information – for example, information about trips you make including precise locations, travel times, and travel mode.
- Demographics – for example, race, ethnicity, and household income to ensure that the survey sample reflects the population of the study region.
- Your household makeup – for example, who is in your household and your relationship to these household members including marital status and gender identity of you and your spouse/partner.
- Contact information – for example, an email address or phone number for the purposes of sending survey reminders and distributing gift cards.
- Computer settings – for example your computer's or device's IP address.

RSG is committed to protecting the confidentiality, integrity, and security of your personal information. We will not disclose or share personal information we collect from you except as required by law or described in this Project-Specific Documentation and our privacy policy.


How are we using this data?



Privacy is ever in the news!

MIT News

ON CAMPUS AND AROUND THE WORLD

 [SUBSCRIBE](#)

3 Questions: The price of privacy in ride-sharing app performance

JTL Urban Mobility Lab researchers examine the effects of protecting user data privacy on the efficiency and service quality of ride-sharing applications.

Kelley Travers | MIT Energy Initiative
October 14, 2020

POPULAR SCIENCE

[SCIENCE](#) [TECH](#) [DIY](#) [REVIEWS](#) [SUBSCRIBER LOGIN](#)

[NEWSLETTER SIGN-UP](#) 

Smartphone location data still poses a real security risk for the military and its personnel

Commercially available data can provide a scary level of information.

BY KELSEY D. ATHERTON MAY 08, 2021

[TECHNOLOGY](#)

[MILITARY](#)

The New York Times

Opinion | [THE PRIVACY PROJECT](#)

You Should Be Freaking Out About Privacy

Nature Scientific Reports Articles Article

[nature](#) [scientific reports](#) [articles](#) [article](#)

[Open Access](#) | [Published: 25 March 2013](#)

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel

The New York Times

[Scientific Reports](#) **3**, Article number: 1

35k Accesses | **604** Citations | **196**

Opinion | [THE PRIVACY PROJECT](#)

The Loophole That Turns Your Apps Into Spies

Just by downloading an app, you're potentially exposing sensitive data to dozens of technology companies, ad networks, data brokers and aggregators.



Problem objective

To understand the trade-off between different techniques for improving participants' privacy while maintaining the usability, accuracy, and precision of the data products within the context of a differential privacy framework.

The techniques of interest include:

- Data perturbation
- Data switching
- The addition of synthetic data
- Spatial aggregation
- Or some combination of these

Open to other mechanisms that might be needed for transportation-specific data.

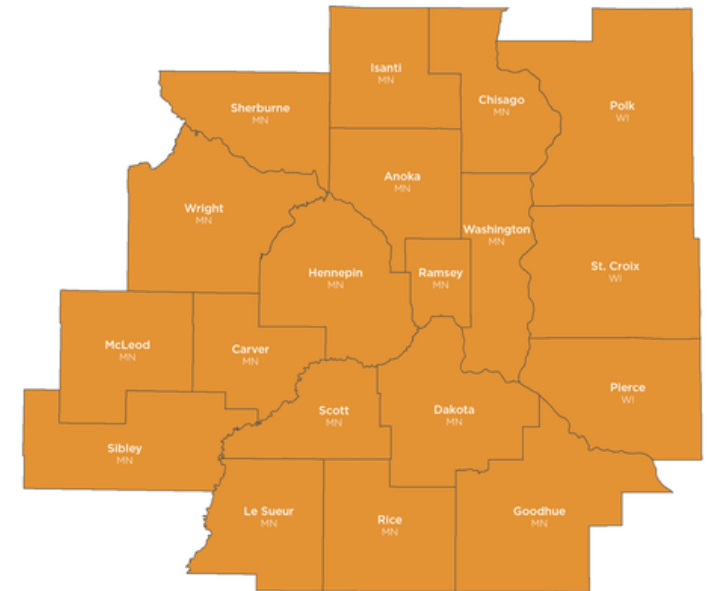




Introduction to our data

About the data

- Survey fielded from **October 1, 2018 through September 30, 2019.**
- Study region was the **19 counties surrounding Minneapolis/St Paul**
- **Smartphone participants completed a 7-day travel diary.**
- **Online and call center participants completed a 1-day travel diary.**
- Same questionnaire was used for smartphone, online, and call center participants.
- Survey was available in English, Spanish, Karen, Oromo, Somali, and Hmong.



Data counts

DATA TABLE	RECORD COUNT
Household	7,837 households
Person	16,152 persons completing one or more days
Vehicle	13,431 vehicles
Day	84,582 days, of which 76,415 are complete travel days
Trip	351,177 trips, of which 315,272 are on complete travel days
Location	6,830,105 trip location/GPS points (latitude/longitude/timestamp)



Household Data Subset

	hh_id	income_detailed	num_people	num_adults	home_bg_perturbed	home_county_perturbed	home_state_perturbed	residence_type	num_vehicles
1	18112062	5	5	2	270370614012	27037	27	1	3
2	18114244	1	6	2	270530022001	27053	27	1	0
3	18114304	6	1	1	551091202021	55109	55	1	1
4	18114536	5	2	2	271410305042	27141	27	1	2
5	18114695	6	4	2	271711007021	27171	27	1	2
6	18115411	10	2	2	271711008022	27171	27	1	2
7	18115821	4	1	1	271230334001	27123	27	1	2
8	18116352	4	1	1	270531039001	27053	27	4	1
9	18116453	5	2	2	270531013002	27053	27	1	2
10	18116766	4	1	1	550959611003	55095	55	1	1
11	18117100	7	1	1	271230352002	27123	27	4	1
12	18117111	7	4	2	270530259053	27053	27	1	2
13	18117117	5	1	1	271230355002	27123	27	1	1
14	18117219	2	1	1	270531091001	27053	27	3	0
15	18118424	4	2	2	270530003002	27053	27	1	1
16	18118923	5	2	2	271390803021	27139	27	4	2
17	18119034	999	1	1	270530259061	27053	27	2	1
18	18119342	7	4	2	271310709012	27131	27	1	3
19	18119982	9	4	2	270530265103	27053	27	1	4
20	18120420	4	1	1	270530240032	27053	27	4	1
21	18120511	9	5	2	270030508135	27003	27	1	2
22	18121027	4	1	1	271230357002	27123	27	4	1
23	18121364	3	1	1	271230416022	27123	27	-9998	1



Trip Data Subset

	trip_id	o_bg_perturbed	d_bg_perturbed	d_purpose_category_imputed	mode_type	depart_time	arrive_time
1	1811206201001	270370614012	270490805003	2	8	2018-12-06 18:30:00	2018-12-06 19:00:00
2	1811206201002	270490806001	270370614012	1	8	2018-12-07 04:00:00	2018-12-07 04:30:00
3	1811206203001	270370614012	270370611022	4	2	2018-12-06 11:45:00	2018-12-06 12:45:00
4	1811206203002	270370611022	270370614012	1	2	2018-12-06 20:00:00	2018-12-06 21:00:00
5	1811206204001	270370614012	270370611023	4	2	2018-12-06 16:45:00	2018-12-06 17:15:00
6	1811206204002	270370611023	270370614012	1	2	2018-12-06 19:30:00	2018-12-06 20:00:00
7	1811206205001	270370614012	270370611023	4	2	2018-12-06 16:45:00	2018-12-06 17:15:00
8	1811206205002	270370611023	270370614012	1	2	2018-12-06 19:30:00	2018-12-06 20:00:00
9	1811424401001	270531023001	270530001014	6	9	2018-12-10 20:36:23	2018-12-10 20:51:24
10	1811424401002	270530001013	270530214002	6	9	2018-12-10 23:29:33	2018-12-10 23:43:26
11	1811424401003	270530214002	270530022001	1	9	2018-12-11 00:00:49	2018-12-11 00:13:14
12	1811424401004	270530022001	270530078012	8	9	2018-12-13 22:19:52	2018-12-13 22:41:43
13	1811424401005	270530078012	270531023001	1	9	2018-12-14 02:00:08	2018-12-14 02:14:16
14	1811424401006	270530022001	271230344002	8	9	2018-12-14 15:32:58	2018-12-14 16:21:32
15	1811424401007	271230344002	271230428001	8	9	2018-12-14 17:02:20	2018-12-14 17:10:10
16	1811424401008	271230428001	271230352001	6	9	2018-12-14 17:12:35	2018-12-14 17:23:06
17	1811424401009	271230352001	270530022001	1	9	2018-12-14 17:48:22	2018-12-14 18:08:58
18	1811424401010	270531023001	270530001013	6	9	2018-12-14 20:37:45	2018-12-14 20:52:15



Key Variables

Table	Variables of interest	Variable names
household	Household identifier Home location Home location aggregated	hh_id home_<lat/lon>_perturbed home_<bg/county/state>_perturbed
person	Person identifier Work/School location Work/School location aggregated	person_id <work/school>_<lat/lon>_perturbed <work/school>_<bg/county/state>_perturbed
trip	Trip identifier Trip origin and destination location Trip origin and destination location aggregated Trip departure time Trip arrival time	trip_id, <o/d>_<lat/lon>_perturbed <o/d>_<bg/county/state>_perturbed depart_time_imputed arrive_time,
location (trip traces)	Trip identifier Location of the collected point Timestamp of the collected point	trip_id <lat/lon>_perturbed collected_time

In addition, there is a detailed codebook that gives more information on the data's meaning.



Primary Keys

TABLE NAME	VARIABLE(S) TO JOIN TO OTHER SURVEY DATA TABLES
Household	hh_id
Person	hh_id, person_id
Vehicle	hh_id
Day	hh_id, person_id, day_num
Trip	hh_id, person_id, day_num, trip_id
Location	trip_id



Data perturbation

To protect privacy and to be able to share data with the team:

- All location data in a household was shifted in a random direction and a random distance
- This maintains the integrity of the trip traces and distances and durations
- However, plotting on a map will be slightly erroneous



Questions?





Contacts

www.rsginc.com

Jeffrey Dumont

Senior Data Scientist

jeff.dumont@rsginc.com

Rachel Schmidt

Senior Analyst

rachel.schmidt@rsginc.com