A Data Science Framework to Measure Vehicle Miles Traveled by Mode and Purpose with Location-Based Service (LBS) Data

Shangqing Cao and Marta C. González

Agenda

1. Comparison between Location-Based Service (LBS) data and Call Detail Records (CDRs)
2. COVID-19 and change in VMT
3. COVID-19 and change in trip purpose
4. COVID-19 and change in residential locations
5. Mobility policies and change in Vehicle Usage Rate (VUR)

Questions: (1) Is Location Based Service (LBS) data able to capture human mobility patterns at the same quality level as Call Detail Records (CDRs)? **(2)** How can we use LBS data to measure change in human mobility behaviors in California due to COVID-19?

1. LBS and CDR Data

Background

- There is a large body of work from the past decades validating the use of CDRs to estimate human mobility patterns, but CDRs are expensive and difficult to obtain
- LBS data has recently become more widely available and is at a much higher resolution than CDRs.
- But, LBS data remains to be as thoroughly validated as CDRs

Question: Is Location Based Service (LBS) data able to capture human mobility patterns at the same quality level as Call Detail Records (CDRs)?

Berkeley
UNIVERSITY OF CALIFORNIA

CALIFORNIA
AIR RESOURCES BOARD

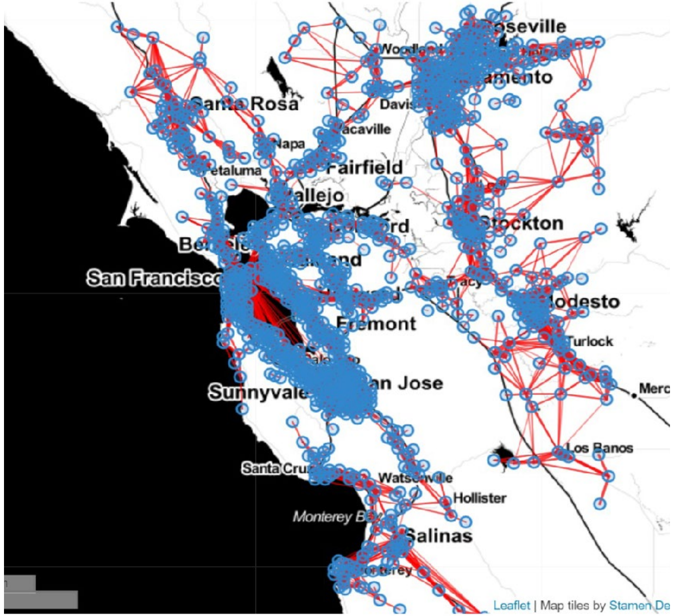
Datasets in the Comparative Study

Call Detail Records (CDRs)

- **Dataset:** 7 months of anonymized individual registers for SMS, voice calls, and data traffic.
- **Spatial Resolution:** records are associated with cell towers

Location Based Services (LBS)

- **Dataset:** 6 months of point locations for anonymized users in CA.
- **Spatial Resolution:** point data with accuracy ranging from 10m to 500m



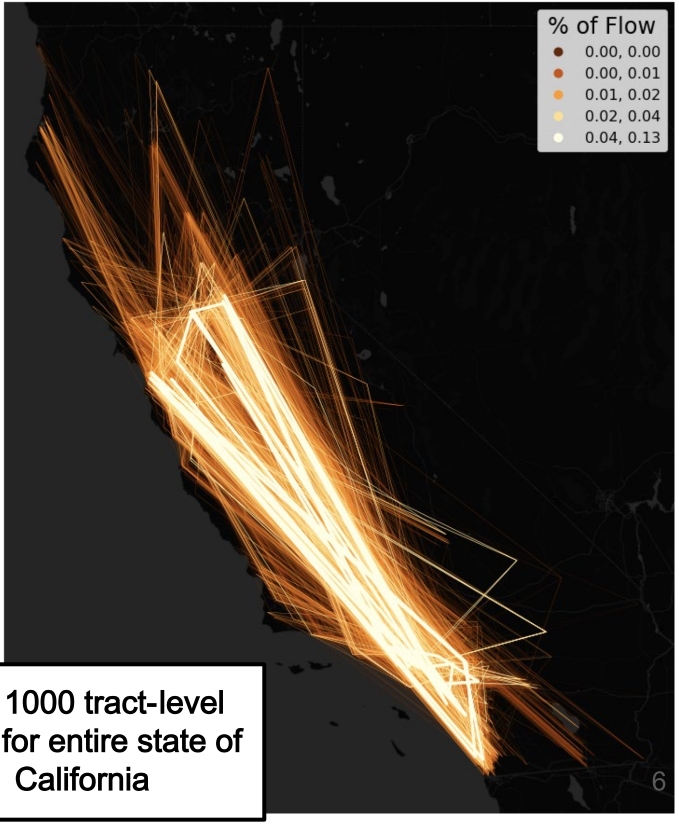
Validation Set: Census Transportation Planning Package

Census Transportation Planning Products Program (CTPP) generates flows between home and work locations based on the ACS and the Census

Most recent dataset: 2012-2016 based on the 2016 American Community Survey

- Tract to Tract
- Census Designated Place (CDP) to CDP
- County to County
- TAZ to TAZ

We can use these datasets for validating our flow estimations



Top 1000 tract-level flows for entire state of California

Methodology: Overview

Stay Detection → Work and Home detection → Population Expansion

Methodology: Stay Detection

Stay detection is the process of converting raw records (which are noisy) into meaningful stays, or instances where a user spent a significant amount of time.

Stay detection requires that thresholds be set for:

- Spatial resolution (tessellation)
- Minimum time duration (20 minutes)

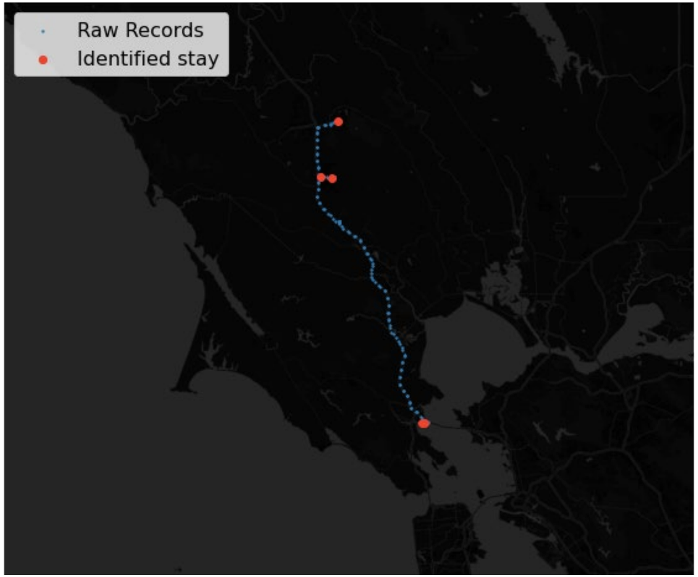


Figure: Example raw records and detected stays for a sample user for a sample day in the LBS dataset (H3 tessellation)

Methodology: Stay Detection

Tessellating CDRs

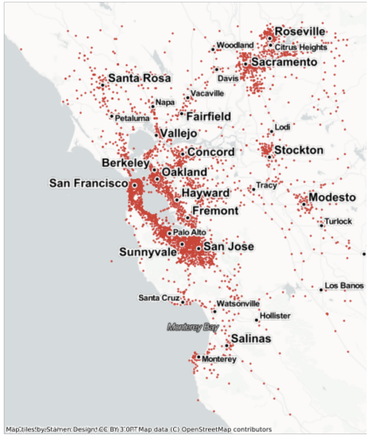
- To estimate cell tower coverage we use a **voronoi tessellation** to create coverage polygons
- Stays will be associated with these cell tower voronois

Tessellating LBS

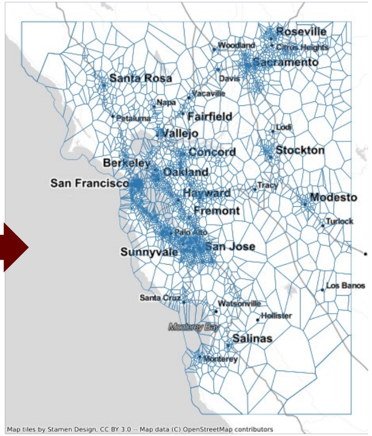
- We use **H3 library** from Uber, size 9 (approx. 0.10 km^2 per hexagon)

Defining stay using different in time between consecutive records and the difference in distance between consecutive records

Cell Tower Locations



Voronoi Tessellation



Berkeley UNIVERSITY OF CALIFORNIA | CALIFORNIA AIR RESOURCES BOARD

Methodology: Work and Home Detection

Home Identification:

- The most frequently visited night time location for a user
  - Must meet minimum visitation frequency requirements

Work Identification:

- Maximizes distance from identified home and visit frequency during work hours
  - Must meet minimum visitation frequency and minimum distance from home

Algorithm 2 Estimating Individual Commute Pattern

Step 1: Home Detection
for *user* $\in$ *Users* **do** ▷ Loops through each user in the stays
  Take only records between 8 p.m. and 7 a.m.
  Count number of visits to each location, determining most frequently visited nightly location, $geoid_{home}$, and the number of visits to that location, n_{home}
  if $n_{home} <$ once per week **then**
    Remove the user's records
  end if
end for
return selected data

Step 2: Work Detection
for *user* $\in$ *Users* **do** ▷ Loops through each user with identified home
  Take only records between 8 a.m. and 7 p.m. on weekdays
  for *geoid* $\in$ *Geoids* **do** ▷ Loops through all unique locations visoted by user
    Calculate distance d_{geoid} between *geoid* and user's home
    Number of visits to this geoid, n_{geoid}
  end for
  Find geoid, *work*, that maximizes $n_{geoid} \times d_{geoid}$
  if $d_{work} < 0.5$ miles OR $n_{work} <$ once a week **then**
    Discard *work* for this user
  end if
end for
return home work pairs for remaining users

Methodology: Population Expansion

In order to scale our data to population level, we must determine the expansion coefficient (expansion factor) we must use for each Census Designated Place

- Once we expand our dataset to commuter populations, we can compare to CTTP flow estimates

Expansion factor for CDP ***tr***

$$f_{tr} = \frac{p_{tr}}{h_{tr}}$$

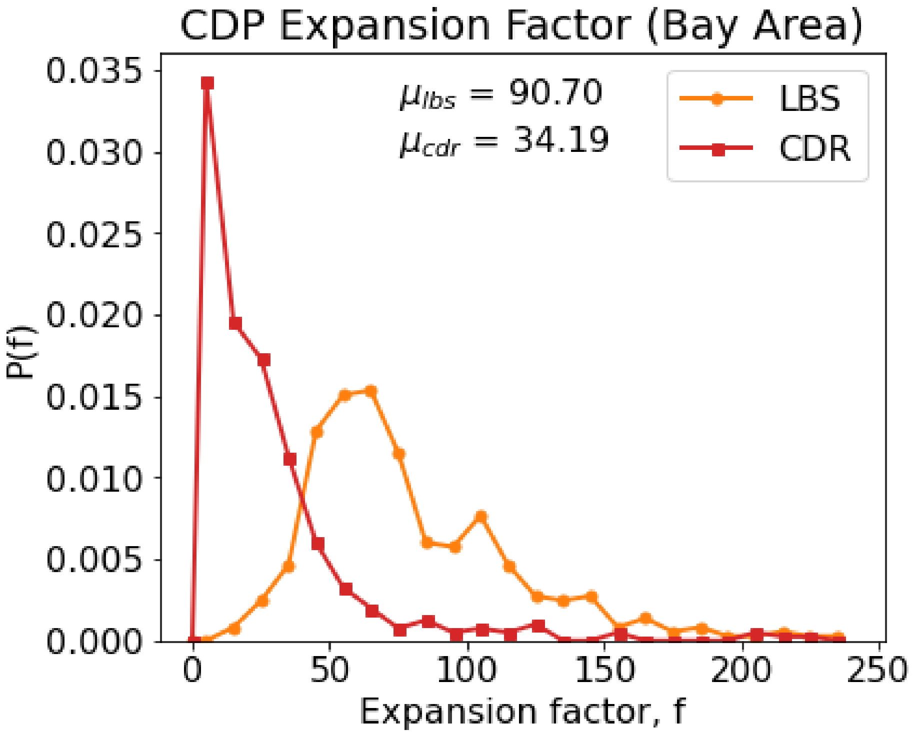
Total population of CDP (from American Community Survey)

Number of homes identified in tract from our data

Methodology: Population Expansion

The distribution of expansion factors shows us on average how much we need to scale our data.

$$f_{tr} = \frac{p_{tr}}{h_{tr}}$$



Pre-Expansion Population Estimates

CDR:

- Must visit identified home location at least once per week (n = 12 for a 3 month analysis)
- Work locations are the stop that are most frequently visited during work hours (8 am to 7 pm) and are the furthest from home

176,729 commuters (home and work identified) out of 1,641,401 users with home (~10%)

LBS

- Initial filtering based on user activity (minimum number of days present and minimum average number of daily records)
- Work locations are the stop that are most frequently visited during work hours (8 am to 7 pm) and are the furthest from home

129,707 commuters (home and work identified) out of 488,414 users with home (~26%)

Pre-Expansion Population Estimates

Description	CDR	LBS
Unique user IDs in raw data	64,889,141	4,520,038
Users with homes in Bay Area	1,641,401	488,414
Users with identifiable works in Bay Area	554,385	205,589
Users with identifiable works that can be assigned to CDPs in the Bay Area	**176,729**	**129,707**

Validating the Expansion Process

CDR Population Expansion

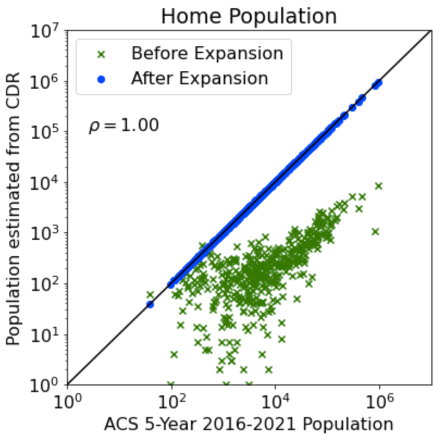
ρ = 0.85 for work population expansion

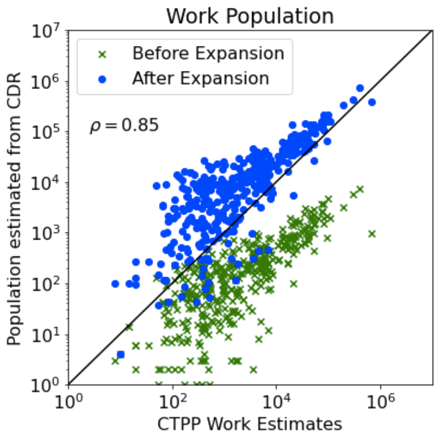
Bay Area • 382 places

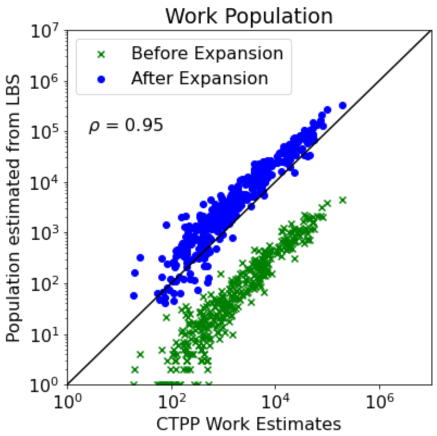
LBS Population Expansion

ρ = 0.95 for work population expansion

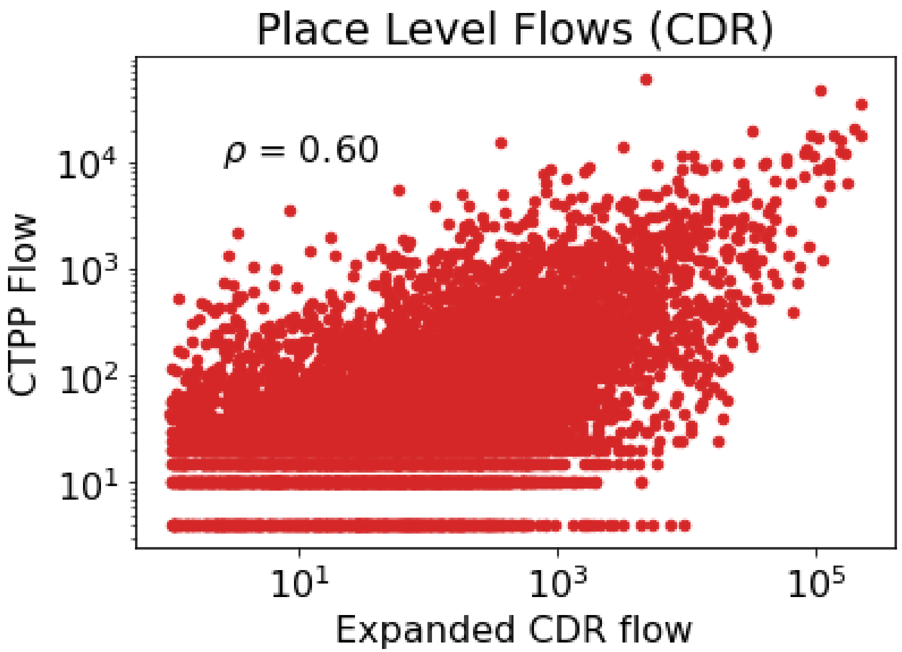
Bay Area • 353 places

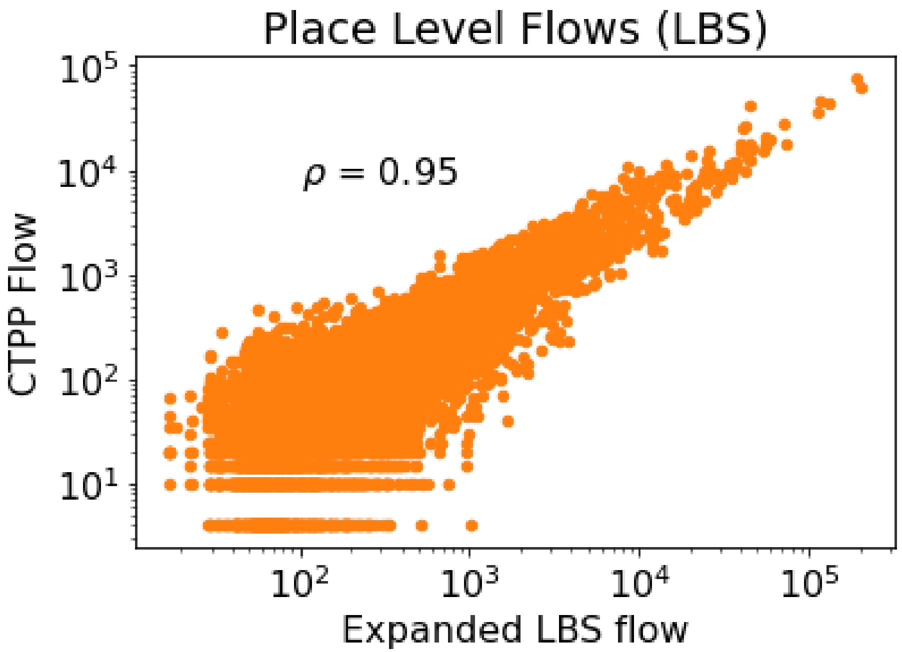






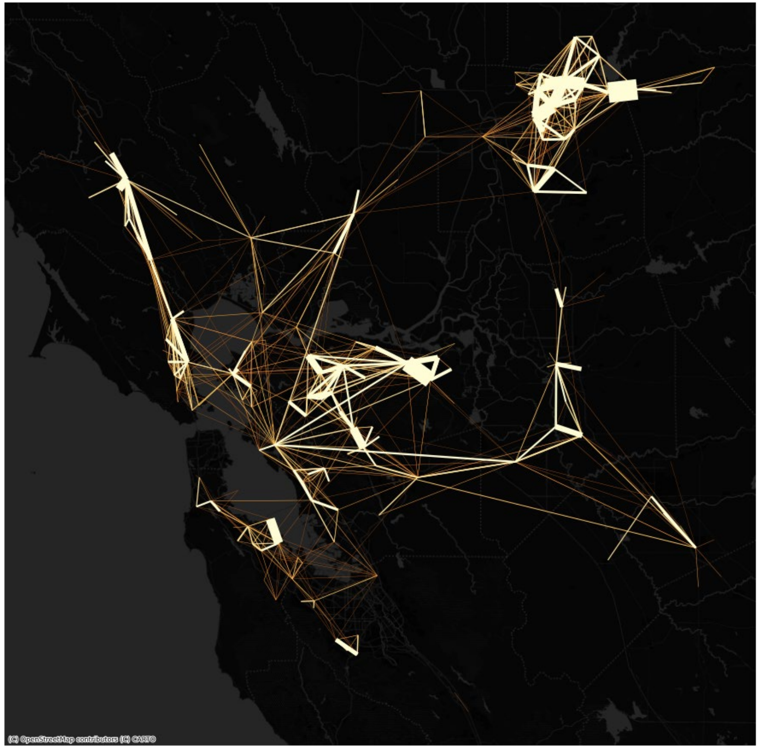
Validating Commuter Flows





Validating Commuter Flows

Top 10% Expanded CDR



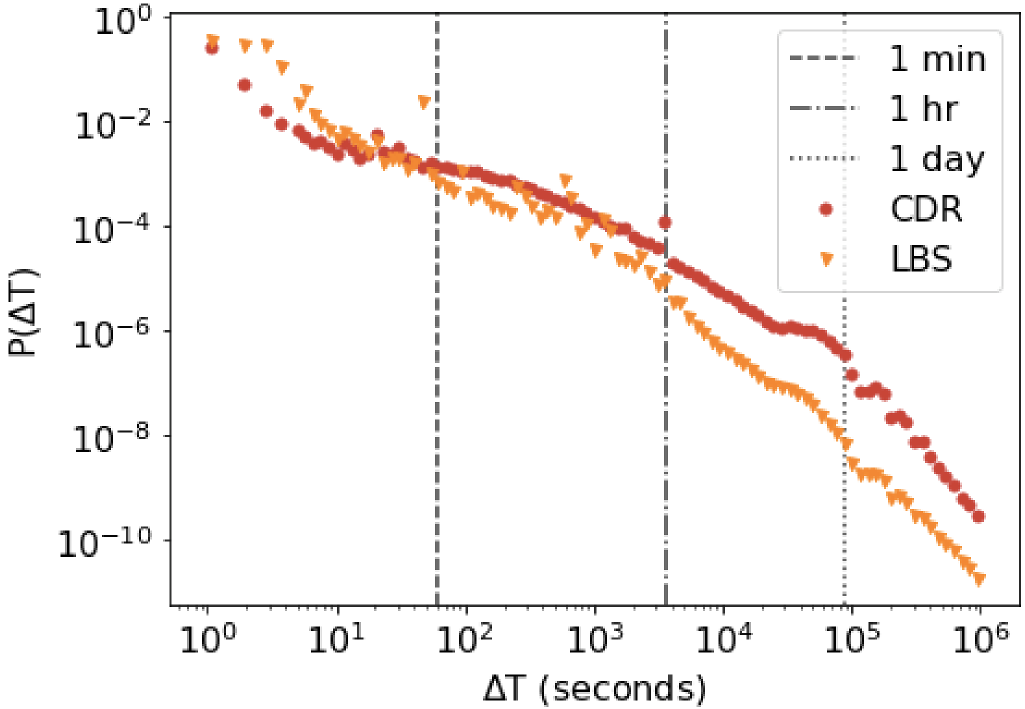
Top 10% CTPP Flows

Validating Commuter Flows

Top 10% Expanded LBS Flows

Top 10% CTPP Flows

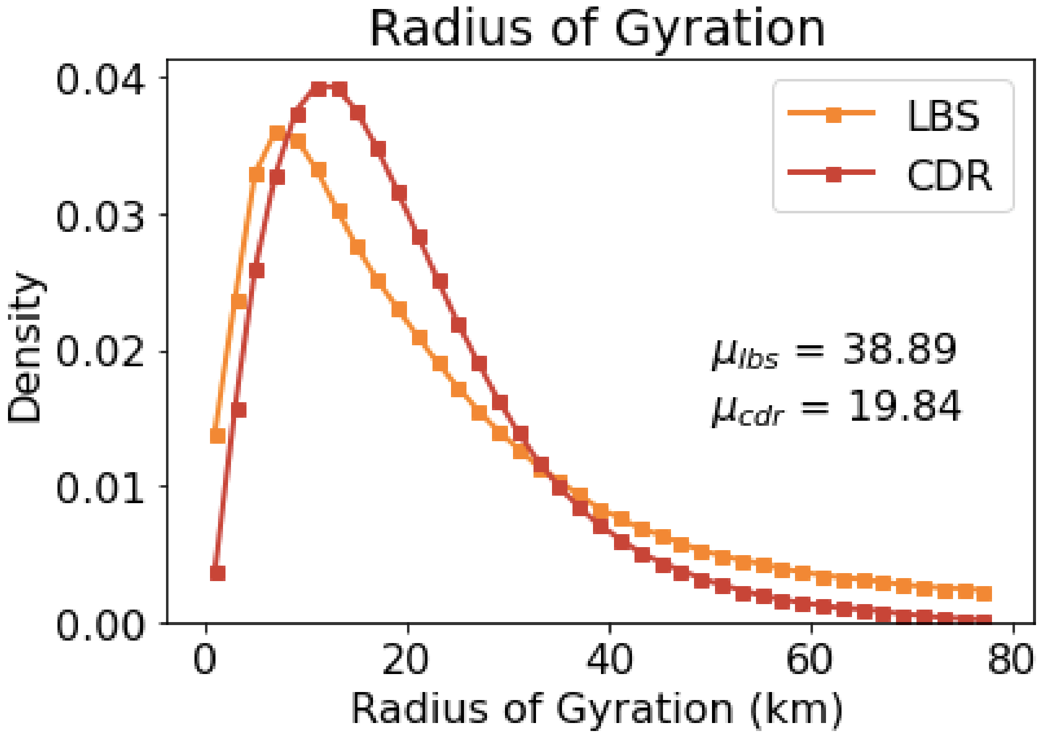
Comparing the Burstiness of Raw Data



Burstiness is the time difference between consecutive records.

Here we show the probability distribution of time deltas for both datasets, finding remarkable similarity.

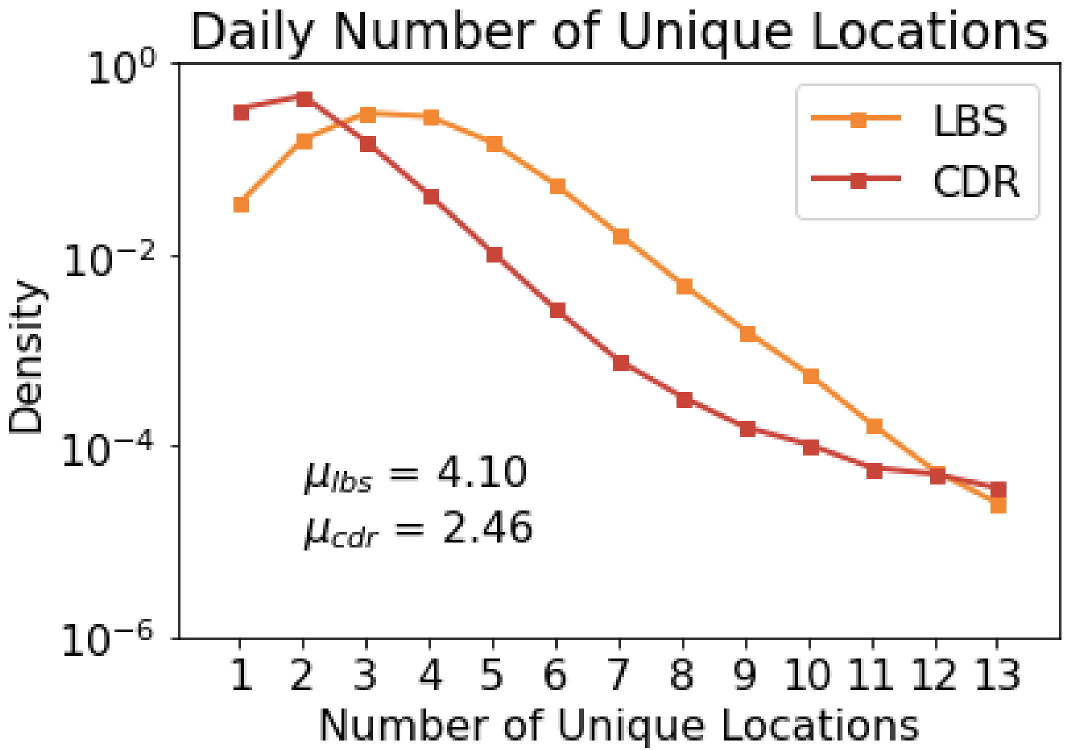
Comparing Individual Mobility Patterns



Radius of gyration rg(*t*) indicates the characteristic distance (in meters) travelled by a user over some period of time *t*

$$r_g(t) = \sqrt{\frac{1}{N(t)} \sum_{i=1}^{N(t)} (\mathbf{r} - \mathbf{r}_{\mathbf{cm}})^2},$$

Comparing Individual Mobility Patterns

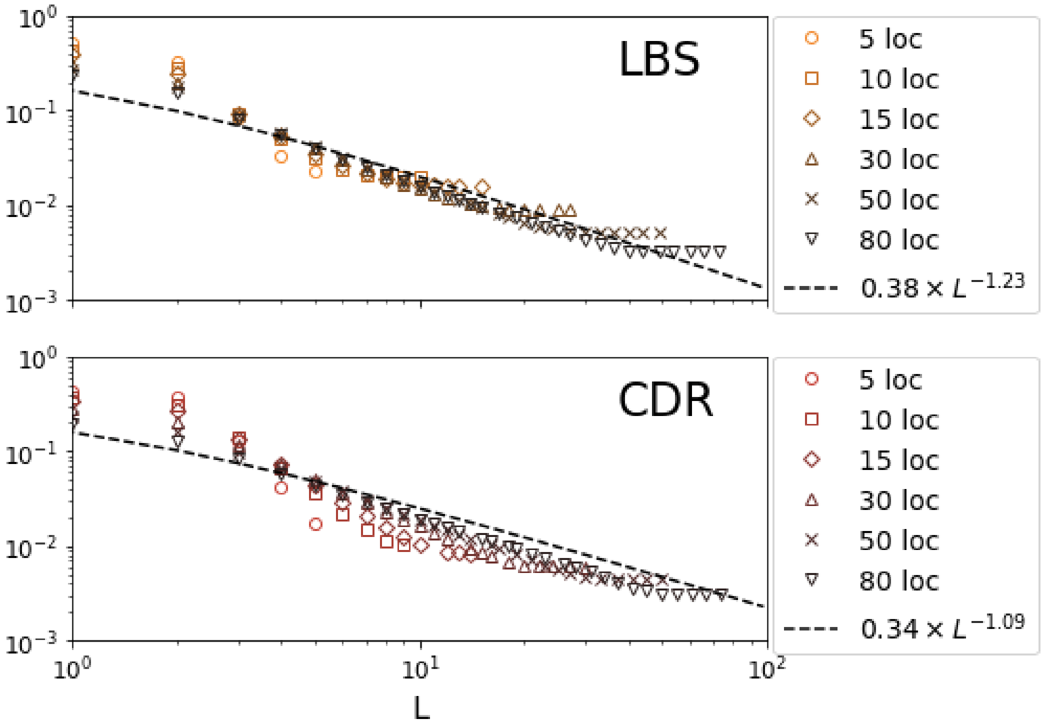


Number of daily unique locations indicates how many unique locations each individual visits per day, on average.

We can see that on average, LBS users visit 4 locations per day and CDR users visit 2.5

Observation Period: 3 months

Comparing Preferential Returns

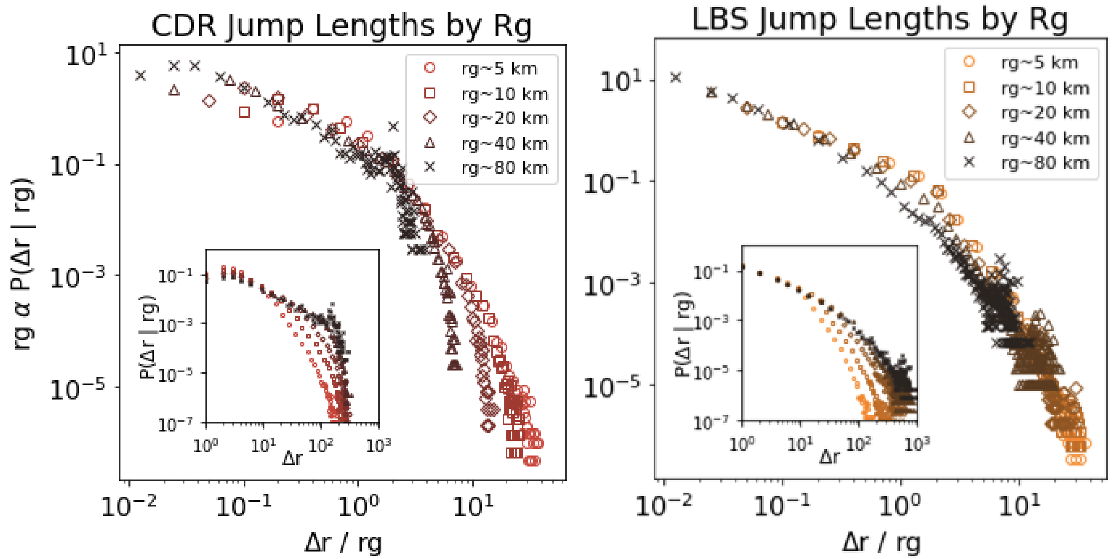


Preferential returns* show that individuals tend to return to locations they have visited before.

We are able to create a distribution of the probability of visiting a given location given its frequency rank, L for different user groups.

*Song, C., Koren, T., Wang, P. et al. Modelling the scaling properties of human mobility. Nature Phys 6, 818–823 (2010). https://doi.org/10.1038/nphys1760

Conditional Jump Length Distributions

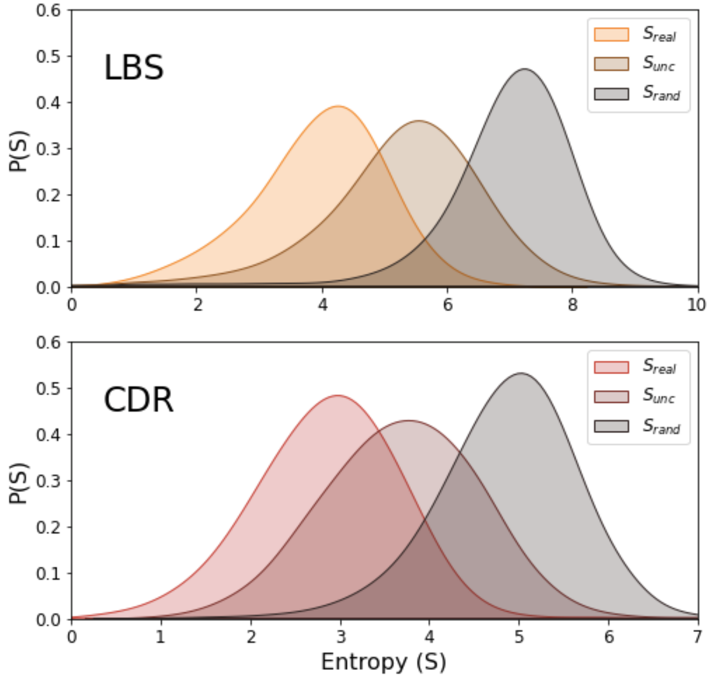


Conditional Jump Lengths* represent the likelihood of a user traveling a distance based on their radius of gyration.

.

*González, M., Hidalgo, C. & Barabási, AL. Understanding individual human mobility patterns. Nature 453, 779–782 (2008). https://doi.org/10.1038/nature06958

Comparing Entropy Estimates



Entropy* estimates the predictability of a user's behavior.

We show the distributions of individual entropy for random, uncorrelated, and "real" entropy for CDR and LBS users

$$S_i^{rand} = \log_2 N_i$$

$$S_i^{unc} = -\sum_{j=1}^{N_i} p_i(j) \log_2 p_i(j)$$

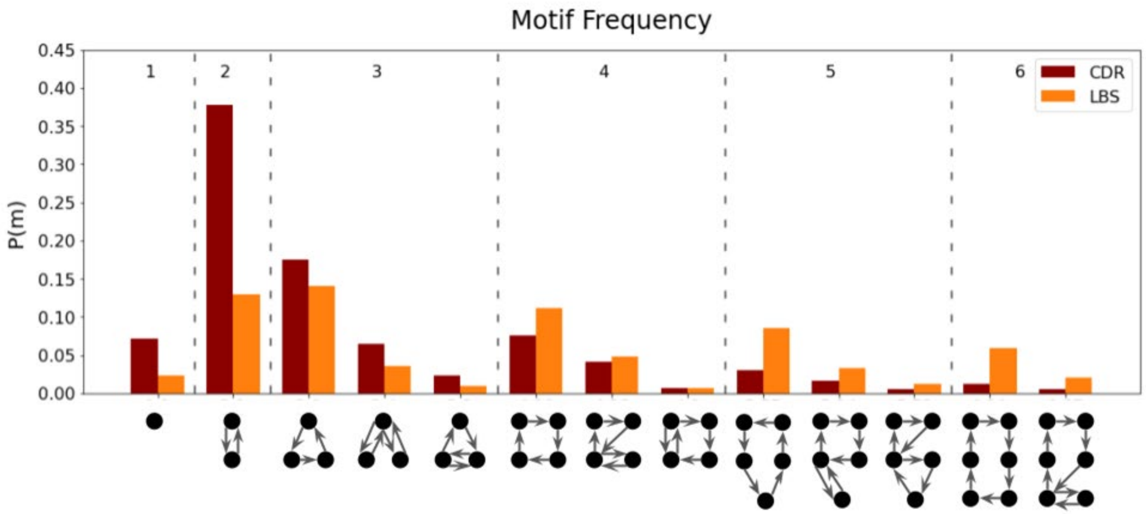
Where $p_i(j)$ is the historical probability that location j was visited by the user i

$$S_i = -\sum_{T_i' \subset T_i} p(T_i') \log_2 [p(T_i')]$$

Where $p(T'_i)$ is the probability of finding a particular time-ordered subsequence T'_i in the trajectory T_i

*Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. "Limits of Predictability in Human Mobility." Science 327, no. 5968 (February 19, 2010): 1018–21. https://doi.org/10.1126/science.1177170.

Comparing Daily Mobility Networks (Motifs)

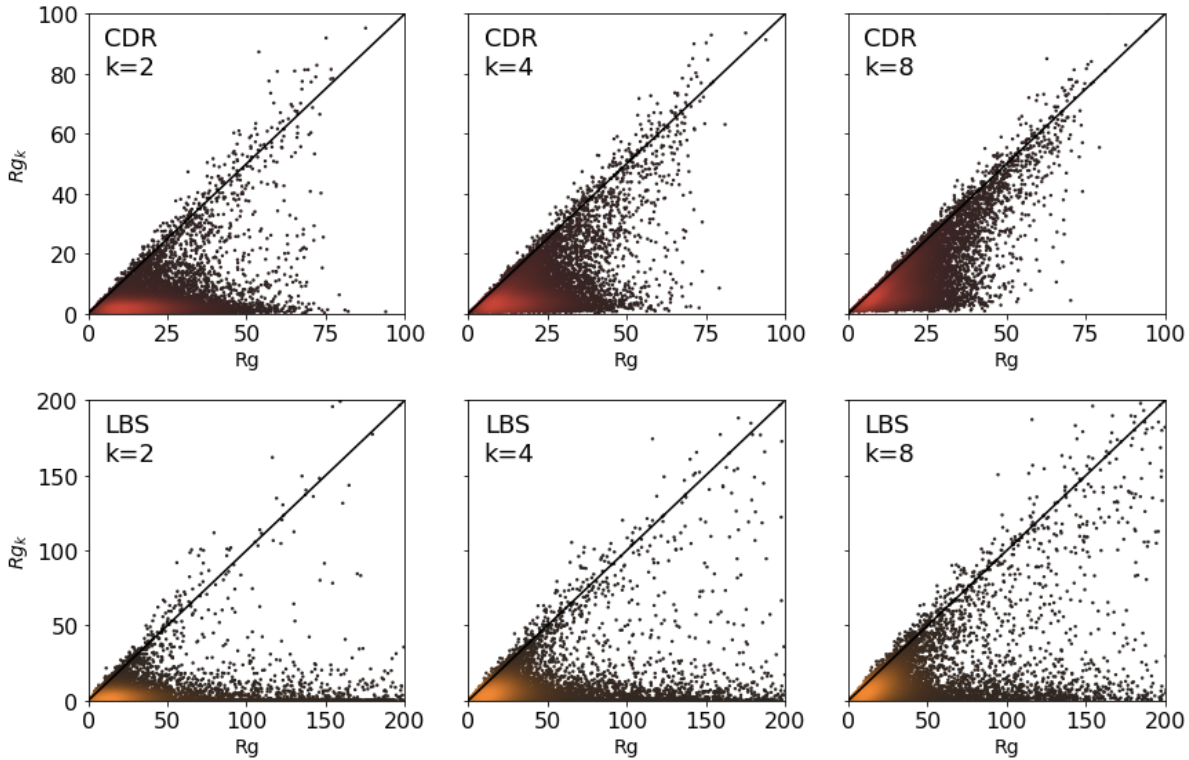


Motifs, or daily mobility networks*, are abstract representations of a users daily travel behavior.

We find that a statistically small number of motifs (here 13) represent at least 80% of travel behavior for LBS and CDR users

*Schneider Christian M., Belik Vitaly, Couronné Thomas, Smoreda Zbigniew and González Marta C. 2013 Unravelling daily human mobility motifs J. R. Soc. Interface.102013024620130246 http://doi.org/10.1098/rsif.2013.0246

Returners vs. Explorers



Individuals can be described as **returners or explorers*** depending on how many of their most visited locations are needed to accurately describe their mobility

$$r_g^{(k)} = \sqrt{\frac{1}{N_k}\sum_{i=1}^{k} n_i \left(\mathbf{r}_i - \mathbf{r}_{cm}^{(k)}\right)^2}$$

*Pappalardo, L., Simini, F., Rinzivillo, S. et al. Returners and explorers dichotomy in human mobility. Nat Commun 6, 8166 (2015). https://doi.org/10.1038/ncomms9166

Conclusion

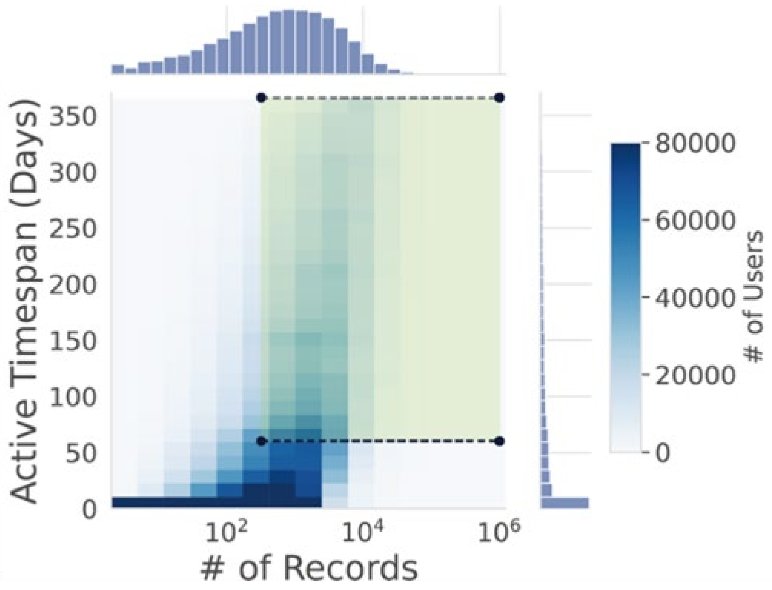
With proper processing steps, LBS data can be used to estimate similar mobility metrics to CDRs

- LBS provides a higher spatial resolution than CDRs
- LBS is easier to obtain

Mobility Dataset for Analyzing the Impact of COVID-19

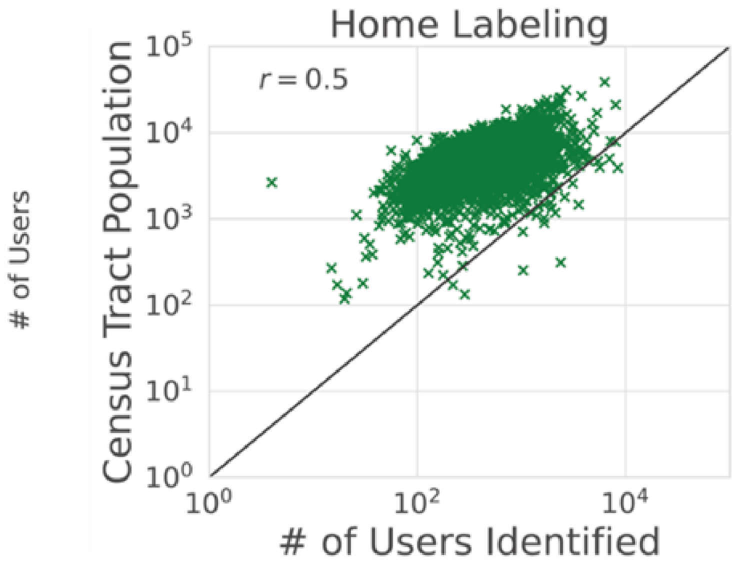
LBS Dataset (2019-2022)

- **Dataset:** 4 years of LBS data provided by Spectus, in trajectory format
- **Spatial Resolution:** Trajectories are defined by their starting census block group and ending census block group
- **Data Quality Control:** Selected active users defined by their number of records (>10^2.5) and timespan (>60 days)



Home and Work Detection

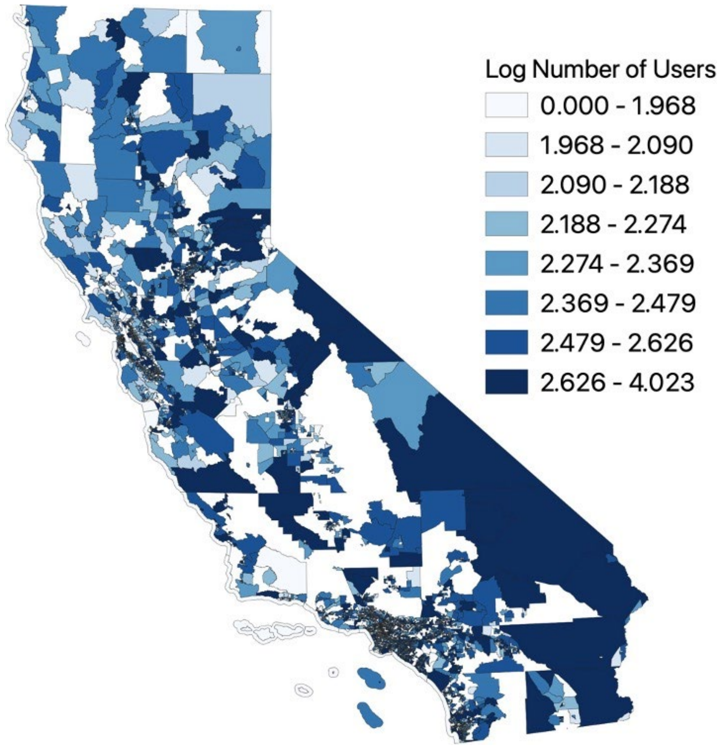
- **Home Detection:** Most frequently visited census block group between 7pm and 7am
- **Work Detection:** Most frequently visited census block group between 7am and 7pm on **weekdays**
- **Threshold:** >10 visits to both home and work locations in each year



Final Dataset

- Trajectories of users with home and work found are used in the analysis

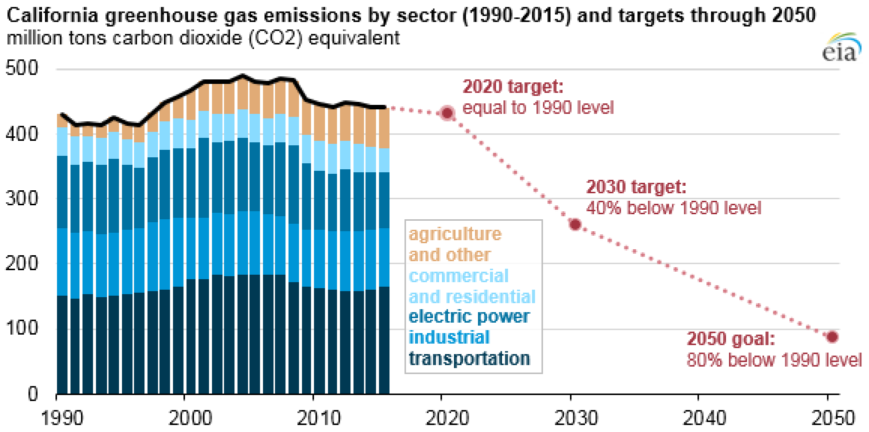
Year	# Users	# High-Quality Users	# Users with Home and Work Found
2019	9,410,380	3,482,574	861,167
2020	5,912,373	2,396,990	431,190
2021	5,222,416	2,036,110	465,311
2022	5,618,760	2,582,405	702,847



2. COVID-19 and Change in VMT

Background

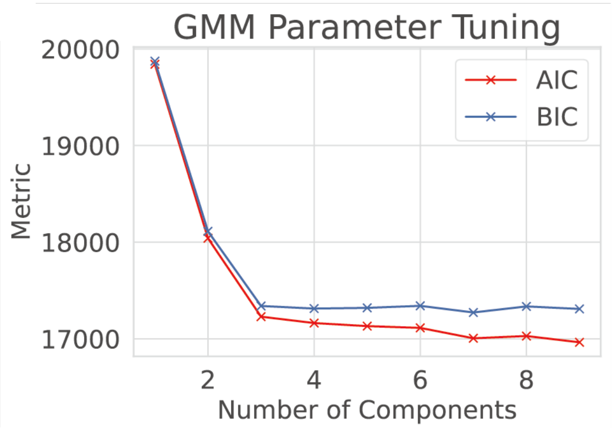
- In 2018, California set an ambitious target: to reduce the state's greenhouse gas emissions to 40% below the 1990 level by 2030.
- The emergence of the COVID-19 pandemic brought about substantial limitations and changes to people's mobility.

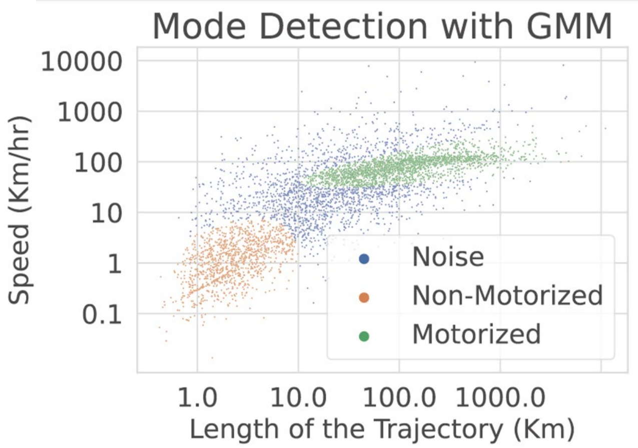


Question: How do we leverage LBS data to detect mode changes?

Mode Detection Algorithm

- **Unsupervised Learning:** Enable mode detection for large datasets at a low cost (with no labels)*.
- **Gaussian Mixture Model:** Clustering algorithm that assumes each observation belongs to a gaussian mixture, characterized by a mean vector and a covariance matrix.
- **Three Clusters:** Motorized, non-motorized, and noise.





*R. A. Hasan, H. Irshaid, F. Alhomaidat, S. Lee, and J.-S. Oh, "Transportation Mode Detection by Using Smartphones and Smartwatches with Machine Learning," *KSCE J Civ Eng*, vol. 26, no. 8, pp. 3578–3589, Aug. 2022, doi: 10.1007/s12205-022-1281-0.

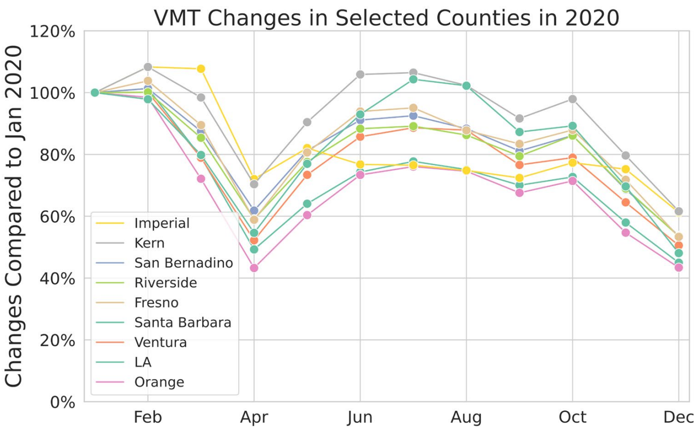
Radius of Gyration

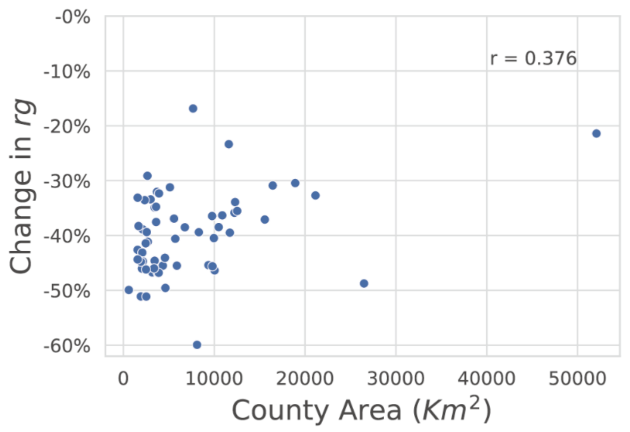
- Radius of gyration is an easy-to-compute statistic that measures the spread of a user's activity
- Higher radius of gyration suggests more vehicle use
- Lower radius of gyration suggests less vehicle use

$$r_g(u) = \sqrt{\frac{1}{n_u}\sum_{i=1}^{n_u} dist(r_i(u) - r_{cm}(u))^2}$$

Statewide Change in VMT

- **State-wide Reduction in VMT:** Starting in March 2020 we observe state-wide reduction in VMT, rapid recovery in the summer months, and more reduction in the winter as a new wave of COVID-19 affected the state
- **Urban vs. Rural:** Urban counties tend to experience larger reduction in VMT compared to more rural counties.

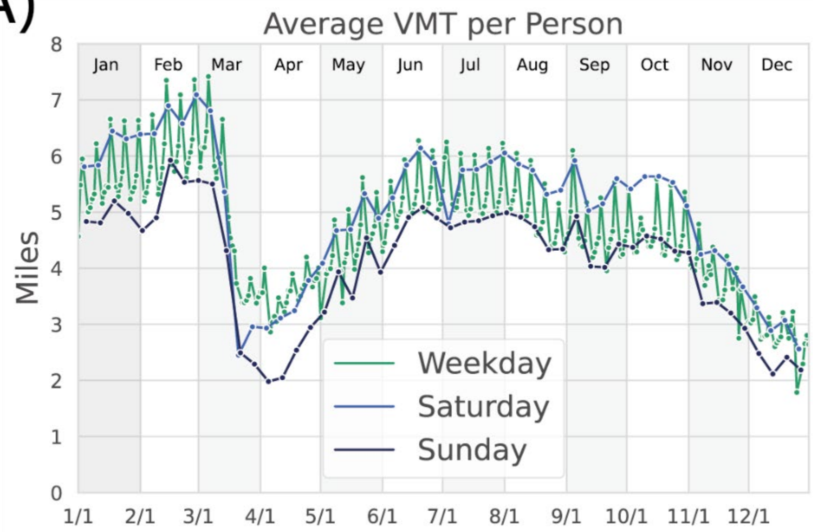




Change in VMT and Day of the Week

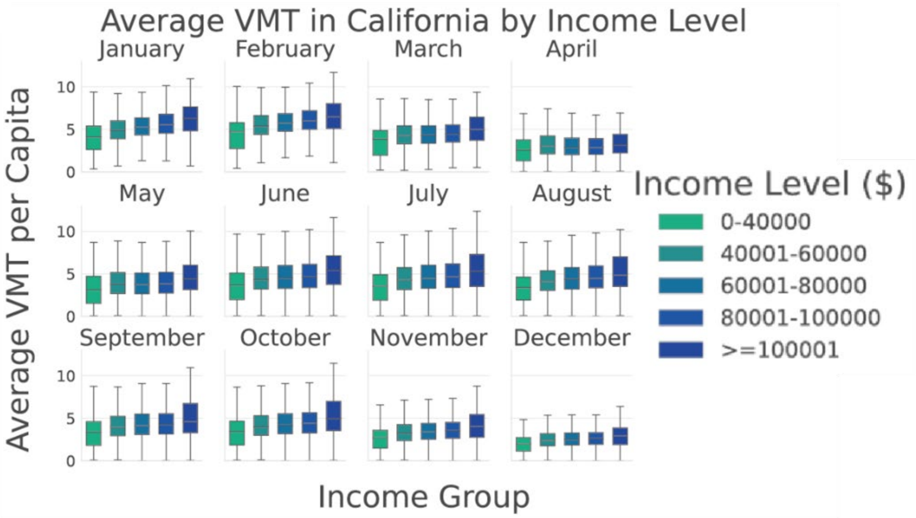
- **Day of Week Variation:** COVID-19 did not change the within-the-week variation in VMT. The weekly patterns are preserved.
- **Larger Reduction on Weekends:** At the beginning of the lockdown, we observe a larger reduction in VMT on weekends compared to weekdays.

A)



Change in VMT and Income Levels

- No notable differences across different income groups in VMT reduction
- Quicker rebound in VMT for tracts of higher income
- Potentially due to greater flexibility and feasibility of remote work



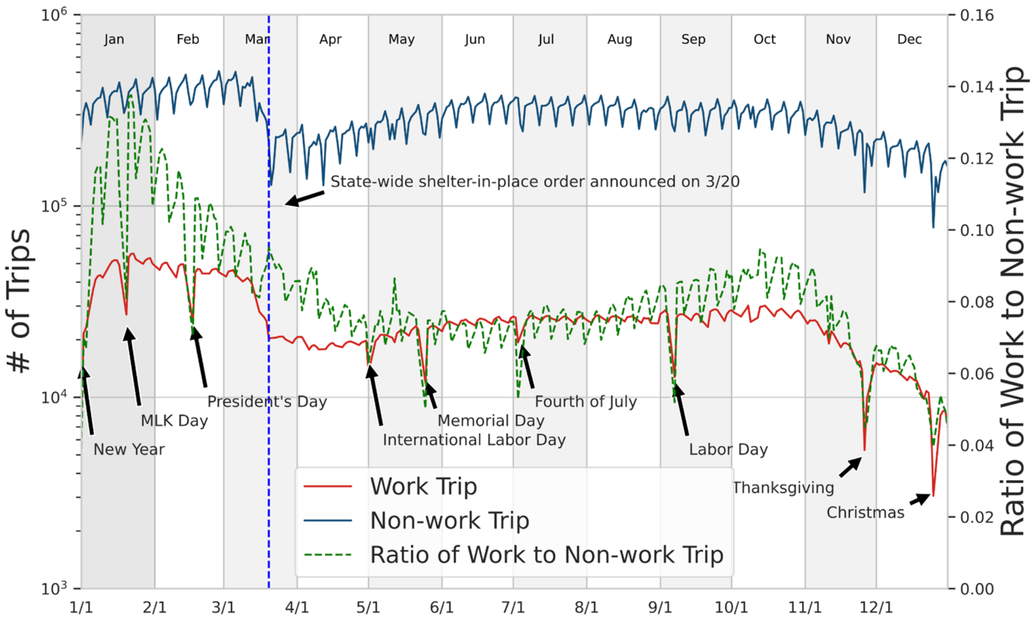
3. COVID-19 and Change in Trip Purpose

Commute Networks

- **Commute:** A trip that start at home and ends at work or vice versa.
- **Commute Networks:** A network whose nodes denote the census tracts and the weight on the edge connecting two nodes represent the amount of commute between the two tracts.
- **Community:** A subset of nodes in a network that are closely connected with each other rather than with other nodes not in the subset.
- **Modularity:** A degree to which a network can be partitioned into subsets
- **Louvain Method:** A method of detecting communities in a network by maximizing modularity

State-Wide Trends

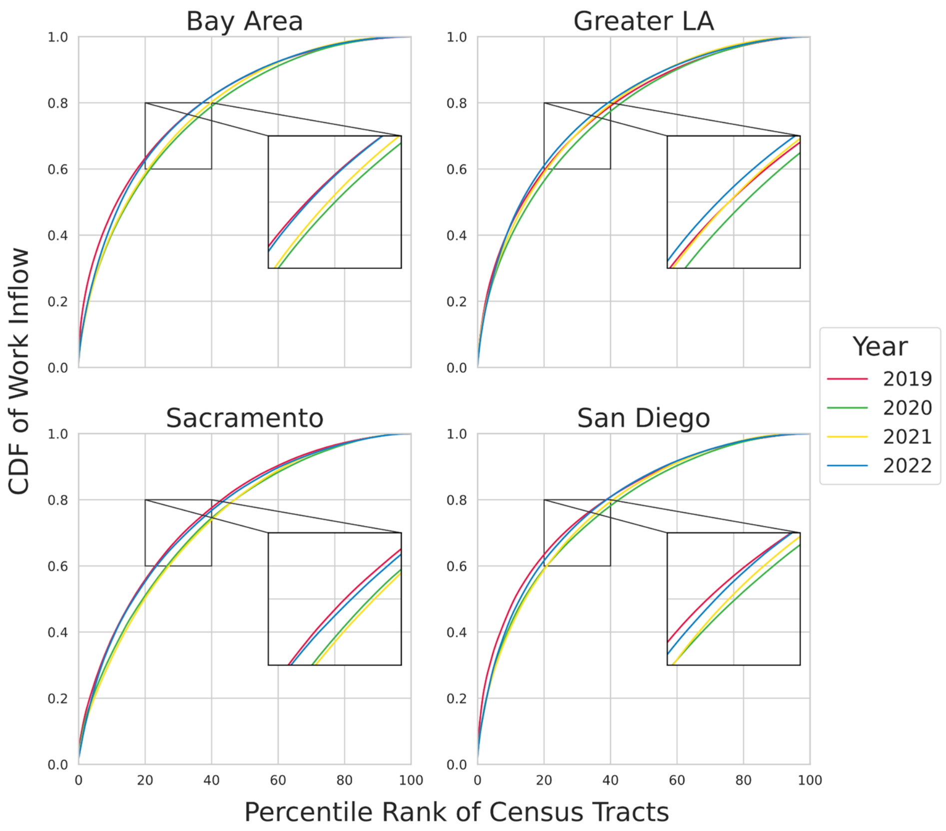
- We observe sharp reduction in both commute and non - commute as a result of the SIP order.
- Non-commute trips recover at a faster rate than commute, as many jobs became remote



Methods
Results

State-Wide Trends

- Work locations are less concentrated during COVID-19 across all four regions in California.
- Returned to 2019 level in 2022.
- Transition to remote work affected areas with more offices.



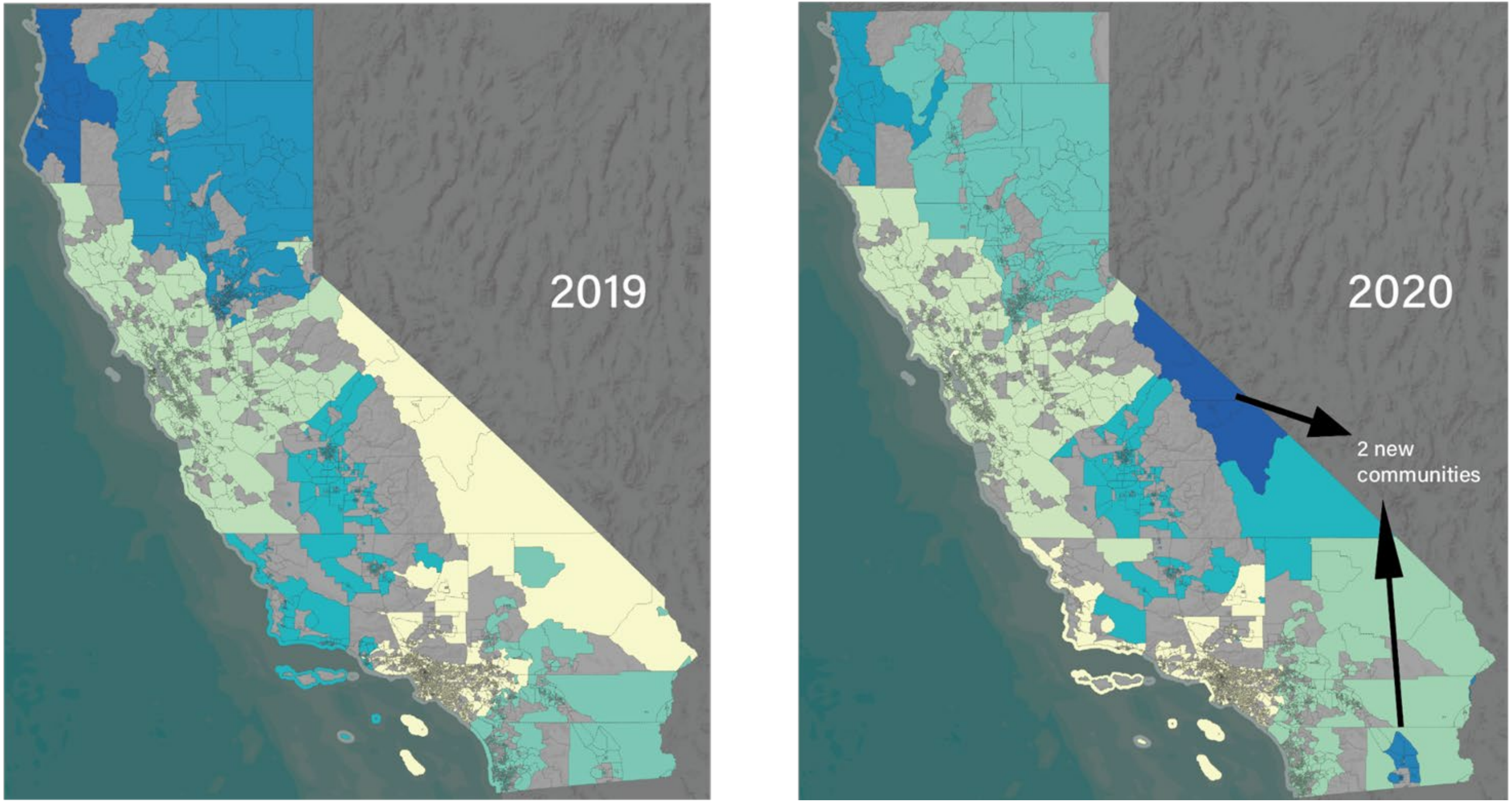
A Network Science Perspective

- **Significant reduction in the number of edges:** Disappearance of census tract pairs that had commutes.
- **Increased Number of Communities:** A more fragmented commute network in which people tend to commute locally to go to work, if they go to work at all.
- **Higher Modularity:** Less flow among the different communities

	# Nodes	# Edges	# Communities	Modularity
2019	8,033	145,838	6	0.628
2020	8,026	89,382	8	0.653
2021	8,009	73,401	8	0.648
2022	8,019	111,652	8	0.666

Methods

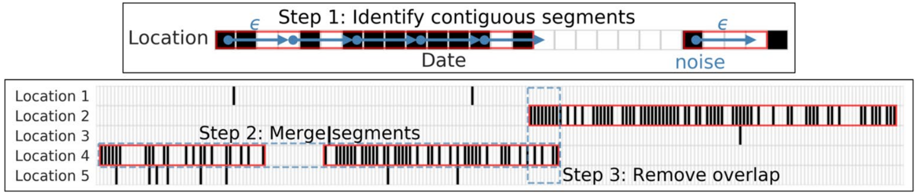
Results



4. COVID-19 and Change in Residential Locations

Background

- Change in residential locations is limited to unsupervised learning
- Existing methods work well with inter-region home changes over long period of time *



Question: How do we detect home changes over short periods of time and short distances with unsupervised learning?

* G. Chi, F. Lin, G. Chi, and J. Blumenstock, "A general approach to detecting migration events in digital trace data," PLoS ONE, vol. 15, no. 10, p. e0239408, Oct. 2020, doi: 10.1371/journal.pone.0239408.

* S. Isaacman, V. Frias-Martinez, and E. Frias-Martinez, "Modeling human migration patterns during drought conditions in La Guajira, Colombia," in *Proceedings of the 1st ACM SIGCAS Conference on*

UNIVERSITY OF CALIFORNIA Berkeley
CALIFORNIA AIR RESOURCES BOARD

Problem Formulation

- Home change detection is different from home detection
  - Unknown date of move – make it difficult to use frequentist home detection algorithms
  - User might still pay frequent visits to the original home after the move
- The goal is to select a move date c, such that the spatial -temporal uncertainty is minimized
- d(*) is a function of the distance between a record (x,y) and the center of the cluster

$$STU(c) = \sum_{t}^{c} \sum_{i=1}^{m} d(x_i, y_i, t) + \sum_{t=c}^{max(t)} \sum_{j=1}^{n} d(x_j, y_j, t)$$

2-Step "Pseudo"-Unsupervised Learning

- K-means clustering is used to create pseudo-labels
  - Standardized latitude, longitude, and time (# of days)
  - Assume 2 clusters (k=2)
  - Use the assigned clusters as pseudo-labels
- Linear Soft Margin SVM
  - Train and predict on the pseudo-labels
  - Used for regularization, with a very small penalty for mislabeled data
- Heuristics for selecting moves

$$\|\mathbf{d}\|_2 = \sqrt{\mathbf{d}^T\mathbf{d}} = \sqrt{\alpha^2\mathbf{w}^T\mathbf{w}} = \frac{|\mathbf{w}^T\mathbf{x}+b|}{\sqrt{\mathbf{w}^T\mathbf{w}}} = \frac{|\mathbf{w}^T\mathbf{x}+b|}{\|w\|_2}$$

margin

misclassification

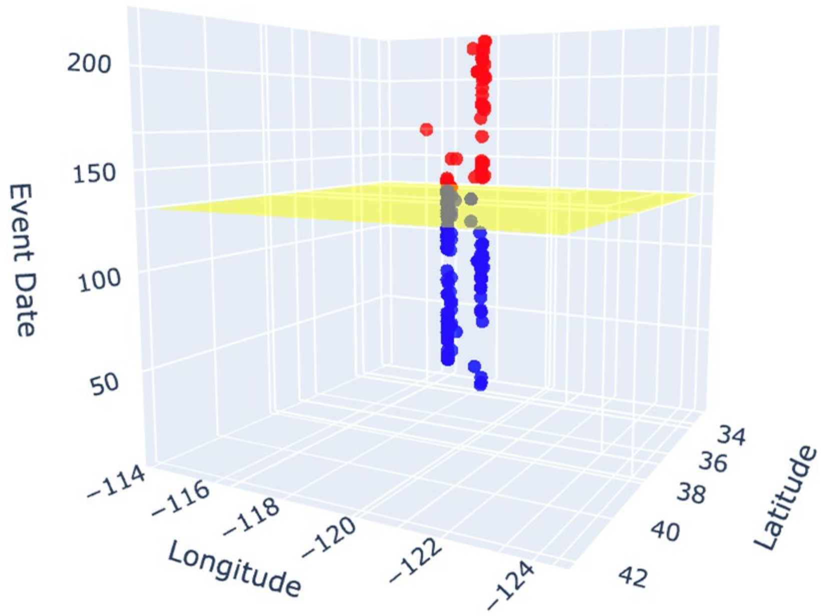
$$\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ \forall i\ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$
$$\forall i\ \xi_i \geq 0$$

Berkeley UNIVERSITY OF CALIFORNIA

CALIFORNIA AIR RESOURCES BOARD

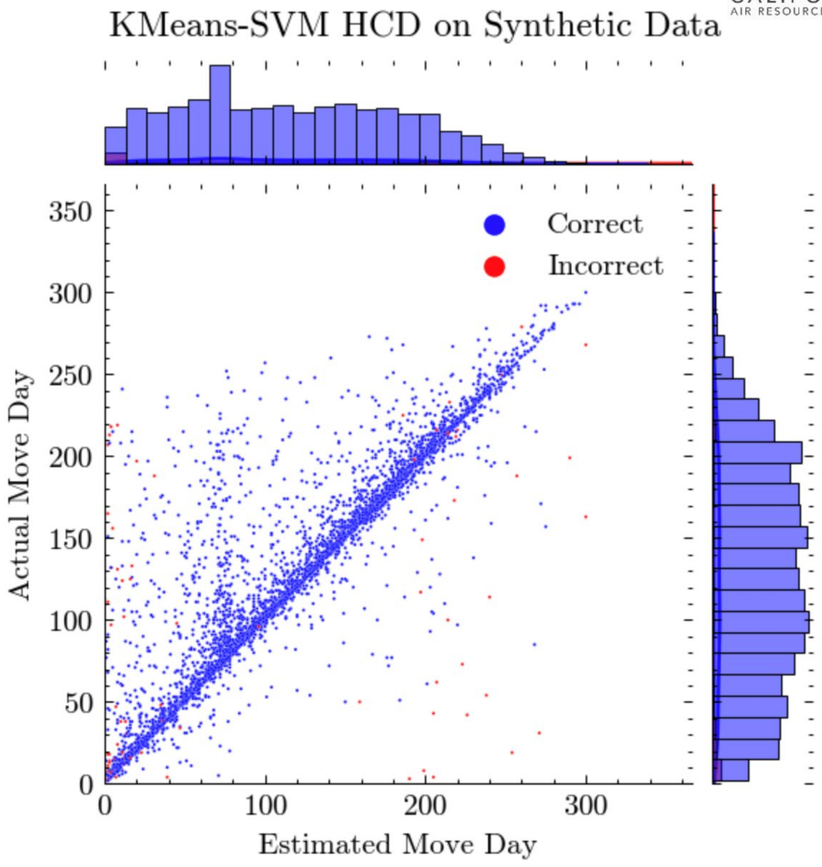
Validation Using Synthetic data

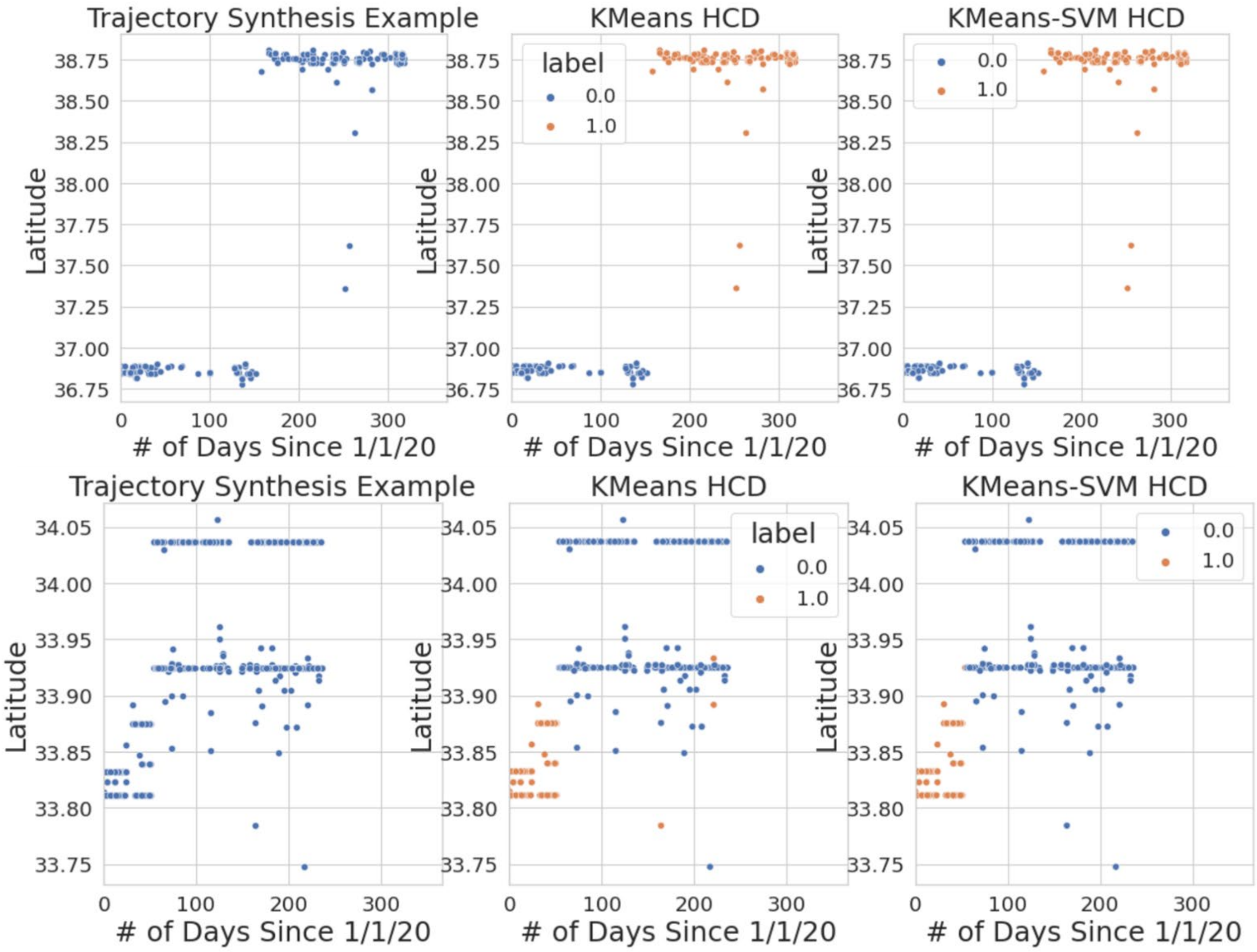
- Select a sample stationary users
  - Same home location in every month
  - Must have a span > 200 days
- Record recombination:
  - Randomly select a day between:
    - max(user1_min, user2_min) and
    - min(user1_max, user2_max)
  - Combine the two sets of records
- 23260 users with no home change
- 5000 synthetic home change

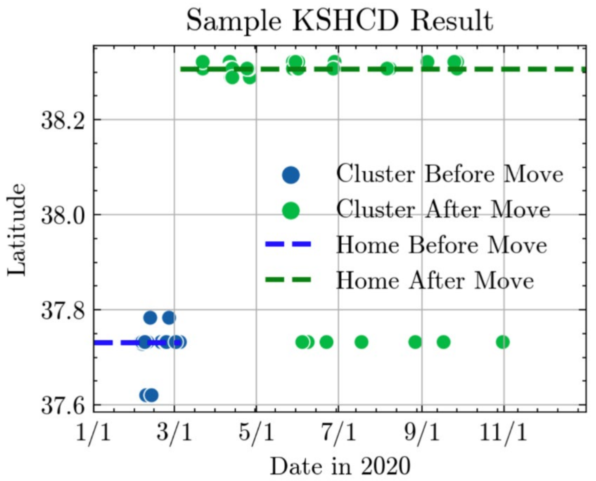


Validation Results

- 447 users with no home change but detected - Type I error (1.6%)
- 4425 (out of 5000) users with home change detected
- 4321 (out of 4425) users are accurately labelled
- Overall accuracy for those with home change (86.4%)

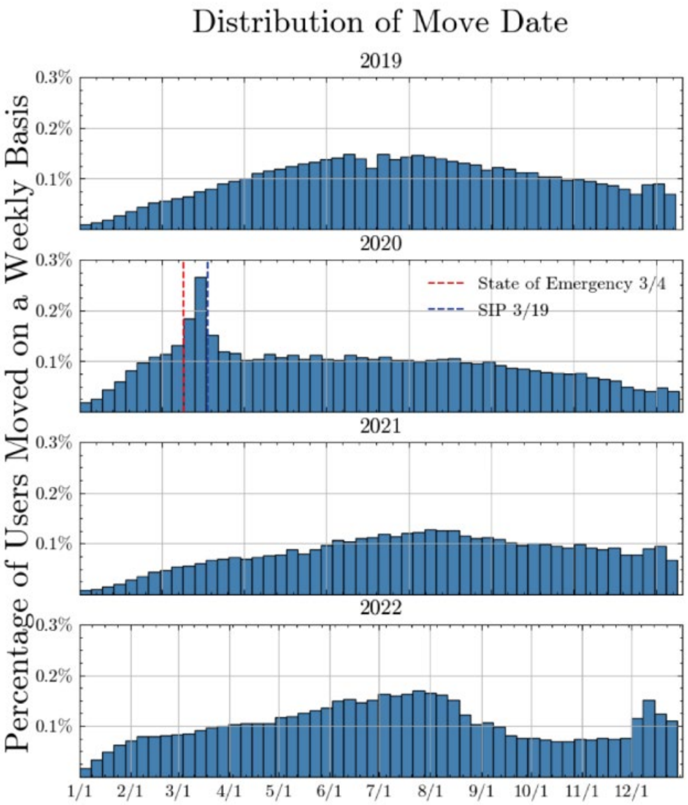



Trajectory Synthesis Example
KMeans HCD
KMeans-SVM HCD
label
0.0
1.0
Latitude
# of Days Since 1/1/20
38.75
38.50
38.25
38.00
37.75
37.50
37.25
37.00
36.75
34.05
34.00
33.95
33.90
33.85
33.80
33.75
0
100
200
300


Sample KSHCD Result
Cluster Before Move
Cluster After Move
Home Before Move
Home After Move
Latitude
38.2
38.0
37.8
37.6
1/1
3/1
5/1
7/1
9/1
11/1
Date in 2020

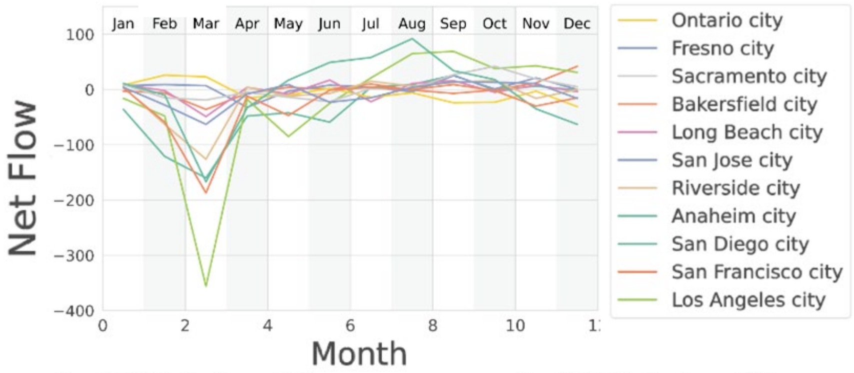
Relocation During COVID-19

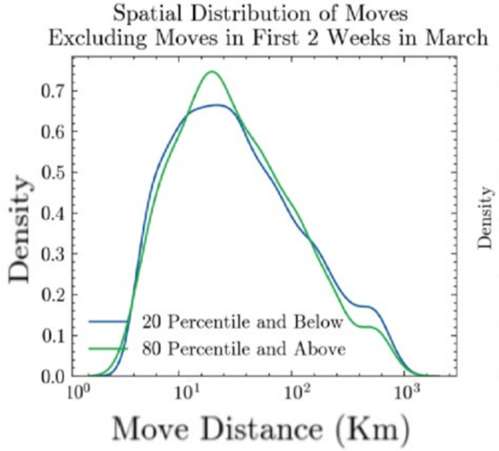
- More moves are observed in the 2-week period between the declaration of the state of emergency and the announcement of the SIP order
- The distribution of 2020 is visibly different, yet the pattern at the beginning and end of the period is skewed due to data anonymization.

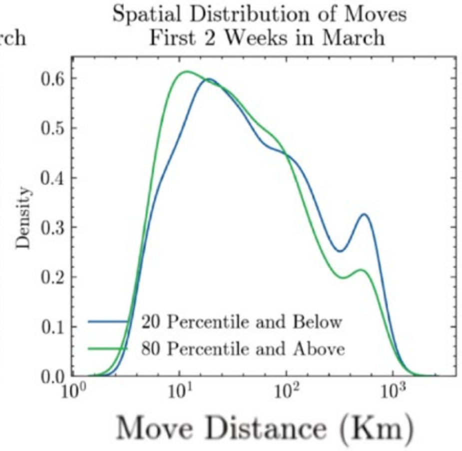


Relocation During COVID-19

- Large urban areas experience large net outflow
- Immigration and emigration in cities in the Central Valley such as Bakersfield and Fresno remain stable
- More moves over longer distance in the 2-week period. The second peak corresponds to the distance between Southern and Northern California







4. Mobility Policies and Change in Vehicle Usage Rate

Vehicle Usage Rate Calculation

- **GMM-Based Mode Detection:** Using the GMM mode detection algorithm discussed in (2), we can detect the mode of each trip.
- **VUR:** We define VUR as the percentage of all trips that are “motorized”.
- **Bootstrap Sampling:** We create distribution of VUR using bootstrap samples.
- **Mobility Policies:** We evaluate the effectiveness of mobility policies by comparing the bootstrapped VUR in the month before the launch of the initiative and in the month of the launch of the initiative.

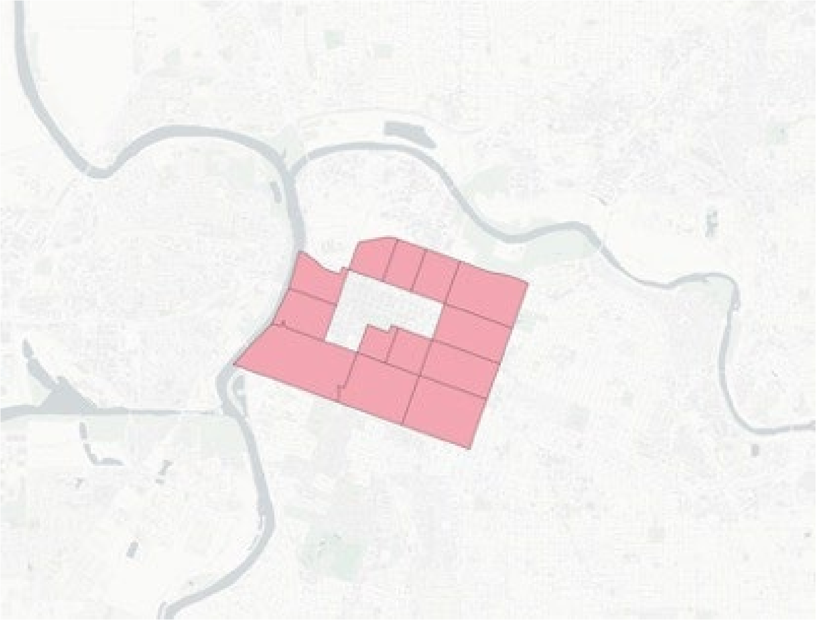
Sacramento Case Study

- 4 mobility initiatives in selected census tracts in Sacramento

Time	Event	Hypothesized Impact
February, 2019	JUMP released electric scooters	Decrease in vehicle usage
March, 2019	GIG Car Share released shared-vehicles	Increase in vehicle usage
June, 2019	JUMP increased its electric bike fleet	Decrease in vehicle usage
September, 2019	Sacramento Rapid Transit launched a new transit program SacRT Forward	Increase in vehicle usage

Sacramento Case Study

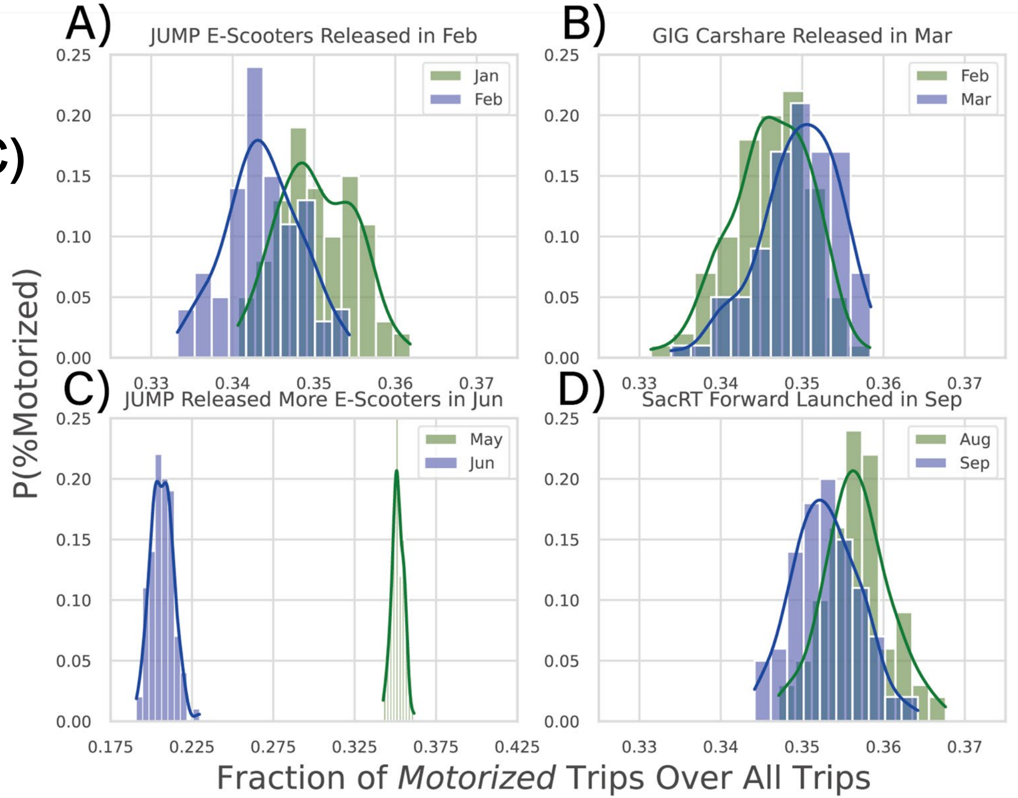
- 4 mobility initiatives in selected census tracts in Sacramento



	# Trips	# Users
January, 2019	192,426	34,636
February, 2019	190,643	36,203
March, 2019	248,462	46,121
May, 2019	270,871	50,633
June, 2019	277,004	36,203
August, 2019	256,785	45,665
September 2019	257,476	46,948

Sacramento Case Study

- Release and increase in E-scooters decreased VUR **(A,C)**
- Release of GIG car share increased VUR **(B)**
- SacRT Forward decreased VUR **(D)**
- Cannot be used to establish causality (observation study rather than controlled experiment)



Conclusions

- We show that both LBS and CDR data sources can be used to extract travel behavior despite CDRs having four time more users
- We observed significant decline in VMT during COVID-19 lockdown with regional disparities. VMT in urban counties decreased up to 55% and 20 -30% in rural counties.
- Commute trips recovered at a slower pace compared to non -commute trips in 2020, signifying a lasting change to remote work.
- We developed a novel home change detection algorithm and found an increase in both the number and the distance of relocations in the first two weeks in March 2020.
- We developed an unsupervised mode detection model and found that JUMP’s increase in fleet size in June 2019 decreased the overall VUR.

Thank you!

Any questions?