Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data

Susan Athey, David Blei, Robert Donnelly, Francisco Ruiz and Tobias Schmidt

January 2018

Restaurant Choice

- Where should a restaurant be located?
- What is the best type of restaurant for a location?
- Who are a restaurant's competitors?
- How far will consumers travel to a restaurant they like?

These are examples of product design, location, and quality questions.

Modeling Consumer Choice with Panel Data

- Seeing many consumers and items helps us learn about restaurant characteristics, even if the matrix is sparse
- Old literature on "product maps" **?**
- Large literature on estimating consumer choice with panel data with random coefficients on observed attributes, unobserved product quality; see **?** for a survey
- New literature on consumer choice with matrix factorization approach to latent factors:
  - Shopping for many independent categories in parallel, with heterogeneity in mean utilities and price coefficients: **?**
  - Estimating pairwise substitution/complement parameters for all items in the grocery store without prior category information: **?**
  - Estimating multi-stage decision model for consumer grocery shopping: **?**
- Estimating travel time preferences from cross-sectional school choice data using traditional approaches: **?**

Travel Time Factorization Model (TTFM) of User Choice

$$U_{uit} = \underbrace{\lambda_i}_{\text{popularity}} + \underbrace{\theta_u^\top \alpha_i}_{\text{customer preferences}} - \underbrace{\gamma_u^\top \beta_i \cdot \log(d_{uit})}_{\text{distance effect}}$$
$$+ \underbrace{\mu_i^\top \delta_{w_{ut}}}_{\text{time-varying effect}} + \underbrace{\epsilon_{uit}}_{\text{noise}},$$

Covariates x_i affect mean of prior of α_i and β_i.

MNL Comparison: λ_i is constant across restaurants, α_i is observable characteristics of items, θ_u is constant across users, δ_w is omitted, and $\gamma_u \cdot \beta_i$ is constant across users and restaurants.

Dataset

Base Data

- SafeGraph, which aggregates locational information from consumers who have opted into sharing their location through mobile applications.
- “pings” from consumer phones: device id; timestamp; latitude, longitude
- January through October 2017, San Francisco Bay Area

Constructed Data

- “Typical” morning location of the consumer, defined as the most common place the consumer is found from 9:00 to 11:15 a.m. on weekdays.
- Most morning pings in morning location
- South San Francisco to San Jose, excl. mountains/coast
- Lunch restaurant visit: observed at least two pings more than 3 minutes apart during the hours of 11:30 a.m. to 1:30 p.m. in a location that we identify as a restaurant.
- Restaurants are identified using data from Yelp that includes geo-coordinates, star ratings, price range, restaurant

Summary Statistics

Table: Summary Statistics.

User-Level Statistics					
Variable (Per User)	Mean	25%	50%	75%	% Missing
Total Visits	11.63	4.00	7.00	13.00	—
Distinct Visited Rest.	7.25	3.00	5.00	9.00	—
Distinct Visited Categories	11.60	6.00	10.00	15.00	—
Median Distance (mi.)	3.06	0.89	1.86	3.79	—
Weekly Visits	0.39	0.15	0.25	0.47	—
Weeks Active	31.14	22.00	33.00	41.00	—
Mean Rating of Visited Rest.	3.29	3.00	3.33	3.61	1
Mean Price Range of Visited Rest.	1.55	1.33	1.53	1.75	0.6
Restaurant-Level Statistics					
Variable (Per Restaurant)	Mean	25%	50%	75%	% Missing
Distinct Visitors	13.53	5.00	10.00	19.00	—
Median Distance (mi.)	2.39	0.93	1.72	2.94	—
Weeks Open	42.17	44.00	44.00	44.00	—
Weekly Visits (Opens)	0.54	0.17	0.37	0.72	—
Weekly Visits (Always Open)	0.52	0.16	0.34	0.68	—
Weekly Visits (Closes)	0.53	0.15	0.34	0.67	—
Price Range	1.56	1.00	2.00	2.00	10.66
Rating	3.38	2.89	3.53	4.00	14.52

Estimation Details

- Bayesian Estimation with Hierarchical Prior
- Gaussian prior over latent char's, shifted by x_i:

$$p(\alpha_i \mid H_\alpha, x_i) = \frac{1}{(2\pi\sigma_\alpha^2)^{k_1/2}} \exp\left\{ -\frac{1}{2\sigma_\alpha^2} ||\alpha_i - H_\alpha x_i||_2^2 \right\},$$

$$p(\beta_i \mid H_\beta, x_i) = \frac{1}{(2\pi\sigma_\beta^2)^{k_2/2}} \exp\left\{ -\frac{1}{2\sigma_\beta^2} ||\beta_i - H_\beta x_i||_2^2 \right\}.$$

- Latent matrices H_α and H_β, of sizes $k_1 \times k_{\mathrm{obs}}$ and $k_2 \times k_{\mathrm{obs}}$ respectively, which weigh the contribution of each observed attribute on the latent attributes.
- Mean-field variational inference-approximate posterior with independent Gaussians and find parameters that minimize distance
- Stochastic gradient descent

Model Fit

Model	MSE	Log Likelihood	Precision@1	Precision@5	Precision@10
Training Sample					
TTFM	0.00025	-3.59	31.8%	59.4%	70.3%
MNL	0.00031	-6.58	2.8%	10.7%	16.7%
Held-out Test Sample					
TTFM	0.00028	-5.19	20.5%	35.5%	42.2%
MNL	0.00031	-6.55	3.1%	11.4%	17.5%

Figure: Goodness of Fit Measures by User Decile

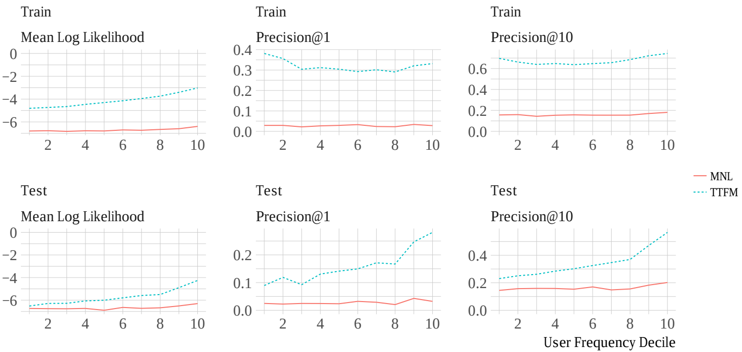


Figure: Goodness of Fit Measures by Restaurant Visit Decile

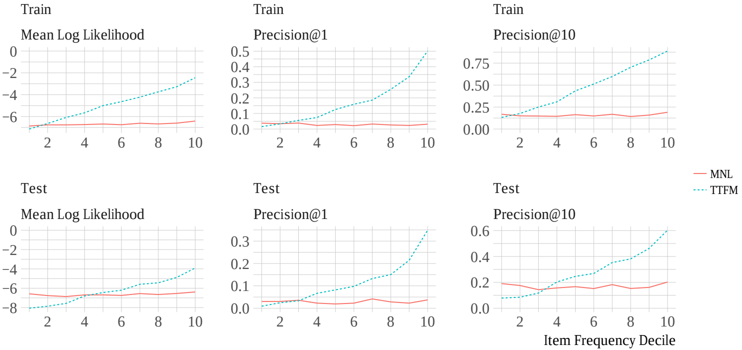


Figure: Goodness of Fit Measures by Distance

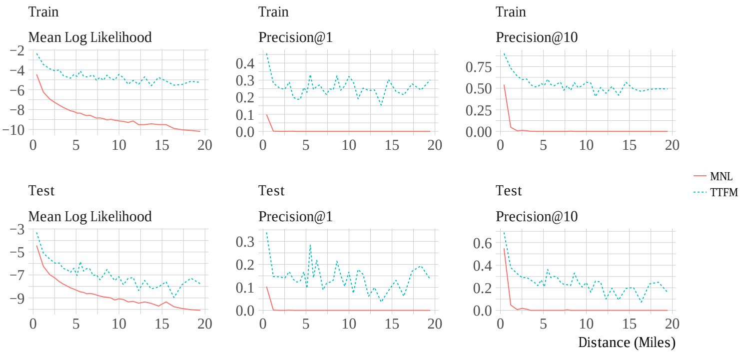


Figure: Predicted Versus Actual Shares By Distance

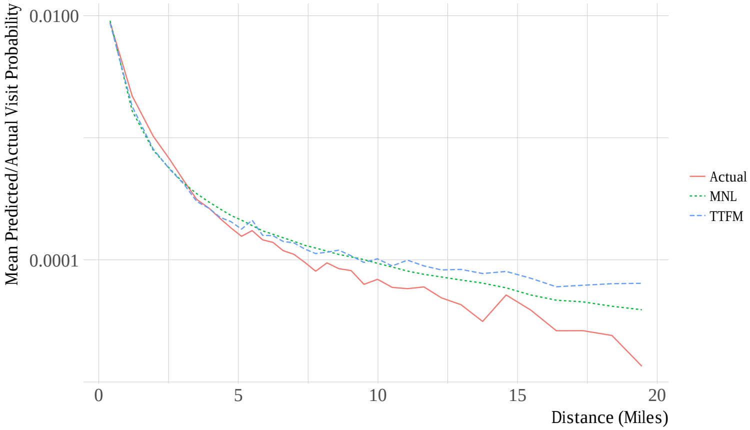


Figure: Actual v. Predicted Visits by Restaurant Visit Decile

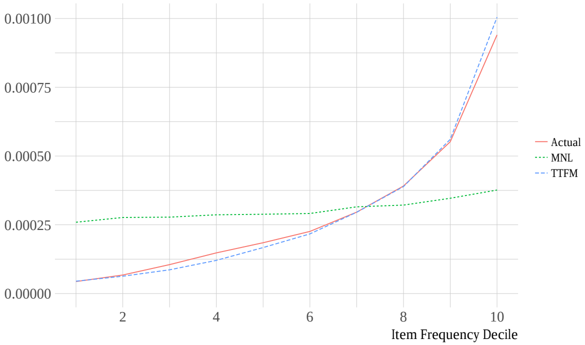


Figure: Distribution of Elasticities

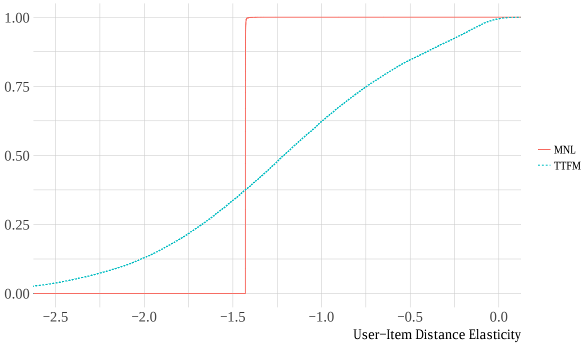


Table: Average Elasticities by Restaurant Characteristics, TTFM model.

Characteristic	Mean	se	25 %	50 %	75 %	N
All restaurants	-1.411	0.0001	-1.585	-1.408	-1.203	4924
Most popular category: Mexican	-1.499	0.0004	-1.664	-1.491	-1.285	694
Most popular category: Sandwiches	-1.435	0.0006	-1.602	-1.441	-1.235	522
Most popular category: Hotdog	-1.403	0.0007	-1.570	-1.390	-1.216	377
Most popular category: Coffee	-1.390	0.0008	-1.563	-1.404	-1.178	365
Most popular category: Bars	-1.370	0.0009	-1.546	-1.362	-1.161	352
Most popular category: Chinese	-1.353	0.0009	-1.517	-1.378	-1.176	350
Most popular category: Japanese	-1.320	0.0011	-1.472	-1.336	-1.140	276
Most popular category: Pizza	-1.497	0.0010	-1.649	-1.481	-1.307	260
Most popular category: Newamerican	-1.323	0.0019	-1.540	-1.351	-1.117	181
Most popular category: Vietnamese	-1.328	0.0020	-1.541	-1.327	-1.155	156
Most popular category: Other	-1.411	0.0002	-1.582	-1.406	-1.189	1391
Price range: 1	-1.446	0.0001	-1.607	-1.435	-1.245	2091
Price range: 2	-1.368	0.0001	-1.542	-1.371	-1.162	2165
Price range: 3	-1.320	0.0026	-1.506	-1.353	-1.108	122
Price range: 4	-1.449	0.0178	-1.664	-1.496	-1.289	21
Price range: missing	-1.474	0.0006	-1.648	-1.455	-1.225	525
Rating, quintile: 1	-1.427	0.0003	-1.605	-1.414	-1.209	842
Rating, quintile: 2	-1.392	0.0003	-1.557	-1.397	-1.187	842
Rating, quintile: 3	-1.364	0.0003	-1.532	-1.366	-1.169	842
Rating, quintile: 4	-1.385	0.0004	-1.571	-1.370	-1.180	842
Rating, quintile: 5	-1.438	0.0003	-1.603	-1.438	-1.250	841
Rating, quintile: missing	-1.475	0.0004	-1.653	-1.464	-1.232	715

Table: Average Elasticities by City, TTFM model.

Characteristic	Mean	se	25 %	50 %	75 %	N
All restaurants	-1.411	0.0001	-1.585	-1.408	-1.203	4924
City: Daly City	-1.105	0.0019	-1.331	-1.150	-0.959	165
City: Burlingame	-1.119	0.0030	-1.327	-1.194	-1.018	110
City: Millbrae	-1.130	0.0049	-1.418	-1.240	-0.954	80
City: San Bruno	-1.132	0.0035	-1.398	-1.216	-0.987	101
City: South San Francisco	-1.187	0.0021	-1.413	-1.232	-0.999	135
City: San Mateo	-1.243	0.0012	-1.454	-1.284	-1.101	268
City: Foster City	-1.318	0.0070	-1.506	-1.397	-1.163	44
City: San Carlos	-1.321	0.0026	-1.479	-1.350	-1.195	95
City: Palo Alto	-1.330	0.0013	-1.519	-1.342	-1.171	234
City: Brisbane	-1.332	0.0139	-1.455	-1.344	-1.181	15
City: Belmont	-1.334	0.0047	-1.500	-1.374	-1.212	58
City: Redwood City	-1.362	0.0012	-1.530	-1.389	-1.217	214
City: Cupertino	-1.365	0.0018	-1.532	-1.386	-1.174	169
City: East Palo Alto	-1.374	0.0142	-1.521	-1.393	-1.229	13
City: Los Gatos	-1.391	0.0026	-1.583	-1.437	-1.219	106
City: Los Altos	-1.406	0.0043	-1.564	-1.394	-1.236	60
City: Menlo Park	-1.407	0.0031	-1.570	-1.428	-1.287	87
City: Mountain View	-1.422	0.0013	-1.592	-1.429	-1.233	213
City: Santa Clara	-1.442	0.0009	-1.681	-1.456	-1.238	355
City: San Jose	-1.451	0.0002	-1.635	-1.464	-1.278	1858
City: Campbell	-1.482	0.0015	-1.640	-1.493	-1.317	144
City: Saratoga	-1.497	0.0059	-1.628	-1.481	-1.394	40

Figure: Model Predictions of the Effect of Restaurant Openings and Closings Controlling for Other Changes.

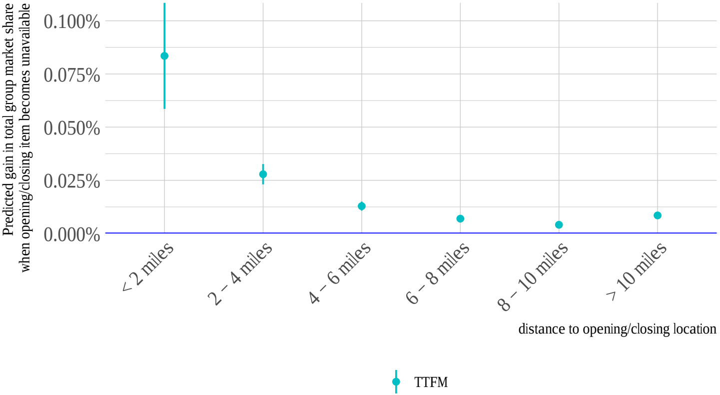


Table: Share of demand redistributed by distance, TTFM model

	Distance from opening/closing restaurant (mi.)					
	< 2	2 - 4	4 - 6	6 - 8	8 - 10	> 10
share	51 %	23 %	10 %	6 %	3 %	6 %
cum. share	51 %	74 %	84 %	90 %	94 %	100 %

Figure: Model Predictions Compared to Actual Outcomes for Restaurant Openings and Closings.

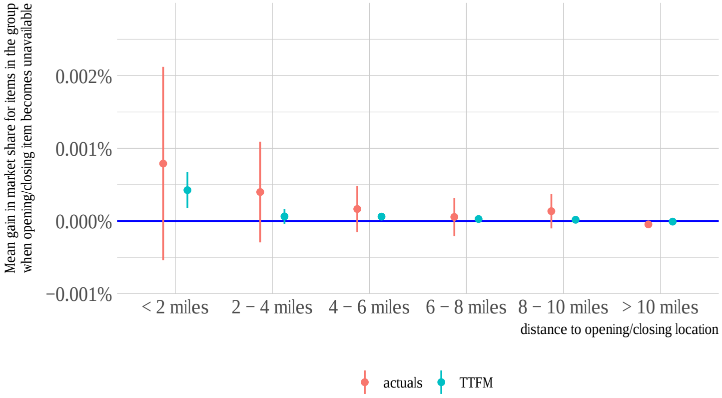
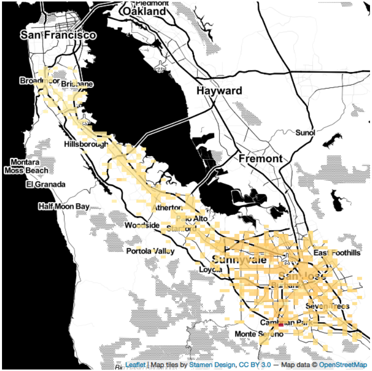


Table: Alternative Restaurant Characteristics for Opening and Closing Restaurants

Mean Predicted Demand	Closing	Opening
Actual Opening/Closing Restaurant	10.33 (0.83)	12.10 (1.14)
Alternative from Same Category	10.08 (0.12)	10.53 (0.11)
Alternative from Different Category	9.09 (0.08)	9.71 (0.08)

Figure: Best Locations for Restaurant Category

Cafes



Chicken Wings

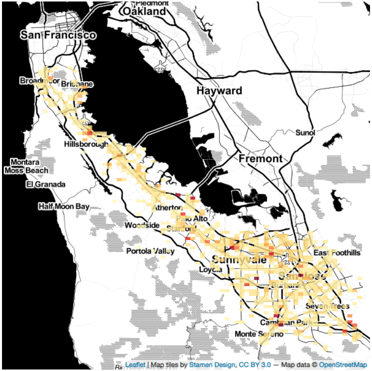
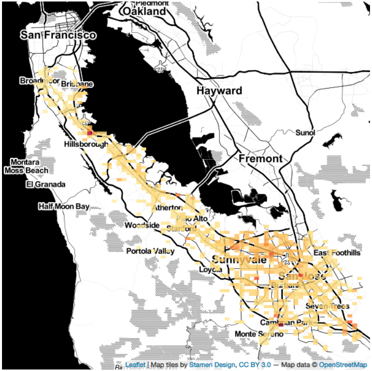


Figure: Best Locations for Restaurant Category

Filipino Restaurants



Sandwiches

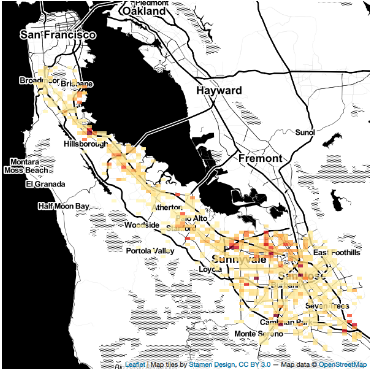
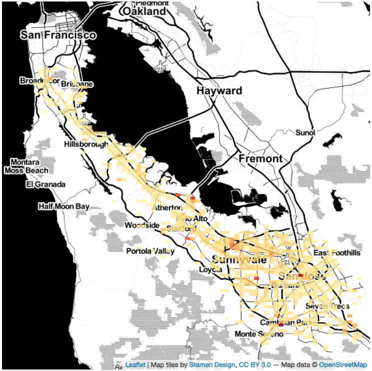


Figure: Best Locations for Restaurant Category

Vegetarian



Vietnamese Restaurants

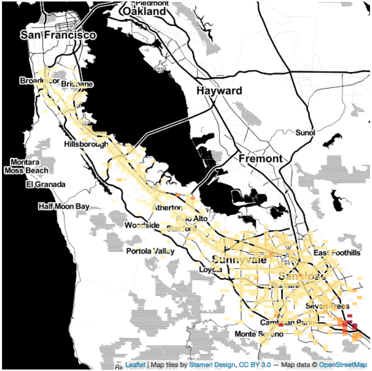
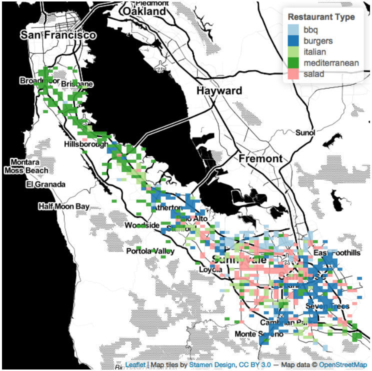


Figure: Best Restaurant Category for Locations

Mid-Priced ($$) Western Cuisine



Mid-Priced ($$) Asian Cuisine

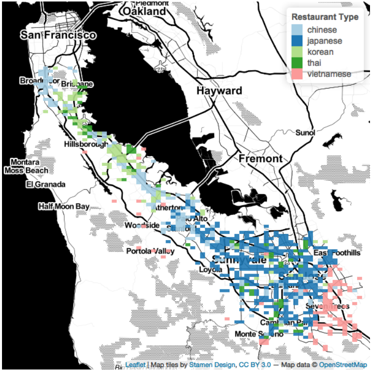
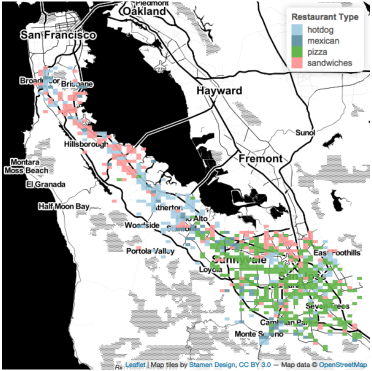
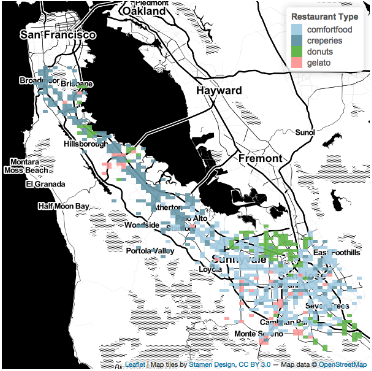


Figure: Best Restaurant Category for Locations

Cheap ($) Fast Food



Cheap ($) Treats



Conclusions

- To analyze product location choices, need a good model of consumer demand in characteristics space and physical location
- Modern panel datasets provide individual-level data that enables learning models with rich heterogeneity
- Computational approaches from ML make these models tractable to estimate
- Rich models do a better job with personalization and counterfactuals
- Understanding travel time preferences is important input for urban planning