Tracking Consumers: The Trade-off between the Value of Granular Data and Consumers' Privacy

Unnati Narang [1] Fernando Luco [2]

November 28, 2022

[1]University of Illinois at Urbana-Champaign

[2]Texas A&M University

Consumer tracking potentially valuable but creates privacy concerns

- Most mobile apps collect granular tracking information on consumer movement
  - 70% brands use (and share) GPS data collected through apps (NY Times 2018)
  - 3.22 trillion miles driven on U.S. roads per year (DoT 2018)
  - $38.7 B spending on mobile location and targeting (eMarketer 2018)
- Firms benefit from investing in data and analytics (e.g., Muller et al. 2018, Berman and Israeli 2022) and real-time tracking (e.g., Luo et al. 2014, Fong et al. 2015)
- Granular tracking data potentially valuable BUT also lead to privacy concerns among consumers and regulators (e.g., Rader and Slaker 2017, Ghose et al. 2022)

Value-privacy tradeoff: $\downarrow$ value of tracking for firms if less granular?

TECHNOLOGY

The New York Times

PLAY THE CROSSWORD

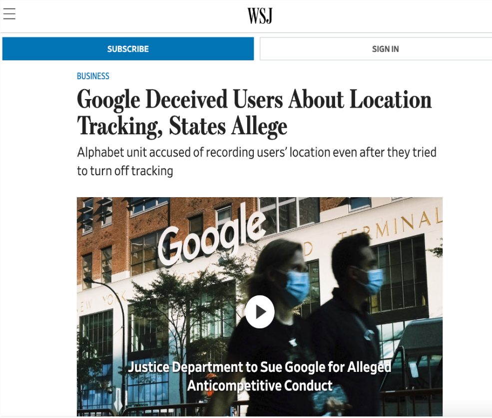
Account

The Battle for Digital Privacy Is Reshaping the Internet

As Apple and Google enact privacy changes, businesses are grappling with the fallout, Madison Avenue is fighting back and Facebook has cried foul.

Give this article

348



WSJ

SUBSCRIBE

SIGN IN

BUSINESS

Google Deceived Users About Location Tracking, States Allege

Alphabet unit accused of recording users' location even after they tried to turn off tracking

Justice Department to Sue Google for Alleged Anticompetitive Conduct

Research Questions

- What is the value of granular consumer tracking?
  - Context: Driving behavior
  - Application: Predicting future retail visits
- Is there a value-privacy trade-off in practice?
  - Policy counterfactuals: 1/2, 1/3 tracking granularity
- Is there any heterogeneity in the value of granularity by firm type?
  - Chain vs. non-chain restaurants

Preview of Results

- What is the value of granular consumer tracking?
  - $\uparrow$ 21.4% prediction accuracy with granular tracking data relative to only demographic and behavioral information on past visits
- Is there a value-privacy trade-off in practice?
  - $\downarrow$ 4.9% loss in prediction accuracy but still a significant improvement over models that do not use tracking data
- Is there any heterogeneity in the value of granularity by firm type?
  - Value of granularity heterogeneous across restaurants and higher for non-chains

Related Literature and Contribution

- Intersection of literature on value of data, tracking, and privacy

- Value of aggregate data investments (Muller et al. 2018, Berman and Israeli 2022)
  - Less explored: Individual tracking data (except Netzer et al. 2019)

- Location tracking improves targeting (Luo et al. 2014, Fong et al. 2015, Ghose et al. 2019)
  - Less explored: Consumer driving trajectories & privacy

- Privacy regulations impact firms (asymmetrically) & consumers (Goldfarb and Tucker 2011, Johnson et al. 2021, Laub et al. 2021, Rafieian and Yoganarasimhan 2021, Zhao et al. 2021)
  - Less explored: Varying levels of granularity; potentially more privacy-preserving for users

Outline

Data and Empirical Approach

Results

Application: Optimal Targeting Policy

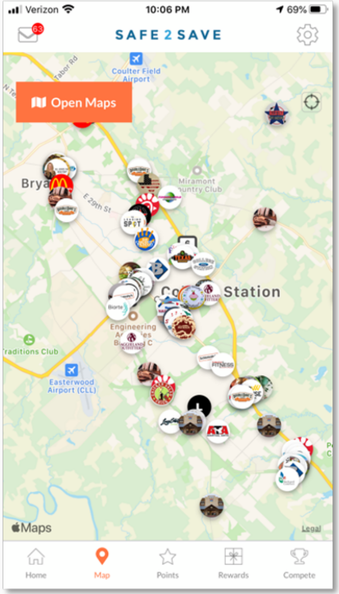
Data and Empirical Approach

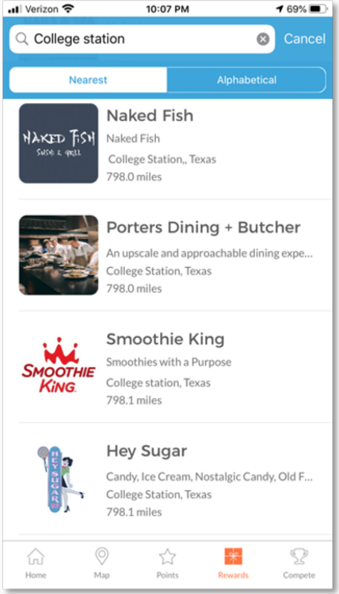
Data

- Proprietary data for 2018-19 from a safe-driving app
  - 200,000 users in Texas
  - Current application: 31,530 users
- Individual-level data Summary stats
  - User demographics (e.g., age, gender)
    - Average age 32, 44% female
  - Driving behaviors and trajectories (450 million+ GPS datapoints)
    - 406 miles, 239 stops, 9 restaurant visits
- Restaurant-level data Summary stats
  - SafeGraph: Polygons for each restaurant in Texas created via satellite imagery
  - Yelp: Restaurant ratings, price levels, categories
  - Pricelisto: Menu prices

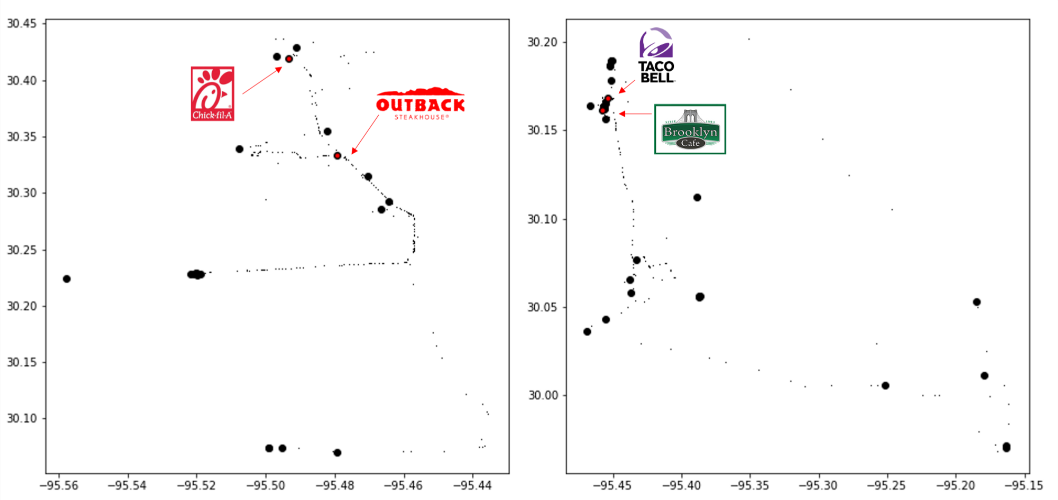
App Preview


Verizon
10:08 PM
68%
63
SAFE 2 SAVE
Is it really worth it?
Drive Safe to Save Lives and Money
Home
Map
Points
Rewards
Compete


Verizon
10:06 PM
69%
63
SAFE 2 SAVE
Open Maps
Coulter Field Airport
Station
Engineering
Traditions Club
Easterwood Airport (CLL)
Maps
Legal
Home
Map
Points
Rewards
Compete

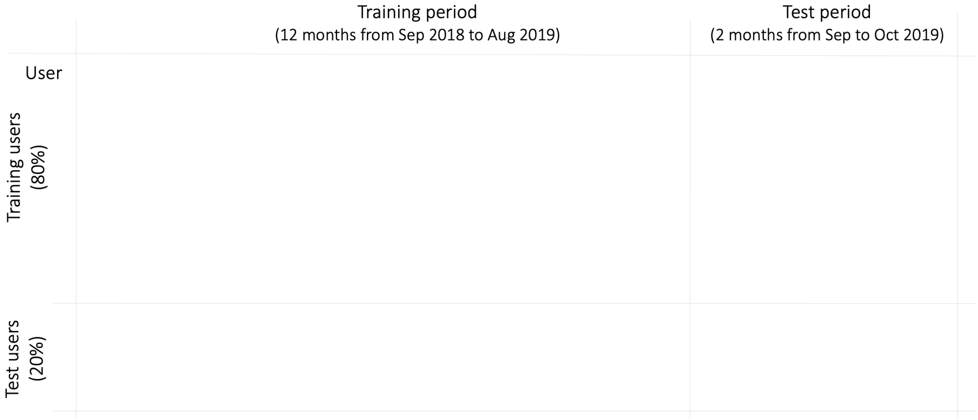

Verizon
10:07 PM
69%
College station
Cancel
Nearest
Alphabetical
Naked Fish
Naked Fish
College Station,, Texas
798.0 miles
Porters Dining + Butcher
An upscale and approachable dining expe...
College Station, Texas
798.0 miles
Smoothie King
Smoothies with a Purpose
College station, Texas
798.1 miles
Hey Sugar
Candy, Ice Cream, Nostalgic Candy, Old F...
College Station, Texas
798.1 miles
Home
Map
Points
Rewards
Compete

Can we better predict customers' visit to a restaurant using tracking data?

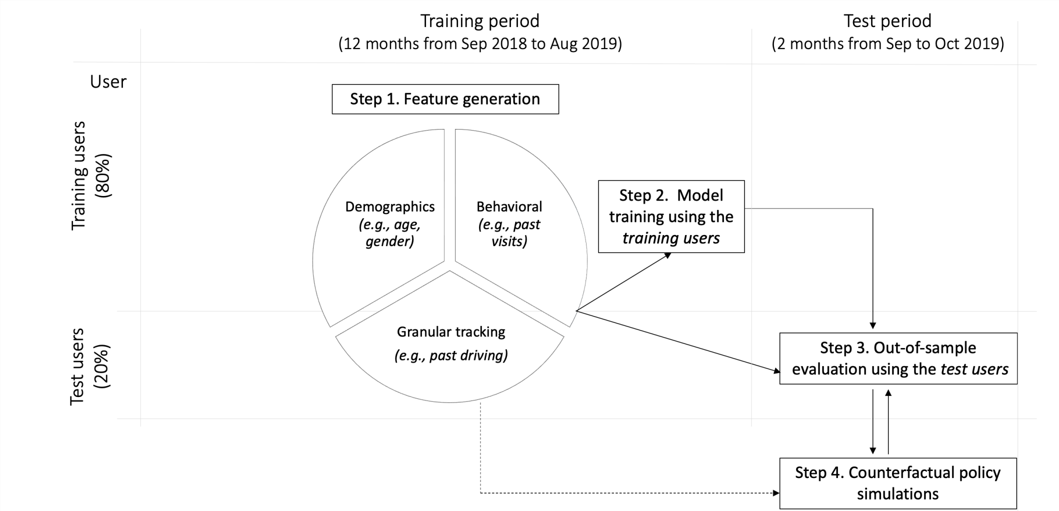


Driving trajectories of different consumers are heterogeneous and inform choices.

Approach: ML Framework for Evaluating Value of Granular Tracking

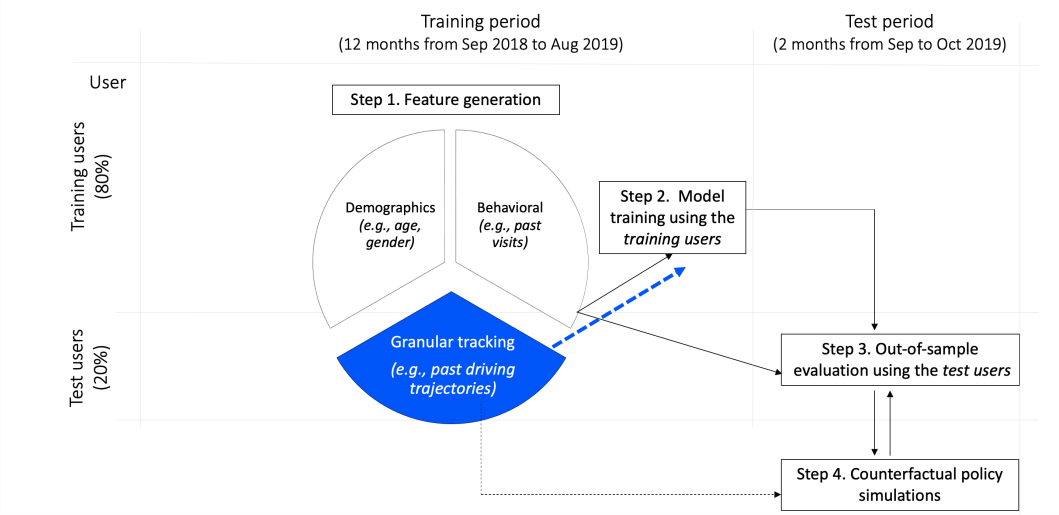


Approach: ML Framework for Evaluating Value of Granular Tracking



Example features

Approach: DL Framework for Evaluating Value of Granular Tracking



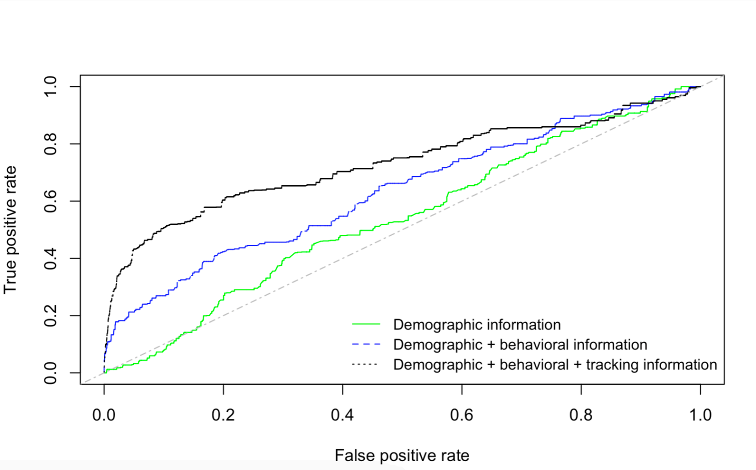
Results

↑ 21.4% prediction accuracy with granular tracking information

Information set	Accuracy	Precision	Recall	F1
Demographic information	52.76% (0.71%)	59.78% (0.76%)	57.39% (0.46%)	56.16% (0.54%)
Demographic + behavioral information	57.93% (0.43%)	61.95% (0.30%)	70.22% (0.63%)	63.37% (0.43%)
Demographic + behavioral + tracking information	70.31% (0.55%)	70.16% (0.55%)	89.85% (0.21%)	77.08% (0.34%)

Model: Lasso

ROC curves: Higher predictive performance with tracking information



Reduced granularity $\implies$ 4.95% $\downarrow$ in accuracy but much better than with no tracking

Data set	Accuracy	Precision	Recall	F1
Complete granular tracking	70.31%	70.16%	89.85%	77.08%
	(0.55%)	(0.55%)	(0.21%)	(0.34%)
Data at 1/2 frequency	66.83%	67.15%	88.59%	74.50%
	(0.81%)	(0.52%)	(0.82%)	(0.38%)
Data at $1/3^{\text{rd}}$ frequency	66.39%	66.28%	86.77%	73.57%
	(0.36%)	(0.41%)	(0.08%)	(0.25%)
Data at 1/2 frequency at random	66.34%	66.72%	89.20%	74.41%
	(0.48%)	(0.41%)	(0.24%)	(0.30%)

Notes: Predictive Performance of Lasso by Counterfactual

Value of granular tracking is heterogeneous across restaurants

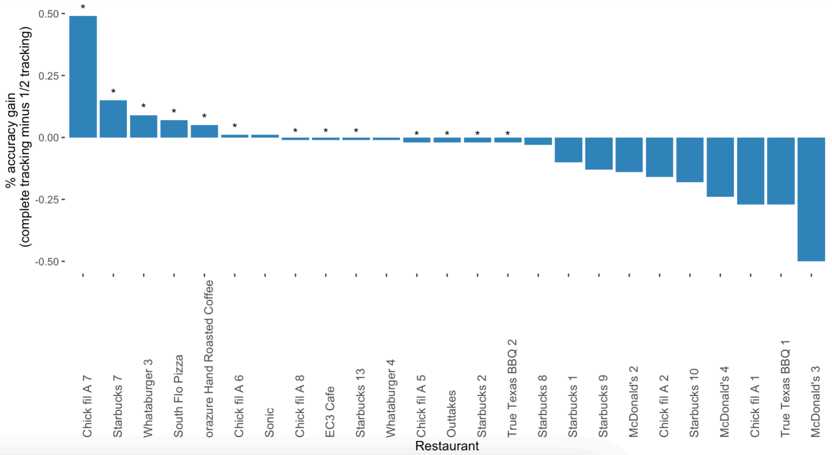


Figure 2: % drop in accuracy compared to data with 100% granularity ($p < 0.05$)

Value of granular tracking is higher for non-chain restaurants

	Model 1	Model 2	Model 3	Model 4
Chain	-0.041*	-0.056***	-0.054***	-0.048***
	(0.022)	(0.012)	(0.010)	(0.004)
Past visits		-0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)
Rating		-0.009	-0.007	-0.009*
		(0.015)	(0.013)	0.005
Price		0.045*	0.028	0.039
		(0.019)	(0.021)	(0.028)
Baseline accuracy				0.448***
				(0.076)
Category fixed effects	Yes	Yes	Yes	Yes
City fixed effects	No	No	Yes	Yes
R^2	0.184	0.193	0.206	0.255

Note: The dependent variable is the difference in prediction accuracy between the full- and halve-granularity data.

Additional result: Driving trajectories improve accuracy by 9.4%

Input type	Accuracy	Precision	Recall	F1
Driving summaries (Lasso)	76.48%	75.48%	82.82%	78.27%
Driving summaries (Transformers)	83.11%	95.60%	77.51%	83.70%
Driving trajectory (Transformers)	90.90%	92.22%	99.52%	95.62%

Application: Optimal Targeting Policy

Setting: Targeted Notifications

Application: Optimal Targeting Policy

- Firms may be interested to use granular tracking data to better target customers (e.g., Ascarza 2017, Hitsch and Misra 2018).
- Can we evaluate a variety of targeting policies suggested by granular tracking models (relative to a firm's default targeting policy) using our framework?
- Assume that a policy $\mathrm{T_k}$ targets $\mathrm{N_k} \subset \mathrm{N}$, where N is the population. Then,

$$\mathrm{E}[\pi|\mathrm{T_k}] = \sum_{\mathrm{i}=1}^{\mathrm{N_k}} \mathrm{E}[\pi_{\mathrm{i}}|\mathrm{T_k}] + \sum_{\mathrm{i}=\mathrm{N_k}+1}^{\mathrm{N}} \mathrm{E}[\pi_{\mathrm{i}}|\mathrm{NT_k}],$$

Application: Optimal Targeting Policy

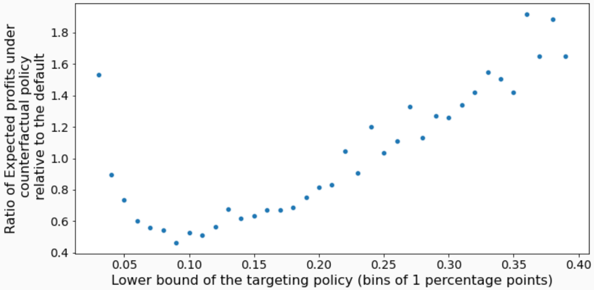
- Assuming π_i has three parts: revenues, marginal costs, and the cost of targeting c_k, we can re-write this as:

$$E[\pi|T_k] = \overbrace{\sum_{i=1}^{N_k} E[\pi_i|V, T_k] P_i(V|T_k)}^{\text{Targeted people who visit}} - \underbrace{c_k \sum_{i=1}^{N_k} P_i(NV|T_k)}_{\text{Targeted people who don't visit}} + \overbrace{\sum_{i=N_k+1}^{N} E[\pi_i|V, NT_k] P_i(V|NT_k)}^{\text{Non-targeted people who visit}} .$$

- Assuming $E[R_i|V,T] = E[R_i|V,NT] = E[R_i|V]$, and using the Law of Total Probabilities, we can simplify the above term and find the condition when between two policies T_k and T_j, a manager will choose T_k if $E[\pi|T_k] > E[\pi|T_j]$.

Results: Optimal Targeting Policy

- Default targeting policy: Push notifications nudging visits to specific restaurants
  - 625 of 31,530 users under default targeting in our data period
- Proposed targeting policy: Target those who are "at the margin" based on unconditional probability of visit from our model



Summary and Implications

- Mobility data provide rich information, are rarely exploited by firms and researchers, and pose unique modeling challenges.
- This paper: what can researchers and firms learn from consumers driving behavior?
    - Accuracy of prediction algorithms improves by 21.4% with granular tracking data relative to models that use only demographic and behavioral information on past visits
    - Accuracy reduces by 4.9% when the granularity of tracking is halved, but this is still a significant improvement over models that do not use tracking data
    - Losses from granular tracking heterogeneous across restaurants
- Implications
    - For managers, tracking data allow firms to better predict consumers' future behavior and to target consumers better compared with default targeting policies.
    - For researchers, tracking data are informative for identifying consumer types based on their driving and observing their choices in varying contexts.
    - For regulators, managing policy pushbacks and privacy law implications.

Thanks!

Email: unnati@illinois.edu; fluco@tamu.edu

Driving Summary

Driving behaviors	Mean	St Dev
Radius of gyration	405.61	740.88
Entropy	9.27	2.16
No. of Stops of 60+ min	158.20	140.7
No. of Stops at restaurants of 60+ min	15.02	5.63
Unique driving days	71.28	56.97
Morning trips	0.33	0.13
Evening trips	0.33	0.13

Notes: N = 31,530. Driving behaviors are measured for Aug' 2018 to Aug' 2019.

Back to main

Data Summary

Covariates	Mean	St Dev	Min	Max
Demographics				
Age	31.88	13.6	14	90
Gender (female)	0.44	0.5	0	1
Driving behaviors				
Radius of gyration	404.59	42.2	0.6	199.5
Entropy	9.25	1.1	3.1	11.9
No. of Stops	158.41	49.1	6.0	466.0
No. of Stops at restaurants	5.74	3.8	0.0	50.0
Max distance	2415.56	271.8	2.0	3,197.6

Notes: N = 31,530

Back to main

Restaurant Summary

Table 1: Restaurant Characteristics: Summary Statistics

Characteristic	Mean
Past visits	81.10
Chain	0.77
Rating	2.86
Price	0.37

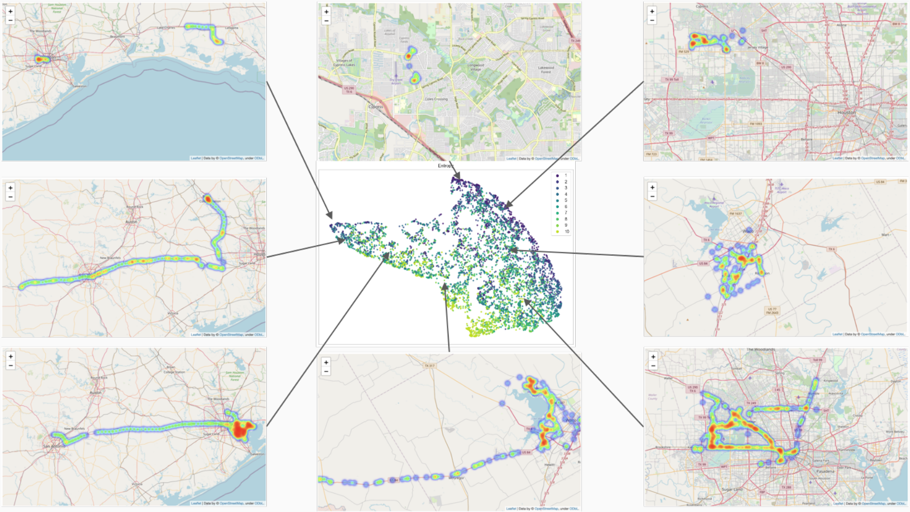
Notes: Past visits are the number of visits by users to restaurants in the training period.

Back to main

Modeling Challenges and Our Approach

- Extracting driving features
  $\rightarrow$ Recover each user's driving behavior (e.g., entropy, Pappalardo and Simini 2018)
- Inferring visits from GPS data
  $\rightarrow$ Map geolocation to satellite images of retailer (polygons)
- Spatial correlations and dynamic temporal patterns
  $\rightarrow$ Deep learning

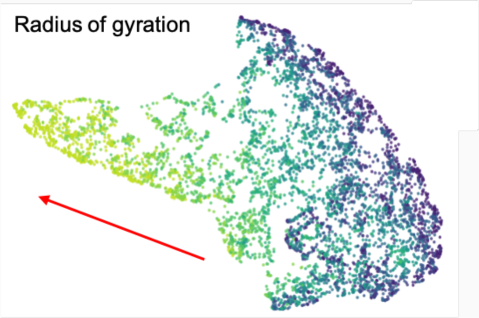
1. Extracting Driving Features

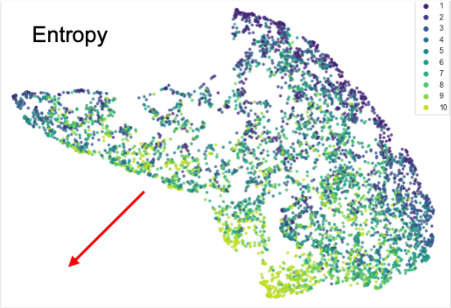


Back

1. Extracting Driving Features: Radius of Gyration (contd.)

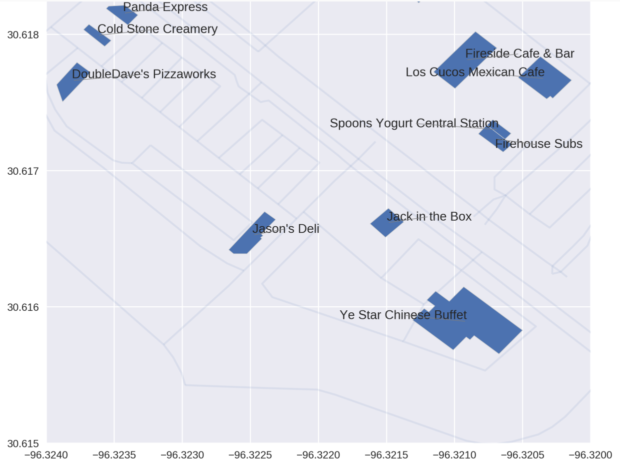
- Driving points spatial distribution of displacements over all users is well approximated by a truncated power-law (Gonzalez et al. 2008) with random walk pattern of step size Δr
- $\mathrm{P}(\Delta \mathrm{r}) = (\Delta \mathrm{r} + \Delta \mathrm{r}_0)^{-\beta} \exp(-\Delta \mathrm{r}/\mathrm{k})$
- where β = 1.75 $\pm$0.15,$\Delta \mathrm{r}_0$=1.5km
- What does this mean? Human motion follows a truncated Levy flight (random walk with a probability distribution that is heavy-tailed)
- Radius of gyration = the characteristic distance travelled by user a when observed up to time t


Radius of gyration


Entropy
1
2
3
4
5
6
7
8
9
10

2. Inferring Visits from GPS data

- GPS data do not identify visits
- Merge with polygons for each location



Do granular tracking data add significantly more value? A bootstrapping procedure

- Compute accuracy for an ML model (e.g., Lasso) for a given specification, say model M1 (e.g., driving + demographics, $\mathrm{acc}_{\mathrm{m1}}$) and model M2 (e.g., only demographics, $\mathrm{acc}_{\mathrm{m2}}$)

- Conduct N (e.g., 100) bootstrap iterations by resampling from original data. For each bootstrap, estimate for model 1 and model 2. Save the corresponding accuracies to get a distribution of 100 estimates for each ($\mathrm{acc}_{\mathrm{sample=k,m1}}$ and $\mathrm{acc}_{\mathrm{sample=k,m2}}$).

- From each distribution, compute the standard error $\mathrm{stderror}_{\mathrm{overall,m1}} = \mathrm{stddev}_{\mathrm{overall,m1}}/\sqrt{\mathrm{N}}$, and $\mathrm{stderr}_{\mathrm{overall,m2}} = \mathrm{stddevoverall, m2}/\sqrt{\mathrm{N}}$.

- Compute test statistic:

$$\frac{[\mathrm{acc}_{\mathrm{overall,m2}} - \mathrm{acc}_{\mathrm{overall,m1}}]}{[\sqrt{\mathrm{stderror}^2_{\mathrm{overall,m2}} + \mathrm{stderror}^2_{\mathrm{overall,m1}}}]} \tag{1}$$

Do granular tracking data add significantly more value? A bootstrapping procedure

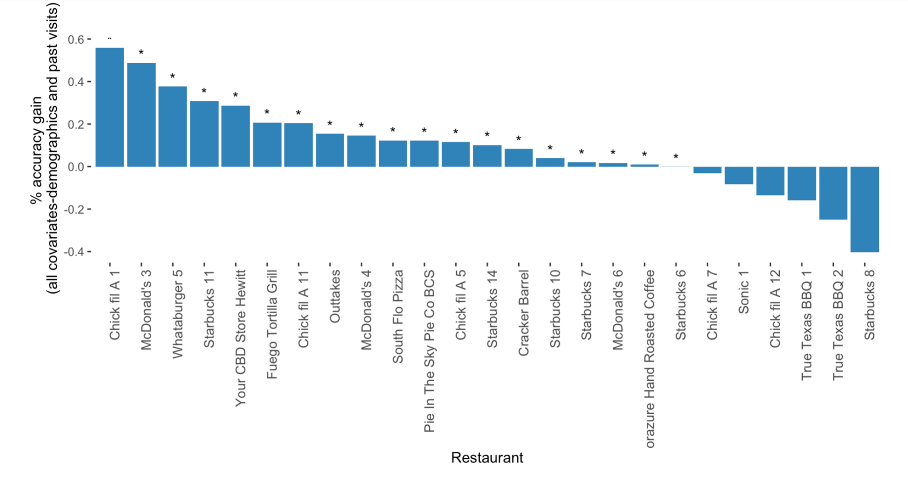


Figure 3: Accuracy changes by restaurant (* $p < 0.05$)

Variable importance