

Using Default Recommendations in Demand Estimation

Jonathon McClure*

Mitch Daniels School of Business

Purdue University

January 2026

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 numerous discussants at the Purdue Economics summer seminar and Purdue Marketing brown bag 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 in order to accurately understand consumer behavior and model counterfactual outcomes. In standard approaches, a key factor to this is the availability of data on product differentiation: both spatial product-space (Pinkse, Slade, and Brett (2002); Pinkse and Slade (2004)) and characteristics-space (McFadden (1978); Berry, Levinsohn, and Pakes (1995), henceforth BLP) approaches require data on product characteristics that capture differentiation. In practice, however, some amount of the true set of characteristics are unobserved: often the data on product attributes is insufficient to properly capture product differentiation or is unavailable.¹ For example, it is not immediately clear what the full set of relevant quantifiable characteristics is when choosing entertainment (a book or movie), fashion, or a vacation destination. This creates a hindrance to researchers who attempt to estimate flexible models of demand in these markets.

In this paper, I demonstrate using unstructured data from online platforms’ recommended alternatives to each product to provide information about product differentiation and estimate substitution patterns in the absence of data on product characteristics. I treat the ranking of recommended alternatives as information about the similarity of products’ mean utilities. These ordinal measures of similarity are used to construct a continuous vector representation (an “embedding”) of the latent characteristic space, where recommended alternatives to a given product are located closer to each other.² The latent characteristics can be included in a typical random-coefficients logit model to estimate demand and recover substitution patterns using typical variation in price and quantity market data. This method requires no data on observable product characteristics, allowing it to complement existing data or substitute for it when it is not available.

¹The 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.

²While the method does not require specifying the data-generating process of the recommendations, some intuition is useful. One interpretation is that 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 we would treat the default recommendations as aggregates of the consumer choices that platforms observe to recover substitutes.

To demonstrate how these embeddings can be taken to data in environments when the observable data are limited, I simulate data drawn from a mixed-logit data generating process. In this example, the practitioner is only able to observe a subset of the product characteristics, but has access to ranked recommendations. Using the t-Distributed Stochastic Triplet Embedding (tSTE) algorithm (Van Der Maaten and Weinberger (2012)) over ordinal measures of similarity implied by the recommendations, I form an embedding to replace the unobserved characteristic space and estimate demand. Based on an analysis of the alternatives with the highest estimated substitution, I find that incorporating the information from recommendations produces better results about consumer preferences than using the limited observed product characteristics.

I then present an empirical example with an application to the hotel sector, where the typical mixed logit approach faces challenges owing to the lack of information about salient characteristics and consumer preferences; the usual demographics approach displayed by Nevo (2001) is not suitable as the customer 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 can be generalized to the collection of many other types of consumer products. 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, class (a measure of hotel quality based on average daily rates), and branding. These patterns are observable in a 2-dimensional embedding, which clusters hotels by proximity and quality.

Applying these approach to data, I use a monthly panel of hotel-level prices and quantities provided by STR LLC to estimate the demand system. I produce results using a coarse set of observable characteristics (the distance of the hotel to key points, as well as a categorical quality indicator) as well as using a four-dimensional embedding produced from the recommendations. Analysis of the results shows that while overall levels of substitution are similar, the top three substitutes by diversion ratio $-\frac{\partial q_k}{\partial p_j} / \frac{\partial q_j}{\partial p_j}$ follow more intuitive patterns in the embedding specification: they are more likely to be nearby, and of the same quality tier, when compared to the characteristics-based specification. These findings indicate that when observable characteristics are limited (or unavailable), using recommendation-based embeddings can improve the estimation of key post-estimation statistics like substitution patterns, which in turn inform counterfactuals like the price effects of mergers or new products.

Following these results, I discuss several topics that provide guidance to practitioners in using this method. A first key result is that the method does not require a large number of recommendations to work, so long as the recommendations form a connected set: 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 further demonstrate that biased recommendations—which may be driven by platform preferencing products with higher rates of return—cause only a small reduction in the model’s fit unless the majority of recommendations are biased. Last, I outline rules of thumb for selecting the dimension of the embedding, showing that pre-estimation statistics do well in determining dimensions where post-estimation results are stable. As a final note, I summarize some discussion of bias correction due to potential mismeasurement of the embedding from finite data.

This paper adds to the growing body of work in economics and marketing that focuses on the use of auxiliary data to augment the estimation of demand systems and substitution patterns. This is not a new issue: for example, Nevo (2001) and Petrin (2002) use consumer demographics alongside aggregate price and quantity data to aid in the estimation of substitution patterns. More specifically, this paper contributes to two main areas: the use of machine learning to infer latent characteristics from auxiliary data, and the incorporation of consumer search and second choice information in demand estimation.

The inferring of latent product and market characteristics from data is an established literature in economics and marketing (Elrod and Keane (1995)). Methodologically, this paper is most similar to the literature which has used machine learning to form *embeddings* of latent characteristics or preferences from consumer data.³ For example, Ruiz, Athey, and Blei (2020); Kumar, Eckles, and Aral (2020); and Gabel and Timoshenko (2022) use data on consumer purchases and search, while Armona, Lewis, and Zervas (2024) use consumers’ online search history to construct a Bayesian Personalized Ranking. Outside of search, several papers have used deep learning models with descriptive product data: Han, Schulman, Grauman, and Ramakrishnan (2021) use fonts’ shapes, while Bajari, Cen, Chernozhukov, Manukonda, Vijaykumar, Wang, Huerta, Li, Leng, Monokroussos, and Wang (2025) and Compiani, Morozov, and Seiler (2025) recover latent attributes from products’ images and text (reviews and descriptions). Outside of platform data, Magnolfi, McClure, and Sorensen (2025) directly survey consumers about which products they see as most similar to each other. Each of these latter cases relies on unstructured data, leveraging large volumes

³For an overview of machine learning approaches in demand estimation, see Bajari, Nekipelov, Ryan, and Yang (2015).

of information without asserting an underlying data-generating process. In Section 2.2, I summarize several of these approaches. This paper proposes a new source of suitable data which is easily collected and does not rely on proprietary data, pre-trained machine learning models, or the cost of designing and implementing a survey.

This paper’s treatment of platform data is related 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); Kim, Albuquerque, and Bronnenberg (2016) use aggregate search data in combination with choice data to estimate consumer preferences. This paper treats platform recommendations as containing information on the aggregate outcome of search behavior as platforms aim to help solve consumers’ choice problem. Other papers have made use of search data more generally to aid demand estimation. Amano, Rhodes, and Seiler (2022) use browsing data to identify sparse choice sets and in turn estimate consumer preferences, while Abaluck, Compiani, and Zhang (2024) propose a class of models that allow for demand estimation when consumers must search for products. Several papers have also attempted to recover the platform’s objectives from a model of search (Hodgson and Lewis (2023); Donnelly, Kanodia, and Morozov (2023)), which provides valuable context for why the information gathered for how the information gathered in this paper can be interpreted.

This method is also conceptually similar to the idea of identifying second (or alternative) choices. Second-choice data is often used to discipline substitution patterns: for example, Berry, Levinsohn, and Pakes (2004) and Grieco, Murry, and Yurukoglu (2021) obtain consumer responses on second choices in the automobile market, using this information to construct additional moments for estimation. Conlon, Mortimer, and Sarkis (2023) use similar data and target surveyed second-choice probabilities with a semiparametric linear model. In this context, the platforms’ recommendations can be interpreted as an ordered list of utilities (and hence second-choice probabilities). The embedding therefore captures information about the covariances between products’ utilities across consumers, which is central to the estimation of substitution patterns.

In Section 2, I outline the methodological approach, discuss several similar approaches in the literature, and illustrate its performance in simulated data. Section 3 presents the empirical application, describes the market and recommendation data, and summarizes the demand model and estimation strategy. In Section 4, I report estimation results and discuss substitution outcomes when comparing observable characteristics to the recommendation

method. Finally, in Section 5, I discuss guidance for the practitioner, with attention to various robustness checks and hyperparameters of the model. Section 6 concludes.

2 Methodology

The usual setting in many empirical contexts is to estimate a discrete-choice model of demand, using a mixed logit specification where utility is a function of product j 's characteristics and consumer i 's preferences in market t . Utility in this system can be written as

$$u_{ijt} = \alpha_i p_{jt} + \beta_i [x_{jt}^o, x_{jt}^e] + \xi_{jt} + \epsilon_{ijt}, \quad (1)$$

given preferences (α_i, β_i) , prices p_{jt} , unobserved product quality ξ_{jt} , an idiosyncratic shock $\epsilon_{ijt} \sim EVT1$, and where product characteristics $x_{jt} \equiv [x_{jt}^o, x_{jt}^e]$ are comprised of observable characteristics x_{jt}^o and unobservable aspects x_{jt}^e .⁴ These latter aspects may lack data, or in practice, be those that are hard to meaningfully quantify (e.g. elements of style, taste, or other abstract characteristics).

The unobserved product characteristics pose a problem: regardless of how flexibly x_{jt}^o is incorporated into the model, omitting x_{jt}^e hinders the practitioner in correctly estimating substitution patterns, particularly if these unobserved elements have a large impact on utility.⁵ As a result, we might under-predict substitution to close substitutes, or over-predict it to dissimilar products, particularly when market shares of these products are high (ending up closer to the outcomes of a simple logit specification). This in turn has ramifications for the estimation of markups, the price effects of a merger, or consumer welfare. The goal of the method proposed here is to estimate x_{jt}^e in order to obtain improved estimates of consumer substitution. By collecting unstructured data from platforms on the recommended alternatives to each product (denoted \mathcal{R}_j), I construct a low-dimensional vector representation (an *embedding*) which flexibly captures implied product similarity. These vectors can be included in the demand specification as characteristics and interacted with random coefficients in the typical mixed logit/BLP models of demand.

An important strength of this method is that the practitioner can remain agnostic as to

⁴I thank an anonymous reviewer for clarifying this setup.

⁵The parametric structure of Equation 1 is not the primary limitation: even semi-parametric approaches to estimating $f(\beta_i)$ such as Fox, Kim, Ryan, and Bajari (2011) are constrained by relying solely on x_{jt}^o .

the exact data-generating process of the recommendations \mathcal{R}_j and the platform’s strategy in offering recommendations.⁶ By constructing an embedding (an estimate of x_{jt}^e) and including it in the demand specification, it becomes an empirically-answerable question as to whether the information in the embedding helps explain the substitution patterns in the data. What matters is how these auxiliary data aid in capturing the covariance in utility between products from variation in the data itself: the mixed logit/BLP model can treat these data as noise (i.e. assign $\beta_i = 0$) if they are not relevant. This minimizes a question of whether we learn solely about product differentiation or about consumer preferences (i.e. x or $x\beta$), as each enters the utility function linearly and our focus is on the whole value of utility. If platform bias produces an \mathcal{R}_j that is not useful for recovering patterns from the market data, then the model need not assign weight to characteristics generated from \mathcal{R}_j .

2.1 Recovering Embeddings from Recommendations

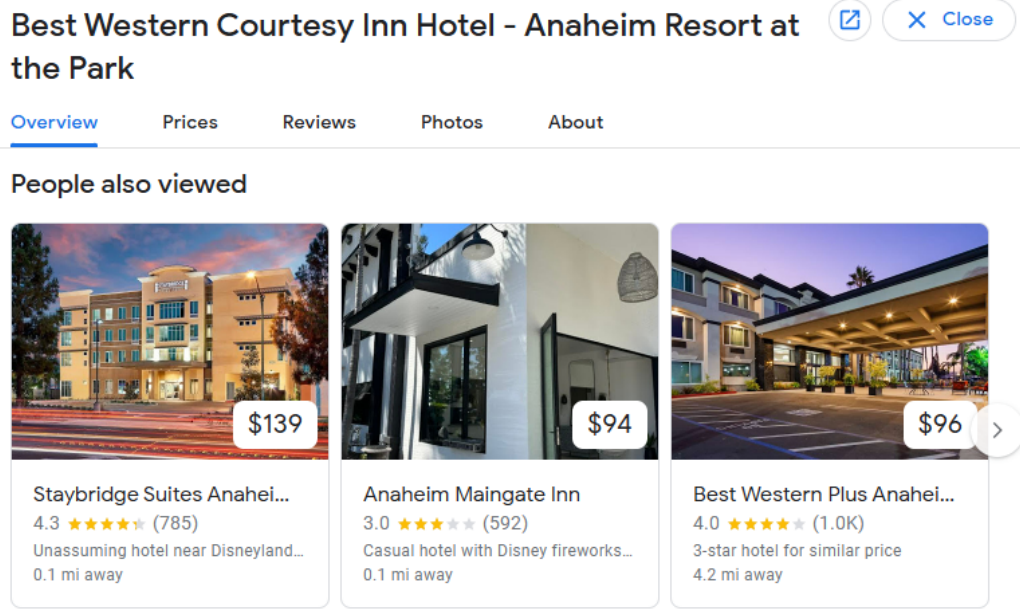
The first step of the proposed methodology is to collect \mathcal{R}_j , the set of recommended substitutes to each product. Many platforms provide this information: Figure 1 gives an example from Google Travel for one hotel. Here, a search for “Best Western Courtesy Inn Hotel - Anaheim Resort at the Park” gives a set of 6 (3 displayed) alternative suggestions that were searched.⁷ The information provided here can be parsed in two ways. A weaker assumption is that recommended alternatives are more similar to the reference product than non-recommended alternatives. For example, if k is recommended alternative to j and ℓ is not, then j is more similar to k than it is to ℓ . A stronger assumption implies ordering: a top-recommended alternative is more similar to the reference than a lower-recommended alternative (and both are more similar than a non-recommended alternative).

The above data provides inequalities of the similarity between products, which I frame as triplets: inequalities that state “product A is closer to B than it is to C” based on the

⁶A simple interpretation of recommendations is that the platform suggests alternatives that are the closest substitute on average: e.g. they have the highest mean second-choice probability, or the highest diversion ratio. In this case, recommendations form an ordered list of second-choice alternatives. However, many other strategies for platforms exist, either in helping the consumer learn the broader product space (as in [Hodgson and Lewis \(2023\)](#)), biasing towards higher-profit alternatives for the platform, or based on other unobserved incentives.

⁷Different 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 products with correlated utility so long as they are not stated to be advertisements.

FIGURE 1: Example of Default Recommendations



Note: Figure presents a screenshot of recommended alternatives at Google Travel. The top 3 alternatives are shown, followed by an additional 3 if the user scrolls right.

similarities implied by the recommendations. I collect these comparisons 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.⁸ The tSTE algorithm proceeds as follows. Formally, given a set of products $j = 1, \dots, J$, we want to find a set of vectors $\mathbf{x} \equiv \{x_1, \dots, x_J\} \in \mathbb{R}^m$ that represent the products in m -dimensional space.⁹ 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

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

⁸The choice of the tSTE approach is largely based on its suitability to ordinal measures of similarity: While this paper only makes use of tSTE, other machine learning tools could also be used to incorporate the same data.

⁹The selection of the hyperparameter m (the number of dimensions in the embedding) is a matter of researcher choice. I discuss this decision further in Section 5.3.

The result is a continuous vector representation in \mathbb{R}^m of the differentiation implied by the triplets. The coordinates of each product can be treated as characteristics, and included in Equation 1 in place of the characteristics x_{jt} . Alternatively, 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 conditional on the variation in x_{jt}^o .

2.2 Comparisons to Existing Approaches

The literature on auxiliary data for demand estimation and embeddings is expanding, and other papers have used similar tools and sources of data to estimate substitution patterns. In this section, I summarize several similar recent approaches and briefly compare this paper’s methodology and applicability to each.

Other papers have highlighted additional uses of platform data when constructing latent characteristics and/or embeddings. When proprietary data on consumer searches through a platform are available, search patterns can reveal which products are considered more similar in utility. [Armona et al. \(2024\)](#) use a Bayesian Personalized Ranking given search data for consumers on Expedia: if products A and B are searched but not C , then the consumer must expect more utility from A and B than any non-searched option C . These inequalities provide similar information as the method in this paper, but do so with more granular (and proprietary) data and a micro-founded approach.

Alternatively, a major source of information for consumers provided by platforms is product descriptions (both text and images) and reviews. Consumers are able to rapidly parse this information to determine product similarities; [Compiani et al. \(2025\)](#) use pre-trained deep learning tools to systematically process these data and obtain embeddings of product similarity. These data can capture intangible elements of product design and broad consumer impressions, and are easily collected from the platform. However, the approach relies on the black box of pre-trained models, which may not be well-suited to capture all dimensions of product differentiation. It is also worth noting that when product descriptions are lacking, large language models (LLMs) can be used to construct descriptions for this method.

Beyond platforms, another method of leveraging consumer knowledge of the product space

is to simply ask them via surveys. In a similar spirit to the above papers, [Magnolfi et al. \(2025\)](#) survey consumers as to which products are more similar to a reference product, eliciting triplets (“ A is more similar to B than it is to C ”), fitting these inequalities into an embedding. The authors similarly make use of the tSTE algorithm to fit these into an embedding. This approach does not rely on any ML interpretation of the unstructured data, but requires a (potentially costly) survey and is only suitable for markets where consumers are well-informed about products. This latter condition may fail if goods are primarily experience goods or the product space is opaque, such that using the aggregated information from a platform (as a consumer might) is more suitable.

2.3 Simulated Example

To illustrate the application and its impact on common post-estimation results, consider a simulated example. Consumers have mixed-logit utility given by Equation 1, over a set of $J = 50$ products randomly offered across $T = 1000$ markets. Importantly, not all product characteristics are observed: $X = (1, x_1^o, x_1^e, x_2^e, x_3^e)$ such that only one attribute is observable to the practitioner.¹⁰ However, the practitioner also has access to data on the top $R = 10$ recommended substitutes for each product. I estimate the demand system and obtain two common post-estimation results: (i) diversion to the nearest products, and (ii) the price effects of a merger. As (x_1^e, x_2^e, x_3^e) are unobserved, I compare two feasible-yet-misspecified approaches: using either the sole observable characteristic x_1^o , or using a set of four embedding dimensions $(x_1^r, x_2^r, x_3^r, x_4^r)$ constructed from simulated recommendations. The full details of the Monte Carlo simulation and recommendation construction are laid out in Appendix B.

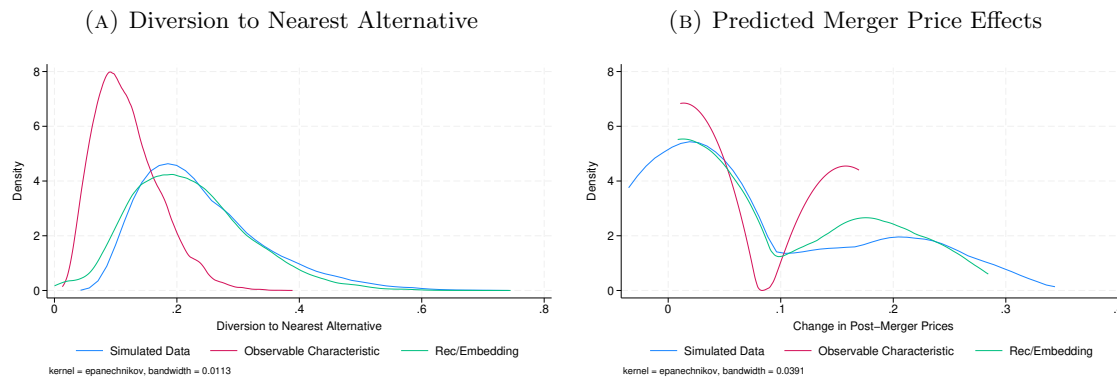
The practitioner in this case is unable to correctly specify the demand model, facing a choice between using the incomplete characteristics data x^o or attempting to recover x^e via the embedding. Using only the incomplete x^o , which has limited information on the utility covariates, results in underestimating substitution to nearby products; Panel A of Figure 2 shows the distribution of diversion ratios to the true nearest alternative product for all product-market observations.¹¹ When using recommendations in place of x^o , the

¹⁰These are constructed as rankings of the nearest alternatives by their true diversion ratios.

¹¹Both of the models are able to estimate sensible mean price sensitivities, as this relies primarily on the cost shifter instrument. However, the x^o specification is not able to identify the random coefficient on price sensitivity. Given a ground truth of $\alpha = -2$ and $\Sigma_\alpha = 0.2$, the x^o specification estimates -1.773 (0.049) and 0.000 (15.083), while the recommendation specification obtains -2.007 (0.057) and 0.354 (0.082).

specification is better able to recover accurate diversion to the closest substitutes. Panel B shows the distribution of average product-level price changes due to a simulated merger, where the recommendation specification also produces more accurate results. The simulated merger increases prices for merging products by an average of 3.75%: the x^o specification predicts 3.29%, while the recommendation specification predicts 3.78%.

FIGURE 2: Comparison of Post-Estimation Results



Note: Panel A shows the diversion from each product-market combination to the nearest alternative (the alternative with the highest diversion in the true data). Panel B shows the average impact of a 5 \rightarrow 4 merger on mean product-level prices predicted from each specification.

3 Empirical Application

3.1 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.¹² The data cover a period of 2017 through 2023. I observe hotel prices and quantities linked to anonymized hotel identifiers. Separately, 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,

¹²I 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)).

airport, etc). I am able to combine non-identifying attributes of hotels with the performance data through the data provider. 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.

TABLE 1: Summary of Hotel Performance Data

	Obs	Hotels	Average Daily Rate (ADR)			Occupancy %		
			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. 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.¹³ These recommendation rankings are converted to triplets and used to estimate an 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.

¹³Not 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 Appendix A, I discuss the importance of recommended hotels forming a connected set with regards to the formation of the embedding.

3.2 Variation in Observed Recommendations

In order to better validate the idea that the default recommendations are presenting—on average—useful information about product attributes, I summarize some descriptive facts about the recommendations relative to their reference products. I find that the reference product is more similar to recommended alternatives than non-recommended options in terms of several dimensions: location, quality, and management. These differences are summarized in Figure 3.

The first observation (Panel A) 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. The mean distance to recommended hotels increases monotonically through ranks 1-6, and recommendations are on average closer than unrecommended properties.¹⁴ 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.¹⁵ 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.¹⁶ 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.

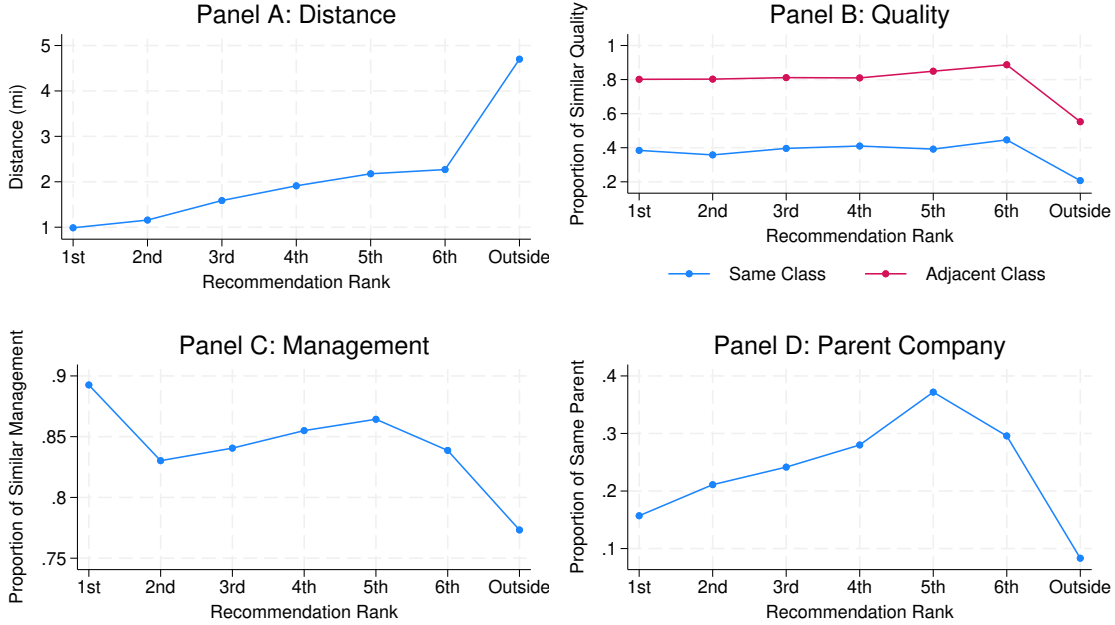
Lastly, I show whether recommendations reflect similarities in terms of management: whether

¹⁴Mean 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.

¹⁵STR 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.

¹⁶For 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 3: Similarity of Ranked Recommendations



Note: Panels present the mean value of each ranking. Panel A outside option contains hotels of the same class and in the same market. Panels B-D consider outside sets of hotels within 3 miles.

the hotels are both branded or independent (Panel C) or whether the hotels have the same parent company (Panel D). 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, and are 77.3% for the outside group. We can reject the mean of the outside group—the set of all non-recommended properties within 3 miles—being equal to that of the recommendation set with $t = 6.59$.

Likewise, we can reject that recommended and non-recommended hotels are similarly distributed with respect to the reference hotel’s brand ($t = 19.2$ that means differ). 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. Further-

more, 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. 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.

Lastly, I examine how recommendations vary across different qualities of hotel, as they attract consumers with different preferences: 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 :

$$\mathbb{1}\{k \in \mathcal{R}_j\} = \Phi(d(x_j, x_k)) \text{ where } x = \{\text{location, quality, management, parent}\} \quad (3)$$

Table 2 presents the results of estimating Equation 3 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. 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.

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

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

3.3 Constructed Embeddings

I construct the embeddings by grouping recommendations into ranks 1, ranks 2-3, and ranks 3-6, all of which are treated as more similar than non-recommended alternatives.¹⁷ These form a set of triplets for all products.

An useful feature of the embedding approach is that the market can be visualized in two dimensions as a sanity check 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 were observed in the recommendation sets. Figure 4 present plots 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 (upper panel), clustering hotels by geographic markets. There is some measure of overlap between each market definition (downtown and the beach, Disneyland and adjacent areas), but no one market is “central”. The correlation between within-market hotel-rival distances in geographic space

¹⁷This is a midpoint between assuming that recommendation ranks are strict over all products, even on the second page of the suggestion carousel, or that assuming that there is no emphasis placed on the higher recommendations that are shown first.

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.¹⁸ Figure 4 also shows a pattern of clustering of hotels by quality (lower panel). Luxury and Upper Upscale hotels are located in close proximity to each other, including across geographic markets which border each other.¹⁹ Economy and Midscale hotels are positioned at the fringes of each market and 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 is 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.²⁰

A final question is whether the embeddings are encoding more information than just using x^o : in other words, examining whether the embedding is a linear transformation of observable data. Note that if the observable data is very good, this is not a negative result as it suggests that the recommendations approach captures product characteristics well. I regress each dimension of a four-dimensional embedding on a large vector of 80 observable hotel characteristics, and summarize the results in Table 3.²¹ These characteristics capture the distance to landmarks: Disneyland, John Wayne Airport (which sits in the middle of the downtown market of Santa Ana/Costa Mesa), and Newport Beach. I also include dummy variables for quality, the geographic market, hotel parent company, and a set of hotel amenities.

Based on the adjusted R-squared, I find that while the linear transformation of observables explains most of the variation in the embedding (roughly 82%), the embeddings still contain additional information.²² The observable variables that appear most statistically important are intuitive: measures of spatial differentiation (distance to key landmarks and geographic market) and of quality differentiation (Upper Upscale quality, along with parking and brand

¹⁸Geographic distances are computed via Stata’s `vincenty` command.

¹⁹While 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.

²⁰Correlations 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.

²¹I use four dimensions for the embedding following a threshold of $m - 1$ dimensions containing less than 95% of the variation: $m = 5$ fails this screen. Further discussion about selecting the dimension of the embedding is included in Section 5.3.

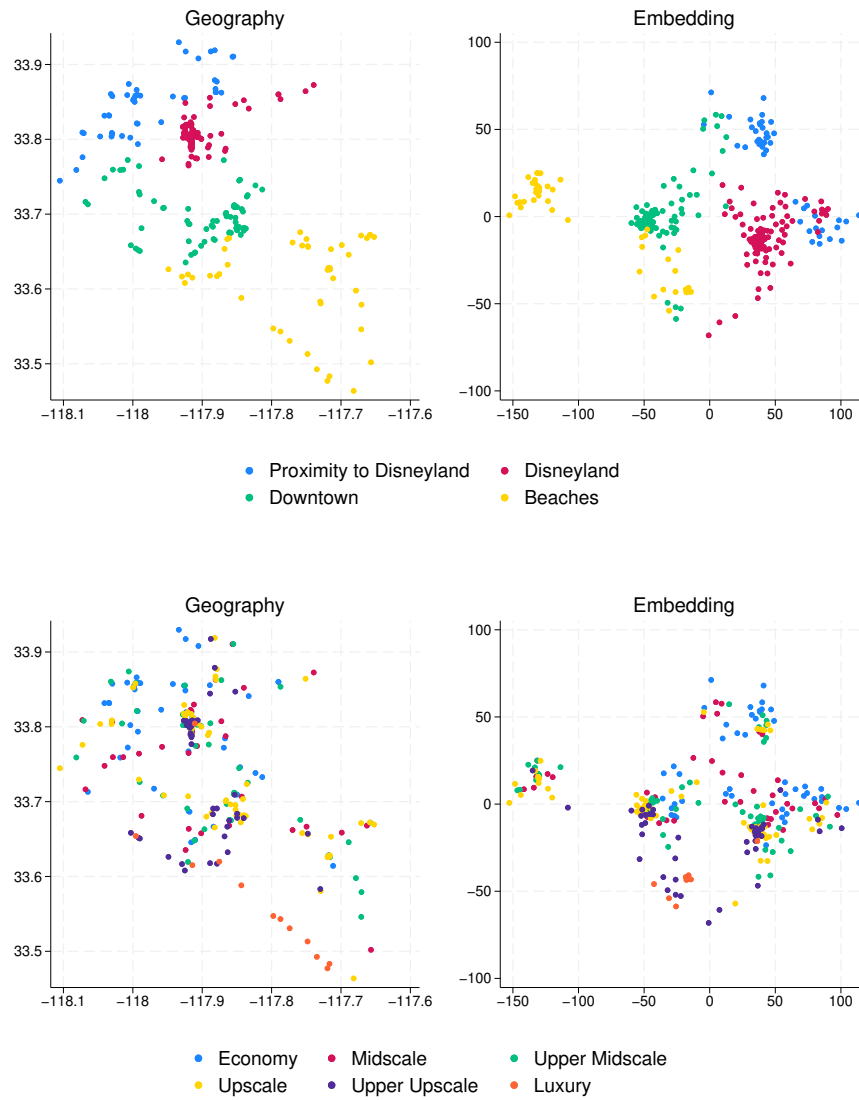
²²Additionally, the variation in the observed data cannot be easily projected in only four dimensions. The 80 observable characteristics include a large number of categorical variables, and it would take 47 principal components to retain 75% of the variation.

TABLE 3: Regression of Embedding Dimensions on Observable Attributes

	x_1	x_2	x_3	x_4
Distance to Disneyland	11.09*** (3.738)	-17.84** (7.900)	-3.974 (5.459)	-20.43*** (4.623)
Distance to Airport	-19.91*** (5.233)	43.40*** (13.29)	23.61** (11.00)	-3.863 (6.087)
Distance to Newport Beach	40.39*** (10.74)	-47.40*** (17.29)	-71.68*** (13.65)	-7.582 (9.119)
Disneyland	5.004 (8.820)	126.4*** (20.65)	4.509 (15.25)	-79.51*** (11.01)
Proximity to Disneyland	-3.677 (8.626)	176.3*** (13.81)	52.68*** (13.90)	-86.04*** (8.675)
Downtown	-0.968 (5.473)	107.4*** (15.26)	60.25*** (13.26)	-40.73*** (7.750)
Upper Upscale Class	-105.8*** (13.13)	121.0*** (17.78)	102.9*** (18.57)	105.6*** (17.78)
Upscale Class	-3.085 (9.486)	12.16 (10.96)	0.128 (13.46)	8.204 (10.83)
Upper Midscale Class	-11.10 (9.808)	10.20 (12.44)	-5.007 (13.21)	10.18 (11.01)
Midscale Class	-21.25* (11.53)	24.34 (15.90)	-12.51 (16.75)	19.82 (13.85)
Economy Class	-16.35 (11.03)	7.201 (14.50)	-8.336 (15.40)	16.67 (12.52)
Observations	244	244	244	244
R-squared	0.815	0.889	0.757	0.814

Note: Distances are included as $\log(1 + mi)$. The full set of included variables includes parent company and market fixed effects, the number of rooms, and a vector of indicators for amenities (swimming pool, bar, restaurant, free WiFi, parking, cable, a fitness center, and a hot tub or spa). Omitted market and class dummy variables are Beaches and Luxury Class. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

FIGURE 4: Two-Dimensional Representation of Hotel Markets



Note: Panel A shows hotel locations by latitude/longitude, while Panel B shows the same set of hotels but arranged in a 2-D embedding. Axes in Panel B have no natural interpretation.

dummies, excluded for brevity).

3.4 Model and Estimation

To demonstrate the incorporation of recommendation data in the typical approach to demand estimation, I employ a standard mixed logit (BLP) demand model estimated via two-step GMM. I model consumers as in Equation 1, where the set of included attributes consists of either a set of observable characteristics \mathbf{x}_{jt}^o or dimensions of the embedding \mathbf{x}_{jt}^r . This yields the typical mixed logit choice probabilities and observed quantities:

$$q_{jt} = M \int_{\beta_i, \alpha_i} \frac{\exp \alpha_i p_{jt} + \mathbf{x}_{jt} \beta_i + \xi_{jt}}{1 + \sum_{k \in \mathcal{J}} \exp \alpha_i p_{kt} + \mathbf{x}_{kt} \beta_i + \xi_{kt}} dV(i)$$

The embedding data consists of the four vectors $\mathbf{x}_{jt}^r = [x_{1,jt}^r, x_{2,jt}^r, x_{3,jt}^r, x_{4,jt}^r]$ introduced in Section 3.3. The characteristics specification takes a set of observables from Table 3: the distances to various landmarks which capture locational preferences, and an indicator for whether the hotel is of high quality (Upscale Class or greater). Linear characteristics are absorbed into hotel-level fixed effects δ_j as they do not vary over time, and seasonality is captured through market-year-month fixed effects. Nonlinear parameters are estimated on price and all elements of \mathbf{x} . 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 $\bar{\alpha} = -0.036$, 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.²³ An implication of the patterns displayed in the embedding (Figure 4) 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.

I construct quadratic differentiation instruments ([Gandhi and Houde \(2019\)](#)) over the m nonlinear terms $l(x_1, \dots, x_m)$.²⁴ 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

²³Taking a similar approach, [Armona et al. \(2024\)](#) use a multiplier of 1.5, while [Farronato and Fradkin \(2022\)](#) use 2.

²⁴There are 4 nonlinear characteristics in the characteristic and embedding specifications aside from price.

instruments $z_{jt} = \left[\sum_k d_{jktl} \times d_{jktl'} \right] \forall l' \geq l$ with market dummies, where $d_{jktl} = x_{l,jt} - x_{l,kt}$. I extend z_{jt} to include interactions with the differences $d_{jk,\hat{p}} = \hat{p}_{jt} - \hat{p}_{kt}$:

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

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.

4 Results

4.1 Demand Model Estimates

Table 4 displays the coefficient results of the logit demand system. 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.²⁵

4.2 Substitution Patterns

The first major post-estimation result of interest is which hotels are close substitutes to each other. Both models produce similar diversion ratios to their closest substitutes, but the set of close substitutes need not be estimated correctly as both specifications are limited. Lacking a ground truth, I instead examine how similar the sets of top substitute hotels are to each other: intuitively, the top substitutes should display similar characteristics that consumers care about. I estimate a probit regression of whether a given rival hotel is one of a reference hotel’s top-3 substitutes by diversion ratio on measures of similarity between the

²⁵Similar work, such as Armona et al. (2024), also encounters these challenges with respect to the identification of non-linear parameters.

TABLE 4: Demand Estimation Results

		Logit	Chars	Embed
β	Price	-0.036	-0.036	-0.036
		-	-	-
Σ	Price	-	0.024 (0.002)	0.025 (0.001)
	x_1	-	0.651 (14.395)	0.307 (1.522)
	x_2	-	0.235 (2.803)	0.766 (2.934)
	x_3	-	0.065 (2.496)	0.330 (2.247)
	x_4	-	0.000 (1.895)	0.627 (2.363)
	Num. Obs.	18,620	18,620	18,620
	Own-price Elasticity	-4.786	-2.748	-2.413
	Outside Diversion	0.503	0.364	0.415
	Inside Diversion	0.006	0.007	0.006
	Markups	0.239	0.355	0.383

Note: For the characteristics specification, x_1 , x_2 , x_3 , and x_4 refer to the natural log of 1 plus the distance in miles to Disneyland, John Wayne Airport, and Newport Beach, and an indicator for upscale-and-higher quality. In the embedding specification, x_1, \dots, x_4 refer to the coordinates of the four-dimensional embedding. Post-estimation statistics are presented as median values of the full sample across products and markets. Markups are computed using hotel parent companies as the ownership structure.

hotels. Table 5 presents results, using the distance between hotels, whether they are of the same (or one-tier removed) quality, and whether they have the same parent company. The top substitutes in the embedding specification see an intuitive and statistically-significant impact of distance (less likely to include) and similar quality (more likely to include), while the characteristics specification—which has coarser measures of product differentiation—does not produce these patterns.

Broadly, both specification identify similar top substitutes: roughly 75% of top-3 substitutes are overlapping between the two specifications. Across classes, these values are 56% (Economy), 66% (Midscale), 76% (Upper Midscale), 88% (Upscale), 91% (Upper Upscale), and 71% (Luxury). The overlap is also higher around key localities (90% for Disneyland hotels) than in broader geographies (just 60% for beach hotels). In Table 6, I present three examples of hotels with substantial non-overlap of identified substitutes. A key observation is that the embedding specification appears to consider different important factors in an intuitive fashion.

TABLE 5: Impact of Factors on Inclusion in Top 3 Substitutes

	Embedding	Characteristics
Distance Between Hotels	-0.030*** (0.005)	-0.007 (0.005)
Same Quality Tier	0.158*** (0.045)	0.088* (0.045)
Adjacent Quality Tier	-0.004 (0.041)	-0.070* (0.040)
Same Parent Company	0.216*** (0.056)	0.214*** (0.056)
Constant	-1.647*** (0.034)	-1.691*** (0.034)
Observations	16,476	16,476
Median Top-3 Diversion	0.032	0.034

Note: Robust standard errors in parentheses.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

In the Economy-class example, the embedding specification does a better job at identifying geographically close substitutes, even if they are of slightly dissimilar quality: given an Economy-class Wyndham, a Midscale Extended Stay located 2.83 miles away is estimated as a top substitute by the embedding specification, while the characteristics specification chooses two alternative Economy-class hotels that are 6.48 and 7.93 miles away. In the Upscale example, the characteristics specification does not identify any other Upscale-class hotels as the closest substitutes, even when there are Upscale options that are very nearby (0.4 miles) or of the same brand: the embedding specification produces both of these as top alternatives. Finally, in the Luxury-class example, the embedding specification finds higher substitution to other Luxury-class alternatives, while the characteristics specification identifies two top substitutes of a different quality tier. These observations suggest that the embedding is better able to weight different characteristics to products where they are most important to respective consumers: proximity for lower-quality, but quality for the higher-quality.

While the true top substitutes are unknown, the patterns discussed above seem intuitive, and the specification that uses a coarse set of observable characteristics does a poorer job at fitting these patterns. However, it needs to be emphasized that the importance of these results is validation (rather than outperformance): if the embedding is able to produce results that pass a sanity check, then that suggests that practitioners can employ this methodology in environments where useful characteristics are completely lacking.

TABLE 6: Sample Top Substitutes by Specification

Hotel	Class	Parent	Substitute	Rival Class	Rival Parent	Distance (mi)
1	Economy	Wyndham	Embed	Economy	Wyndham	1.24
1	Economy	Wyndham	Embed	Midscale	Extended Stay	2.83
1	Economy	Wyndham	Chars	Economy	G6	7.93
1	Economy	Wyndham	Chars	Economy	Wyndham	6.48
1	Economy	Wyndham	Both	Economy	G6	9.22
2	Upscale	Hilton	Embed	Upscale	Marriott	0.40
2	Upscale	Hilton	Embed	Upscale	Hilton	5.70
2	Upscale	Hilton	Chars	Upper Upscale	Hyatt	13.80
2	Upscale	Hilton	Chars	Midscale	Extended Stay	5.27
2	Upscale	Hilton	Both	Upper Upscale	Marriott	11.84
3	Luxury	Marriott	Embed	Luxury	Independent	10.48
3	Luxury	Marriott	Embed	Luxury	Independent	14.71
3	Luxury	Marriott	Embed	Luxury	Hilton	0.45
3	Luxury	Marriott	Chars	Upper Upscale	Hyatt	13.76
3	Luxury	Marriott	Chars	Luxury	Montage	1.38
3	Luxury	Marriott	Chars	Upper Upscale	Marriott	13.32

Note: Hotel A is an Economy-class Wyndham in the proximity of Disneyland. Hotel B is an Upscale Hilton near the beach. Hotel C is a Luxury Marriott near the beach. The specification column indicates whether the rival was flagged as a top-3 substitute by the embedding specification, the characteristics specification, or by both.

5 Discussion

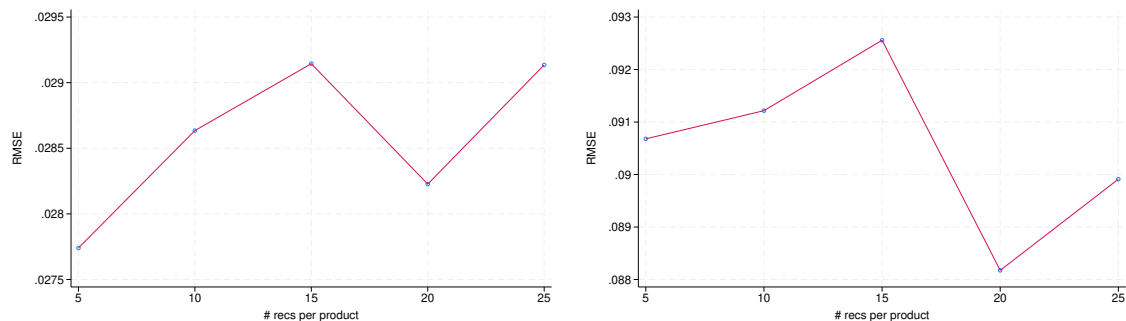
5.1 Number of Recommendations

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. Using the simulated environment from Section 2.3, I compute 4-D embeddings using $R = \{5, 10, 15, 20, 25\}$, and estimate the demand specification to recover diversion ratios.²⁶ Figure 5 plots the RMSE of mean product-rival-level diversion ratios to all products and to the true closest product. There is no clear relationship between the number of recommendations and better capture of substitution patterns: in other words, only a small amount of information about the nearest substitutes is sufficient to construct the embedding. Furthermore, as will be shown in Section 5.3, the magnitude of the differences in RMSE between recommendation counts is substantially smaller than the decision of what

²⁶The simulation in Section 2.3 used $R = 10$ and $m = 4$.

dimension of embedding to use.

FIGURE 5: Diversion RMSE by Number of Recommendations



(A) RMSE of Diversion Ratios

(B) RMSE of Diversion to Closest Substitute

5.2 Biased Recommendations

The practitioner may be concerned that the platform(s) they are using to collect recommendations from are not providing data on the best/closest substitutes. Rather, they may be suggesting products that provide more revenue (for example, a platform may preference its own manufactured products) or are sponsored (consider the case of search results). While empirically the model will treat less informative recommendations as noise if they do not help in estimating the demand system, it is useful to look at how this affects the results. I extend the simulation by selecting five products at random as “privileged”, raising their recommendation ranks. Mechanically, this is done by increasing the diversion ratio of these products by $b = \{0.01, \dots, 0.05\}$ when determining product ranks. Table 7 summarizes the frequency that each privileged product appears in the top 10 substitutes (and is hence one of the 10 recommended alternatives): these products are recommended alternatives less than 20 percent of the time in the true data, but nearly 70 percent at the highest level of bias.

Figure 6 again presents the RMSE of diversion to all products and to the true top product. As the level of bias rises, and hence the set of top products is driven by noise rather than consumer preferences, the model becomes worse at estimating substitution to the top alternative (Panel B). However, this effect is not immediate with the introduction of bias. Privileging some product does not affect the ordinal ranking of all other products, and while the privileged products are a minority of the information from recommendations, the model’s performance holds up. This suggests that small perturbations are not immediately

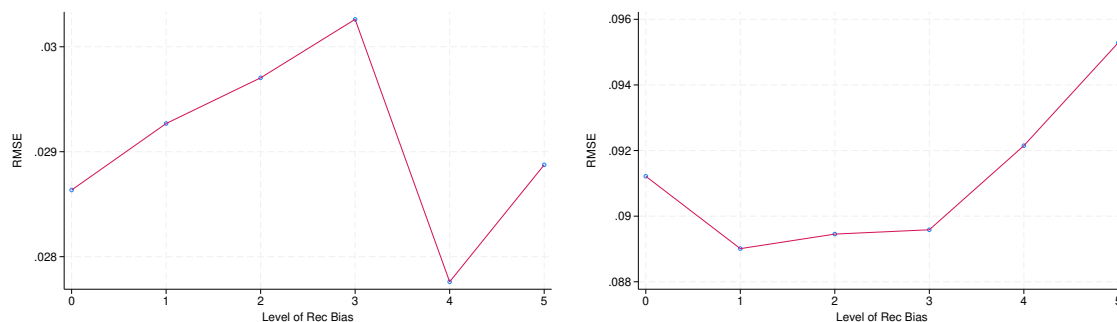
TABLE 7: Frequency of Privileged Products in Top 10 Product Ranks

	Level of Ranking Bias (b)					
	0	0.01	0.02	0.03	0.04	0.05
Product 1	0.18	0.26	0.34	0.50	0.62	0.68
Product 2	0.04	0.06	0.14	0.34	0.44	0.64
Product 3	0.28	0.36	0.44	0.50	0.60	0.72
Product 4	0.42	0.46	0.48	0.56	0.62	0.70
Product 5	0.00	0.04	0.16	0.34	0.50	0.70

Note: Bias of zero indicates the base simulation level. Levels $b = \{0.00, \dots, 0.05\}$ correspond to raising the diversion ratio of the privileged five products by b for the purposes of ranking for recommendations. The embedding includes the top 10 recommended alternatives.

a dealbreaker for the method. When the biased recommendations are the majority of the recommendation information, fit quickly begins to fall. However, this pattern is not observed for diversion ratios to all products (Panel A), as the comparative positioning of all other products with respect to each other is unaffected by privileging a subset of products.

FIGURE 6: Diversion RMSE by Recommendation Bias



(A) RMSE of Diversion Ratios

(B) RMSE of Diversion to Closest Substitute

5.3 Embedding Dimension Selection

One hyperparameter of the model that the practitioner needs to choose is the number of dimensions of the embedding. The underlying problem is a trade-off between reducing the dimensionality of the data for computational reasons versus the potential loss of variation. This question has no obvious answer, and impacts all unstructured data methods. While model selection tools like information criteria or cross-validation are potential solutions, post-estimation model selection may be computationally burdensome. [Compiani and Christensen \(2026\)](#) provide a theoretical basis for dimension selection, while in this section

I suggest some quicker rules of thumb for pre-estimation dimension selection, and compare them to post-estimation substitution results using the simulated environment introduced in Section 2.3.

The spirit of the question is “how few dimensions are needed to capture the variation from the recommendations?” I present two tests. In the first, I evaluate the objective of the tSTE algorithm in fitting Euclidean distances between products. More dimensions allows for better fit to the data: the practitioner can observe when the distances between products settle, such that more dimensions does not improve the fit. Panel A of Figure 7 presents the Froebinus norm between the pairwise product distances of dimension m and $m - 1$. In the second, I check whether other dimension-reduction methods (PCA) can further shrink the characteristics space without sacrificing variation: the smallest dimensionality that cannot be further reduced while retaining a threshold is accepted. Panel B presents the variation retained by $m - 1$ principal components of the m -dimension embedding. In both approaches, the improvement flattens out past six dimensions. The appropriate threshold is an arbitrary decision: embeddings of above four dimensions can be reduced via PCA while retaining over 90% of variation, while embedding of above six can be reduced while retaining 95%. Similarly, the embedding has largely stabilized beyond six dimensions.

FIGURE 7: Dimension Selection Test Results

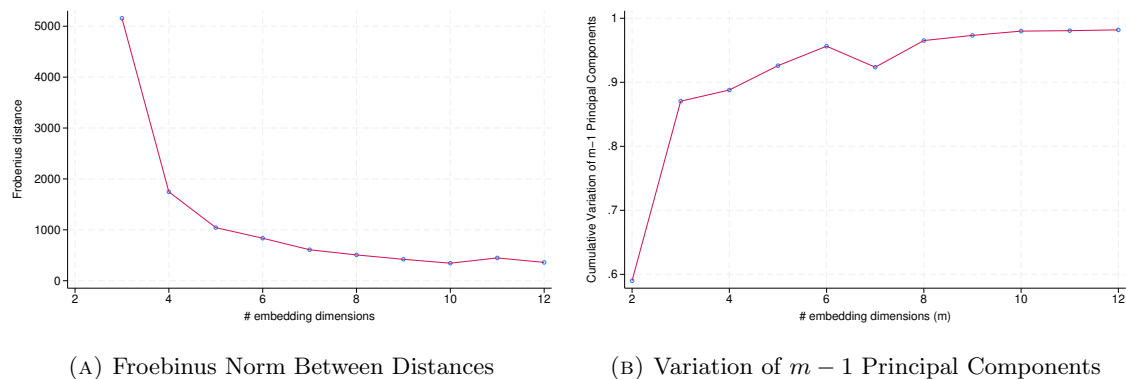
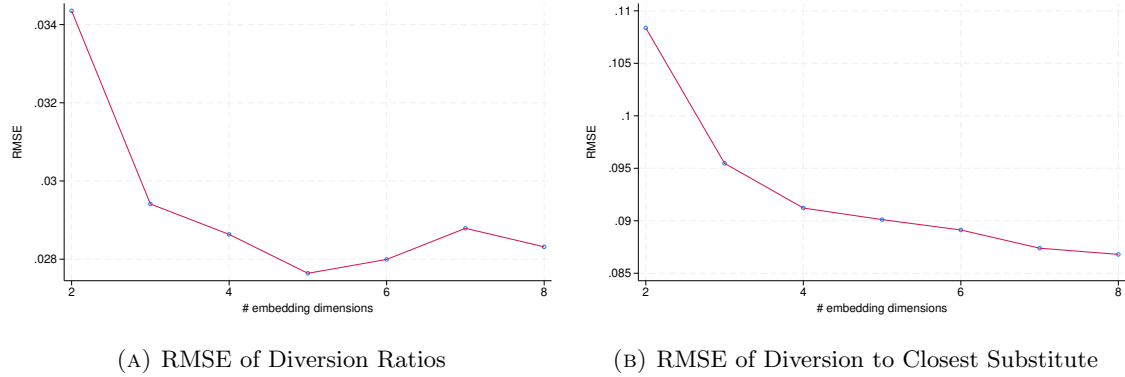


Figure 9 plots the RMSE of estimated average diversion between products across dimensions of the embedding, using the true values from the simulation. I limit the analysis to eight dimensions: the prior figure suggests very little improvement beyond that point. Panel A shows the results across all product pairs, while Panel B looks only at diversion to the true closest substitute. RMSE flattens out at around six dimensions in both cases, suggesting that the pre-estimation measures are doing a reasonable job at showing where dimension

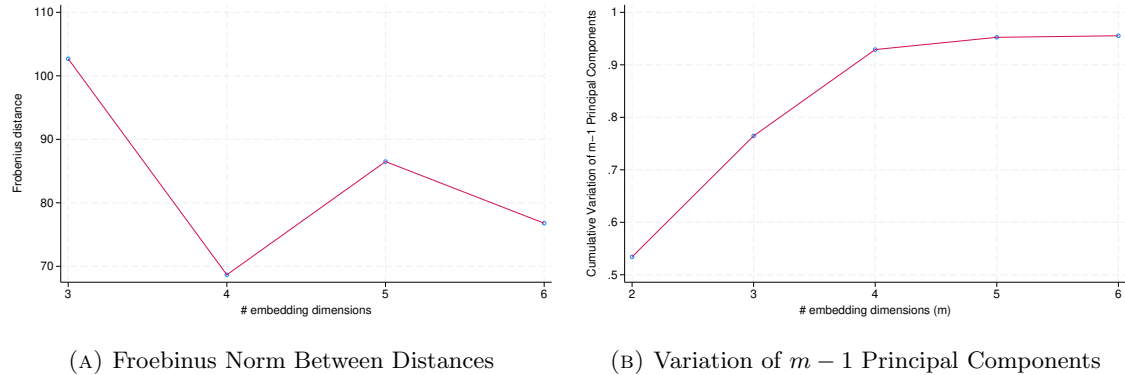
choice is (i) suitable, and (ii) the results are not sensitive to marginal changes in dimension choice.

FIGURE 8: Diversion Ratio Fit by Dimension of Embedding



Lastly, I present similar figures from Table 7 for the hotel data for dimensions up to six. The amount of captured variation appears to level off after four dimensions: subsequent dimensions can be reduced to $m - 1$ principal components while retaining at least 95% of variation. The distance norm is less clear, showing a non-monotonic trend.²⁷ Despite this, the correlation in pairwise distances for dimensions beyond three is approximately 95%, suggesting that the distance function has already settled down and remaining fluctuation is largely noise.

FIGURE 9: Dimension Selection Tests in Hotel Data



²⁷Due to variation in the scale of the distances, I normalize the distances within each m to mean zero, standard deviation 1. This corresponds to the treatment of the characteristics when included in the mixed logit specifications.

5.4 Bias Correction of Unstructured Data

In the methods detailed here, the embedding is treated simply as data, as in any typical method where the product characteristics are taken as given. A point of potential criticism is that the embedding can instead be interpreted as an estimate, with corresponding inference problems if the practitioner ignores any potential bias in estimating the latent characteristics.²⁸ Consider a model where the practitioner wishes to recover γ —the effect of θ on Y —via OLS, but only observes an estimate $\hat{\theta}$. Battaglia, Christensen, Hansen, and Sacher (2024) note that the estimates of parameter $\hat{\gamma}$ are asymptotically normally distributed with the same variance as the true regression, but off-center when the regressor $\hat{\theta}$ is a biased estimate of the regressor θ . In this context, $\hat{\theta}$ is the estimated embedding. They propose a bias-corrected value of the parameter γ :

$$\hat{\gamma}^{bc} = \left(1 + \frac{1}{nC} \frac{\sum_{i=1}^n \hat{\theta}_i (1 - \hat{\theta}_i)}{\sum_{i=1}^n (\hat{\theta}_i - \bar{\theta}_n)^2} \right) \hat{\gamma}, \quad (4)$$

where $\bar{\theta}_n$ is the sample mean of $\hat{\theta}_i$, C is the number of sampled items, and confidence intervals are $\hat{\gamma}^{bc} \pm 1.96$ the OLS standard error.

In a similar fashion, Compiani and Christensen (2026) formalize a diagnostic test of the bias arising from mismeasured characteristics (via unstructured data or from a selection of observables). Given a statistic

$$LM_1 = \|\sqrt{n}\hat{H}^{-1}\hat{S}\|^2, \quad (5)$$

where \hat{S} is the score and \hat{H} the expected Hessian, if LM_1 is below a suitable threshold, then the discrepancy between $\hat{\gamma}$ and γ is sufficiently small.²⁹

²⁸This is not an entirely unreasonable assumption, as any set of chosen product characteristics are typically used to arbitrarily construct a linear function of utility without corresponding inference questions. In this case, the practitioners should be clear about the assumptions made, and that their estimates are conditional on the data.

²⁹ $\hat{S} = \frac{1}{n} \sum_{i=1}^n \sum_{j=0}^J \frac{d_{ij}}{\sigma_{ij}(\hat{\gamma})} \hat{\sigma}_{ij}(\hat{\gamma})$

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 hotels. In 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. 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.
- AMANO, T., A. RHODES, AND S. SEILER (2022): “Flexible Demand Estimation with Search Data,” *Working Paper*.
- ARMONA, L., G. LEWIS, AND G. ZERVAS (2024): “Learning Product Characteristics and Consumer Preferences from Search Data,” *Marketing Science*, 1–18.
- BAJARI, P., Z. CEN, V. CHERNOZHUKOV, M. MANUKONDA, S. VIJAYKUMAR, J. WANG, R. HUERTA, J. LI, L. LENG, G. MONOKROUSSOS, AND S. WANG (2025): “Hedonic prices and quality adjusted price indices powered by AI,” *Journal of Econometrics*, 251.
- BAJARI, P., D. NEKIPELOV, S. P. RYAN, AND M. YANG (2015): “Machine Learning Methods for Demand Estimation,” *American Economic Review: Papers & Proceedings*, 105, 481–485.
- BATTAGLIA, L., T. CHRISTENSEN, S. HANSEN, AND S. SACHER (2024): “Inference for Regression with Variables Generated from Unstructured Data,” *Working Paper*.
- BERRY, S., A. GANDHI, AND P. HAILE (2013): “Connected Substitutes and Invertibility of Demand,” *Econometrica*, 81, 2087–2111.
- 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.
- COMPIANI, G. AND T. CHRISTENSEN (2026): “From Unstructured Data to Demand Counterfactuals: Theory and Practice,” *Working Paper*.
- COMPIANI, G., I. MOROZOV, AND S. SEILER (2025): “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.
- DONNELLY, R., A. KANODIA, AND I. MOROZOV (2023): “Welfare Effects of Personalized Rankings,” *Marketing Science*, 43, 92–113.
- ELROD, T. AND M. P. KEANE (1995): “A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data,” *Journal of Marketing Research*, 32.
- 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.
- FOX, J. T., K. I. KIM, S. P. RYAN, AND P. BAJARI (2011): “A Simple Estimator for the Distribution of Random Coefficients,” *Quantitative Economics*, 2, 381–418.
- 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 (2019): “Measuring Substitution Patterns in Differentiated-Products Industries,” Working Paper 26375, National Bureau of Economic Research.
- GRIECO, P. L., C. MURRY, AND A. YURUKOGLU (2021): “The evolution of market power in the US auto industry,” Working Paper 29013, National Bureau of Economic Research.
- HAN, S., E. SCHULMAN, K. GRAUMAN, AND S. RAMAKRISHNAN (2021): “Shapes as Product Differentiation: Neural Network Embedding in the Analysis of Markets for Fonts,” *Working Paper*.
- HODGSON, C. AND G. LEWIS (2023): “You Can Lead a Horse to Water: Spatial Learning and Path Dependence in Consumer Search,” Working Paper 31697, National Bureau of Economic Research.

- KIM, J. B., P. ALBUQUERQUE, AND B. J. BRONNENBERG (2010): “Online Demand Under Limited Consumer Search,” *Marketing Science*, 29, 1001–1023.
- (2016): “The Probit Choice Model Under Sequential Search with an Application to Online Retailing,” *Management Science*, 63, 3911–3929.
- KUMAR, M., D. ECKLES, AND S. ARAL (2020): “Scalable bundling via dense product embeddings,” *arXiv preprint arXiv:2002.00100*.
- MAGNOLFI, L., J. MCCLURE, AND A. SORENSEN (2025): “Triplet Embeddings for Demand Estimation,” *American Economic Journal: Microeconomics*, 17, 1–26.
- 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 semi-parametric 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 Path-Connection of Recommendations

A caveat to this paper’s approach is that for products to be placed in the same embedding, it is necessary that they be path-connected by the set of recommendations, else no unique distance exists between them.³⁰ Two products j and k are path connected if there exists some pattern of recommendations $j \rightarrow \dots \rightarrow k$ and/or $k \rightarrow \dots \rightarrow j$. Conceptually, this implies a consumer can search from j to k solely by following recommendations. Products not being in a connected set may imply that the products should be in separate markets, or that the recommendations provide insufficient data to characterize similarity over products. In this paper’s application, hotels in the Orange County sample are connected, and hence I will use more traditional geographic market definitions to subdivide the product space.

I define recommendation spaces as topological spaces formed by *path-connected* products. All products within a recommendation space \mathcal{S} are path-connected in recommendations. To construct the recommendation space in the data, 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 $\mathcal{R} \times S' \rightarrow \mathcal{S}$ 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. Separate recommendation spaces \mathcal{S}_1 and \mathcal{S}_2 are *disconnected* if no hotels in set \mathcal{S}_1 contain a recommendation for, or are recommended by, any hotel in set \mathcal{S}_2 (i.e. all products in \mathcal{S}_1 are *path-disconnected* from all products in \mathcal{S}_2).

Appendix B Monte Carlo Details

Consumers in the simulated data have utility given by Equation 1, which is reproduced below:

$$u_{ijt} = \alpha_i p_{jt} + \beta_i [1, x_{jt}^o, x_{jt}^e] + \xi_{jt} + \epsilon_{ijt},$$

There are $J = 50$ products in total. Products have $K = 4$ characteristics aside from the constant, which are distributed as

³⁰This is similar to the property of connected substitutes in [Berry, Gandhi, and Haile \(2013\)](#). I thank a referee for noticing this similarity.

$$\begin{bmatrix} x_1^o \\ x_1^e \\ x_2^e \\ x_3^e \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -0.8 & 0.3 \\ 0 & -0.8 & 1 & 0.3 \\ 0 & 0.3 & 0.3 & 1 \end{bmatrix} \right),$$

such that the first characteristic is observable and the latter three are unobserved. The first characteristic is independent of x^e , while x^e are more impactful covariates of utility, further constraining the information provided by the observed data. Products are assigned randomly to $F = 5$ firms such that each firm has ownership of 5 products. Firm costs are linear with $c_{jt} = 3 + 2w_{jt} + \omega_{jt}$, where w_{jt} is an observable cost shock distributed uniformly on $[0, 1]$ and $\omega \sim N(0, 1)$ is an unobserved cost shock. I draw $T = 1000$ markets, where each market offers a random set of 25 products to provide for choice set variation.

Consumer preferences for $N = 1000$ per market are drawn with an unobservable product-market quality shock $\xi \sim N(0, 1)$, and individual i preferences

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \Sigma v_i,$$

where coefficients on the attributes $(1, p, x_1^o, x_1^e, x_2^e, x_3^e)$ have means of $(5, -2, 0, 0, 0, 0)$ and diagonal elements of $\Sigma = (0.2, 0.2, 1, 2, 2, 2)$.

The practitioner observes a set of recommendations for each product. These are simulated by ranking products by their true average diversion to alternatives, and listing the top 10 alternative products. These ranks are treated strictly ordinal, such that the reference product is deemed closer to the rank 1 alternative than the rank 2 alternative, closer to the rank 2 alternative than the rank 3 alternative, etc. These triplets are then mapped into a 3-dimensional embedding via the tSTE algorithm. The choice of dimension is ad hoc, combining considerations about dimensionality with a desire to capture suitable variation. I select a 4-dimensional embedding to match the dimension of the true $[x^o, x^e]$. This also fits a 90% rule of thumb: the 4-dimensional embedding cannot be reduced to three principal components while retaining 90% of its variation (88.8%). On the other hand, a 5-dimensional embedding could be reduced to four principal components and retain 92.6% of variation, suggesting that the improvement from added dimensions is marginal.





Estimation of the demand system uses either the observable x_1^o or latent characteristics from an embedding $(x_1^r, x_2^r, x_3^r, x_4^r)$. Linear effects are concentrated out using product-level fixed effects, while random coefficients are estimated on the constant, price, and each characteristic. Dimensions of the embedding x^r are normalized to a mean of zero and standard deviation of one in order to improve the initial step of computation. I use a 2-step GMM approach for estimation via `pyBLP` (Conlon and Gortmaker (2020)). As instruments, I include the cost shifter w , as well as differentiation instruments suggested by Gandhi and Houde (2019): I incorporate the quadratic interactions of distances across all attributes ℓ (the cost shifter along with observable attributes) for rival ($k \notin J_{ft} \setminus j$) and nonrival ($k \in J_{ft} \setminus j$) products:

$$z_{jt} = \left\{ w, \sum_{k \in J_{ft} \setminus \{j\}} d_{jkt\ell} \times d_{jkt\ell'}, \sum_{k \notin J_{ft} \setminus \{j\}} d_{jkt\ell} \times d_{jkt\ell'} \right\} \quad \forall \ell' \geq \ell$$

Appendix C Additional Tables and Figures

APPENDIX FIGURE 1: Recommendations at Booking.com

Travelers who viewed **Marriott Marquis Chicago** ended up booking these properties Show more

 <p>Hampton Inn Chicago McCor... 8.3 Very Good 📍 2 miles from center 🌱 Travel Sustainable property Starting from \$458</p>	 <p>Home2 Suites By Hilton Chica... 8.4 Very Good 📍 2 miles from center 🌱 Travel Sustainable property Starting from \$441</p>	 <p>Luxéry Stay Chicago - ACROS... 9.2 Wonderful 📍 1.9 miles from center Starting from \$500</p>	 <p>Hilton Chicago 8.1 Very Good 📍 0.7 miles from center 🌱 Travel Sustainable property Starting from \$544</p>
---	--	--	---

APPENDIX FIGURE 2: Recommendations by Classes

