AEA Papers and Proceedings 2018, 108: 64–67
https://doi.org/10.1257/pandp.20181031

ECONOMIC APPLICATIONS OF MACHINE LEARNING[‡]

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

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

Where should a new restaurant be located? What type of restaurant would be best in a given location? What defines a geographical market? These are examples of questions about product design and product choice. While there is extensive literature on consumer response to prices, there is relatively little attention to firm choices about physical location and product characteristics. Recent trends in digitization have led to the creation of many large panel datasets of consumers, which in turn motivates the development of models that exploit the rich information in the data and provide precise answers to these questions.

Answering many of these questions requires a model that incorporates individual-level heterogeneity in preferences for product attributes and travel time, as these characteristics might vary substantially even within a city; and understanding individual heterogeneity in travel preferences is a key input for urban planning. To this end, we develop an empirical model of consumer choices over lunchtime restaurants, the travel-time factorization model (TTFM). We apply the model to a dataset derived from mobile phone locations. The personalized predictions of TTFM for individuals and restaurants are more accurate than existing methods, especially for high-activity individuals and restaurants.

TTFM can answer counterfactual questions such as, what would happen if a restaurant with a given set of characteristics opened or closed in a given location? Using data about several hundred openings and closings, we compare TTFM's predictions to actual outcomes.

TTFM incorporates recently developed approaches from machine learning for estimating models with a large number of latent variables. It uses a standard discrete choice framework to model each user's choice over restaurants, inferring the parameters of the users' utility functions from their choice behavior. TTFM differs from traditional models in the number of latent variables; it incorporates a vector of latent characteristics for each restaurant as well as latent user preferences for these characteristics. In addition, it incorporates heterogeneous user preferences for travel distance, which vary by restaurant. These distance preferences are represented as the inner product of restaurant-specific factors and user willingness to travel to restaurants with those factors. Finally, TTFM is a hierarchical model, where observable restaurant characteristics affect the distribution of latent restaurant characteristics. We use a Bayesian approach to

[‡] *Discussants:* Michael Luca, Harvard Business School; Marshall Burke, Stanford University; Greg Lewis, Microsoft Research; Shane Greenstein, Harvard Business School.

* Athey: Stanford University, 655 Knight Way, Stanford, CA 94305 (email: athey@stanford.edu); Blei: Department of Computer Science, Columbia University, New York, NY 10027 (email: david.blei@columbia.edu); Donnelly: Stanford University, 655 Knight Way, Stanford, CA 94305 (email: rodonn@stanford.edu); Ruiz: Department of Computer Science, Columbia University, New York, NY 10027, and University of Cambridge (email: fr2392@columbia.edu); Schmidt: Stanford University, 655 Knight Way, Stanford, CA 94305 (email: tobiass@stanford.edu). The authors are listed in alphabetical order. We are grateful to SafeGraph and Yelp for providing the data, and to Paula Gablenz, Renee Reynolds, Tony Fan, and Arjun Parthipan for exceptional research assistance. We acknowledge generous financial support from Microsoft Corporation, the Sloan Foundation, the Cyber Initiative at Stanford, and the Office of Naval Research. Ruiz is supported by the EU H2020 programme (Marie Skłodowska-Curie grant agreement 706760).

[†] Go to https://doi.org/10.1257/pandp.20181031 to visit the article page for additional materials and author disclosure statement(s).

inference, where we estimate posterior distributions over each user's preferences and each restaurant's characteristics. We rely on stochastic variational inference to approximate the posterior distribution with a stochastic gradient optimization algorithm.

Our approach builds on a large literature in economics and marketing on estimating discrete choice models of consumer behavior (e.g., Keane 2015). It also relates to a literature in marketing on inferring "product maps" from panel data (Elrod 1988). Our estimation strategy is drawn from approaches developed in Athey et al. (2017) and Ruiz, Athey, and Blei (2017), both of which considered the problem of choosing items from a supermarket, and it also relates to Wan et al. (2017). A few authors have estimated consumer preferences for travel time, e.g., Neilson (2013).

I. Empirical Model

We model the consumer's choice of restaurant conditional on deciding to go out to lunch. We assume that the consumer selects the restaurant that maximizes utility, where the utility of user u for restaurant i on her t-th visit is

$$U_{uit} = \lambda_i + \theta_u^\top \alpha_i + \mu_i^\top \delta_{w_{ut}} - \gamma_u^\top \beta_i \log(d_{ui}) + \epsilon_{uit},$$

where w_{ut} denotes the week in which trip t happens, and d_{ui} is the distance from u to i. This gives a parameterized expression for the utility: λ_i is an intercept term that captures a restaurant's popularity; θ_u and α_i are latent vectors that model a user's latent preferences and a restaurant's latent attributes; β_i is a vector that captures a restaurant's latent factors for travel distance and γ_u is a user's latent preferences of willingness to travel to restaurants with those factors; δ_w and μ_i are latent vectors of week/restaurant time effects (this allows us to capture varying effects for different parts of the year); and ϵ_{uit} are error terms, which we assume to be independent and identically Gumbel distributed. We specify a hierarchical model where observable characteristics of restaurants, denoted by x_i, affect the mean of the distribution of latent restaurant characteristics α_i and β_i. This hierarchy allows restaurants to share statistical strength, which helps to infer the latent variables of low-frequency restaurants. We estimate the posterior over the latent model parameters using variational inference. See online Appendix A.A3 for details.

For comparison, we also consider a simpler model, a standard multinomial logit model (MNL), which is a restricted version of our proposed model: the term λ_i is constant across restaurants, α_i is set to be equal to the observable characteristics of items, θ_u is constant across users, δ_w is omitted (including it created problems with convergence of the estimation), and $\gamma_u \cdot \beta_i$ is restricted to be constant across users and restaurants.

II. Data, Estimation, and Model Fit

The dataset is from SafeGraph, a company that collects anonymous, aggregate locational information from consumers who have opted into sharing their location through mobile applications. The data consists of "pings" from consumer phones; each observation includes a unique device identifier that we associate with an anonymous consumer, the time and date of the ping, and the latitude, longitude and accuracy of the ping over a sample period from January through October 2017.

Using this data, we construct the approximate "typical" morning location of the consumer, defined as the most common place the consumer is found from 9:00 AM to 11:15 AM on weekdays. We restrict attention to consumers whose morning locations are consistent over the sample period, and for which these locations are in a subset of the San Francisco Bay Area. We determine that the consumer visited a restaurant for lunch if we observed at least two pings more than three minutes apart during the hours of 11:30 AM to 1:30 PM in a location that we identify as a restaurant. Restaurants are identified using data from Yelp that includes geo-coordinates, star ratings, price range, and restaurant categories. We narrow the dataset to a subset of restaurants that appear sufficiently often in the data, and to consumers who visit a sufficient number of restaurants. This results in a final dataset of 106,889 lunch visits by 9,188 users to 4,924 locations. Online Appendix A.A2 gives details and summary statistics.

We divide the dataset into three parts, 70.6 percent training, 5.0 percent validation, and 24.4 percent testing. We use the validation dataset to select parameters such as the length of the latent vectors α_i and β_i, 80 and 16, respectively), while we compare models and evaluate performance in the test dataset.

Across several measures evaluated on the test set, TTFM is a better model than MNL (see Table A2 for details). For example, Precision@5 is the percentage of times that a user's chosen restaurant is in the set of the top five predicted restaurants. It is 35 percent for TTFM and 11 percent for MNL. Further, as shown in Figures A4 and A5, TTFM predictions improve significantly for high-frequency users and restaurants, while MNL does not exhibit that improvement. This highlights the benefits of personalization: When given enough data, TTFM learns user-specific preferences.

Figure A7 illustrates that both TTFM and MNL fit well the empirical probability of visiting restaurants at varying distances from the consumer's morning location. But Figure A8 shows that TTFM outperforms MNL at fitting the variation in visit rates across restaurants.

III. Parameter Estimates

The distributions of estimated elasticities from TTFM are summarized in Table A4 and Figure A9. The elasticities in MNL vary only because the baseline visit probabilities vary. TTFM elasticities are more dispersed, reflecting the personalization capabilities of the TTFM model. The average elasticity across consumers and restaurants (weighted by trip frequency) is -1.41. Thus, distance matters substantially, consistent with the fact that roughly 60 percent of visits are within two miles of the consumer's morning location. Across users and restaurants, the standard deviation of elasticities in the TTFM model is −0.68, while the average within-user standard deviation of elasticities is −0.30 and the average within-restaurant standard deviation of elasticities is −0.60.

Willingness to travel is lower for low-priced restaurants (elasticity −1.45 for price range $ (under $10) versus −1.37 for price range $$ ($11–$30)); lower for Mexican restaurants and pizza places than for Chinese and Japanese restaurants (elasticities of −1.50 and −1.50 versus −1.35 and −1.32, respectively). Cities with many work locations nearby retail districts, including San José, Sunnyvale, and Mountain View have a lower willingness to travel than cities that are more spread out like Daly City, Burlingame, San Bruno, and San Mateo. Online Appendix A.A5 provides further analysis of latent factors and model estimates.

IV. Counterfactuals

The TTFM model can predict how market share will be redistributed when restaurants open or close, and these predictions can be compared to the actual changes. We focus on 221 openings and 190 closings where, both before and after the change, there were at least 500 restaurant visits by users with morning locations within a 3 mile radius of the relevant restaurant.

One challenge of analyzing market share redistribution is that for any given target restaurant that opens or closes, we would expect some baseline level of market share changes of competing restaurants due to changes in the open status of neighboring restaurants. We address this in an initial exercise where we hold the environment fixed in the following way. For each target restaurant that changed status, we first construct the predicted difference in market shares for each other restaurant between the "closed" and "open" regime (irrespective of which came first in time), and then subtract out the predicted change in market share that would have occurred for each restaurant if the target restaurant had been closed in both periods. We then sum the changes across restaurants in different groups defined by their distance from the target restaurant. Figure A11 illustrates the average of the latter statistic across target restaurants. The TTFM model estimates imply that just over 50 percent of the market share impact of a closure accrues to restaurants within 2 miles of the target.

Figure A12 compares the actual changes in market share that occurred against the predictions of the TTFM model. Our model's predictions match well the actual changes that occurred, but there is substantial variation in the changes that occurred in the actual data.

Our final exercise considers the best choice of restaurant type for a location. For the set of restaurants that open or close, we look at how the demand for the restaurant that changed status (the "target restaurant") compares to the counterfactual demand the model predicts in the scenario where a different restaurant in our sample (as described by its mean latent characteristics) is placed in the location of the target restaurant. For each target, we consider a set of

Table 1—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)

200 alternative restaurants, 100 from the same category as the target restaurant and 100 from a different category. We then compare the target restaurant's estimated market share to the mean demand across the set of alternatives. Table 1 illustrates that both the restaurants that opened and those that closed on average have higher predicted demand than either group of alternatives (standard errors in parentheses). However, the restaurants that opened appear to be in more valuable locations, since for the 200 alternative restaurants, we predict higher average demand if they were (counterfactually) placed at the opening locations than at the locations of closing restaurants. As a further comparison, we split the set of alternatives into groups based on whether or not they are in the same broad category as the restaurant that opened or closed. We find alternatives from the same category as the target would perform better on average than alternatives from a different category. Next, we consider the match between restaurant characteristics and locations. In gridcells of roughly 0.7 × 0.35 miles, we select one restaurant location at random and use the TTFM model to predict what total demand would have been if a different restaurant had been located in its place. Figure A13 shows, for example, that the most attractive location for Vietnamese restaurants is in a dense region southeast of San José. Figure A14 illustrates considerable spatial heterogeneity in which restaurant category is predicted to perform best in each location.

Beyond these questions, the TTFM model can be used to evaluate a wide variety of counterfactual scenarios, and the results can also contribute to urban planning debates about zoning and transportation. In future work we may consider a larger set of cities in order to enable more precise comparisons of the model's predictions versus actual outcomes for the impact of restaurants opening and closing.

REFERENCES

Athey, Susan, David M. Blei, Robert Donnelly, and Francisco J. R. Ruiz. 2017. "Counterfactual Inference for Consumer Choice across Many Product Categories." Unpublished.

Elrod, Terry. 1988. "Choice Map: Inferring a Product-Market Map from Panel Data." *Marketing Science* 7 (1): 21–40.

Keane, Michael P. 2015. "Panel Data Discrete Choice Models of Consumer Demand." In *The Oxford Handbook of Panel Data*, edited by Badi H. Baltagi, 549–83. Oxford, UK: Oxford University Press.

Neilson, Christopher. 2013. "Targeted Vouchers, Competition among Schools, and the Academic Achievement of Poor Students." Unpublished.

Ruiz, Francisco J. R., Susan Athey, and David M. Blei. 2017. "SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements." *arXiv* 1711.03560.

Wan, Menting, Di Wang, Matt Goldman, Matt Taddy, Justin Rao, Jie Liu, Dimitrios Lymberopoulos, and Julian McAuley. 2017. "Modeling Consumer Preferences and Price Sensitivities from Large-Scale Grocery Shopping Transaction Logs." Paper presented at the International World Wide Web Conference, Perth, Australia.