Using Default Recommendations in Demand Estimation

Jonathon McClure*
Mitch Daniels School of Business
Purdue University

January 2025

Abstract

Online recommendation platforms aid consumers in making decisions amid large choice sets by suggesting commonly-chosen alternatives to explored products. Using default recommendations for alternative hotel choices from Google Travel, I rank products on similarity and construct an embedding of the latent preference space for the mean consumer for hotels in Orange County, CA. I show via simulation and an empirical exercise that these data can be used to augment the estimation of a flexible demand model incorporating heterogeneous preferences for differentiated products. This approach is viable even in the absence of observed product characteristics and requires no proprietary platform data.

*Assistant Professor, Department of Economics, Mitch Daniels School of Business, Purdue University. Email: *mcclur47@purdue.edu*. I would like to thank Giovanni Compiani, Lorenzo Magnolfi, Ralph Siebert, and discussants at the Purdue Economics summer seminar series for comments and feedback. Derek DeVito provided valuable RA support. Data was provided by Duane Vinson at STR.

1 Introduction

A challenge central to many studies in empirical IO is the estimation of demand models and the recovery of substitution patterns. Typical approaches—product-based (e.g. Deaton and Muellbauer (1980); Pinkse, Slade, and Brett (2002)) or characteristics-based (e.g. Berry, Levinsohn, and Pakes (1995), henceforth BLP; Nevo (2001))—require data demonstrating how products are differentiated, which is usually paired with either assumptions on aggregated preferences and/or combined with data on local consumers in order to add richness to the distribution of consumer preferences. This in turn creates two challenges for the practitioner: first, whether the data on product attributes is sufficient to properly capture product differentiation, or whether it is available at all.[1] Second, while distributional assumptions of consumer preferences are often paired with demographic information to characterize variation in local consumers' tastes, these data are not applicable to spaces where the consumer base of local products is undefined. This makes it more challenging to consider how products relate in *preference* space rather than simply product space, potentially weakening the identification of the parameters of the demand model.

In this paper, I show how information from product recommendation systems—specifically, the publicly-shown default recommendations of alternatives to a chosen product—can be incorporated to augment the estimation of demand systems in the absence of detailed product characteristics. In contrast to approaches which use consumer responses to obtain second-choice data and construct additional moments to match (Berry, Levinsohn, and Pakes (2004), Conlon, Mortimer, and Sarkis (2023)), I use rankings over recommendations to construct a continuous vector representation (an "embedding") of the latent preference space, where more frequently-chosen substitutes to a given product are located closer to it in a metric space. I use the resulting embedding to construct demand estimates using a distance-based product-space approach in the vein of Pinkse et al. (2002), as well as a random-coefficients logit model where the coordinates of the embedding reflect variation in characteristics *and* the representative consumer's preferences for the product.

I illustrate this method with an application to the hotel sector, where the typical mixed logit approach faces challenges owing to the lack of information about consumer preferences; the usual demographics approach displayed by Nevo (2001) is not suitable as the customer

[1] The scalability of IO research from market studies to broader trends is often hindered by the lack of collectible data on product characteristics: see e.g. De Loecker, Eeckhout, and Unger (2020) and Syverson (2019) for discussion.

base is not local. I recover recommended alternatives for 310 reference hotels in Orange County, CA from public searches through Google Travel. These data contain no proprietary information: the method which can be generalized to the collection of many other types of consumer products. Given an assumption that platforms aim to maximize the probability that a searching user makes a purchase, the ranked order of alternative recommendations for a product j can be interpreted as a descending ordinal ranking of choice probabilities, conditional on the user expressing interest in product j, such that the set of displayed suggestions maximizes the probability of a selection being made. This is similar to the treatment of platform data by Kim, Albuquerque, and Bronnenberg (2010), where product search data on Amazon.com is treated as aggregation of individual searches: here I treat the default recommendations as aggregates of the consumer choices that platforms observe to recover substitutes.

The observed recommendations demonstrate a number of intuitive patterns that suggest the information is sensible: alternative hotels are more likely to be recommended when they are closer in distance and class (a measure of hotel quality based on average daily rates). I find that the alternatives to luxury hotels are least sensitive to distance but most sensitive to quality similarity, suggesting strong preferences for quality that outweigh spatial concerns for guests interested in the luxury experience. Conversely, economy-class hotels place the most weight on distance when selecting recommended alternatives. Lastly, I find that hotels under the same parent company are more likely to be recommended: a facet of either consumer brand preferences or of platform preference towards brands which provide greater revenue. These recommendations form triplets: ordinal measures of distance stating "i is closer to j than it is to k" based on their recommendation rank (or whether the alternative is not recommended at all), which are then used to construct an embedding using the t-Distributed Stochastic Triplet Embedding (tSTE) algorithm (Van Der Maaten and Weinberger (2012)). The patterns in a 2-dimensional embedding continue to reflect the above trends, clustering hotels by proximity and quality beyond their initial physical locations. I merge the coordinates of these embeddings with monthly price and quantity data for hotels in Orange County, CA, a market environment with a high density of spatially-differentiated products and where the estimation of hotel-level substitution patterns would be challenging with the limited observed characteristics and consumer demographics available.

To demonstrate the use of the embedding, I construct several Monte Carlo tests incorporating different forms of consumer heterogeneity. In a simple example of random-coefficients

logit, I test both a distance-based approach similar to Pinkse et al. (2002) and a more conventional BLP approach. I find that a BLP specification using coordinates of the embedding in place of characteristics is able to produce closer estimates of diversion to the outside option and markups. In a more comprehensive example where unobserved consumer heterogeneity results in variation that poorly identifies the demand system, I show that a specification using coordinates of the embedding produces lower RMSE in terms of estimates of diversion, markups, out-of-sample fit, and merger profit and welfare predictions when compared to a specification using the full set of true characteristics. Results are further improved when using both sets of data via a mixed embedding, suggesting the complementary of the approach in appropriate settings.

Applying this approach to data, I then estimate a BLP demand system in monthly Orange County hotel data using four specifications—using observed characteristics (latitude, longitude, and quality), recommendations, both, and neither—and recover hotel-level diversion ratios and markups. I find that specifications including the recommendations estimate higher median diversion to inside products and lower median markups. Additionally, while all specifications other than simple logit exhibit trends of diversion being higher to hotels which are closer in physical and quality space, these trends are smoother in specifications which incorporate the recommendations. Lastly, I note that the recommendation specifications estimate higher median diversion to Upper Upscale hotels: the 2-D embedding suggests that Upper Upscale properties are centrally clustered in each market, suggesting they are more uniformly recommended as alternatives. These findings indicate that the proposed method is easy to implement and is scalable to other environments where platform data are available, with particular value when data on demand-relevant product characteristics are difficult to collect: a concern highly relevant as many such platforms operate in less-studied digital markets.

This paper complements three areas of the literature. First, I contribute to the the rapidly growing literature on the use of auxiliary data to enhance demand estimation and pin down more accurate substitution patterns. The use of auxiliary data is not new: Nevo (2001) and Petrin (2002) use consumer demographics to aid in the estimation of substitution patterns. The method is also conceptually similar to the idea of identifying second (or alternative) choices: survey data has often been used for this purpose (Berry et al. (2004), Grieco, Murry, and Yurukoglu (2021), Conlon et al. (2023)). Survey data has also been used to construct embeddings of the product space for demand estimation, as in Magnolfi, McClure,

and Sorensen (2025) and Compiani, Morozov, and Seiler (2023). In the context of this paper, which is generalizable to other settings where consumers lack knowledge of the full product space, surveys are infeasible: it is unlikely that the survey respondents' preferences are complete over a large number of hotels in a given city, particularly as knowledge of a hotel's characteristics or utility are hard to discern without prior research or experience. The platform instead pools the information of consumers who have already searched: as the platform sees what consumers search for and eventually select, it can summarize these outcomes as recommendations for a consumer currently engaging in search. A remaining problem in this space is, however, how the utilization of these data—which Battaglia, Christensen, Hansen, and Sacher (2024) refer to as "unstructured data"—affects inference given that they are the result of an algorithmic construction; I treat the computed embedding as data and propose estimation of demand *given* said input, rather than asserting that there exists a true latent preference space which I aim to recover.

Secondly, I add to studies of platforms and consumer behavior. My approach is similar in concept to work which makes use of platform search and clickthrough data, which has been used to recover the product space and consumer preferences or otherwise learn about consumer search patterns (Kim et al. (2010), De Los Santos, Hortacsu, and Wildenbeest (2012), and Hodgson and Lewis (2024)). These approaches have seen prior applications to hotels, as in Armona, Lewis, and Zervas (2024), who use search data from Expedia to construct a Bayesian Personalized Ranking for consumers to learn latent product attributes and Kaye (2024), who examines the effects of personalized recommendations on consumer welfare. Related papers using embeddings built from data on consumer search and purchases to estimate demand are Ruiz, Athey, and Blei (2020), Kumar, Eckles, and Aral (2020), and Gabel and Timoshenko (2022). However, many of these approaches rely on the availability of micro-data on searches or purchases: an advantage of the method proposed by this paper is how it can be generalized to new settings, and the convenience of public-facing information which is easy to collect.

Finally, I contribute to empirical studies of the hotel sector and the methodologies by which it can be examined. Estimating a differentiated-products demand system for hotels—where consumers' types are unobserved and where demand is highly spatially volatile—is traditionally challenging.[2] Prior work has often sidestepped the problem: Lewis and Zervas

[2] Berry and Jia (2010) demonstrate an example of estimating different consumer types for airline flights, where they encounter similar issues in unobserved local customer bases. However, identification relies on observing variation in ticket prices for the same flights, which is information not often available in daily

(2019) estimate a simplified monopoly problem, while Farronato and Fradkin (2022) and McClure (2024) estimate demand across aggregated market segments, each in order to examine aspects of the supply side of the market. Armona et al. (2024) provide the most similar exercise, as discussed above, by using search data to recover latent preferences. The approach presented in this paper demonstrates an additional source of information which can illuminate which hotels are close substitutes without the estimation of a demand system, information which can be used to understand how (unobserved) consumers compare hotels or used to subsequently estimate a demand model.

In Section 2, I discuss the methodology of the paper: the available market data, collection of the recommendations, and formation of the embedding. Section 3 details the recommendations themselves, summarizing how they may be generated by the platform from observed consumer choices and providing reduced form and visual evidence for what information the recommendations contain. Following this, in Section 4 I show via simulation how data from a stylized platform can be incorporated in demand estimation and improve common post-estimation statistics. Section 5 explains the empirical application and summarizes the results, comparing model performance. Section 6 concludes.

2 Methodology

2.1 Data

This paper relies on two sources of data. The first is a panel of hotel-level monthly average daily rates (ADR) and occupancy rates from Orange County, CA, provided by STR LLC. The data cover a period of 2017 through 2023. Aside from prices and quantities, I observe hotels' names, addresses, brand affiliation, size, and a number of general characteristics of hotels: their quality tier (class) from Luxury to Economy, their rough number of rooms (allowing occupancy rates to be converted to quantities of sold rooms), and their categorical location (downtown, airport, etc). I normalize hotel-month quantities to the average daily number of rooms sold in the month. Hotels are assigned to one of four geographic markets: Disneyland (Disneyland and Anaheim), proximity to Disneyland (Orange County Northwest/Fullerton), downtown (Santa Ana/Costa Mesa), or beach (Newport Beach/Dana Point). Table 1 summarizes the performance data for hotels in the sample.

hotel market data.

I choose to observe data at the monthly level to relax issues related to stockouts. In higher-frequency (i.e. daily) hotel data, finite capacity results in the presence of corner solutions, which impede inverting the demand system and identifying parameters as the unconstrained quantities demanded are unobserved. Several approaches to resolving this issue have been proposed, such as using micro-data to estimate the latent choice sets or estimating over the various observed choice sets (Conlon and Mortimer (2013), Agarwal and Somaini (2022)). I instead sidestep the problem through aggregation to the monthly level.

TABLE 1: Summary of Hotel Performance Data

			Average Daily Rate (ADR)			Occupancy %		
	Obs	Hotels	5	50	95	5	50	95
Luxury	786	10	226.75	431.66	1100.91	20.25	72.00	98.67
Upper Upscale	3,116	41	100.25	167.86	389.85	15.16	77.50	99.28
Upscale	4,647	60	88.51	143.16	246.17	22.22	78.74	99.43
Upper Midscale	3,809	50	77.29	119.09	196.10	22.60	72.97	99.00
Midscale	3,040	39	65.37	92.03	147.35	25.76	76.99	98.50
Economy	3,314	44	58.35	78.98	137.24	18.75	69.53	97.78
Disneyland	7,060	94	68.45	123.73	267.53	17.52	77.59	99.19
Near Disneyland	3,138	41	60.05	97.76	175.28	20.00	72.82	99.02
Downtown	5,144	66	66.19	115.86	249.31	23.08	74.53	99.24
Beaches	3,370	43	76.71	152.17	645.47	22.97	72.97	98.85
Total	18,712	244	66.57	121.00	320.62	20.59	75.00	99.15

Source: STR hotel data. ADR and Occupancy values are presented as the 5th, 50th, and 95th percentiles for the hotels in the sample. Values for hotels are monthly averages of daily performance data.

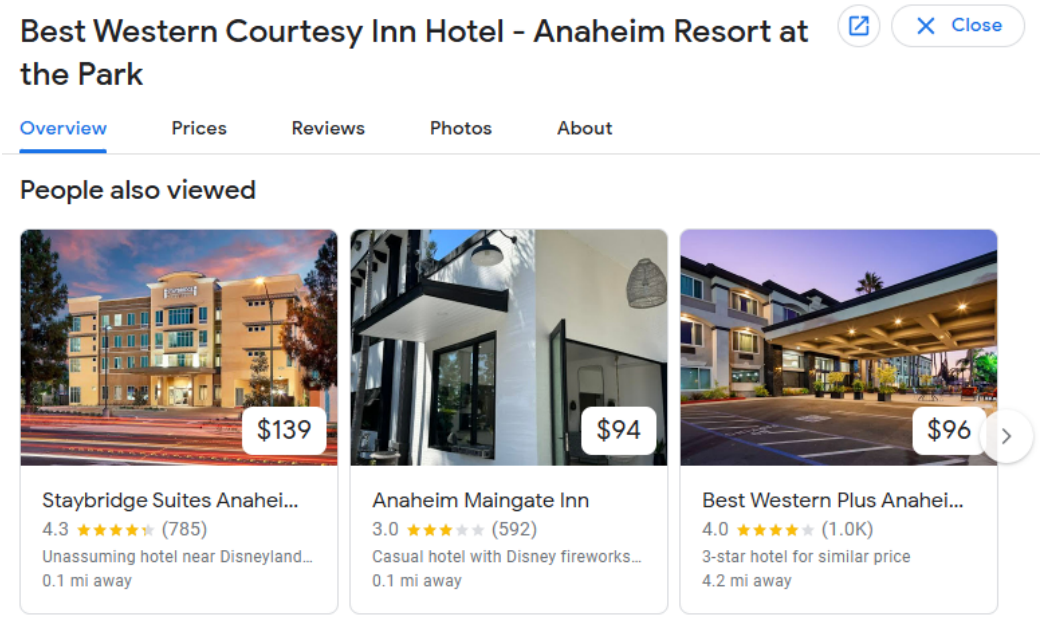
My second source of data is a set of recommendations collected from Google Travel. I collect up to 6 alternate-product recommendations—described as hotels that "People also viewed"—for hotels in Orange County, presented in a panel as displayed in Figure 1. Here, a search for "Best Western Courtesy Inn Hotel - Anaheim Resort at the Park" gives a set of 6 (3 displayed) alternative suggestions that were searched.[3] In total, I collect recommendations for 310 hotels: when limiting recommendations to hotels which appear in the STR data, I recover a connected set of 268 hotels, of which 244 appear in the price/quantity data.[4] These recommendation rankings are converted to triplets and used to estimate an

[3] Different platforms present this information in different ways. For example, Booking.com states "Travelers who viewed [this hotel] ended up booking these properties" when presenting alternatives, which suggests a more intuitive link to second-choices (see Appendix Figure 1). As the objective function of the platform is obfuscated, I treat these recommendations as "similar products by preference" so long as they are not stated to be advertisements.

[4] Not all hotels which are profiled by STR provide performance data. Hotels which have recommendations and are placed in the embedding but lack performance data are simply excluded from demand analysis (i.e. they enter the outside option). In Section 2.2.1, I discuss the importance of recommended hotels forming a

embedding using the tSTE algorithm. In the following sections, I summarize key facts about the contents of each hotel's set of recommendations and the resulting embedding.

FIGURE 1: Example of Default Recommendations



2.2 Triplets and Embeddings

By scraping hotel recommendations, I construct an ordered list of substitutes to each product for the default consumer. These recommendations are collected for nights a minimum of six months in the future, using short, incognito searches in order to capture the recommendations presented without bias for search history. I use this to construct triplets: inequalities that state "product A is closer to B than it is to C" using the ranking of products suggested when searching for each product $j \in \mathcal{J}$. I then employ the t-Distributed Stochastic Triplet Embedding (tSTE) algorithm proposed by Van Der Maaten and Weinberger (2012) to compute a continuous vector representation of the products' mean utility in a low-dimensional latent space. This exercise is similar to Armona et al. (2024), who consider that if consumers search products j_1, j_2 in order, then the products mush have related attributes and be more similar than products which were not clicked. However, they

connected set with regards to the formation of the embedding.

make use of the consumers' search data from the platform itself: a feature of my method is that I do not require data beyond what platforms display to consumers.

The tSTE algorithm proceeds as follows. Formally, given a set of products $j = 1, \ldots, J$, we want to find a set of vectors $\mathbf{x} \equiv \{x_1, \ldots, x_J\} \in \mathbb{R}^m$ that represent the products in m-dimensional space.[5] Letting $\mathcal{T}$ be the set of triplet comparisons in our data, each one indicating that some product i is closer to j than it is to k, tSTE solves

$$\max_{\mathbf{x}} \sum_{(i,j,k)\in\mathcal{T}} ln(\pi_{ijk}) \qquad \text{where} \qquad \pi_{ijk} = \frac{\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\|x_i - x_j\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\|x_i - x_k\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}$$

The result is a continuous vector representation of the differentiation implied by the triplets. Furthermore, if observed data contains desirable information (for example, if a key counterfactual hinges on an observed attribute), columns of the embedding can be fixed at the level of the observed data, allowing the "mixed embedding" to fit the remaining dimensions of the embedding which incorporating the differentiation of the included data.

Previous studies have made use of survey or recommendations data with the explicit goal of construction the *product space* (Magnolfi et al. (2025)), allowing for a more natural interpretation of demand coefficients and for variation in the data to reveal consumer preferences. By contrast, this paper's method is less clear about whether it recovers the product or *preference space*: these triplets may represent differentiation in utility rather than purely the characteristics space, and hence encode consumer-based information on both characteristics *and* preferences. This blurs the assumptions made in logit models where the inputs X solely reflect product characteristics, giving the estimated parameters a sensible economic interpretation. In this application, the model acts more like approaches in machine learning, where inputs of the model are used to best fit substitution patterns demonstrated by the data, without clean interpretations for the model's parameterization.

[5] The selection of the hyperparameter m (the number of dimensions in the embedding) is a matter of researcher choice. Magnolfi et al. (2025) discuss several rules of thumb: a simple approach is to examine whether the variation in the embedding can be reflected in fewer dimensions through principal component analysis. If $m - 1$ components in PCA capture over some threshold of variation, reject m and proceed to testing $m - 1$.

In the context of hotels, the embedding reflecting elements of consumer preferences may be a feature rather than a bug in one regard. A known challenge is that consumer attributes cannot be incorporated by typical methods such as census data, as local households are not the consumer base for hotels, and so recovering information on consumer preferences is useful as this may be all the information available. What is not entirely clear is the how the difference between these two spaces matters for the interpretation of the results. It also constrains the set of counterfactuals: an analysis of product entry would be extremely limited without making substantial assumptions over the location of the product in the latent space.

2.2.1 Path-Connection of Recommendations

The pattern of which products are recommended to one another forms a set of observable links between products. I define recommendation spaces as topological spaces formed by *path-connected* products. Two products j and k are path connected if there exists some pattern of recommendations $j \to \ldots \to k$ and/or $k \to \ldots \to j$. Conceptually, this implies a consumer can search from j to k solely by following recommendations. All products within a recommendation space $\mathcal{S}$ are path-connected in recommendations.[6] Separate recommendation spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are *disconnected* if no hotels in set S_1 contain a recommendation for, or are recommended by, any hotel in set S_2 (i.e. all products in $\mathcal{S}_1$ are *path-disconnected* from all products in $\mathcal{S}_2$).

It is necessary for products to be path-connected in recommendations for the tSTE embedding to present a unique distance between them. Consider hotels (A, B, C, D), where (A, B, C) and D are disconnected. As D is never recommended when searching for any of (A, B, C), all information relating D to (A, B, C) shows that D is the further component of any triplet, and no unique position in the embedding for D exists as all distances further than the distances between any of (A, B, C) fit. Hence, no sensible distance metric—or measure of differentiation—between (A, B, C) and D exists.

The separability of path-disconnected products allows for an assumption about their sub-

[6] I define a matrix of recommendations $\mathcal{R}$, where $\mathcal{R}_{ij} = 1$ denotes that hotel j is recommended when a user searches for hotel i. Then the matrix $S = \mathcal{R} \times \mathcal{R}'$ updates matrix $\mathcal{R}$ to include a single level of connections, and $\mathcal{R} \times S' \to \mathcal{S}$ iterates through levels of connection to a matrix of recommendation spaces, where each value $\mathcal{S}_{ij} > 0$ denotes that hotels i and j are in the same recommendation space and $\mathcal{S}$ is symmetric in the positioning of non-zero elements. Appendix A summarizes this in more detail.

stitution patterns:

Assumption 1. *If products j and k are path-disconnected in recommendations, then demand shocks ξ_j do not affect s_k, and diversion between the two products is zero such that they can be considered separable.*

Given Assumption 1, products which are not path-connected and are thus in separate recommendation spaces can be treated as in separate markets, i.e. a consumer searching for one will never be directed to the other by any chain of recommendations, which implies no *diversion* from one to the other due to exogenous shifts. The practitioner may augment their understanding of the product space with this information: for example, separating hotels by market, or identifying which products are separable in utility. Assumption 1 is sufficient but not necessary: products may be connected in recommendations yet still divided into separate markets based on domain knowledge of the practitioner. In this paper's application, I will show that hotels in the Orange County sample are connected, and hence I will use more traditional geographic market definitions to subdivide the product space.

2.3 Using Embeddings in Demand Models

Two straightforward applications of the constructed latent space are (i) using the computed distance metrics between products in a product-space distance-based demand model, and (ii) using the coordinates vectors for each product as latent characteristics and employing a characteristics-space logit demand model. I describe both below.

2.3.1 Product-space Approaches

A straightforward way to incorporate continuous and observed measures of differentiation is a linear distance-based approach (Pinkse et al. (2002), Pinkse and Slade (2004)), even when the distance between products is an abstraction rather than literal distance. In this case, the dimensionality issue of the typical product-based approach to demand estimation is eased: rather than estimate J^2 cross-price elasticity parameters, substitution is recovered via estimating a function $f(\cdot)$ of observed differentiation between product attributes:

$$\log(q_{jt}) = \alpha_0 + \alpha_1 \log p_{jt} + \sum_{k \neq j} f(d_{jk}; \beta) \log p_{kt} + e_{jt}, \tag{1}$$

for some function f over observed distances d_{jk} between products j and k, estimating cross-price elasticities through a small number of parameters β. The distances between products are, as discussed in Section 2.2, distances in preference space, and hence the parameters β on the distance function are a scaling of distances between utilities rather than a strict preference measure of the sensitivity of substitution to spatial competition.

This model is attractive for its ease of estimation and interpretability. However, it has several limitations. First, the model is not micro-founded and so lacks welfare interpretations: one solution to this is the Almost-Ideal Demand System (AIDS) of Deaton and Muellbauer (1980), where the distance-based approach is applied similarly to discipline the estimation of cross-price elasticities. Second, the model makes extremely strong assumptions on the structural errors in the demand system as discussed in Berry and Haile (2021), with one error per equation. Lastly, a common metric of interest to researchers and practitioners is the diversion ratio between products. In the log-linear model, these values are not naturally bounded by $(0, 1)$ and are biased by the ratio of sales quantities.[7]

2.3.2 Characteristics-space Approaches

The second—and more common—approach is a characteristics-based mixed logit model in the style of Berry et al. (1995). To define hotel "characteristics," I make use of the coordinates of the embedding, expressed in K dimensions. The coordinates of the embedding—as mentioned earlier—provide differentiation in *preference* space rather than strictly product space. I write consumers as making a discrete choice over hotels j in area-month market t:

$$\begin{aligned} u_{ijt} &= \alpha_i p_{jt} + x_{jt}\beta_i + \xi_{jt} + \epsilon_{ijt} \\ (\alpha_i, \beta_i) &= (\alpha, \beta) + \Sigma v_i \end{aligned} \tag{2}$$

Hotel average monthly prices are denoted p_{jt}, and hotel exogenous characteristics are captured in the vector x_{jt}. Consumers have heterogeneous preferences over observable characteristics, reflected by the random coefficients α_i and β_i with variances denoted by the diagonal matrix Σ. ξ_{jt} reflects an unobserved demand shock. The error term ϵ_{ijt} is distributed as extreme value type I. Instruments in this context can be constructed following practices described as "BLP instruments" or differentiation instruments (Gandhi and Houde

[7] Given own-price elasticity α_j and cross-price elasticity $f(d_{jk})$, diversion $\mathcal{D}_{jk} = \frac{f(d_{jk})}{\alpha_j}\frac{q_k}{q_j}$.

(2023)).

There is no natural interpretations of the random coefficients on the parameters relating to these characteristics. For example, it is not clear how a consumer with a strong preference for Marriott hotels exhibits this through the parameters of the model, as preferences and characteristics for Marriott do not correspond to any single dimension. One interpretation is mechanical: as the embedding is constructed to fit Euclidean distances between product utilities, random coefficients allow for flexible scaling of utility on each dimension of the embedding, and hence relaxes the assumption of Euclidean distances between products. For some values of the parameters β_{ik}, the hypothetical Marriott enjoyer's preferences are captured by weighting the distances between Marriott hotels as smaller across the respective dimensions of the embedding.

2.3.3 Incorporation as Second-Choice Data

Separate to this paper's discussion of the continuous representation of the preference space, a natural implementation of default recommendations is to use them as a form of second-choice data combined with traditional methods of demand estimation. This takes a similar place to more traditional survey-based approaches for revealing the explicit second choice of a consumer (Berry et al. (2004)): if the consumer indicated preference for j, their second choice would most likely be one of the closely-recommended alternatives, and so preferences over the characteristics of j and its alternative are correlated. In such a context, few recommendations per product are necessary as their role is to define the top substitute(s) rather than identify the local choice set or overall product space. While I consider this the most straightforward way of incorporating recommendations, it is not the focus of this paper: I focus on the use of embeddings—discussed in the following section—to create vector representations of the utility space.

As an example of one possible method for how to apply these data in the context of the estimation of aggregated demand systems, recommendations can be used to construct additional moments to discipline the estimates of the demand system. This is similar to the approach of Conlon et al. (2023), who choose parameters of the model to match observed first- and second-choice probabilities, minimizing the least squares error to the estimated first- and second-choice probabilities. In my case, I have sets of recommended alternatives but no observed choice probabilities, and so instead one can choose parameters of the model

such that the estimates of substitution patterns select one of the recommendations as the top alternative to each product. For each product j in the product set$\mathcal{J}$ with a set of platform-recommended alternatives R_j, and with estimated mean product-level diversion ratios $D(\theta)j$ at the parameter draw θ, define the moment g as:

$$g(\theta)_j = \lambda \mathbb{1}\{k \notin \mathcal{R}_j\} \text{ where } \mathcal{D}(\theta)_{jk} = \max \mathcal{D}(\theta)_j,$$

where λ is a penalty function that increases with how far product k is from the top recommended alternatives $r \in \mathcal{R}_j$.

3 Information from Recommendations

The most critical question for this method is whether the observed recommendations are actually conveying the expected information, as the platform's objective function cannot be identified solely from default recommendations. To address this concern, in this section I first present a sketch of platform behavior and assumptions under which the obtained information is what we might expect. Second, I assess the recovered recommendations and produced embeddings to consider whether they match observed measures of similarity between hotels.

3.1 A Model of Platform Behavior

Consider the following environment. There are $\mathcal{J}$ differentiated products in the market indexed by $j = 1, \ldots, J$. Consumers have rational preferences over these goods, but are not fully aware of the choice set, and so make use of a platform to identify choices.

Assumption 2. *Consumer behavior can be reflected by a model of discrete choice. Consumer utility for individual i for good j is $U_{ij} = \beta_i x_j + \alpha_i p_j + \epsilon_{ij}$, giving logit choice probabilities $s_{ij} = \frac{e^{\beta_i x_j + \alpha_i p_j}}{1 + \sum_{k \in \mathcal{J}} e^{\beta_i x_k + \alpha_i p_k}}$.*

Assumption 2 is not particularly strong, and simply allows the rest of this section to focus on the case of logit demand.

Next, I make assumptions regarding the knowledge and behavior of the platform. Because the platform has extensive observed data on consumers' search paths and selections, it is

aware of what consumers choose and what they ultimately choose as alternatives when rejecting an option.

Assumption 3. *The platform is aware of all products $\mathcal{J}$, consumers' choice probabilities in aggregate $s_j = \int s_{ij}\, dG(i)$, and substitution probabilities from any product j to any other product $k \neq j$ for the mean consumer $D_{j \to k} = \int \frac{s_{ik}}{1 - s_{ij}} \cdot \frac{s_{ij}}{s_j}\, dG(i)$.*

Assumption 3 implies that the platform is aware of the choice probabilities s_j for all consumers, and the conditional choice probabilities $s_{k|j}$ for consumers who initially chose j and selected k instead, and so conditions on a stated preference for j. This suggests that the platform can answer the question "if you clicked on j and did not purchase it, what are you most likely to purchase instead?" What the assumption does not require is that the platform observe or record any information about individual consumers or their preferences: in this paper I focus solely on aggregate preferences so as to abstract away from learning about the consumer and tailoring recommendations. This distinguishes the approach from studies that use search paths to study consumer preferences (Armona et al. (2024), Kaye (2024)).

When a user selects an option j on the platform, the revenue for the platform can be written as the revenue for their choice weighted by the respective choice probabilities:

$$\Pi_j = t_j s_j + \sum_{k \in \mathcal{J}} t_k s_{k|j} \text{ for } k \neq j \tag{3}$$

where t is a vector of revenues earned when a booking is made for a product. However, consumers' knowledge of the product space is limited, and so the platform presents ν alternative recommendations to j to steer them towards making a purchase. Given Assumptions 2 and 3, the platform's revenue-maximizing behavior for serving default recommendations for product j can be written as:

$$\Pi_j = t_j s_j + g(s_{k|j},\, t,\, \nu) \tag{4}$$

where g is the revenue expected from substitution to the recommended alternatives and hence $R_j = \arg\max_{k \in \mathcal{J}} g(s_{k|j},\, t,\, \nu)$ is the set of default recommended alternatives for product j which the platform chooses to serve the mean consumer. The set R generalizes a number of platform strategies given consumer behavior, which are often the result of

an unknown recommendation algorithm.

Assumption 4. *If the vector of revenues t does not change the order of expected revenue versus expected choice, i.e. $t_k s_{k|j} > t_\ell s_{\ell|j} \Leftrightarrow s_{k|j} > s_{\ell|j}$, then R_j is the set which maximizes second-choice likelihood from j: $R_j = \arg\max_{k \in \mathcal{J}} \sum_k^\nu s_{k|j}$.*

Under Assumption 4, the presented recommendations are a selection of conditional second choices for the mean consumer who selected the reference product. Products which are recommended are hence "more similar" in terms of utility for said consumer than those which are not: a stronger assumption is that the ordering of recommendations additionally indicates rank-ordering of conditional second choice probabilities.

Next, I discuss the implications of various assumptions on the interpretation of R.

Example 1. *Products are recommended by characteristic similarity instead of preference similarity. Given (some basic assumptions on logit, preferences are linear on X), $||X_j, X_k|| < ||X_j, X_\ell|| \Leftrightarrow ||u_j, u_{k|j}|| < ||u_j, u_{\ell|j}||$. Hence, Assumption 4 is satisfied.*

This may be untrue if unobserved consumer preferences are particularly idiosyncratic and closeness in characteristics does not correlated with closeness in utility. In these cases, recommendation by conditional choice is more suitable, but recommendations by similarity are no worse than the typical approach of having the characteristics themselves.

Example 2. *Consumers care about ordered presentation: If consumers inspect recommended alternatives in order and probabilities are independent of one another's inclusion, then the options within R_j are a rank-order list by descending conditional choice probability $s_{k|j}$, which is consistent with Assumption 4.*

This fits sequential strategies such as search in order of expected utility (Abaluck, Compiani, and Zhang (2024)), which in turn is robust to strategies of consumer search (also cases such as satisficing, directed cognition, and full information). This is the easiest interpretation but a stronger assumption, as position in preference space is closer to the reference for the mean consumer of j based on recommendation rank.

Example 3. *Consumers don't observe order: If consumers inspect the recommendation set as a whole and decide on either j or an option within R_j without any direction, and if the conditional choice probabilities are independent of one another's inclusion, then Assumption 4 holds.*

Consistent with solving simultaneous search for the mean consumer, and a weaker assumption than ordering. In this case, it can only be assumed that recommended products are closer than those not recommended, but not that any given recommendation is a better option than another.

Lastly, it is important to discuss the cases under which the above assumptions can fail. The assumptions of ordered conditional utility potentially overlook how the objectives of the platform may bias the results that they provide to their consumers. Many platforms are upfront with the fact that their recommendations are not unbiased in optimizing a revenue-maximization process: they may prioritize certain hotels for which they receive higher revenue due to price, contract terms, or other factors. Kaye (2024) discusses in more detail the underlying trade off of match quality versus price competition using clickthrough data, while Hodgson and Lewis (2024) explores the conditions under which a platform may prefer to recommend similar products (consumer finds the best local alternative) versus using recommendations to steer towards product discovery (consumer gets a wider picture of the product space). For example, if purchases of a certain product gave proportionately higher benefit to the platform, and the platform aimed to steer consumers towards this product as a result, it should be placed consistently closer to rivals than it otherwise would be. Outside of hotels, Christensen and Timmins (2022) provides one such example where recommendations for real estate are systematically biased to steer minorities towards less-desirable neighborhoods.

Additional concerns about the unobservability of the platform's behavior arise from how users interact with the platform. The platform may—rather than assuming users simply browse linearly—attempt to tailor the menu of displayed options to induce a selection by showing less desirable or otherwise-extreme options. The platform may also be in a non-equilibrium state of continually learning from consumers' choices, who in turn make choices based on the platform and subsequently feed back into the platform's data. Further work on the differences in portrayed default recommendations across platforms can help inform researchers on what to take away from the displayed options.

3.2 Sample Recommendations

In order to better validate the idea that the default recommendations are presenting—on average—products that are most similar to the reference product in the eyes of the consumer,

I summarize some descriptive facts about the recommendations relative to their reference products that suggest that the recommendations are capturing aspects of observable product differentiation.

The first observation is that recommended properties are (i) located closer than the mean non-recommended property, suggesting that similar products are being recommended, and (ii) closer to the reference product the earlier they are presented as recommendations, indicating rank-ordering. Figure 2 displays the similarity of the recommendations to the reference product on several dimensions. Panel A presents the average distance in miles to the first six recommended properties, as well as to non-ranked properties which are of the same class and located in the same market. The mean distance to recommended hotels increases monotonically through ranks 1-6, and recommendations are on average closer than unrecommended properties.[8] We can reject that the mean distance to the outside group is equal to the mean distance of the inside group with $t = -23.9$.

Recommendations are also more similar to their reference products in terms of quality than the mean, which captures similarity in both the consumer experience and in terms of average price.[9] Panel B presents the average proportion of recommendations which are of the same quality tier or within one quality tier of the reference, compared to the average of all other non-recommended properties within 3 miles.[10] Approximately 40% of recommendations are of the same quality and 80% are within one quality tier for recommendations, while the average for properties within 3 miles is 20.7% and 55.3% respectively. We can reject the hypothesis that the means of the outside set and recommendation set are equal with t-values of 15.2 (same quality) and 18.7 (adjacent quality). Appendix Figure 2 shows the proportion of recommendations by class of reference and recommendation: Luxury and Upper Upscale hotels focus closely on similar quality options, while lower quality hotels see more variation.

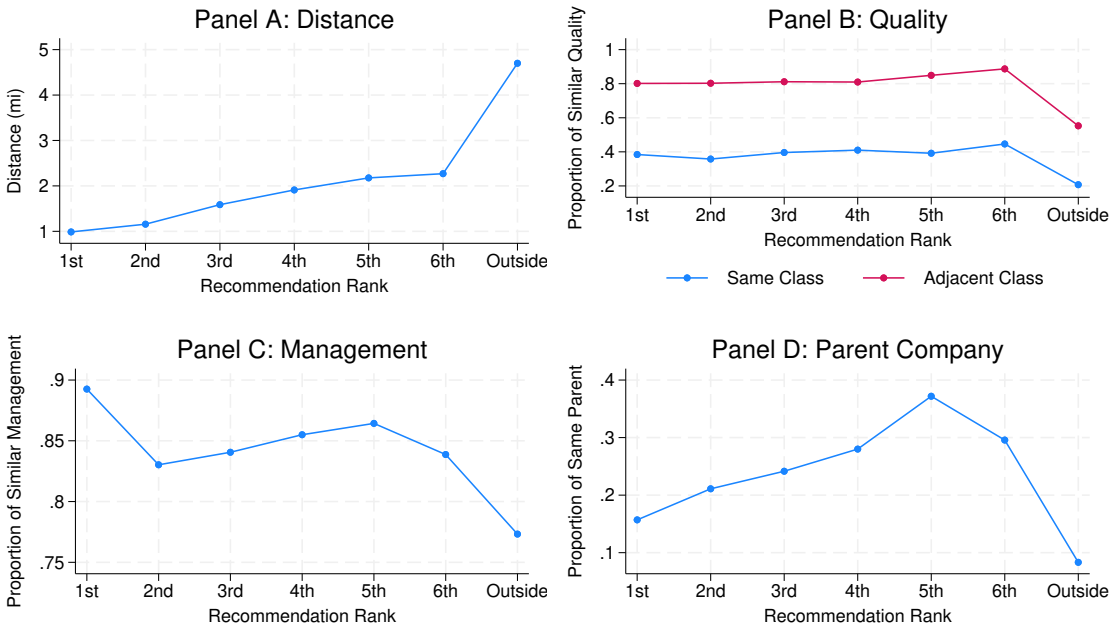
Another dimension of consumer preference for lodging is choosing branded establishments versus independent hotels. Panel C compares whether recommendations share the same management structure as the reference hotel: specifically, whether they are independent

[8] Mean distances for ranks 1-6 are 0.97, 1.16, 1.59, 1.91, 2.18, and 2.27 miles. Mean distance to the within-market average hotel of the same class is 4.56 miles.

[9] STR groups hotel chains by Chain Scales primarily based on their average room rates within the market. Independent hotels are assigned a class (equivalent to chain scale) based on where their average room rates would place them.

[10] For example, the first value considers the proportion of recommendations for an Upscale hotel which are of the Upscale class, while the second considers the proportion which are Upper Upscale, Upscale, or Upper Midscale.

FIGURE 2: Similarity of Ranked Recommendations



or branded. While not as stark as the previous comparisons, recommended properties are consistently more likely to share the same management structure: i.e. a consumer is more likely to be recommended another independent hotel if they searched for an independent hotel, relative to the average proportion. These rates are as similar as 89.3% for the top recommendation, falling to 83.9% for the 6th recommendation, and are 77.3% for the outside group. We can reject the mean of the outside group—again the set of all non-recommended properties within 3 miles—being equal to that of the recommendation set with $t = 6.59$.

Finally, panel D performs the same analysis as above but for whether the recommendation is a hotel under the same parent company. Hotel parent companies (Marriott, Hyatt, etc. with independent hotels treated as unique entities) provide non-compete agreements to their franchises, making it unlikely to that nearby hotels are licensed with the same chain in the short run, but they also operate a range of chains which allow for greater market penetration and product differentiation. Furthermore, consistent placement of co-branded properties highly in the recommendation set may be evidence of systematic bias in terms of booking revenue, though this cannot be easily disentangled from genuine consumer brand preferences. The results show that recommended properties are much more likely to be of the same parent than the outside set ($t = 19.2$ that means differ). However, the

top recommendations are the least likely to share the brand, suggesting that location and quality preferences dominate brand preferences for ranking recommendations and casting some doubt that the top recommendations are biased by platform-brand incentives.

An alternative source of preference variation is across the types of consumers who prefer different classes of hotel: for example, we might intuit that luxury customers are more quality-sensitive and hence quality similarity is a bigger driver of recommendations. I estimate a probit model of whether hotel k is in hotel j's recommendation set $\mathcal{R}_j$ based on its distance to the reference hotel on measures of differentiation d:

$$\mathbb{1}\{k \in \mathcal{R}_j\} = \Phi(d(x_j, x_k)) \text{ where } x = \{\text{location, quality, management, parent}\} \quad (5)$$

Table 2 presents the results of estimating Equation 5 by the class of the reference hotel, including only hotels in the reference hotel's market to avoid biasing the distance results with hotels far across Orange County. In all cases, closer hotels are more likely to be recommended, and recommendations prioritize hotels of the same class more than those of an adjacent class, which in turn is more likely than a hotel with very different quality. The one exception is Economy hotels, which see a larger effect from a hotel being Midscale. A possible interpretation is that as Economy hotel recommendations are most sensitive to distance, guests are most interested in location and willing to adjust quality so long as it remains cheap. By contrast, Luxury hotel recommendations are least sensitive to distance and most sensitive to similarity in quality: an intuitive conclusion given the prospective tastes of luxury hotel guests who prioritize the quality of their stay. There are no statistically-significant effects of hotels being independent or non-independent on inclusion in the recommendation set when also including whether the hotel is operated by the same parent company, however, outside of Luxury hotels, being branded under the same parent company is a substantial positive factor for being recommended.

3.3 Constructed Embeddings

I begin by presenting visualizations of a two-dimensional embedding based on the recommendations alongside the true geographic locations of hotels in order to visually assess measures of similarity: market definitions and quality. Here, we can assess whether the statistical structure of the tSTE algorithm preserves the similarities between properties that

TABLE 2: Impact of Factors on Recommendation Inclusion by Reference Hotel Class

	(1) Luxury	(2) Upper Upscale	(3) Upscale	(4) Upper Midscale	(5) Midscale	(6) Economy
Dist to Hotel (mi)	−0.130***	−0.222***	−0.248***	−0.227***	−0.244***	−0.333***
	(0.027)	(0.022)	(0.062)	(0.041)	(0.032)	(0.028)
Same Class	1.936***	0.731***	0.641***	0.677***	0.607***	0.722***
	(0.095)	(0.095)	(0.170)	(0.165)	(0.156)	(0.184)
Adjacent Class	1.348***	0.630***	0.365***	0.441***	0.389*	0.817***
	(0.229)	(0.064)	(0.123)	(0.132)	(0.229)	(0.154)
Same Mgmt	0.067	0.129	0.296*	0.101	0.258	0.102
	(0.255)	(0.108)	(0.161)	(0.139)	(0.160)	(0.122)
Same Parent	0.774	0.581***	0.514***	0.708***	0.779***	0.644***
	(0.489)	(0.115)	(0.128)	(0.073)	(0.087)	(0.115)
Constant	−2.165***	−1.784***	−1.714***	−1.578***	−1.514***	−1.369***
	(0.299)	(0.174)	(0.268)	(0.433)	(0.311)	(0.341)
Observations	1,034	4,509	5,557	4,505	3,228	3,531

Note: Adjacent class does not nest the same class. Management refers to branded versus independent, while parent refers to the hotel chain's parent company. Standard errors (presented in parenthesis) are clustered by geographic market to account for underlying market-level consumer patterns which act as treatment effects. *** p<0.01, ** p<0.05, * p<0.1

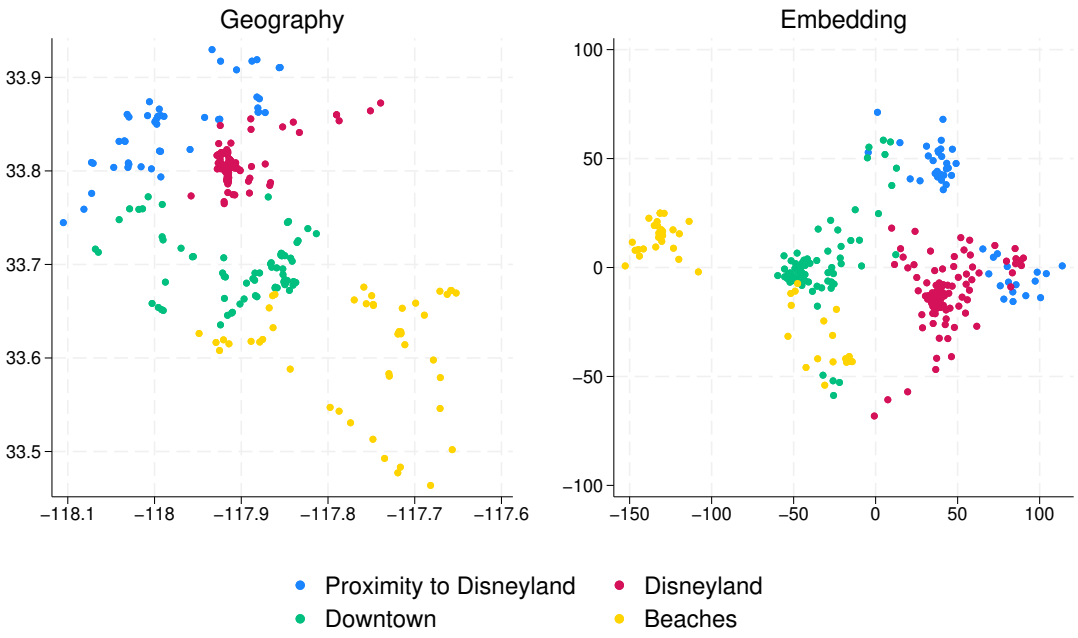
were observed in the recommendation sets.

Figure 3 presents a plot of the sample hotels by latitude and longitude (left) versus a two-dimensional embedding (right), highlighting the market categorization of the hotels. The embedding succeeds in capturing geographic dispersion, clustering hotels by geographic markets. Several interesting features emerge: there is some measure of overlap between each market definition, evidenced by the overlap in placement of hotels at the fringes of each market. Some beach hotels are similar to downtown hotels, while others form an isolated cluster which is more differentiated. There is also considerable overlap in hotels at Disneyland and those in the proximity of Disneyland. However, the market lacks any "central" product which is equidistant in consumers' tastes to all other hotels, and hence embedding forms a ring. The correlation between within-market hotel-rival distances in geographic space and the distances in the embedding is 0.727, suggesting that the embedding is fitting some within-market variation that is not reflected purely by geographic distances.[11]

Figure 4 displays the same graph, instead grouping hotels by their quality (class). Compared to the graph by geography, the embedding shows greater clustering by quality. An implication of this is that the embedding is attempting to reflect product differentiation

[11] Geographic distances are computed via Stata's vincenty command.

FIGURE 3: Two-Dimensional Representation of Hotel Markets



that arises from hotel quality (and hence, product attributes). Luxury and Upper Upscale hotels are located in close proximity to each other, including across geographic markets which border each other.[12] Economy and Midscale hotels are positioned at the fringes of each market are exhibit the most dispersion. The correlation between hotel-rival distances in geographic space and the distances in the embedding for pairs of the same class 0.695 when limiting to pairs in the same market. As before, this is likely due to the embedding fitting more information: hotels are more clustered by class where this was not present in the geography.[13]

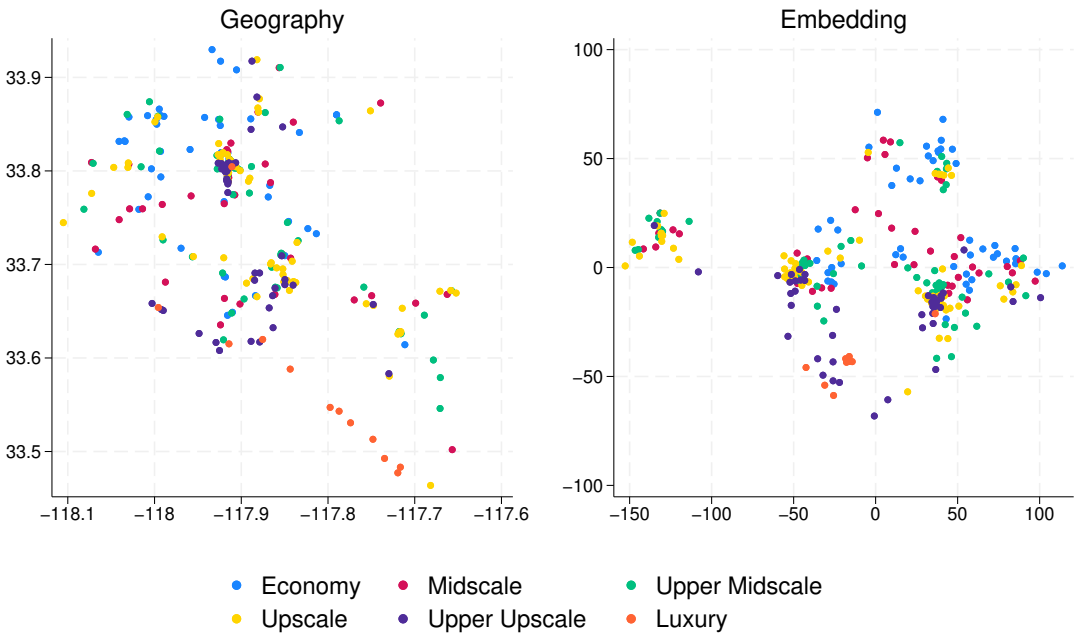
Some observable patterns of Figures 3 and 4 may be due to the limitations of fitting the triplets in only two dimensions. In the empirical application in Section 5, I use a three-dimensional embedding, which cannot be reduced to $m-1$ principal components without falling below 90% cumulative variation.[14]

[12] While this paper imposes geographic market definitions regardless, the ability of the embedding to consider within-group and cross-group differentiation continuously is a valuable aspect when products would otherwise need to be assigned discrete nests.

[13] Correlations between geographic and embedding distances when not limiting to the same market are 0.884 for all hotels and 0.913 for hotels of the same class. However, the higher correlation is likely driven by long distances between hotels in different markets in both the geography and the embedding.

[14] A 4-D embedding of the sample hotels can be reduced to 3 principal components and retain 92.9% of

FIGURE 4: Two-Dimensional Representation of Hotel Classes



4 Simulation Performance

In this section, I demonstrate through Monte Carlo tests using the embedding for demand estimation and compare results to approaches using observed characteristics. In the first example, preferences are straightforward but characteristics are only partially observed. In the second example, characteristics are observed but consumers' preferences are highly heterogeneous in an unobserved fashion. In each case, variation in the data does not perfectly identify the estimated models and so the test's goal is to that the RMSE of key post-estimation statistics (diversion ratios, markups, and out-of-sample fit) decreases when incorporating recommendation data versus the case where no product-space data are available, or when using product-space data when variation in the data poorly identifies the parameters of the model.

Recommendations in this context are simulated by taking product-level rankings of closeness to substitutes. As a proxy for conditional choice probabilities from simulated consumers, for each product j I rank products $k \neq j$ in descending order of their true diversion ratios

variation, while a 5-D embedding can be reduced to 4 principal components and retain 95.2% of variation. Hence, a 90% threshold selects $m = 3$, while a 95% threshold would select $m = 4$.

$\mathcal{D}$, aggregated to the level of $\mathcal{D}_{jk}$.[15]

$$\underbrace{\sum_i \pi_i \frac{s_{ik}}{1-s_{ij}} \cdot \frac{s_{ij}}{s_j}}_{\text{Mixed logit choices}} \approx \underbrace{\frac{s_k}{1-s_j}}_{\text{Diversion ratios}}$$

This captures a similar concept of what the closest-preferred alternative to product j is in the data: which is the most likely alternative chosen if j was no longer selected. The assumed behavior of the platform is therefore to recommend products in order of these rankings. The econometrician, however, does not observe these true diversion rankings—or the exogenous characteristics—and only sees prices, quantities, a product-market-level cost shifter, and the recommendation rankings.

4.1 Random Preferences

I first consider an environment where a large number of products are highly differentiated, with utility modeled using common assumptions of normally-distributed consumer preferences. This environment allows me to explore the performance of the proposed method when true characteristics are effective for estimating the demand system and the variation in the data is well-understood. I construct embeddings of $m = \{2, \ldots, 12\}$ dimensions using the ordinal rankings of products, incorporating recommendations of the top 5, 10, 25, or 50 products (indexed by R), as well as approaching the problem without any recommendation data as a baseline.[16] To select K, I apply a rule of thumb from Magnolfi et al. (2025), assessing whether the $m-1$ principal components of the m-dimension embedding capture at least some threshold of the variation, and rejecting m if so.[17] I find that a threshold of 75% would reject $m = 3$, 90% would reject $m = 7$, and 95% would fail to reject $m = 8$. I proceed with the 90% threshold and thus each level of R selects $m = 6$.

Data are simulated from a mixed-logit data-generating process with $J = 100$, $T = 1000$, and $F = 10$, with utility taking a BLP framework as in Equation 2 and firms competing via Bertrand-Nash. Products $j \in \mathcal{J}$ have a constant, a price, and six exogenous characteristics

[15] I take the quantity-weighted average of product-market-level diversion $\mathcal{D}_{jkt}$ to form product-level diversion ratios $\mathcal{D}_{jk}$.

[16] All embeddings use the tSTE algorithm with a convergence threshold of $1e-7$.

[17] Panel A of Appendix Figure 3 displays the values across $m = \{2, \ldots, 12\}$.

generated $N(0,1)$ i.i.d. Each of these eight characteristics has both a linear and nonlinear coefficient in simulation. The nonlinear coefficient matrix Σ has no non-zero off-diagonal values. The integral over consumers' preference draws v_i is simulated using 1000 Halton draws. Full details of the DGP are included in Appendix B. The outlined specification results in a mean inside share of 0.67, with $[5, 95]$ percentile bounds on prices and shares at [6.28, 10.28] and [0.002, 0.017].

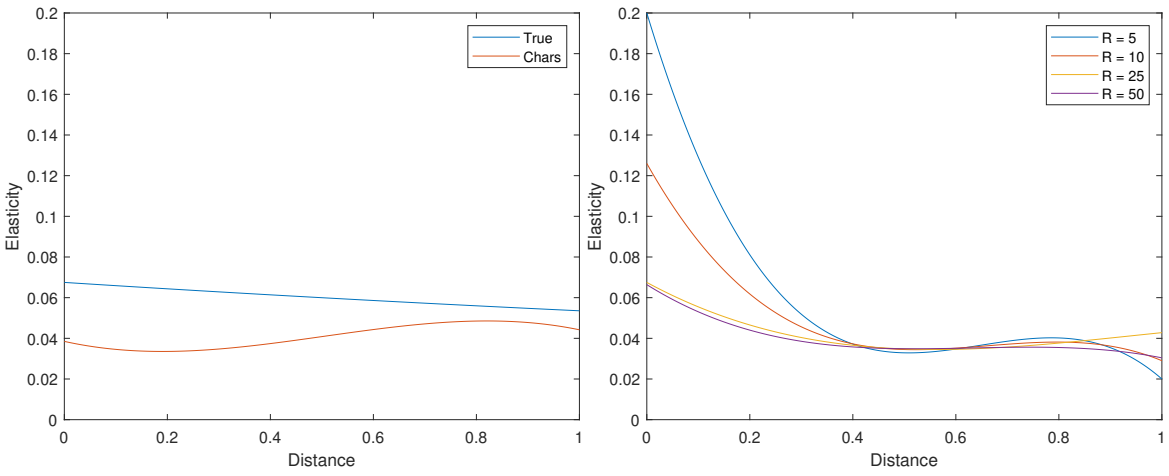
4.1.1 Impact of the Number of Recommendations

A first question is to what degree having more or fewer recommendations matters for the results. Intuitively, more information is helpful. However, even a limited set of recommendations can help identify the local choice set for consumers of those products. This is analogous to the discussion of how models of discrete choice can be estimated using just a subset of the choice model (McFadden (1978), Fox (2007)). I examine the relationship between estimated elasticities and distances between products using a distance-based log demand setup (Pinkse et al. (2002)). While this specification cannot reproduce the discrete choice data-generating process (Jaffe and Weyl (2010)) and imposes strict restrictions on the structural errors of the demand system (Berry and Haile (2021)), it is simple to compute and demonstrative of the relationship between distances and substitution patterns. I write the demand system as in Equation 1.

Figure 5 plots $f(d_{jk})$ for each value of R and when using the observed characteristics, versus estimating $f(d_{jk})$ from the true cross-price elasticities. The observed characteristics result in a non-monotonic function, suggesting that they are not well-suited to capturing substitution via the (inherently misspecified) log-log model. By contrast, the estimated $f(d_{jk})$ with values of $R = \{25, 50\}$ produce closer patterns to the true relationship. The lower values of R, which use a smaller set of close substitutes, result in overestimating the elasticities of the closest substitutes.

A more common application of the embedding coordinates is using them as inputs for a characteristics-based approach. A second question is thus how the embedding performs across different numbers of recommendations, and when compared to specifications incorporating the true characteristics. Using the coordinates of these embeddings $(\tilde{x}^r_{j1}, \ldots, \tilde{x}^r_{j6})$ as exogenous characteristics, I estimate the mixed-logit demand system. No fixed effects are included as these would be collinear with instruments given the invariant choice sets -

FIGURE 5: Estimated Cross-Price Elasticity Function



in practice, product-level fixed effects are sensible. As instruments I include the cost shifter w_{jt}, as well as differentiation IVs based on the nonlinear characteristics.

Table 3 lists the error in estimated diversion to inside and outside options across values of $R = \{0, 5, 10, 25, 50\}$, with the estimates using the true characteristics as a comparison. As expected, having the true characteristics provides the best fit - however, in practice the "true" characteristics are at least partially unknown.[18] Furthermore, this method is applicable to cases where characteristics are unavailable or unquantifiable in a useful way, such as with highly varied or stylistic consumer products. Thus, the relative close fit of substitution patterns is a useful indicator. As R increases, the RMSE of both the outside and inside estimated diversion falls, most noticeably for diversion to the outside option.

TABLE 3: Estimated Results by Number of Recommendations

		Recommendations				
	True x_j	0	5	10	25	50
Inside RMSE	0.000	0.021	0.011	0.005	0.005	0.005
Outside RMSE	0.003	0.072	0.076	0.030	0.007	0.003
Markups RMSE	0.009	0.888	0.679	0.019	0.011	0.011

Estimates utilize 2-step GMM, followed by iterating the 2-step problem using the approximation to the optimal instruments (Reynaert and Verboven (2014)). All specifications include linear coefficients on the constant, price, and embedding coordinates $\tilde{x}$. In the $R = 0$ case, there are no $\tilde{x}$. In the x_j case, the true characteristics are used. The diversion statistics are medians of the product-level $\mathcal{D}_{jk}$.

[18] Consider the challenge of capturing in finite dimensions all aspects of product differentiation in a fashion item.

4.1.2 Comparison to Partial Characteristics

Table 4 compares estimated outcomes when looking at four cases: when the researcher observes product characteristics, recommendations ($R = 25$), both, or neither. As including all six true characteristics would closely recover the exact DGP, I assume the researcher only observes partial true characteristics, and does not observe x_4, x_5, x_6, a plausible scenario where aspects of utility are difficult to observe in data. The key comparison is columns (2) and (3): relative to having some measure of true data, incorporating recommendations does slightly worse in estimating the median diversion to inside products, but notably better in estimating diversion to the outside option and markups. Incorporating a mixed embedding of characteristics and recommendations further improves estimates relative to only having recommendations. The results suggest that researchers should—unsurprisingly—use as much correctly-specified data as possible, but adding data from recommendations can improve the scaling of estimated utility such that outside diversion is better estimated, with implications for markups and counterfactuals.

TABLE 4: Comparative Performance of Data Sources

		RMSE			
	TRUE	(1)	(2)	(3)	(4)
Inside Diversion	0.008	0.021	0.003	0.005	0.005
Outside Diversion	0.107	0.072	0.012	0.007	0.004
Markups	0.278	0.888	0.021	0.011	0.011
Partial True Characteristics			X		X
Recommendations				X	X

Partial true characteristics are x_1, x_2, and x_3. All specifications include all X in the linear specification with no fixed effects. Columns (1) and (3) are equivalent to the specifications shown for $R = 0$, 25 in Table 3.

4.2 Unobserved Consumer Heterogeneity

A second—and more relevant—environment is one where unobserved consumer heterogeneity impedes the identification of the demand model even when data on the product space is available. Prior work such as Nevo (2001) and Backus, Conlon, and Sinkinson (2021) often makes use of consumer demographics to aid in identifying substitution patterns; environments such as the hotel sector have no data on consumers, creating challenges for mixed logit demand estimation. I hence create an extreme example where the parameters on the

true characteristics are poorly identified, and hence they are of limited use in estimating substitution.

I simulate a DGP with $J = 100, T = 1000$, and $F = 10$, where each market includes a random set of 50 products and firms compete via Bertrand-Nash. Consumers vary in terms of product-specific demographics (bliss points):

$$u_{ijt} = x_{jt}\beta + \alpha p_{jt} + \lambda d_{ijt} + \xi_{jt} + \epsilon_{ijt} \quad \text{where } d_{ijt} = \left(\sum_{kt} (B_{ikt} - x_{jkt}^2)^2 \right)^{0.5} \tag{6}$$

given $\alpha,\ \lambda < 0$ and $\epsilon \sim EVT1$. Bliss points are drawn from a multivariate Gamma distribution $(B_{i1}, B_{i2}, B_{i3}) \sim \Gamma(2, 0.5)$. As consumers weigh the distance to the square of (normally-distributed) product attributes, products which are far apart in the product space may be very close in the preference space for given consumers. Appendix B includes the full details of the DGP.

As the researcher does not observe the consumer demographics (i.e. the bliss point values), utility is modeled as the typical random-coefficients logit equation:

$$u_{ijt} = x_{jt}\beta_i + \alpha_i p_{jt} + \xi_{jt} + \epsilon_{ijt},$$

using typical assumptions that the random coefficients are normally distributed. In this case, I include all of the observed X_{jt} in the model which uses observed characteristics, and compare to a model using the coordinates of an 8-dimension embedding using the top 25 recommendations.[19] All specifications include product-level fixed effects. Instruments are the same: the cost shifter and differentiation IVs.

Table 6 displays the results for three cases: with recommendations, with true characteristics, or with neither. In this case, the results in Column (2) reflect that the variation in the data, owing to large unobserved consumer heterogeneity, poorly identifies the demand system and leads to substantial errors in estimated results. Column (3) shows that using an embedding based on recommendations in place of having *all* of the true characteristics produces lower RMSE on all four examined metrics: diversion to products and to the outside option,

[19] Panel B of Appendix Figure 3 shows the cumulative variation of principal components: I select $m = 8$ to fit a 95% threshold.

markups, and predicted out-of-sample market shares.[20] When both sets of information are available, incorporating both in a mixed embedding further reduces errors.[21]

Table 5: Comparative Performance of Data Sources: Case 2

		RMSE			
	TRUE	(1)	(2)	(3)	(4)
Inside Diversion	0.009	0.007	0.006	0.005	0.005
Outside Diversion	0.498	0.260	0.148	0.146	0.135
Markups	0.183	0.031	0.027	0.027	0.025
Shares Out-of-Sample	0.007	0.002	0.002	0.001	0.001
True Characteristics			X		X
Recommendations				X	X

All specifications include product-level fixed effects in the linear specification, with random coefficients on all non-linear X terms and prices. Diversion and markups are compared at the product level. Out-of-sample shares are compared at the product-market level.

The estimates of the demand system are most relevant in the context of the question they are used to answer: I construct a merger simulation in the data and compare profit and welfare change predictions in each specification. When comparing the RMSE of estimated percentage changes in profits and welfare, the recommendations specification outperforms the true characteristics—and are further outperformed by a mixture of both sources of data—but by an economically insignificant degree. Regardless, this demonstrates that recommendations can substitute for characteristics in such a context when the latter are not available, both in terms of predicting substitution patterns and their implications for relevant counterfactual analysis.

Table 6: Simulated Merger Results

	TRUE	(1)	(2)	(3)	(4)
Change in Profits	0.279%	0.599%	0.270%	0.313%	0.273%
RMSE (pct pt)		0.444	0.148	0.147	0.141
Change in Consumer Surplus	−0.416%	−0.259%	−0.452%	−0.442%	−0.444%
RMSE (pct pt)		0.613	0.081	0.072	0.069
True Characteristics			X		X
Recommendations				X	X

Percentage change displayed is the mean of product-market-level profits and market-level consumer surplus. Merger simulation is a small $10 \to 9$ merger across all simulated markets.

[20] The holdout sample consists of 10 markets with all 100 products, and new draws of $N = 1000$ consumers per market. When applying sample estimates to the holdout sample, I assume $\xi = 0$.

[21] The mixed embedding uses the 3 observable characteristics, and 6 dimensions freely chosen by the tSTE algorithm.

5 Empirical Application

To demonstrate an application of the method in data, I estimate a simple demand system for hotels using four specifications: using characteristics, recommendations via an embedding, using both, and using neither (plain logit).

5.1 Model and Estimation

I model consumers as in Equation 2, where the linear characteristics X_j are absorbed into hotel-level fixed effects δ_j as they do not vary over time, and where seasonality is captured through market-year-month fixed effects. The mean price parameter is calibrated as in Armona et al. (2024): the lack of hotel-level cost shifters makes identification of this parameter challenging. I use a value of $\bar{\alpha} = -0.036$, loosely targeting their average mean own-price elasticity of -2.3 estimated by the specifications which include random coefficients. Markets are defined as the four geographic areas $\times$ year-months for the 2017-2023 period. I define each market's size as a constant equal to 2 times the highest total room sales in that market across all months.[22] An implication of the patterns displayed in the embedding (Figure 3) is that these market definitions may be more exclusionary than anticipated, as there is substantial overlap between properties on the "fringes" of each market definition in the embedding. To keep the models easily comparable, I keep the same market definition in each specification.

While the mean price parameter is calibrated, I estimate hotel and market-year-month fixed effects as well as nonlinear coefficients on price, and on additional exogenous characteristics. The nonlinear characteristics include longitude, latitude, and an indicator for whether the hotel is of upscale class or better (characteristics specification), the coordinates of a 3-dimensional embedding (embedding specification), or a mixed embedding including longitude, latitude, and three free dimensions (mixed).[23]

I construct quadratic differentiation instruments (Gandhi and Houde (2023)) over the m

[22] Taking a similar approach, Armona et al. (2024) use a multiplier of 1.5, while Farronato and Fradkin (2022) use 2.

[23] I use three dimensions for the embedding following a threshold of $m-1$ dimensions containing less than 90% of the variation: $m = 4$ fails this screen.

nonlinear terms l $(x_1, \ldots, x_m)$, where $d_{jktl} = x_{l,jt} - x_{l,kt}$:[24]

$$z_{jt} = \left[\sum_k d_{jktl} \times d_{jktl'} \right] \quad \forall \; l' \geq l$$

Variation in these instruments is primarily driven by entry and exit of rivals. I also incorporate a measure of the exogenous variation in price $\hat{p}_{jt} = E[p_{jt}|x_{jt}, z_{jt}]$, using hotel and market-year-month fixed effects and interacting the instruments z_{jt} with market dummies, and extend z_{jt} to include interactions with the differences $d_{jk,\hat{p}} = \hat{p}_{jt} - \hat{p}_{kt}$:

$$z_{jt}^{\text{full}} = \left[\sum_k d_{jk,\hat{p}}^2, \;\; \sum_k d_{jk,\hat{p}} \times d_{jktl}, \;\; \sum_k d_{jktl} \times d_{jktl'} \right] \quad \forall \; l' \geq l$$

The column vectors of the instruments z_{jt}^{full} are subsequently normalized to mean zero, standard-deviation 1. Following the typical 2-step generalized method of moments procedure, I take the approximation to the optimal instruments (Reynaert and Verboven (2014)) and solve the updated problem. Estimation makes use of `pyBLP` (Conlon and Gortmaker (2020)), using 1000 Halton draws to simulate the normal distribution of consumer preferences.

In this context, the estimation—and hence comparison across models—of substitution patterns is challenging owing to the lack of variation in the product space.[25] Hotel entry and exit is infrequent and hence I do not observe the same hotel places against a varying set of rivals as is more common in contexts such as retail scanner data. Additionally, I lack spatial data on local demand shocks that would allow for capturing drivers of demand for a given hotel beyond hotel and year-month level fixed effects.

5.2 Results

Table 7 displays the coefficient results of the logit demand system. The nonlinear parameters on the heterogeneous preferences can be interpreted as the degree of dispersion in consumer

[24] There are 3 nonlinear characteristics in the characteristic and embedding specifications, and 5 in the mixed specification.

[25] Consumer-level variation in the observed choice sets, a feature driven by capacity constraints, is unfortunately unobserved and aggregated away by monthly data. Agarwal and Somaini (2022) discuss a method by which this variation can be modeled and identified if the appropriate data exist.

sensitivities to the dimensions of the product space: rather than a literal heterogeneity in preference for longitude, variation suggests that some consumers are more sensitive to geographic distances on that axis. Standard errors are unsurprisingly large as the lack of variation in choice sets—a natural feature of hotel markets, as products are locked to specific geographies and entry is infrequent—limits the identifying variation in the data and instruments.[26] Consistent with the findings in Section 4, the specifications which include embeddings estimate higher substitution between products, measured by median diversion to inside options. The characteristics specification estimates particularly low median diversion between hotels and higher median markups.[27]

TABLE 7: Demand Estimation Results

		Logit	Chars	Embed	Mixed
β	Price	−0.036	−0.036	−0.036	−0.036
		–	–	–	–
Σ	Price	–	0.027	0.026	0.025
			(0.002)	(0.045)	(0.002)
	x_1	–	0.009	0.046	0.000
			(6.019)	(73.407)	(5.369)
	x_2	–	3.875	1.409	2.156
			(4.673)	(364.199)	(4.842)
	x_3	–	3.346	1.296	0.206
			(4.871)	(144.605)	(1.246)
	x_4	–	–	–	1.942
					(8.651)
	x_5	–	–	–	0.000
					(3.046)
Num. Obs.		18620	18620	18620	18620
Own-price Elasticity		−4.786	−2.482	−2.517	−2.673
Outside Diversion		0.503	0.311	0.369	0.353
Inside Diversion		0.006	0.002	0.006	0.006
Markups		0.239	0.425	0.372	0.369

Note: For the characteristics specification, x_1, x_2, and x_3 refer to latitude, longitude, and an indicator for upscale-and-higher quality in the characteristics specification. In the embedding specificaiton, these refer to the coordinates of the three-dimensional embedding. In the mixed specification, x_1 and x_2 are latitute and longitude, while $x_3, \ldots, x_5$ are three dimensions of a mixed embedding. Post-estimation statistics are presented as median values of the full sample across products and markets.

To compare the validity of the estimates of substitution, I examine the diversion ratios

[26] Similar work, such as Armona et al. (2024), also encounters these challenges with respect to the identification of non-linear parameters.

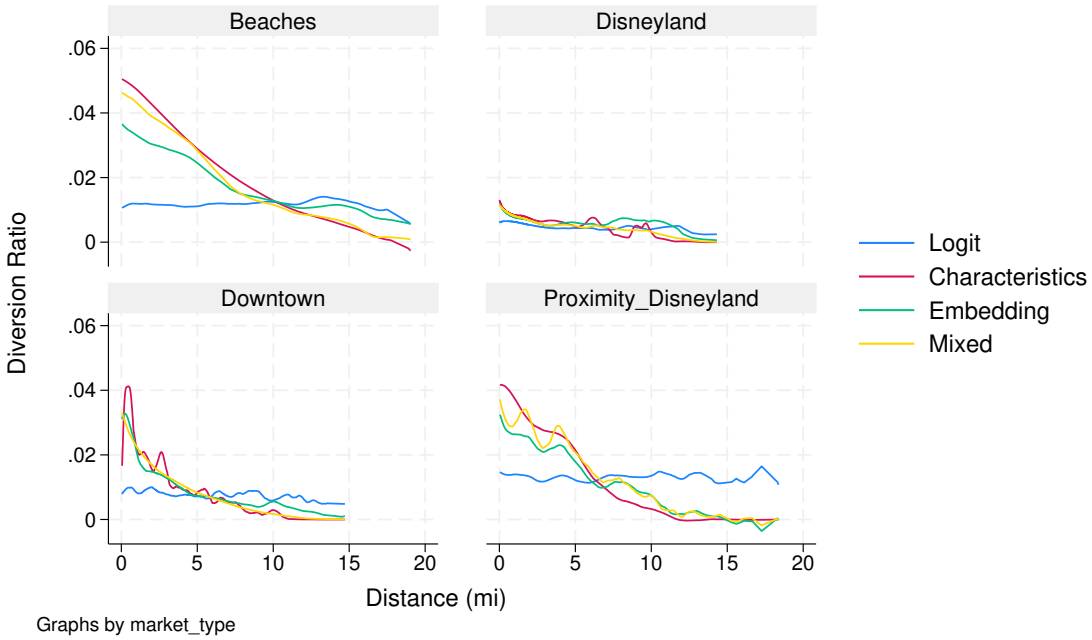
[27] I show the correlation between estimated diversion ratios in Appendix Table 2. The embedding and mixed specification are most similar with a correlation of 0.87. The characteristics specification has correlation of diversion ratios of 0.67 to the embedding specification and 0.78 to the mixed specification.

$-\frac{\partial q_k}{\partial p_j} \Big/ \frac{\partial q_j}{\partial p_j}$ between products across known dimensions of differentiation. An intuitive expectation is that the diversion ratios from hotel j to rival k—the proportion of customers who switch to k from j due to an increase in p_j—should be higher when j and k are more similar. Figure 6 plots, for each specification, the estimated mean diversion ratio from a kernel regression between hotels as a function of distance between hotels in miles.[28] The logit specification produces—unsurprisingly—diversion ratios that are on average constant across distances. Each of the other three specifications produces a generally monotonically-decreasing function in terms of distance. Despite the characteristics specification directly incorporating latitudes and longitudes, it has regions of notable non-monotonicity in the Downtown market, while the embeddings specification—despite not directly incorporating location in the specification—produces a smoother relationship. Given that all three specifications beyond simple logit are fairly simple and that the characteristics specification contains useful spatial and quality information, the purpose of this test is not to necessarily demonstrate the superiority of the embeddings specification. Instead, it is to show that in an environment without product characteristics (the hypothetical case where only recommendations are used), the quality of the results is at least similar to those which employ observable data on product characteristics, and that adding recommendations to the mix may also be a viable approach.

The pattern of greater diversion to alternatives which are more similar is also observable when considering quality as the measure on which to gauge hotel similarity. Table 8 presents the estimated diversion ratios by the average diversion from a hotel in one class to a hotel in another class. In general, all models estimate that within-type diversion is higher than diversion across classes, with some exceptions. Substitution between far types, such as luxury to lower quality or economy to high quality, is estimated to be extremely low, as expected. As suggested by the recommendations, luxury hotels exhibit extreme drop-off in substitution when the alternative quality is below upper upscale, as shown best by the embeddings specification. Another observation is that the characteristics model, which includes quality as two discrete types, results in more structured substitution to within-bracket types and outside types: there is a stark decrease in substitution between types in the upper (Luxury through Upscale) and lower (Upper Midscale through Economy) types due to the model specification that is smoother in the embeddings and mixed specifications. Whether this is an accurate pattern, and hence an aspect that should be directly included

[28] Estimates on the effect of distance on diversion from a nonparametric kernel regression on the full sample are 0.0001 (logit), −0.0008 (characteristics), −0.0010 (embeddings), and −0.0007 (mixed). Distances are taken as straight-line distances.

FIGURE 6: Mean Kernel Estimates of Diversion by Distance



in the model rather than allowing variation in the data to estimate it flexibly, is a question left open.

6 Conclusion

As the amount of data relating to consumer preferences expands, IO practitioners have continually developed new methods for leveraging these data to estimate more flexible models. In this paper I discuss a generalizable approach for the collection and incorporation of publicly-available and easily-collected data on default recommendations for demand estimation, relevant to both linear distance-based demand and more complex mixed logit approaches. This method allows practitioners to make use of the information provided by default recommendations to place products in utility space, even when the researcher does not have access to useful data on product characteristics or consumer preferences (e.g. search, second choice, etc).

To demonstrate the usefulness of this approach, I use an embedding constructed from the ranked recommendations in simulation and an empirical exercise using Orange County

TABLE 8: Mean Diversion by Class

	Luxury	Upper Up.	Upscale	Upper Mid.	Midscale	Economy
Luxury	0.011	0.017	0.005	0.001	0.002	0.000
	0.032	0.023	0.002	0.001	0.001	0.000
	0.026	0.013	0.005	0.002	0.002	-0.002
Upper Upscale	0.001	0.015	0.013	0.000	0.000	0.000
	0.004	0.010	0.008	0.006	0.003	0.004
	0.002	0.010	0.008	0.005	0.003	0.003
Upscale	0.001	0.015	0.016	0.000	0.000	0.000
	0.000	0.012	0.010	0.008	0.004	0.005
	0.001	0.010	0.009	0.007	0.004	0.004
Upper Midscale	0.000	0.001	0.001	0.013	0.010	0.012
	0.000	0.010	0.008	0.007	0.005	0.005
	0.001	0.009	0.008	0.006	0.004	0.005
Midscale	0.000	0.000	0.000	0.010	0.010	0.009
	0.000	0.007	0.006	0.005	0.005	0.005
	0.001	0.006	0.005	0.004	0.004	0.005
Economy	0.000	0.000	0.000	0.011	0.009	0.013
	0.000	0.007	0.006	0.006	0.005	0.007
	0.001	0.006	0.006	0.006	0.004	0.007

Note: Cells denote the median diversion from a hotel of the row class to a hotel of the column class. Within-class diversion cells are highlighted. The values (in vertical order) refer to the characteristics, embedding, and mixed specification respectively.

hotels. In two Monte Carlo experiments, I show that using the embedding in place of a product space can improve key post-estimation results of interest. This is most relevant in cases where data on the product space are not readily available and recommendations can enable demand estimation where it would otherwise be infeasible, or where unobserved heterogeneity in preferences results in variation that poorly identifies a demand system using the true characteristics. Results are further improved by using the recommendations to form an embedding in tandem with observed characteristics which are relevant to consumer utility and demand. Taking these observations to data, I estimate a BLP demand specification for hotels in Orange County, CA and recover substitution patterns and markups using observed characteristics and the embedding. I show that reasonable substitution patterns can be estimated with or without observed hotel characteristics when recommendations are available.

Beyond this application, this approach suggests promise in settings where characteristics are challenging to obtain, and more comprehensive data collection methods are infeasible.

Large product spaces or experience goods may make survey-based approaches inappropriate, and the proprietary nature of some data limits the resources available to researchers. The expanding digitization of consumer engagement with markets provides continually more cases where search tools such as platforms operate: in these environments, this approach is low-cost in terms of data acquisition, providing a useful alternative for practitioners.

References

Abaluck, J., G. Compiani, and F. Zhang (2024): "A Method to Estimate Discrete Choice Models that is Robust to Consumer Search," *Working Paper*.

Agarwal, N. and P. J. Somaini (2022): "Demand Analysis under Latent Choice Constraints," Working Paper 29993, National Bureau of Economic Research.

Armona, L., G. Lewis, and G. Zervas (2024): "Learning Product Characteristics and Consumer Preferences from Search Data," *Marketing Science*, Forthcoming.

Backus, M., C. Conlon, and M. Sinkinson (2021): "Common ownership and competition in the ready-to-eat cereal industry," Tech. rep., National Bureau of Economic Research.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024): "Inference for Regression with Variables Generated from Unstructured Data," *Working Paper*.

Berry, S. and P. Jia (2010): "Tracing the Woes: An Empirical Analysis of the Airline Industry," *American Economic Journal: Microeconomics*, 2, 1–43.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 841–890.

——— (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112, 68–105.

Berry, S. T. and P. A. Haile (2021): "Foundations of Demand Estimation," *Handbook of Industrial Organization*, 4, 1–62.

Christensen, P. and C. Timmins (2022): "Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice," *Journal of Political Economy*, 130, 2110–2163.

Compiani, G., I. Morozov, and S. Seiler (2023): "Demand Estimation with Text and Image Data," *working paper*.

Conlon, C. and J. Gortmaker (2020): "Best practices for differentiated products demand estimation with pyblp," *The RAND Journal of Economics*, 51, 1108–1161.

Conlon, C. and J. Mortimer (2013): "Demand Estimation under Incomplete Product Availability," *American Economic Journal: Microeconomics*, 5, 1–30.

Conlon, C., J. Mortimer, and P. Sarkis (2023): "Estimating preferences and substitution patterns from second choice data alone," *Working paper*.

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 135, 561–644.

De Los Santos, B., A. Hortacsu, and M. Wildenbeest (2012): "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior," *American Economic Review*, 102, 2955–2980.

Deaton, A. and J. Muellbauer (1980): "An almost ideal demand system," *The American Economic Review*, 70, 312–326.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice RAndom Coefficients Demand Estimation," *Econometrica*, 80, 2231–2267.

Farronato, C. and A. Fradkin (2022): "The Welfare Effects of Peer Entry in the Accommodation Market: The Case of Airbnb," *American Economic Review*, 112, 1782–1817.

Fox, J. T. (2007): "Semiparametric estimation of multinomial discrete-choice models using a subset of choices," *The RAND Journal of Economics*, 38, 1002–1019.

Gabel, S. and A. Timoshenko (2022): "Product choice with large assortments: A scalable deep-learning model," *Management Science*, 68, 1808–1827.

Gandhi, A. and J.-F. Houde (2023): "Measuring Substitution Patterns in Differentiated-Products Industries," *Working Paper*.

Grieco, P. L., C. Murry, and A. Yurukoglu (2021): "The evolution of market power in the US auto industry," *NBER Working Paper*.

Hodgson, C. and G. Lewis (2024): "You Can Lead a Horse to Water: Spatial Learning and Path Dependence in Consumer Search," *Working Paper*.

Jaffe, S. and E. G. Weyl (2010): "Linear demand systems are inconsistent with discrete choice," *The BE Journal of Theoretical Economics*.

Kaye, A. P. (2024): "The Personalization Paradox: Welfare Effects of Personalized Recommendations in Two-Sided Digital Markets," *Working Paper*.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010): "Online Demand Under Limited Consumer Search," *Marketing Science*, 29, 1001–1023.

Kumar, M., D. Eckles, and S. Aral (2020): "Scalable bundling via dense product embeddings," *arXiv preprint arXiv:2002.00100*.

Lewis, G. and G. Zervas (2019): "The Supply and Demand Effects of Review Platforms," *Working Paper*.

Magnolfi, L., J. McClure, and A. Sorensen (2025): "Triplet Embeddings for Demand Estimation," *American Economic Journal: Microeconomics*, 17, 1–26.

McClure, J. (2024): "Markups and Costs under Capacity Constraints: the Welfare Effects of Hotel Mergers," *Working Paper*.

McFadden, D. (1978): "Modeling Choice of Residential Location," in *Spatial interaction theory and planning models*, ed. by A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, Amsterdam: North Holland.

Nevo, A. (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69, 307–342.

Petrin, A. (2002): "Quantifying the benefits of new products: The case of the minivan," *Journal of Political Economy*, 110, 705–729.

Pinkse, J. and M. E. Slade (2004): "Mergers, brand competition, and the price of a pint," *European Economic Review*, 48, 617–643.

Pinkse, J., M. E. Slade, and C. Brett (2002): "Spatial price competition: a semiparametric approach," *Econometrica*, 70, 1111–1153.

Reynaert, M. and F. Verboven (2014): "Improving the performance of random coefficients demand models: the role of optimal instruments," *Journal of Econometrics*, 179, 83–98.

Ruiz, F. J., S. Athey, and D. M. Blei (2020): "Shopper: A probabilistic model of consumer choice with substitutes and complements," *The Annals of Applied Statistics*, 14, 1–27.

Syverson, C. (2019): "Macroeconomics and market power: Context, implications, and open questions," *Journal of Economic Perspectives*, 33, 23–43.

Van Der Maaten, L. and K. Weinberger (2012): "Stochastic triplet embedding," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, IEEE, 1–6.

Appendix A Recommendation Space Formation

To determine which products are connected via recommendations and hence can be mapped into a single embedding, I construct the recommendation space $\mathcal{S}$, which contains separate disconnected recommendation spaces $\mathcal{S}_1$, $\mathcal{S}_2$, ... Recommendation data is often limited: recommendations may be non-reciprocal ($A \to B$ while $B \nrightarrow A$) or second-order ($A \to B \to C$ but $A \nrightarrow C$), but these products are still considered connected. This section outlines a simple method for mapping recommendation data into spaces.

Consider a simple market of four products (A, B, C, D) with recommendations given by

$$\mathcal{R} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

such that the element of row i and column j indicates whether product j is recommended as an alternative to product i. Thus, B is recommended for A, C is recommended for B, and both C and D have no alternatives recommended. (A, B, C) form one recommendation set, while D forms another. Updating the matrix $\mathcal{R}$ to form the recommendation set involves iterating:

$$S_1 = \mathcal{R} \times \mathcal{R}' = \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{j \text{ recommended to } i} \times \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{i \text{ is recommended to } j} = \underbrace{\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{i \text{ and } j \text{ are connected}}$$

S_1—read the same way as $\mathcal{R}$—includes links between products which are one-degree-connected via recommendations: B is connected to A ($S_{1,(2,1)} > 0$) as $A \to B$ despite $B \nrightarrow A$, and likewise C and B. A and C are not yet linked as their connection $A \to B \to C$

takes more than one step, and so via iteration:

$$S_2 = \mathcal{R} \times S_1' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 1 & 0 \\ 1 & 3 & 2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

S_2 includes in the 3rd column the link between A and C. Iterating a third time forms the connection between C and A, which is third-order as it relies on the second-order connection between A and C:

$$S_3 = \mathcal{R} \times S_2' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 2 & 1 & 0 & 0 \\ 3 & 3 & 1 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 4 & 1 & 0 \\ 4 & 5 & 2 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

At this point, further iteration will incorporate no new connections. $\mathbb{1}\{S_{3,ij} > 0\}$ is symmetric, and its non-zero elements form the recommendation space $\mathcal{S}$, which place (A, B, C) $\in \mathcal{S}_1$ together and $D \in \mathcal{S}_2$ separately.

Appendix B Monte Carlo Construction

In the first environment, consumer utility is given by $u_{ijt} = x_{jt}\beta_i + \alpha_i p_{jt} + \xi_{jt} + \epsilon_{ijt}$, with errors $\epsilon \sim$ EVT1 and i.i.d. Random coefficients $(\beta_i, \alpha_i) = (\beta, \alpha) + \Sigma v_i$, where sigma is a diagonal matrix and v_i a vector of 1000 Halton draws from a normal distribution. The $F = 10$ firms each hold 10 products and compete via Bertrand-Nash. Product costs are given by $c_{jt} = \gamma x_j + 2w_{jt}$, where w_{jt} is a uniformly-distributed random variable in $[0, 1]$ which is observed as a cost shifter. Table 1 lists the true parameters of the model:

The second environment constructs consumer utility as $u_{ijt} = 5 - p_{jt} - 2\left(\sum_{k=1}^{3}(B_{ikt} - x_{jkt}^2)^2\right)^{0.5} + \xi_{jt} + \epsilon_{ijt}$, given $\xi \sim N(0, 0.2)$ and EVT1 errors ϵ. $N = 1000$ consumers are simulated per market with bliss points drawn from a Gamma distribution: $(B_{i1}, B_{i2}, B_{i3}) \sim \Gamma(2, 0.5)$. $J = 100$ products owned by $F = 10$ firms are generated with $K = 3$ characteristics:[29]

[29] The distribution of characteristics in the product space is taken from Dubé, Fox, and Su (2012).

Appendix Table 1: Simulation 1 True Parameters

	Constant	Price	x_1	x_2	x_3	x_4	x_5	x_6
β	1	-0.5	0	0	0	0	0	0
Σ	5	0.075	0.5	0.5	0.5	0.5	0.5	0.5
γ	5	-	0.1	0.1	0.1	0.1	0.1	0.1

The outlined specification results in a mean inside share of 0.67. The [5, 95] percentile bounds on prices and shares are [6.28, 10.28] and [0.002, 0.017].

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 & 0.3 \\ -0.8 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix} \right)$$

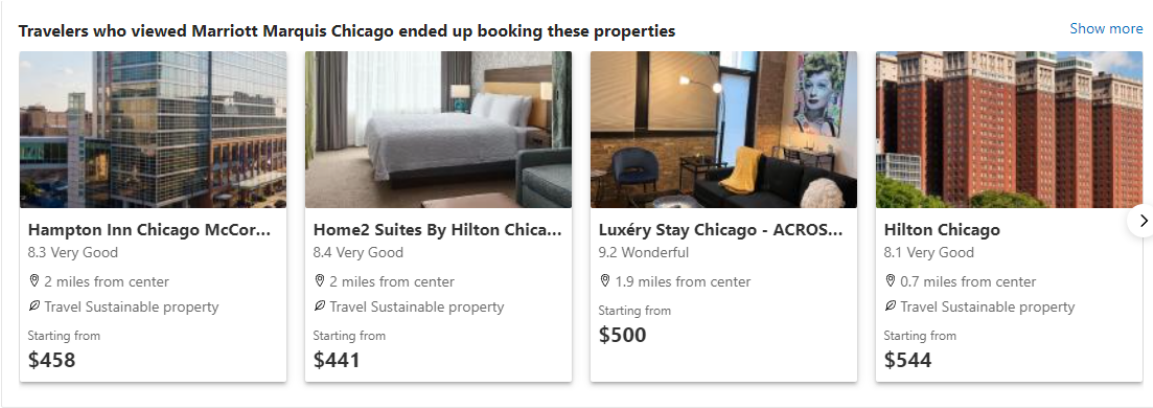
Marginal costs are $4 + w$, where $w \sim U[0, 2]$. Firms compete via Bertrand-Nash. In each scenario, the equilibrium prices are solved for by iterating towards the fixed point that solves the Bertrand-Nash first-order conditions:

$$p - c = \left(-\frac{\partial s(p)}{\partial p} \cdot \Omega \right)^{-1} s(p) \tag{7}$$

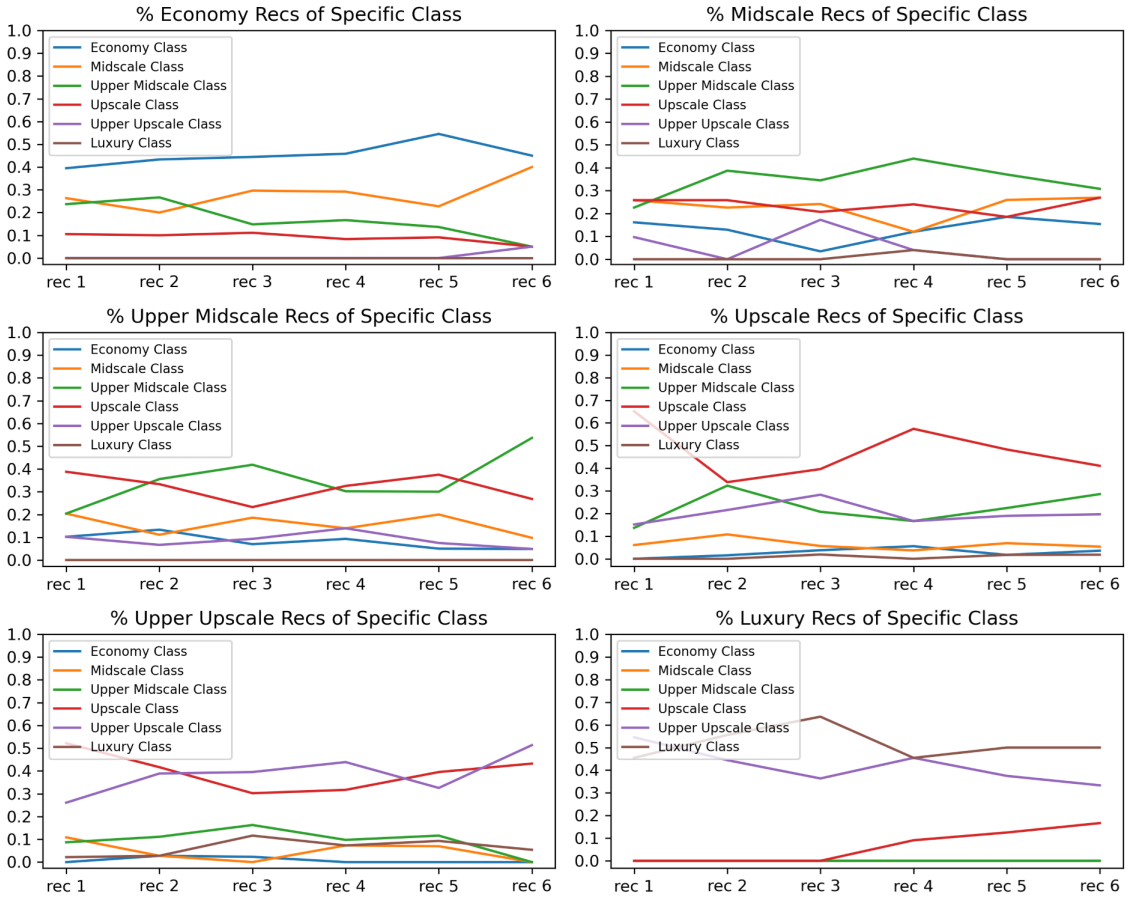
given a $J \times J$ matrix of firm ownership Ω.

Appendix C Additional Tables and Figures

Appendix Figure 1: Recommendations at Booking.com



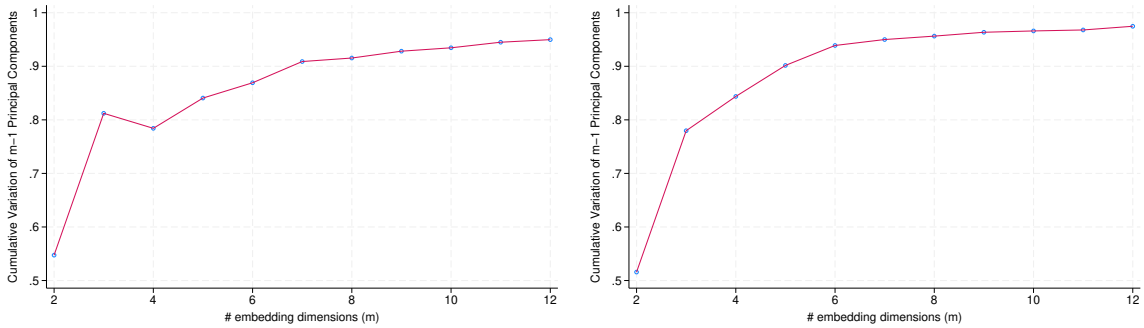
Appendix Figure 2: Recommendations by Classes



Appendix Figure 3: Cumulative variation of $m-1$ Principal Components

(A) Monte Carlo 1: $R = 5$

(B) Monte Carlo 2: $R = 25$



Appendix Table 2: Correlation in Estimated Diversion

	Logit	Chars	Embed	Mixed
Logit	1.0000			
Chars	0.2206	1.0000		
Embed	0.3141	0.6735	1.0000	
Mixed	0.3050	0.7781	0.8694	1.0000