Using Online Default Recommendation Data for Differentiated Product Demand Estimation

Jonathon McClure*
Mitch Daniels School of Business
Purdue University

August 2024

Abstract

Online recommendation platforms aid consumers in making decisions amidst large choice sets by suggesting commonly-chosen alternatives to a given product. Using observed recommendations for hotels, I construct an embedding of the latent preference space for the mean consumer. Using this information, I estimate typical distance-based linear and mixed logit demand models and attempt to recover substitution patterns in the absence of observed characteristics or consumer demographics. Monte Carlo tests suggest that in environments where unobserved consumer heterogeneity results in poor identification of demand system parameters, using coordinates of the embedding in place or in addition to true characteristics improves estimates of substitution, markups, and merger effects. Using data for hotels in downtown Chicago, I estimate hotel-level diversion ratios and find initial evidence that Upper Upscale-class hotels receive the highest mean diversion ratios from rival hotels, suggesting that the quality tier may be most represented in the choice sets of downtown consumers. Future work aims to compare the price and welfare implications of mergers across alternative data and models.

*Assistant Professor, Department of Economics, Mitch Daniels School of Business, Purdue University. Email: *mcclur47@purdue.edu*. I would like to thank discussants at the Purdue Economics summer seminar series for comments and feedback.

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 or characteristics-based—require data on product differentiation, which is usually paired with either inherent assumptions on aggregated preferences 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. Second, while Census data is often used to characterize local consumers, these data might not be relevant to the actual consumer base, or otherwise insufficient to capture richness of consumer heterogeneity, resulting in poor 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 on alternative products—can be incorporated to augment the estimation of demand systems. 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, Slade, and Brett (2002), as well as a random-coefficients logit model (Berry, Levinsohn, and Pakes (1995), henceforth BLP) where the coordinates of the embedding reflect variation in characteristics *and* the representative consumer's preferences for the product.

I demonstrate 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 base is not local. The recommendations are recovered from public searches on Booking.com, and contain no proprietary information: a 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.

I merge these recommendations with monthly price and quantity data for hotels in downtown Chicago, a market environment with a very 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. I show that the platform data can aid the definition of market sets through the connectivity of recommendations, allowing the downtown market to be separated from confounding collected data on O'Hare hotels. The recovered embedding—visualized in two dimensions—demonstrates clear patterns of grouping luxury hotels closely together, separate from a large cluster of varied hotels likely driven by geographic proximity.

To demonstrate the use of the embedding, I first 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.

Second, I obtain results from a BLP demand system in the Chicago hotel data using the embedding in place of unobserved characteristics and consumer demographics. In this data-limited context, I am able to recover substitution patterns and markups in an environment with over a hundred competing products (hotels). I find that upper upscale hotels capture the highest average diversion from rivals, while within-quality-class diversion is typically second-highest. This might suggest that the upper upscale quality class is better-dispersed across the market, or that the platform skews consumers towards these higher-cost options.

Future work will build on these results with more detailed data and model comparisons.

This paper complements 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 (2023) 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.

This paper's approach is also 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 (2021), 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. A remaining problem 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.

2 Method

2.1 Recommendation Rankings and Recovering Triplets

Consider a consumer who approaches a market unsure of their purchase decision and who searches for options on an online platform. When consumers investigate an option (i.e. click on it to learn more without making a purchase decision), they are presented with a list of similar or recommended products from which to also choose. Part of the platform's product is easing the search problem that consumers face, which in turn makes it more likely that the consumer makes a purchase. As platforms have extensive data on consumers' purchase options and search paths, they are able to tailor these recommendations towards what the consumer is likely to pick if they reject the initially-selected option in favor of an alternative. As the behavior of the platform is in part a black box without evidence from internal data, I make a simplifying assumption as to their objective function.

Assumption 1. *The objective of a recommendation platform is to maximize the probability that an arriving consumer makes a purchase on the platform.*

Assumption 1 is fairly strong and requires further study to wholly validate, but provides an intuitive generalization. While plausible, it is not immediately provable that this is an accurate heuristic for platform behavior in all respects: platforms may approach recommendations in terms of boosting product discovery versus maximizing consumer search precision in terms of local utility maximization, or their recommendations may lean towards steering consumers towards options which provide higher revenue to the platform (Hodgson and Lewis (2024)). A deeper question is what underlying data drives the recommendations that platforms make, and whether these data are in turn generated in part by the outcomes of the platform's recommendations. In this study, as some assumption on platform behavior is necessary, I will treat Assumption 1 as true, noting that my results hinge on its validity. I discuss this topic further in Section 2.5.

The platform's behavior under Assumption 1 is such that for any number of displayed recommendations R and given an initially-clicked option j, the platform wishes to maximize the choice probabilities across displayed options $r \in R$ conditional on j. Assuming consumers focus their consideration on recommendations that they are provided, this in turn maximizes the likelihood of a discrete choice being made across a distribution of consumer

i preferences $g(\theta)$:

$$\begin{aligned} R(j) &= \arg\max_r \sum_{r \in R(j)} Pr(r|j) \\ &= \int \frac{\exp u_{ij}}{\sum_{k \in \mathcal{J}} \exp u_{ik}} \sum_{r \in R} \frac{\exp u_{ir}}{\sum_{k \in \mathcal{J} \setminus j_i} \exp u_{ik}} g(\theta) d\theta \end{aligned} \tag{1}$$

Given a mixed logit data generating process, the result of this is that recommendations display a rank-order of aggregate second choices, which Conlon et al. (2023) write as:

$$D_{j \to k} = \sum_i \pi_i \frac{s_{ik}}{1 - s_{ij}} \cdot \frac{s_{ij}}{s_j},$$

given probability weights of π on consumers i and market shares s.[1]

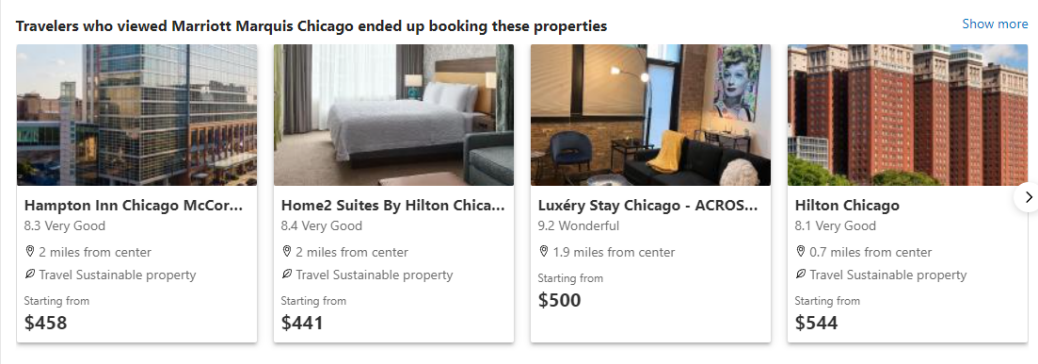
Based on Assumption 1, if product k is the top recommended substitute for product j, then k is the most-likely alternate choice for the representative consumer who searched for j. Hence, $\mu_{k|j} > \mu_{\ell|j} \;\; \forall \ell \neq j, k \in \mathcal{J}$ given mean utility μ for the products: k provides the highest choice probability on average, conditional on being interested in searching for j. Extending this to cardinal utility, the implication of Assumption 1 is that if k is the top recommendation for j, then $||\mu_j, \mu_k|| < ||\mu_j, \mu_\ell|| \; \forall \; \ell \neq j, k \in \mathcal{J}$ for some distance metric on mean utility. This logic extends for each of the subsequently-recommended alternatives, such that the platform provides substantial information on which products are considered closest substitutes for the consumer who demonstrates interest for given products (Compiani, Lewis, Peng, and Wang (2022) note that consumers search in order of observed utility).

Figure 1 shows an example of the recommendations when accessing a hotel on Booking.com. Most hotels include a panel which includes up to 7 suggested alternatives which are commonly booked by travelers who viewed the initial hotel. A consumer clicks on a hotel, and is presented with the hotel's details and a list of alternatives:

By scraping hotel suggestions—in this case, performed via manually recording the suggested alternatives when exploring each hotel's page on the platform—from Booking.com, 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

[1] I briefly discuss the incorporation of recommendations as second-choice data in Section 2.4.

FIGURE 1: Example of Recommendations Page



short, incognito searches in order to capture the recommendations presented without bias for search history. I use this to construct triplets: data points of "product A is closer to B than it is to C" using a 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. (2021), who consider that if consumers search products j_1, j_2 in order, then the products mush have related attributes. However, they 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, as Assumption 1 allows the argument that platforms are incentivized to portray accurate information based on past consumer purchase decisions.

2.2 Triplet Embeddings

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. We use the t-distributed Stochastic Triplet Embedding (tSTE) algorithm proposed by Van Der Maaten and Weinberger (2012). 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

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

Previous studies have made use of survey or recommendations data with the explicit goal of construction the *product space* (Magnolfi et al. (2023)), 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 recovers the *preference space*: these triplets 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.

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. Additionally, given that I treat recovered information from the embedding as data, there may be additional questions about inference, as discussed by Battaglia et al. (2024).

I am not aware of strict rules for the selection of the hyperparameter m (the number of dimensions in the embedding). Magnolfi et al. (2023) 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 general, having a wider set of recommendations is better, as it provides more information for the relative positioning of products. 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)). Section 5 suggests, however, that larger sets of recommendations are useful for framing local subsets with respect to each other, improving estimation of diversion to the outside option.

2.3 Path-Connectedness of Recommendations

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, then 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.

I define recommendation spaces as topological spaces formed by reciprocal bonds between hotels when one is recommended from the other. Separate recommendation spaces S_1 and 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. 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 $\mathcal{S} = \mathcal{R}'\mathcal{R}$ is a matrix of recommendation spaces, where each value $\mathcal{S}_{ij} > 0$ denotes that hotels i and j are in the same recommendation space.

Assumption 2. *If hotels 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.*

Hotels which are not path-connected and are thus in separate recommendation spaces can be treated as 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. Assumption 2 allows the practitioner to augment their understanding of the product space with additional information: for example, separating

hotels by market, or identifying which products are separable in utility. In this paper's application, sub-markets are not observed in the data within the large Chicago MSA, and so this method is used to clarify geographic markets.

2.4 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 $r \in \mathcal{R}_j$.

2.5 Platform Bias

In Section 2.1, I detail how I make the assumption that recommendations steer consumers to the most commonly chosen alternatives—the closest in preference space—with the caveat that at this point I am unable to test this hypothesis for platform conduct. This potentially overlooks how the objectives of the platform may bias the results that they provide to their consumers. An open question is whether platform conduct can be identified by the data used in this paper: in other words, in the absence of micro-data that allows for observable individual consumer search patterns.

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 Data and Embedding

In this section, I detail the price and quantity data for hotels as well as the collected recommendations which I use to build the embeddings.

3.1 Product Space

The first source of data is a panel of hotel-level monthly average daily rates (ADR) and occupancy rates from Chicago, provided by STR LLC, a common source for studies of the hotel sector. The data cover a period of 2010 through 2018. Hotels in this sample are anonymized, but listed by a consistent identification code which allows for consistent representation in the data. Additionally, I observe 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.

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.

My second source of data is scraped hotel recommendations from Booking.com. I collect up to alternate-product recommendations from hotels in two markets: the primary sample is downtown Chicago, though I also collect recommendations from hotels near O'Hare. As these are both anonymized and grouped into Chicago jointly in the STR data, it provides a demonstration for recovering separate markets as the recommendation sets are path-disconnected. In total, I collect recommendations for 180 hotels: when limiting to those which have at least six recommended alternatives and which appear in the STR data, this falls to 132 unique hotels. These recommendation rankings are converted to triplets and used to estimate an embedding using the tSTE algorithm.

3.2 Recommendations and Embeddings

Taking the 132 hotels which have at least six recommendations, I identify the connected sets S_1, S_2 which emerge from the patterns of recommendations. As I can observe the location indicator flags for the hotels in the sample, the data reveal that set 1 is entirely downtown hotels while set 2 is wholly airport hotels: the downtown and O'Hare markets are wholly separated in that platforms will never recommend one when searching for the other. In total, my data set separates into 108 hotels in downtown Chicago and 24 hotels near O'Hare.

TABLE 1: Data Summary Statistics

Class	Area	Hotels	ADR	SD	Occupancy	SD
Luxury	1	21	269.86	90.02	72.09	16.06
Upper Upscale	1	46	180.28	49.46	74.93	16.66
Upscale	1	25	165.47	51.04	76.66	15.26
Upper Midscale	1	12	168.23	44.58	78.91	14.35
Midscale	1	3	128.89	35.79	74.87	19.90
Economy	1	1	105.60	25.99	50.39	22.75
Luxury	2	1	135.90	18.42	67.81	14.76
Upper Upscale	2	11	135.33	22.79	70.88	14.25
Upscale	2	8	108.91	19.19	75.49	13.14
Upper Midscale	2	2	99.70	26.56	75.66	12.50
Midscale	2	2	66.28	15.40	73.32	12.85

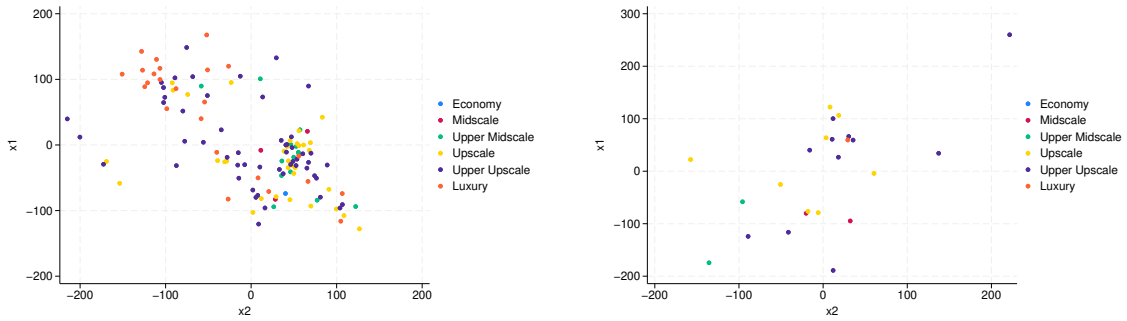
Area 1 corresponds to downtown Chicago, while Area 2 is O'Hare.

Figure 2 displays the embedding on hotel recommendation rankings computed in 2 dimensions for each of the separated markets. This is a highly limiting number of dimensions, but is a useful way of visualizing the output. While individual hotels are not named – and hence their identities and locations are unknown – there are a few intuitive clusters. In the downtown market, there is a concentrated cluster of luxury hotels, as well as a broader, more general close grouping that might indicate geographic proximity. O'Hare hotels are less tightly defined, though a central cluster of upscale and upper upscale hotels suggests a similar grouping of similar properties. Visually, this provides some evidence that the embedding is able to capture similarities between hotels and represent them in a metric space.

As each embedding is independent of the other, it creates a challenge in that the distances or characteristics of each market are not comparable.[2] I proceed by estimating demand

[2]If the set of products was constant across markets, such as in a more typical consumer goods scenario,

FIGURE 2: Plot of Embedding in 2 Dimensions



(A) Embedding of Downtown Chicago Hotels

(B) Embedding of Chicago O'Hare Hotels

solely on the downtown Chicago market where the product differentiation is denser.

A possible criticism is that as preferences include price sensitivity, the embedding vectors X—and hence the representation of the preference space—is correlated with the prices of the hotels themselves. Hotels exhibit clear vertical differentiation in quality, which positions hotels at various price points, and hence price variation might affect the layout of the preference space in ways the embedding restricts.[3] As a defense, I argue that hotel price variation can be understood in two ways: prices that correspond to the quality tier and are part of product differentiation, and price variation linked to short-term supply and demand shocks. The former is exogenous and linked to the product space: hotel quality is defined outside of the model, does not vary over time, and is linked to the hotel's average price.[4] The latter price variation is endogenous and what I aim to use to recover substitution. In other words, short-term price fluctuations drive substitution between hotels, but should not adjust how hotels are perceived respective to each other and hence not impact the embedding. However, where this is a concern to the practitioner, one possibility is to condition the vectors of X on prices by including the average price of each product as a vector in a mixed embedding, allowing each other vector to reflect differentiation that is not price-related.

this would not be an issue.

[3] This contrasts with studies that focus on recovering the product space independently of price: Magnolfi et al. (2023) explore ready-to-eat cereals which are primarily horizontally differentiated in terms of brands.

[4] STR defines the chain scale—my chosen quality metric—of a hotel by "grouping branded hotels based on average room rates."

4 Model and Empirical Strategy

In this section, I outline how the distances computed by the embedding can be used in typical product-based approaches to demand estimation, before moving to using the embedding coordinates for characteristics-based logit approaches. As discussed in Section 2.2, I treat the embedding of the preference space as a suitable representation of the product space given the acknowledgement that the estimated parameters no longer have a clean interpretation of consumer preferences over distances or product attributes. Instead, they act to best fit the model to the data, and so the focus of each model is their estimated substitution patterns or other post-estimation statistics.

4.1 Distance-Based Model

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:

$$\log(q_{jt}) = \alpha_0 + \alpha_1 \log p_{jt} + \sum_{k \neq j} f(d_{jk}; \beta) \log p_{kt} + e_{jt}, \tag{2}$$

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.

I take a simplified function of distance and define $f(d_{jk})$ as:

$$f(d_{jk}) = \frac{\beta}{1 + \frac{d_{jk}}{\max ||d||}} \text{ and } d_{jk} = \left(\sum_{i=1,\ldots,5} (x_{ij} - x_{ik})^2 \right)^{0.5} \tag{3}$$

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.[5]

Identification of the endogenous prices is a challenge regardless of the demand specification: as Armona et al. (2021) note, cost shifters are not readily available in hotel data.[6] I instead use the implicit supply shifter of the presence of hotel capacity constraints. Farronato and Fradkin (2022) discuss the use of this source of identification: the impact of demand shocks is larger when the shock is large relative to available capacity, as hotels ration their finite capacity dynamically.[7] I compare the exogenous variation in quantities—predicting quantities based on hotel and market fixed effects—to $\bar{q}_j$, the capacity of hotel j:[8]

$$z^p_{jt} = \frac{\hat{q}_{jt}}{\bar{q}_j} \text{ where } log(\hat{q}_{jt}) = \hat{\tau}_j + \hat{\tau}_t \tag{4}$$

Hence, controlling for observable demand variation, the interaction of exogenous variation in demand with the excluded number of rooms $\bar{q}$ is itself excluded:

$$E(e' z^p | x_{jt}) = 0 \tag{5}$$

I discuss the validity of the instrument and detail F-test statistics in Appendix A. Rival prices are instrumented by including z^p_k in the same functional form as $log(p_k)$. For simplicity, I estimate a common own-price elasticity α. Equation 2 is estimated by 2-stage least squares, using a vector of hotel and year-month fixed effects for α_0.

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

[6] While Armona et al. (2021) calibrate their logit price parameter to avoid this issue, an alternative solution is proposed by Lewis and Zervas (2019), who jointly model the monopolist's supply-side problem.

[7] See Cho, Lee, Rust, and Yu (2018) for a detailed description of dynamic pricing in the hotel sector.

[8] A meaningful extension is to pair the price and quantity data with additional observable market or hotel characteristics which may predict demand, providing more variation in the instrument.

In environments where the it is challenging to find good instruments for own and rival prices—hotels being one such example—an alternative approach is a spatial autoregressive model (SAR) which has seen other use in real-estate contexts. The spatial endogeneity of prices can be account for through controlling for spatial correlation, constructing instruments based on pairwise connectedness (see Kelejian and Prucha (1998) and Kelejian and Prucha (2010)). Further work will incorporate such a model given the challenge of finding appropriate instruments and natural spatial element in the hotel context.[9]

4.2 Mixed Logit Demand Model

Treating the coordinates of the embedding as latent characteristics, I apply a mixed logit demand model in the style of Berry et al. (1995), writing consumers as taking a discrete choice over hotels j in area-month market t:

$$\begin{aligned} u_{ijt} &= \alpha_i p_{jt} + x_{jt}\beta_i + \xi_{jt} + \epsilon_{ijt} \\ (\alpha_i, \beta_i) &= (\alpha, \beta) + \Sigma v_i \end{aligned} \tag{6}$$

Hotel average monthly prices are denoted p_{jt}, and hotel exogeneous 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.

The outside option—making no purchase, or staying in a hotel not included in the data—is normalized to $u_0 = 0$. I define the market size as in Farronato and Fradkin (2022): the market size M is constant and equal to $2 \times \max_t \sum_j q_{jt}$.

As hotel characteristics do not change over time in my sample, I concentrate non-price linear characteristics into hotel and market fixed effects. Letting $V_{ijt} = \alpha_i p_{jt} + x_{jt}\beta_i + \xi_{jt}$, I can then write quantities as Equation 7, given a distribution over consumer preferences $G(i)$:[10]

$$q_{jt} = M \int \frac{\exp(V_{ijt})}{1 + \sum_k \exp(V_{ikt})} \, dG(i), \tag{7}$$

[9] Siebert and Zhou (2024) provide an example of using a SAR in the context of housing demand.
[10] I simulate the integral over $G(i)$ using 1,000 Halton draws.

To define hotel "characteristics," I make use of the coordinates of the embedding, expressed in $m = 5$ dimensions. Accounting for a constant and price, using the embedding results in a total of 7 nonlinear characteristics with random coefficients. The coordinates of the embedding—as mentioned earlier—provide differentiation in *preference* space rather than strictly product space. 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.

To construct instruments, I use the same price instrument z^p as defined previously as a proxy cost shifter. I also construct quadratic differentiation instruments (Gandhi and Houde (2023)) over the five nonlinear terms l $(x_1, \ldots, x_5)$, where $d_{jktl} = x_{l,jt} - x_{l,kt}$:

$$z_{jt} = \left[z^p_{jt},\ \sum_k d_{jktl} \times d_{jktl'} \right] \ \forall\ l' \geq l$$

I incorporate price variation by constructing a measure of the exogenous variation in price $\hat{p}_{jt} = E[p_{jt}|x_{jt}, z_{jt}]$, and extend z_{jt} to include interactions with differences $d_{jk,\hat{p}} = \hat{p}_{jt} - \hat{p}_{kt}$:

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

The column vectors of the instruments 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)).

5 Monte Carlo Examples

Through two Monte Carlo tests, I provide preliminary evidence that my proposed method improves the estimation of substitution patterns when the product space is unobserved, and when the true demand system involves substantial heterogeneity in preferences. In the following two sections, I detail two examples of mixed logit data-generating processes. 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 $\mathcal{D}$, aggregated to the level of $\mathcal{D}_{jk}$.[11] [12]

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

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.

5.1 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 prefer-

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

[12] See: Conlon et al. (2023).

ences. 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.

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 6 and firms competing via Bertrand-Nash. Products $j \in \mathcal{J}$ have a constant, a price, and six exogenous characteristics 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].

A first question is to what degree having more or fewer recommendations matters for the results. 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 2, where distances are written as:

$$f(d_{jk}) = \sum_{r=0,\ldots,3} \beta_r \left(\frac{d_{jk}}{\max ||d||} \right)^r \text{ and } d_{jk} = \left(\sum_{i=1,\ldots,6} (x_{ij} - x_{ik})^2 \right)^{0.5} \tag{8}$$

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.[13] To select K, I apply a rule of thumb from Magnolfi et al. (2023), 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.[14] 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

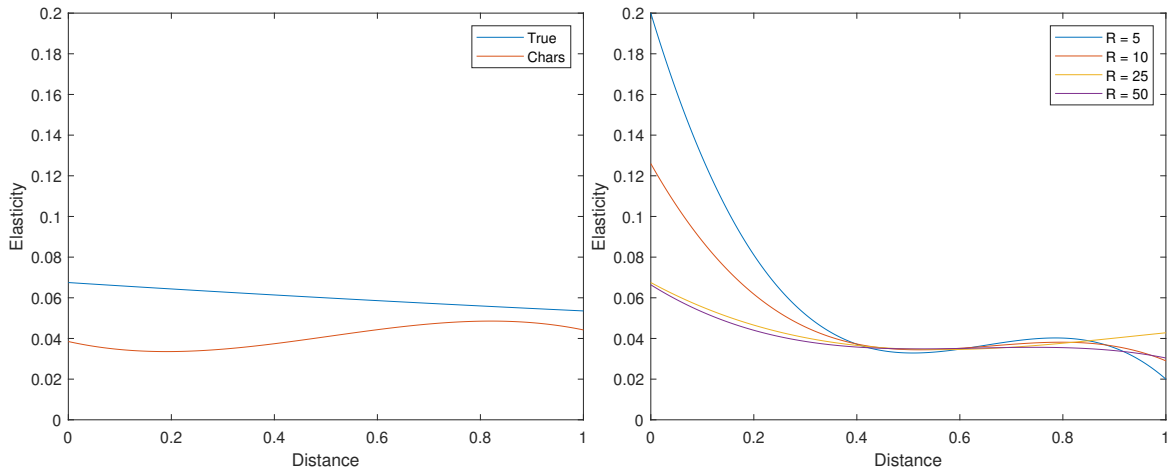
[13] All embeddings use the tSTE algorithm with a convergence threshold of $1e - 7$.

[14] Appendix Figure 2 displays the values across $m = \{2, \ldots, 12\}$.

and thus each level of R selects $m = 6$.

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.

FIGURE 3: Estimated Cross-Price Elasticity Function



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 - 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 2 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.[15] Furthermore, this method is

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

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 2: 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 $\mathcal{D}_{jk}$.

Table 3 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.

5.2 Unobserved Consumer Demographics

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

item.

TABLE 3: 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 2.

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):

$$u_{ijt} = x_{jt}\beta + \alpha p_{jt} + \lambda d_{ijt} + \xi_{jt} + \epsilon_{ijt} \quad \text{where } d_{ijt} = \left(\sum_{kt} (B_{ikt} - x_{jkt}^2)^2 \right)^{0.5} \tag{9}$$

given α, $\lambda < 0$ 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.[16] All specifications include product-level fixed effects. Instruments are the same: the cost shifter and differentiation IVs.

Table 5 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, markups, and predicted out-of-sample market shares.[17] When both sets of information are available, incorporating both in a mixed embedding further reduces errors.[18]

TABLE 4: 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

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

[17] 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 $\xi = 0$.

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

not available, both in terms of predicting substitution patterns and their implications for relevant counterfactual analysis.

TABLE 5: 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 $10 \rightarrow 9$ merger across all simulated markets.

6 Results

6.1 Distance-based Approaches

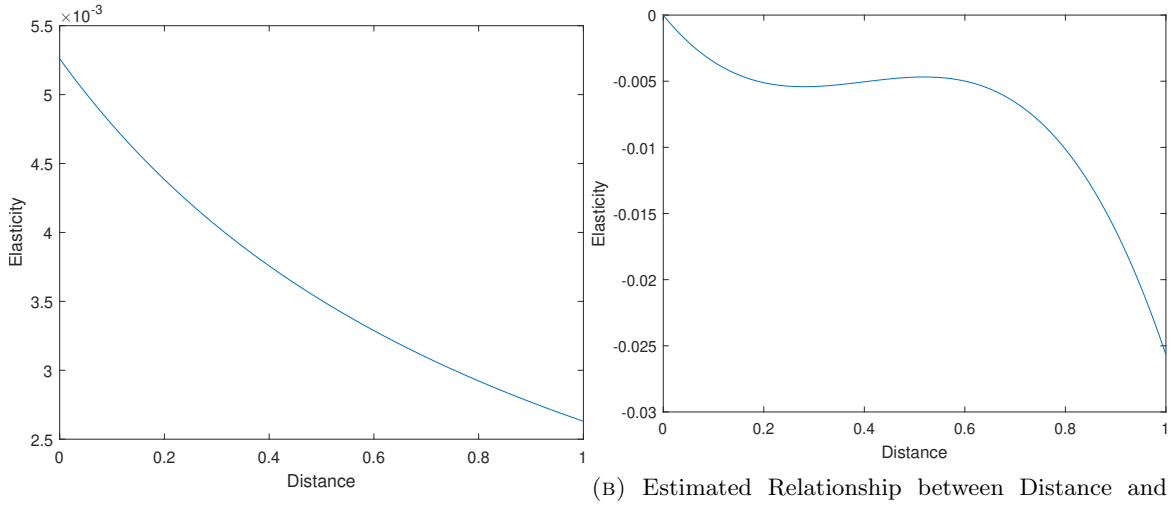
I estimate two specifications of the log-linear approach. The first is the nonlinear form $f(d_{jk}) = \frac{\beta}{1+d_{jk}}$ as in Equation 3, normalizing d_{jk} by the maximum distance between products such that it is bounded by $(0, 1]$. The structure of the elasticities is kept simple (e.g. one single α parameter, one term for $f(\cdot)$) in order to focus on the concept and minimize identification challenges: in practice, more complex methods for finding appropriate specifications for $f(\cdot)$ can be used. The second is a cubic function of distance $f(d_{jk}) = \beta_1 d_{jk} + \beta_2 d_{jk}^2 + \beta_3 d_{jk}^3$. This more flexible form is inhibited by the collinearity of the usual constant term of the cross-price elasticity d_{jk}^0 with the fixed effects. As such, it is solely useful to prove the decreasing relationship between elasticity and distance but not to recover cross-price elasticities. As the estimates rely heavily on the limited variation present in the price instrument, further work will expand on the identification strategy or explore alternative approaches.[19]

Appendix Table 2 shows the parameter results of the two specifications. In each case, the own and cross-price elasticity parameters are statistically significant. Own-price elasticities

[19] See Lewis and Zervas (2019) for one example of modeling the monopolist's supply and demand problem for hotels jointly. Given the spatial context, a SAR may also be relevant in reducing reliance on price instruments.

are not unreasonable—though potentially quite low for a major downtown area—and cross-price elasticities are decreasing in distances between products. The variation of elasticities with distance are presented in Figure 4: Panel A shows the value of $\frac{\beta}{1+d}$ for values of $d \in [0, 1]$. Panel B reflects the change in cross-price elasticity from some absorbed-by-fixed-effects baseline using Specification 2: the function is flat for much of the variation in distance, suggesting that either the linear distance-based model is not doing well at differentiating competitors purely by distance, or the model struggles to identify cross-price elasticities. Given the noted challenges in identifying price effects in hotels and the necessity of instruments for the prices of rival products, this latter case is plausible.

FIGURE 4: Estimated Cross-Price Elasticity Function



(A) Estimated Cross-Price Elasticity Function

(B) Estimated Relationship between Distance and Changes in Cross-price Elasticity

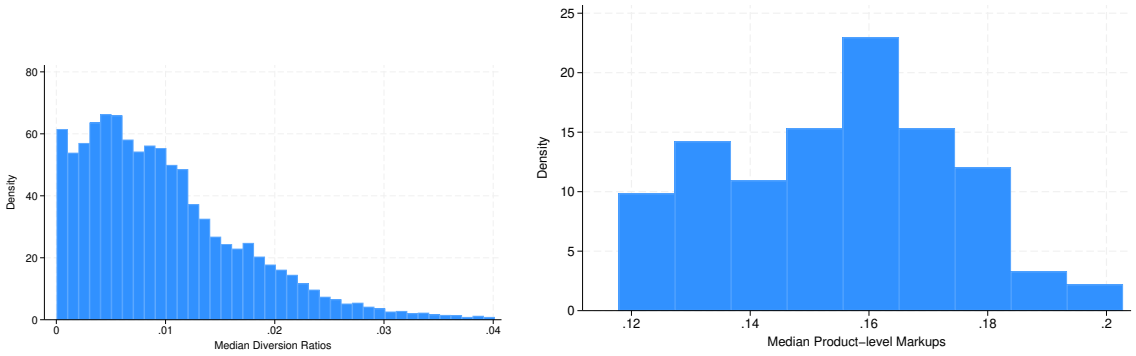
6.2 Using the Embedding as Characteristics

Appendix Table 3 displays the estimated coefficients of the BLP specification. I obtain statistically-significant values for both (β, Σ) on the price parameter. The median own-price elasticity is similar to estimates obtained by Farronato and Fradkin (2022) in other major cities - the focus on a large, downtown area affirms that the highly-elastic demand is sensible. Figure 5 presents distributions of the estimated median markups and diversion ratios.[20] Median diversion to the outside option is 0.07, a particularly low value given the 50%+ share of the outside option, which in turn is a positive sign that the model avoids pitfalls related to IIA substitution patterns. Intuitively, most hotel guests are not

[20] Outliers beyond the $[5, 95]$ percentiles are dropped.

substituting from a hotel choice to not staying at all due to price fluctuation: a confirmed trip leads them to exclude the "no-purchase" outside option entirely, so this low value is reasonable to obtain and fits with the Monte Carlo results that recommendations can improve estimates of outside diversion. Median markups of 0.16 are not easily compared to data: marginal costs recovered by the model are a function of both accounting costs and economic costs stemming from the underlying dynamic process of price-setting (see Cho et al. (2018)). Markups from the model are hence understated relative to industry statistics

FIGURE 5: Estimated Diversion and Markups



(A) Estimated Median Diversion Ratios

(B) Estimated Median Markups

In Table 6 I list the own-price elasticity by hotel class, as well as the mean diversion ratio from a hotel in the row's class to one in the column-labeled class. Mean own-price elasticities are -7.17, which is not overly surprising given the high number of nearby hotels may lead to consumers who can afford to be price-sensitive in the downtown area. The estimated own-price elasticities are decreasing in magnitude with respect to hotel quality, another intuitive conclusion as luxury customers are likely to be less price-sensitive versus consumers of lower-quality hotels.

Another way of validating the recovered substitution patterns is to compare the average product-level diversion ratios across known categories: in this case, hotel quality tiers. In general, the diversion ratios suggest that diversion is primarily to other hotels in the same class, or to those in adjacent classes. Upper upscale hotels have, on average, a diversion ratio of 1.2% to another upper upscale hotel, which is higher than to any other class. The diversion ratio also falls with distance to the hotel's own class. This pattern begins to lose coherence with upper midscale and midscale hotels, though this may have simple explanations: the focus on the high-demand downtown Chicago area suggests that the cheaper hotels may be more scattered and face closer competition from other hotel classes.

The upper upscale class also receives an outsized share of diversion: as the most-represented class in the sample, these hotels may simply see greater representation across the choice sets of consumers.

TABLE 6: Estimated Elasticities and Diversion

	Own Elasticity	Luxury	Upr Upsc	Upscale	Upr Mid	Midscale
Luxury	-3.708	0.018	0.021	0.009	0.003	0.001
Upper Upscale	-7.106	0.003	0.012	0.006	0.002	0.001
Upscale	-7.433	0.005	0.022	0.011	0.004	0.001
Upper Midscale	-7.848	0.009	0.044	0.024	0.009	0.003
Midscale	-8.012	0.029	0.181	0.107	0.037	0.016

Own-price elasticities are the simple average of computed elasticities in that category. Diversion ratios are the average diversion ratio from a hotel in the row-labelled class to a hotel in the column-labeled class. For example, the average diversion between luxury hotels and upper upscale hotels is 2.1%. Economy hotels are omitted as there is only one in the sample.

The primary planned extension to the paper is to further test the results obtained in Section 6 by comparing the post-estimation statistics to alternative demand models. I look to estimate out-of-sample fit, diversion ratios, and markups for hotels in the sample using other common models (i.e. logit, the monopolist model used by Lewis and Zervas (2019), and the aggregated category demand used by Farronato and Fradkin (2022)) to improve the sense of how much the method provides for practitioners. This would also include comparing price and welfare estimates from merger simulation, in order to put the results in the context of common counterfactuals employed in the IO literature.

7 Conclusion

In this paper I discuss a generalizable approach to the collection and incorporation of publicly-available and easily-collected data on default recommendations for demand estimation. Applications of this approach discussed for augmenting common IO demand estimation models, such as linear distance-based demand and more complex mixed logit approaches. I show how this method pertains to two questions: first, can we make use of the information provided by default recommendations in order to help estimate demand, by placing products in utility space when the researcher does not have access to useful data on product characteristics or consumer preferences (e.g. search, second choice, etc)? Second and more specifically, can the aforementioned information help estimate some of the heterogeneity in preferences due to unobserved consumer information by positioning hotels

in utility space?

I provide preliminary evidence as to the applicability of the method: in two Monte Carlo experiments I show that incorporating a preference space constructed from recommendations 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 a set of over 100 hotels in downtown Chicago and recover substitution patterns and markups. Further work will attempt to validate these estimates through additional counterfactuals such as merger simulation, and by comparing the results of these tests to other demand models.

Beyond this application, this approach suggests promise in similar settings where the large product space makes survey-based approaches challenging but existing consumer recommendation and search tools such as platforms operate. Unlike other studies that have made use of machine learning and consumer surveys or search data, this approach is low-cost in terms of data acquisition, providing a useful alternative for practitioners in the field.

References

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 (2021): "Learning Product Characteristics and Consumer Preferences from Search Data," in *Proceedings of the 22nd ACM Conference on Economics and Computation*, 98–99.

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., 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.

Cho, S., G. Lee, J. Rust, and M. Yu (2018): "Optimal Dynamic Hotel Pricing," *Working Paper*.

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., G. Lewis, S. Peng, and P. Wang (2022): "Online Search and Product Rankings: A Double Index Approach," *Working Paper*.

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 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*.

Kelejian, H. H. and I. R. Prucha (1998): "A Generalized Spatial Two-stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances," *Journal of Real Estate Finance and Economics*, 17, 99–121.

——— (2010): "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances," *Journal of Econometrics*, 157, 53–67.

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 (2023): "Triplet Embeddings for Demand Estimation," *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.

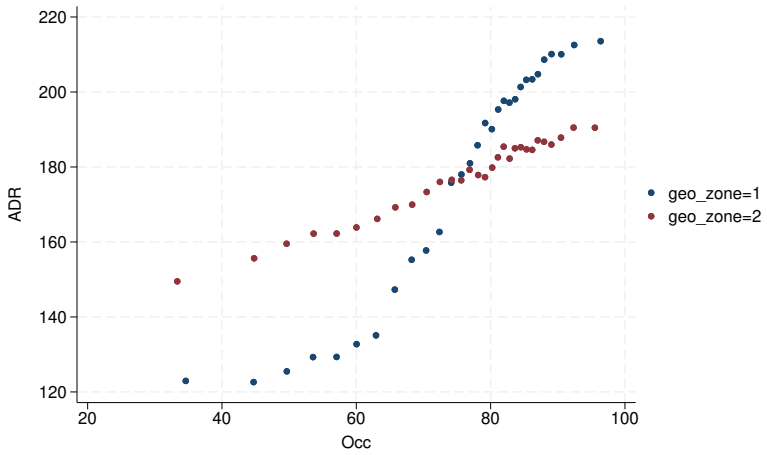
Siebert, R. and X. Zhou (2024): "Spatial Competition: Evidence from the Real Estate Market," *working paper*.

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 Instruments

If the assumption that the relationship between demand shocks and prices steepens when hotels are nearer to full capacity holds, it should be visible in the data. Figure 1 plots a binned scatter-plot of this relationship for hotels in each of the two markets, controlling for hotel-level fixed effects. The steepening relationship is consistent with what is observed in downtown Chicago, but not near O'Hare, suggesting that the identification strategy may be valid downtown but not in the O'Hare market.

Appendix Figure 1: Relationship Between ADR and Occupancy



As a robustness check, I estimate a linear regression of log quantity on log price separately on each market area and employ a weak instrument test for the case of 1 endogenous regressor and 1 instrument:

$$\log q_{jt} = \alpha_0 + \alpha_1 \log p_{jt}, \tag{10}$$

using z^p_{jt} to instrument for p_{jt}. The own-price elasticity coefficient, its standard error, and the Montiel-Pfleunger Effective F-statistic for downtown Chicago are -3.23 (0.18) and 419.4, while near O'Hare they are 26.03 (24.89) and 1.053. Additionally, when incorporating the price instrument in a logit model with one endogenous regressor where $u_{jt} = x_{jt}\beta + \alpha p_{jt} + \epsilon_{jt}$ and making the same exclusion restriction, I obtain an Effective F-statistic of 280.7 for downtown Chicago.

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:

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].

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:[21]

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

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:

$$p - c = \left(-\frac{\partial s(p)}{\partial p} \cdot \Omega\right)^{-1} s(p) \tag{11}$$

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

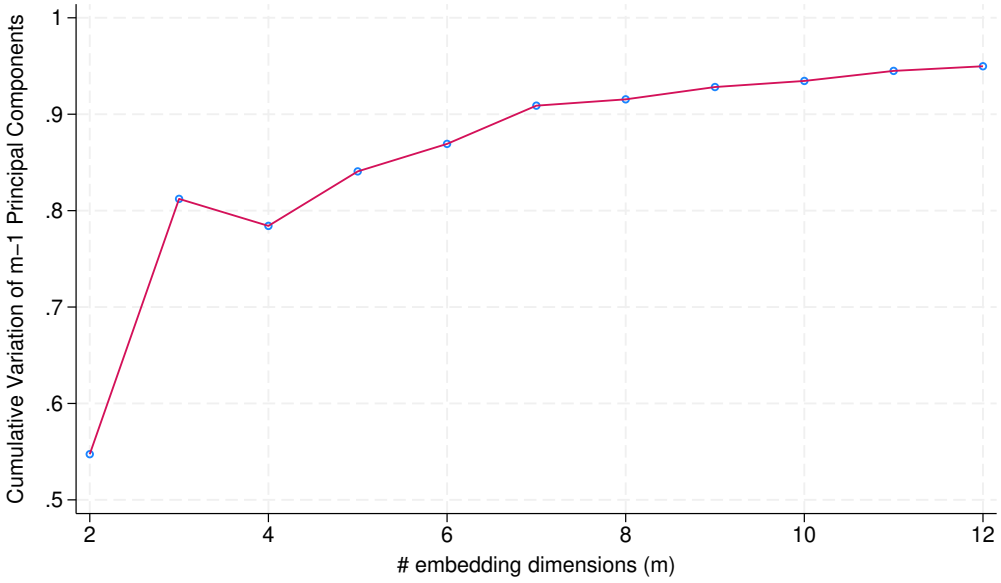
[21] The distribution of characteristics in the product space is taken from Dubé, Fox, and Su (2012).

Appendix C Additional Tables and Figures

Appendix Table 2: Estimated Linear Demand Coefficients

	(1)	(2)
α	-3.310^{***}	-3.275^{***}
	(0.186)	(0.183)
β	0.005***	
	(0.001)	
β_1		-0.047***
		(0.015)
β_2		0.130***
		(0.037)
β_3		-0.108***
		(0.028)
F-Statistic	198.93	100.7
Distance Function	$1/(1+d)$	Cubic
Observations	9,112	9,112
Hotel FE	Yes	Yes
Year-Month FE	Yes	Yes

Appendix Figure 2: Cumulative Variation of $m-1$ Principal Components ($R=5$)



Appendix Table 3: Estimated Demand Coefficients

	β	SE	Σ	SE
Constant			16.369	(9.248)
Price	-0.043	(0.017)	0.026	(0.010)
x_1			0.000	(0.310)
x_2			0.000	(0.417)
x_3			0.549	(0.265)
x_4			0.181	(0.306)
x_5			0.105	(0.531)
Median Outside Diversion			0.069	
Median Own-price Elasticity			-7.166	
Median Cross-price Elasticity			0.056	
Number of Obserations			9,112	

BLP specification includes product and market-year-month level fixed effects to absorb linear components of demands. Estimation utilized 2-step GMM followed by approximation to the optimal instruments and updating of results. The Montiel-Pflueger Effective F-statistic for the logit specification and single price instrument z^p is 280.7.

Appendix Figure 3: Cumulative Variation of $m-1$ Principal Components ($R = 25$)