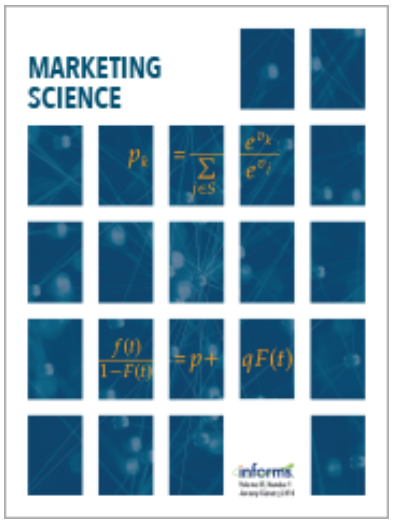
This article was downloaded by: [154.59.124.38] On: 28 June 2021, At: 00:55
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

An Analysis and Visualization Methodology for Identifying and Testing Market Structure

Stephen L. France, Sanjoy Ghose

To cite this article:
Stephen L. France, Sanjoy Ghose (2016) An Analysis and Visualization Methodology for Identifying and Testing Market Structure. Marketing Science 35(1):182-197. https://doi.org/10.1287/mksc.2015.0958

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2016, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

MARKETING SCIENCE
Vol. 35, No. 1, January–February 2016, pp. 182–197
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2015.0958
© 2016 INFORMS

An Analysis and Visualization Methodology for Identifying and Testing Market Structure

Stephen L. France
School of Business, Mississippi State University, Mississippi State, Mississippi 39762, sfrance@business.msstate.edu

Sanjoy Ghose
Sheldon B. Lubar School of Business, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201, sanjoy@uwm.edu

We introduce a method for identifying, analyzing, and visualizing submarkets in product categories. We give an overview of the market structure and competitive submarket literature and then describe a classic model for testing competitive submarkets along with associated extensions. In the era of big data and with the increasing availability of large-scale consumer purchase data, there is a need for techniques that can interpret these data and use them to help make managerial decisions. We introduce a statistical likelihood based technique for both identifying and testing market structure. We run a series of experiments on generated data and show that our method is better at identifying market structure from brand substitution data than a range of methods described in the marketing literature. We introduce tools for holdout validation, complexity control, and testing managerial hypotheses. We describe a method for visualization of submarket solutions, and we give several traditional consumer product examples and in addition give an example to show how market structure can be analyzed from online review data.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2015.0958.

Keywords: market structure; likelihood estimation; brand switching; submarket analysis; market segmentation; cluster analysis; big data; visualization

History: Received: August 26, 2013; accepted: May 26, 2015; Preyas Desai served as the editor-in-chief and John Hauser served as associate editor for this article.

Introduction

Analyzing brand competition and identifying sets of competing brands is a core marketing activity and is an important activity for tasks such as managing product portfolios, determining product assortment, and positioning new brands for the market. Given a modern consumer product environment, with increasing channel diversification (Schoenbachler and Gordon 2002), a high degree of product assortment (Ailawadi and Keller 2004), and large amounts of available consumer data, there is a need for data-driven tools to aid the analysis of brand competition.

The major focus of this paper is on the development of a method for market structure analysis for identifying and testing competitive submarkets. Our method builds particularly on the work of Urban et al. (1984) on testing submarkets. The major contributions of our work are the ability to identify optimal submarket configurations in large modern consumer data sets and the use of visual analytics to help interpret these configurations with respect to prior managerial intuition. We utilize a maximum likelihood-based methodology, which is shown to outperform a range of existing methods on the task of estimating submarket membership.

Background and Literature Review

Market competition for a product category can be characterized as a set of product submarkets, with each brand typically a member of a submarket and brands within a submarket competing with one another. At the lowest level of granularity, a submarket can be defined as a brand. Urban et al. (1984) give an example of submarkets in the car market. They describe two alternative models, a two-submarket model where the car market is partitioned on fuel type (diesel or petrol) to give two submarkets and a model where each car brand is considered to be its own submarket. However, a combination of features is not the only way of defining a submarket partition. If one were to consider a submarket of the overall car market, such as the "small family car" submarket, then a range of both easily quantifiable attributes, such as length, width, price, etc., as well as less easily quantifiable attributes, such as consumers' perceptions of quality and style, could be used to define the submarket. Glushko et al. (2008) note that categorization schemes can come from cultural, individual, and institutional sources. A submarket categorization can come from a combination of these sources.

The overall rationale behind market structure analysis is to provide some managerially useful representation of competition between products in a market. Splitting an overall market into submarkets provides a discrete representation of market structure. A market structure representation can be continuous, discrete, or a hybrid solution containing both continuous and discrete features (Shocker et al. 1990). For example, a product map displays the continuous positions of brands, whereas a partitioning clustering solution or submarket analysis organizes brands into groups using a discrete membership variable.

A typical application of Urban et al. (1984) would be to use empirical brand substitution data to test a submarket partitioning scheme (e.g., splitting cars into petrol and diesel) and then, based on in/out group substitution behavior, evaluate the partition with respect to an unstructured market. An assumption of these submarket models is that if an item in a submarket becomes unavailable, then consumers are more likely to switch to items within the submarket rather than to items outside of the submarket. It is this basic hypothesis that is tested by Urban et al. (1984). It is perfectly possible to have each brand as a separate submarket, but this is rare for consumer products. In frequently purchased consumer product categories such as coffee, toothpaste, and shampoo, complete brand loyalty is rarely assumed. In fact, Ehrenberg et al. (2004) find in a study of the consumer products market that only 12% of consumers are completely brand loyal. Thus, frequently purchased consumer product markets with large numbers of competing brands are liable to have a large degree of substitutability between brands. The idea of market structure is intrinsically related to that of variety-seeking and variety-avoiding behavior. Market structure techniques that model the closeness of brands assume that closeness is a function of brand substitutability or similarity (Lattin and McAlister 1985). Similar or substitutable brands can be considered to be part of the same submarket. There may be some substitutability between submarkets, due to variety-seeking behavior. In fact, some methods of market structure analysis explicitly model variety-seeking behavior. For example, Zahorik (1994), when describing a model of brand-switching behavior, explicitly splits consistency switches and variety-seeking switches and estimates the relative proportions of these switches.

To help motivate our model, we give three different managerial scenarios in which there is a need for a definition of submarkets within a brand category.

Scenario 1. A company may wish to determine whether its existing brands compete within the same submarket. To give an example, Folgers has many different brands of coffee in its portfolio. These brands are ordered by category (regular, decaf, flavored, instant, gourmet, cappuccino, stomach friendly, and blends), by delivery method (ground, instant, K-Cup, filter pack, and single server), and by roast type (mild, medium, medium-dark, and dark). A brand manager may wish to see how the submarket partition defined by different categorical groupings compares with the empirically observed submarket partition, thus determining which attributes are most important for defining submarkets.

Scenario 2. A supermarket with a range of private label brands may wish to see how its private label brands are competing with other brands in the same category. In some markets, such as the United Kingdom, retailers position private label brands at multiple price/quality levels. Geyskens et al. (2010) describe a three-level model incorporating value, quality, and private label brands. There can be a perceived quality difference between national brands (not private label) and private label brands (Steenkamp et al. 2010), so a retailer may wish to be reassured that its premium private label brands are competing with premium national brands and not lower-level national brands.

Scenario 3. A retailer may use submarket information to help prune brands and reduce brand assortment, to help reduce clutter in certain submarkets. The effects of reducing assortment in a product category do not necessarily result in a loss of sales (Boatwright and Nunes 2001), but a large-scale assortment reduction can negatively affect sales and product retention (Borle et al. 2005). The retailer must choose a set of products to remove so that the remaining products still sufficiently cover all product category submarkets.

Each of these scenarios has several common features, including the need to be able to identify sets of competing products and thus a need to be able to define submarkets managerially. After taking managerial action, there is again a need to be able to identify submarkets, to test whether or not the product line changes have been successful.

Defining Submarkets

In this section, we give a detailed description of the Urban et al. (1984) (denoted UJH after the authors initials for brevity) method for testing submarkets. Consider a simplified car market, containing only three cars, the Chevy Aveo, the Honda Accent, and the Lexus. Both the Aveo and the Accent have a market share of 35%, and the Lexus has a market share of 30%. The Aveo and the Accent are low-cost subcompact cars, and the Lexus is a high-end luxury car. If the Accent was removed from the market and its purchasers were split between the Aveo and the Lexus based on the relative market shares of the two remaining cars, then the market share for the Aveo would be $0.35/(1-0.35)=0.538$ and the market share for the Lexus would be $0.30/(1-0.35)=0.462$. These relative market share calculations follow the aggregate constant

ratio method (ACRM) model described by Bell et al. (1975) and Tversky and Sattath (1979). The ACRM is an aggregate version of the choice model described by Luce (1959). In the simple three-car market example, the ACRM model is somewhat unrealistic, because the Accent buyers are much more likely to gravitate to the Aveo than to the Lexus. In fact, Tversky and Sattath (1979) note that if two alternatives share many features, then switching between these alternatives is liable to be higher than that predicted by the ACRM.

The major contribution of UJH is the development of a set of statistical tests for comparing partitions of submarkets with respect to deviation from the ACRM. Here, a "market" consists of the brands making up a product category, and every brand in the market belongs to a single submarket. For example, in a car example given in UJH, a higher-level "diesel versus gasoline" submarket configuration and a "model-specific" submarket configuration are compared. If the ACRM holds within a submarket, the expected market share for a submarket $\mathbf{S}$ given that product i is removed from consideration is given in (1)

$$p_i(\mathbf{S}) = \sum_{j\in\mathbf{S},\, j\neq i} p_i(j) = \frac{1}{1-m_i} \sum_{j\in\mathbf{S},\, j\neq i} m_j, \tag{1}$$

where m_i is the market share for product i. If the ACRM holds, then the proportion of product i's consumers buying in set $\mathbf{S}$ when i is not available, $\hat{p}_i(\mathbf{S})$, should be equal to $p_i(\mathbf{S})$. The actual value of $\hat{P}_i(\mathbf{S})$ is given in (2), where n_i is the number of purchases of product i, and $n_i(\mathbf{S})$ is the number of these purchases that will go to submarket $\mathbf{S}$ if product i is removed

$$\hat{p}_i(\mathbf{S}) = n_i(\mathbf{S})/n_i. \tag{2}$$

If the overall market is split into submarkets, then it is expected that items within a subgroup $\mathbf{S}$ should compete more closely with one another than with items from outside of $\mathbf{S}$. Thus, if an item i in $\mathbf{S}$ is removed from consideration, $\hat{p}_i(\mathbf{S})$ should be larger than predicted by the ACRM, i.e., $\hat{p}_i(\mathbf{S}) \geq p_i(\mathbf{S})$. If item i is taken from outside of $\mathbf{S}$, then $\hat{p}_i(\mathbf{S})$ should be less than that predicted by the ACRM, i.e., $\hat{p}_i(\mathbf{S}) \leq p_i(\mathbf{S})$. The value $p_i(\mathbf{S})$ can be considered to be an aggregate proportion derived from a set of binomial draws, which can be approximated to be normal using the central limit theorem. UJH describe a test for a submarket partition violating the ACRM. A one tailed z-test is used for testing proportions, and aggregate tests are used both for individual submarkets and for the overall market. Assume a partition of the total market $\mathbf{T}$ into r submarkets, so that each product $i = 1, \ldots, N$ is a member of a submarket $\mathbf{S}_k$, where $k = 1, \ldots, r$. The distribution for z statistic for one submarket is given in (3).

$$n(\mathbf{S}_k) \sim N\left[\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k),\ \sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))\right], \tag{3}$$

The distribution for the aggregate z statistic across all submarkets is given in (4).

$$\sum_{k=1}^{r} n(\mathbf{S}_k) \sim N\left[\sum_{k=1}^{r}\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k),\ \sum_{k=1}^{r}\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))\right]. \tag{4}$$

Detailed derivations for (3) and (4) are given in UJH. The test described in UJH and the related tests given in Novak and Stangor (1987) and Kannan and Wright (1991) provide useful measures for testing submarket partitions. However, the techniques described in these papers are primarily testing techniques. Some product categories may have a large number of brands, multiple brand features, and possible ambiguity. In these categories, it may be hard to hypothesize candidate submarket partitions for testing, and thus a technique for identifying significant submarket configurations would be useful. The problem of identifying submarkets is a combinatorial one, and as the number of products increases, the number of possible submarkets rises exponentially. Consider a one tailed z-test with a null hypothesis of no structure and a type I error α. For n items and r submarkets, the number of possible partitions is approximately $r^n/r!$. If one does not know a priori the number of submarkets, then given an upper bound on r, denoted as $r_{\max}$, the number of possible submarket partitions is $\sum_{r=1}^{r_{\max}} r^n/r!$. Given the normal approximation to the binomial, the number of submarket partitions with significant structure is $\alpha\sum_{r=1}^{r_{\max}} r^n/r!$. As the number of items increases, the number of possible submarket partitions grows exponentially. In fact, $\lim_{n\to\infty}\sum_{r=1}^{r_{\max}} r^n/r! = \sum_{r=0}^{r_{\max}} r^n/r! - 1 = e^n - 1$. For a market with 100 possible items and a level of significance of $\alpha = 0.01$, there are an estimated $0.01(\sum_{r=1}^{100} r^n/r!) = 1.415 \times 10^{41}$ significant submarket partitions. Thus, being able to find a significant submarket configuration does not guarantee that the submarket is among the most strongly significant submarket partitions. Attempting to hypothesize all significant submarket configurations manually would not be possible. This suggests that an approach of identifying strongly significant submarkets would be appropriate for categories with large numbers of brands. In practice, it would be hard to find managerial interpretation for submarket partitions with large numbers of submarkets. However, even for a small value of $r_{\max}$, such as $r_{\max} = 10$, a consumer market with a reasonable number of products will have too many significant submarket partitions to be explored manually.

Model

The problem of identifying optimal submarket configurations is essentially a partitioning problem. Many

techniques for partitioning clustering utilize a likelihood formulation to estimate parameters. In fact, the minimization function for k-means clustering is equivalent to a function for maximizing the likelihood of a partitioning function given independent and identically distributed (i.i.d.) data with Gaussian error (MacKay 2003, pp. 300–310). Let $\mathbf{\Phi}$ be an $N \times r$ matrix of product to submarket assignments, let $\mathbf{n}$ be an $N \times 1$ vector of the volume of product sales, and let $\mathbf{Q}$ be an $N \times N$ matrix of product substitution data, where $q_{ij} = n_i(j)$ is the number of purchases of product i that will transfer to product j when product i is removed. The log-likelihood function for the model is derived in the appendix, resulting in (5)–(7), for cases where $\hat{p}_i(\mathbf{S}_k) \geq p_i(\mathbf{S}_k)$

$$LL(\Phi \mid \mathbf{n}, \mathbf{Q}) = \sum_{k=1}^{r} \log\left(\frac{1}{\sqrt{2\pi G_k}}\right) - \sum_{k=1}^{r} \mathbf{1}\left(f\left(\frac{H_k^2}{2G_k}\right) \geq 0\right) f\left(\frac{H_k^2}{2G_k}\right); \quad (5)$$

$$G_k = \sum_{i=1}^{n} \phi_{ik} n_i \left(\sum_{j=1}^{N} \phi_{jk} n_j - n_i\right)\left(\sum_{j=1}^{N} n_j(1-\phi_{jk})\right) \times \left(\sum_{j=1}^{N} n_j - n_i\right)^{-1}; \quad (6)$$

$$H_k = \sum_{i=1}^{N} \phi_{ik} \left[\sum_{j=1}^{N} \phi_{jk} n_i(j) - n_i\left(\sum_{j=1}^{N} \phi_{jk} n_j - n_i\right) \times \left(\sum_{j=1}^{N} n_j - n_i\right)^{-1}\right]. \quad (7)$$

Maximizing the log-likelihood gives the closest approximation to the ACRM. To find optimal submarkets, the worst approximation to the ACRM is required. Thus, we wish to minimize the log-likelihood (and thus maximize $(\hat{p}_i(\mathbf{S}_k) - p_i(\mathbf{S}_k))$ for cases where $\hat{p}_i(\mathbf{S}_k) \geq p_i(\mathbf{S}_k)$). The data in $\mathbf{Q}$ can be derived either from brand purchase data or from perceptual product substitution data. To calculate $\mathbf{Q}$ from brand purchase data, let there be $v = 1$ to M consumers, and let the proportion of user v's sales accounted for by product i be y_{iv}. For any two products i and j, the entry q_{ij} in the product substitution matrix is equal to (8)

$$q_{ij} = \sum_{v:\, y_{iv>0}} n_{iv} \frac{y_{jv}}{1 - y_{iv}}. \quad (8)$$

The resulting estimation problem is to minimize $LL(\Phi \mid \mathbf{n}, \mathbf{Q})$, for cases where $\hat{p}_i(\mathbf{S}_k) \geq p_i(\mathbf{S}_k)$, by fixing Φ. A local search based optimization algorithm is given in the Web appendix (available as supplemental material at http://dx.doi.org/mksc.2015.0958). As per the discussion in the Web appendix, the optimization problem is NP-hard, so any algorithm that can solve the optimization always converges to a good "optimal" solution and the algorithm is heuristic, so it is not guaranteed to be globally optimal. As described in the Web appendix our software gives three solution evaluation criteria. These criteria are LL (the value of $LL(\Phi \mid \mathbf{n}, \mathbf{Q})$), diff (the sum of $\hat{p}_i(\mathbf{S}_k) - p_i(\mathbf{S}_k)$, and z-score (the original UJH z-score). Interpretation of these different criteria is discussed in the Examples section.

Experimentation

We tested our method using a set of Monte Carlo simulation experiments. The rationale behind this experiment was to discover how well our method could recover submarket structure from brand substitution data, relative to a set of techniques described in the marketing literature as being appropriate for partitioning brands using brand-switching/substitution data. Although a whole host of market structure algorithms and techniques have been introduced in the marketing literature, many are customer segmentation algorithms in the tradition of Grover and Srinivasan (1987). However, several published papers describe methods of finding market structure from product similarity/substitution data. Most market partitioning techniques utilize some variant of cluster analysis. Punj and Stewart (1983) give a literature review on the use of cluster analysis in marketing. They describe how the technique of k-means clustering is a good choice when partitioning market data. In fact, the use of k-means clustering for finding market structure and partitioning/clustering brands has been described in several papers (Green et al. 1990, Hruschka and Natter 1999, Iacobucci et al. 2000). The basic k-means clustering algorithm is designed for dimensional data, but variants, such as k-medoids clustering, have been developed for similarity data. We utilized an R implementation of the k-medoids algorithm (Gordon and Vichi 1998), which is designed to work on similarity data, such as brand substitution data.

We utilized several hierarchical clustering techniques from the R "cluster" package. Hierarchical clustering techniques have been used extensively for brand segmentation and positioning (Srivastava et al. 1981, Punj and Stewart 1983, DeSarbo and De Soete 1984, Zhai et al. 2011). For each hierarchical clustering run, we first created the hierarchical tree and then "cut" the tree at the required number of partitions. We ran single linkage, complete linkage, average linkage, and Ward's method variants of the hierarchical clustering algorithms. Single linkage, complete linkage, and average linkage are the "standard" hierarchical clustering algorithms, as defined by Johnson (1967). Ward's method is recommended by Punj and Stewart (1983), as a technique that works well with marketing data.

There have been several pieces of work where optimizing market structure is conceptualized as a problem of maximizing within-segment entropy (Herniter 1973, Carter and Silverman 2004). We implemented a procedure in MATLAB to maximize Shannon entropy as defined for market structure problems by Carter and Silverman (2004) using the same optimization procedure as our log-likelihood-based UJH method. We implemented the block clustering algorithm developed by Hartigan (1972) in MATLAB. This type of clustering has been used for market structure analysis (Carmone et al. 1999) and provides a general method of splitting a matrix directly into partitions or blocks.

We ran the experiment using our log-likelihood based method and the seven comparison algorithms described above. These algorithms are *k*-medoids, single link, average link, complete link, Ward's, entropy, and block clustering. For the experiment, we needed test data sets where the "true" market structure was known. We generated random submarket partitions and then randomly generated error perturbed brand substitution data from the "true" submarket partitions. We then evaluated techniques on the recovery of "true" market structure. The process that we followed for each experimental run is as follows:

1. Select the number of products N, the number of submarkets r, the sales distribution S, and the noise parameter $\lambda \in [0,1]$.

2. For each product $i = 1, \ldots, N$, randomly generate the product sales from the sales distribution S. Store the product sales values in an $N \times 1$ vector $\mathbf{n}$.

3. Randomly assign each of N products to one of r submarkets, giving the underlying submarket partition matrix, Φ.

4. Calculate a product substitution matrix $\mathbf{Q}$ using (9). Assume within submarket homogeneity and that for items i and j in the same submarket k, the value of q_{ij} is the proportion of remaining submarket sales accounted for by j. If i and j are not in the same submarket, then $q_{ij} = 0$. An equation for q_{ij} is given in (9)

$$q_{ij} = \sum_{k=1}^{r} \phi_{ki}\phi_{kj} n_i \frac{n_j}{(\sum_{l=1}^{N} \phi_{kl} n_l) - n_{ij}}. \tag{9}$$

5. Calculate a product substitution matrix $\mathbf{Q}_2$ for a market with a single submarket, i.e., with $r = 1$. Set $\mathbf{Q} = \lambda\mathbf{Q} + (1-\lambda)\mathbf{Q}_2$. If $\lambda = 0$, then $\mathbf{Q}$ does not have any submarket structure, whereas if $\lambda = 1$, then $\mathbf{Q}$ has perfect submarket structure for r submarkets.

6. Add random noise distributed as $X \sim N(0, \mu(\mathbf{Q}))$ to $\mathbf{Q}$.

7. Estimate the market subgroup membership matrix $\hat{\mathbf{\Phi}}$ for each experimental technique. For each $\hat{\mathbf{\Phi}}$, calculate the similarity between $\hat{\mathbf{\Phi}}$ and $\mathbf{\Phi}$ using the adjusted Rand index of clustering similarity (Hubert and Arabie 1985).

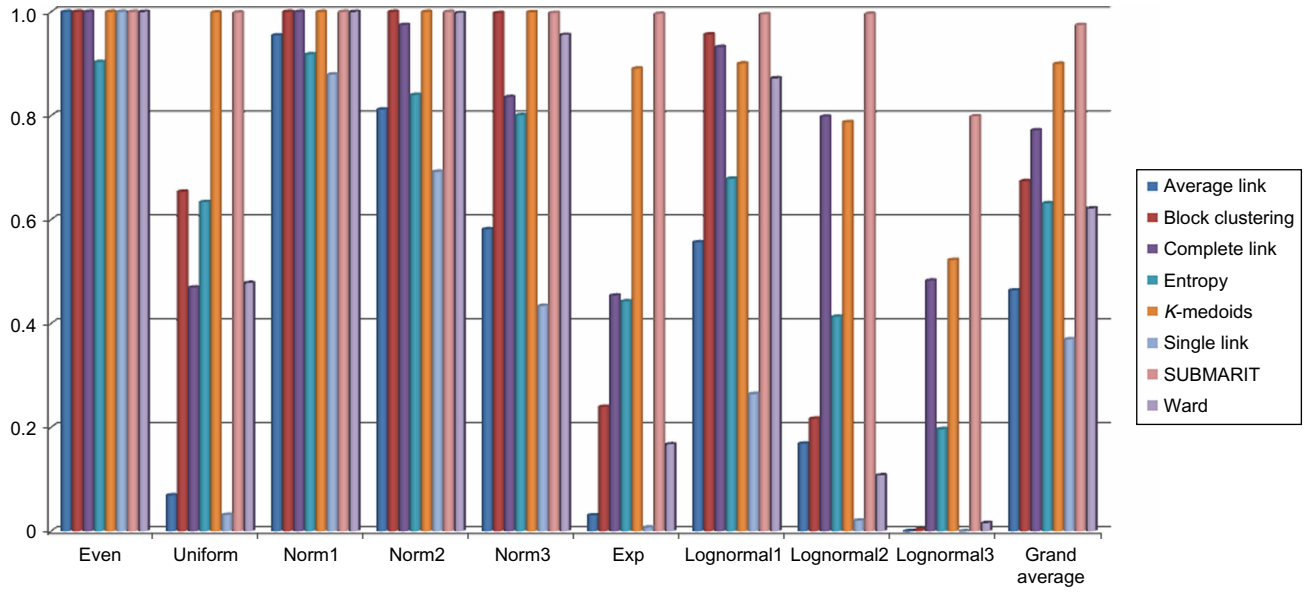
A range of different sales distributions were tested. The rationale behind the use of multiple distributions is that the distribution of sales of products in a category will vary greatly depending on market conditions and the competitive environment (Varian 1980). We included the lognormal and exponential distributions, which have been shown to be useful for modeling market share values (Kohli and Sah 2006, Hisano and Mizuno 2011). We also included the standard normal distribution and the uniform distribution as reference distributions. The distributions used are summarized in Table 1.

We used a full factorial design for the experiment with experimental factors for the partitioning technique (the seven comparison techniques described above and our method), sales distribution (the nine distributions listed in Table 1), the number of products $N \in \{50, 100, 150, 200\}$, the number of submarkets $r \in \{2, 3, \ldots, 6\}$, and the random noise level $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. A value of $\lambda = 0$ gave completely random recovery, and a value of $\lambda = 1$ gave perfect

Table 1 Sales Distributions Used for Experiment

Name	Parameters	Description
Even	$\mu = 50$	All products have sales of 50
Uniform	$a = 0$, $b = 0$	Uniform distribution in the range [a,b]
Norm1	$\mu = 50$, $\sigma = 5$	Normal distribution, truncated at 0, so that all products have sales > 0; in practice, very few values are truncated, as even for $\sigma = 15$, 0 is more than 3 standard deviations away from the mean
Norm2	$\mu = 50$, $\sigma = 10$	
Norm3	$\mu = 50$, $\sigma = 15$	
Exp	$\mu = 50$	The exponential distribution
Lognormal1	$\mu = 3.787024$, $\sigma = 0.5$	The lognormal distribution, so that the distribution X is lognormally distributed if $Y = e^X$ and X is normally distributed with mean μ and standard deviation σ; as σ increases the lognormal distribution becomes more skewed to the left; values of μ have been selected so that $\mu(Y) = 50$ and $\sigma(Y) = 5$, as per Norm1
Lognormal2	$\mu = 3.412024$, $\sigma = 1$	As above; $\mu(Y) = 50$ and $\sigma(Y) = 10$, as per Norm2
Lognormal3	$\mu = 3.910040$, $\sigma = 2$	As above; $\mu(Y) = 50$ and $\sigma(Y) = 15$, as per Norm3

Figure 1 (Color online) Adjusted Rand Index Comparison



recovery for most techniques, so we left these values out of the experiment. We ran five replications for each combination of experimental parameters. In total, this gave $5 \times 8 \times 9 \times 4 \times 5 \times 9 = 64{,}800$ experimental runs. For each experimental run with either our log-likelihood based method, *k*-medoid, or entropy techniques, we ran the selected technique 50 times and chose the run with the best optimization function criterion. The hierarchical clustering algorithms and the block clustering algorithm are completely deterministic and do not require multiple starting solutions, so only required a single algorithm run per experimental run.

We examined the results using the analysis of variance (ANOVA) technique, with the value of the adjusted Rand index as the dependent variable and the categorical experimental parameters as the independent variables. The overall fit of the model was $R^2 = 81.2\%$. All ANOVA experimental factors were significant with $p = 0.000$. Using a Tukey multiple comparisons test, the average value of the adjusted Rand index (aR) for our method (0.9748) was significantly greater than for any of the other techniques. The next highest aR value was for *k*-medoids clustering (0.8997), and was significantly greater than that for the hierarchical clustering, block clustering, and entropy techniques. For parsimony, we summarize the results in Figure 1, which gives aR for each combination of technique and sales distribution. Our log-likelihood based method is named SUBMARIT (submarket identification and testing). One can see that it outperformed the other techniques, with almost perfect recovery on the majority of the sales distributions. The outperformance was particularly strong on the lognormal and exponential distributions. The drop-off in performance for the other techniques seems to be correlated with the number of products likely to have very small values of sales. This number is 0 for the even sales case, small for the normal distributions, moderate for the uniform distribution, and larger for the strongly left-skewed distributions. For all of the techniques, the lognormal3 distribution had the lowest average values of aR. The average value of aR was 0.7989 for our method and was no higher than 0.53 for any of the other techniques. This is because the lognormal3 distribution was the most left skewed of all of the distributions, and the majority of the brands had very low sales values, which were swamped by the random error incorporated into the experiment.

The comparison techniques were implemented on different platforms, with some implemented in *R*, some implemented in MATLAB, and some implemented as executables. Thus, it was not possible to give a fair empirical comparison of runtimes. However, our optimization algorithm ran quickly and efficiently. On a Dell T110 with 16 GB RAM and an Intel X3470 CPU running MATLAB 2010b, each experimental run consisting of 50 optimization runs took 0.574 seconds on average.

Overall, our method gave strong recovery of true submarket partitions. This implies that even with data with significant error or noise, our UJH-based log-likelihood method can identify the underlying market structure better than the other benchmark methods.

Complexity Control and Holdout Validation

The previous section describes an optimization procedure that, given a product substitution matrix and a target number of submarkets, will find an optimal

submarket partition. However, finding an optimal submarket partition does not necessarily give managerial validity or relevance. In this section, we describe a set of tools for testing solution stability and validity, and we describe a procedure that helps control for solution complexity and gives guidance on selecting the number of submarkets.

Several of the tools described in this section utilize the adjusted Rand index (Hubert and Arabie 1985) to compare submarket configurations. Because the adjusted Rand index is rescaled downward for expected random agreement, it cannot be interpreted on a $[0, 1]$ scale in the manner of a correlation efficient. To be interpreted, some statistical measure of significance is required. The adjusted Rand index is based on a contingency table and can be interpreted in terms of a chi-squared coefficient. Unfortunately, the associated chi-squared test is only valid when cluster sizes are equal (Krieger and Green 1999). Several papers (Shuweihdi and Taylor 2014, Qannari et al. 2014) have developed approaches for hypothesis testing for the adjusted Rand index by using a reference empirical distribution of values calculated from randomly generated configurations. This is the approach that we use.

Best k Solution Agreement and LpOCV Solution Validation

As noted previously, partitioning clustering is an NP-complete optimization problem, and no polynomial time algorithm is guaranteed to find a good solution. When running an algorithm, a common tactic is to run the algorithm multiple times and then select the solution with the best value of the optimization criterion. However, if these "best" values of the optimization solution are strongly inconsistent/uncorrelated, then the amount of insight that can be gained from the solutions is limited. By comparing the "best" locally optimal solutions, one can gain an overall view of the stability of the solutions.

We describe a procedure to perform this comparison below.

1. Run the optimization algorithm M times and choose the top k solutions with the best values of the optimization criterion.

2. Calculate the adjusted Rand index (aR) between each of the $k(k-1)/2$ pairs of solutions. Calculate $\bar{a}\bar{R}$ as the average value of aR.

3. Create a comparison empirical distribution for aR by randomly generating N pairs of random partitions and calculating the adjusted Rand index for each of these pairs.

4. As per a bootstrapping procedure (Efron and Tibshirani 1993), resample from the empirical distribution, randomly selecting with replacement T sets of $k(k-1)/2$ values of aR. Calculate the average value of aR for each of these sets and order these values as $S = [\bar{a}\bar{R}_1, \bar{a}\bar{R}_2, \ldots, \bar{a}\bar{R}_i, \ldots, \bar{a}\bar{R}_T]$, where for two values $\bar{a}\bar{R}_i$ and $\bar{a}\bar{R}_j$, $j > i$ implies that $\bar{a}\bar{R}_j > \bar{a}\bar{R}_i$.

5. One-tailed significance values for $\bar{a}\bar{R}$ can be calculated as $p = |\{\bar{a}\bar{R}_t \in S \mid \bar{a}\bar{R} > \bar{a}\bar{R}_t\}|/T$, i.e., the proportion of empirical values below $\bar{a}\bar{R}$.

When fitting statistical models, one must be cognizant of the effect of noise on the data and the potential for overfitting. Overfitting occurs when a statistical model is too sensitive to noise in the data. A common method of testing for overfitting is to utilize a cross-validation procedure such as LpOCV (leave p out cross-validation), which is described by Shao (1993), who notes that LpOCV with p set significantly large is much more stable than the basic LOOCV (leave one out cross-validation) method. A procedure to test solution stability is described below.

1. Take an $N \times N$ brand substitution matrix $\mathbf{Q}$ defined for a set of brands S.

2. Select p items either sequentially or randomly with replacement as the holdout data set. Repeat until H holdout samples have been generated. For each holdout sample $h = 1$ to H, denote the set of holdout brands as S_h^- and the set of remaining brands as $\tilde{S}_h$, where $S_h^- \cup \tilde{S}_h = S$ and $S_h^- \cap \tilde{S}_h = \phi$. Calculate $\mathbf{SQ}_h$, the subset of $\mathbf{Q}$ for $\tilde{S}_h$.

3. For each $h = 1, \ldots, H$, run the LL optimization algorithm on $\mathbf{SQ}_h$ M times and take the best submarket clustering solution $\mathbf{SC}_h$. Now optimize a restricted version of the model, fixing the submarket assignments of the $N - p$ items in the training set to be $\mathbf{SC}_h$, i.e., $\mathbf{C}_h\{\tilde{S}_h\} = \mathbf{SC}_h$. Record the resulting submarket solution $\mathbf{C}_h$. To form an empirical comparison distribution, for each $h = 1, \ldots, H$, create $t = 1, \ldots, T$ solutions where $\mathbf{C}_{h,t}\{\tilde{S}_h\} = \mathbf{SC}_h$ and the remaining items $\mathbf{C}\{S_h^-\}$ are assigned randomly.

4. Calculate the average adjusted Rand index between all $h = 1, \ldots, H$ submarket solutions $\mathbf{C}_h$ as $\bar{a}\bar{R}$. For each $t = 1, \ldots, T$, calculate the average adjusted Rand index between all $h = 1, \ldots, H$ submarket solutions $\mathbf{C}_{h,t}$ as $\bar{a}\bar{R}_t$. Order the values of $\bar{a}\bar{R}_t$ in a set S. One-tailed significance values can be calculated as $p = |\{\bar{a}\bar{R}_t \in S \mid \bar{a}\bar{R} > \bar{a}\bar{R}_t\}|/T$.

Overall, the rationale behind the LpOCV cross-validation procedure is to test how robust the algorithm is to changes in the set of products. If products are removed or added to a category, then the core set of products that are in both the original and revised product sets should have similar submarket configurations in both sets. We implemented code for the LpOCV cross-validation procedure in MATLAB. We also created a procedure to give substatistics for the reliability of individual items. This procedure utilizes the partial adjusted Rand index statistic described in Milligan and Cheng (1996). A full description of this procedure is left out for the sake of brevity.

A Gap Statistic for Complexity Control

When implementing statistical models, there is a need to control for complexity and to prevent overfitting. By sequentially adding additional parameters to likelihood based models, one can continuously increase the fit of the model. However, the additional parameters may not add additional insight to the model. For partitioning methods, the number of parameters is usually controlled by selecting the number of partitions. In fact, Steinley (2006) mentions a range of methods for finding the number of clusters for k-means clustering solutions, including the Akaike information criterion (AIC, Akaike 1974), scree plots, and several pseudo F-tests. The AIC is not directly applicable to the submarket partitioning problem because the data source is a similarity matrix with no intrinsic data dimensionality. Instead, we utilize a variant of the gap statistic described by Tibshirani et al. (2001).

The gap statistic is a flexible method of finding good clustering solutions. The intuition for the statistic is based on the idea of a scree plot. In a scree plot, the number of clusters is plotted on the x axis, and the total within cluster sum of squares error for the optimal solution for each x value is plotted on the y axis. The number of clusters is chosen from the elbow of the graph, i.e., where the angle with respect to the origin between adjacent points goes from greater than 45° to less than 45°. However, the use of this "elbow" criterion can be somewhat arbitrary and unscientific. The gap statistic works by finding the average log of the sum of squared errors when clustering uniformly distributed data in the range of the original data and finding the minimum "gap" between this value and the equivalent value for the original data. The rationale behind the statistic is to remove fit due to noise. The uniform distribution is used as a good "null" reference distribution.

A modified gap statistic for analyzing $r = 2, \ldots, R$ submarkets is given below.

1. Take the original reference customer × sales matrix $\mathbf{Y}$. Derive $\mathbf{Q}$ using (8).

2. For each number of submarkets $r = 2, \ldots, R$, run the optimization algorithm M times, and from the optimal solution calculate W_r using (10)

$$W_r = \sum_{k=1}^{r} \left(\sum_{i \in \mathbf{S}_k} (n_i \hat{p}_i(\mathbf{S}_k)) - \sum_{i \in \mathbf{S}_k} (n_i p_i(\mathbf{S}_k)) \right)^2. \quad (10)$$

3. For each $h = 1$ to H, randomly generate a matrix $\mathbf{Y}_h$, where $\mathbf{Y}_h$ is the same size as Y, and the entries are generated from the uniform distribution on $[\min(\mathbf{Y}), \max(\mathbf{Y})]$.

4. For each $r = 2, \ldots, R$ and each $\mathbf{Y}_h$, carry out Step 2, and calculate W_{rh} using (10). The gap statistic for r submarkets is calculated in (11)

$$\text{Gap}_r = \log(W_r) - \log\left(\frac{1}{H} \sum_{h=1}^{H} W_{rh} \right). \quad (11)$$

Tibshirani et al. (2001) describe a procedure for choosing an exact number of clusters, but we take the approach of plotting the gap statistic against the number of submarkets. This gives more flexibility and allows users to find areas of k with "good" solutions. It also allows the comparison of optimal solutions with hypothesized managerial, as described in the next section.

Testing Managerial Hypotheses

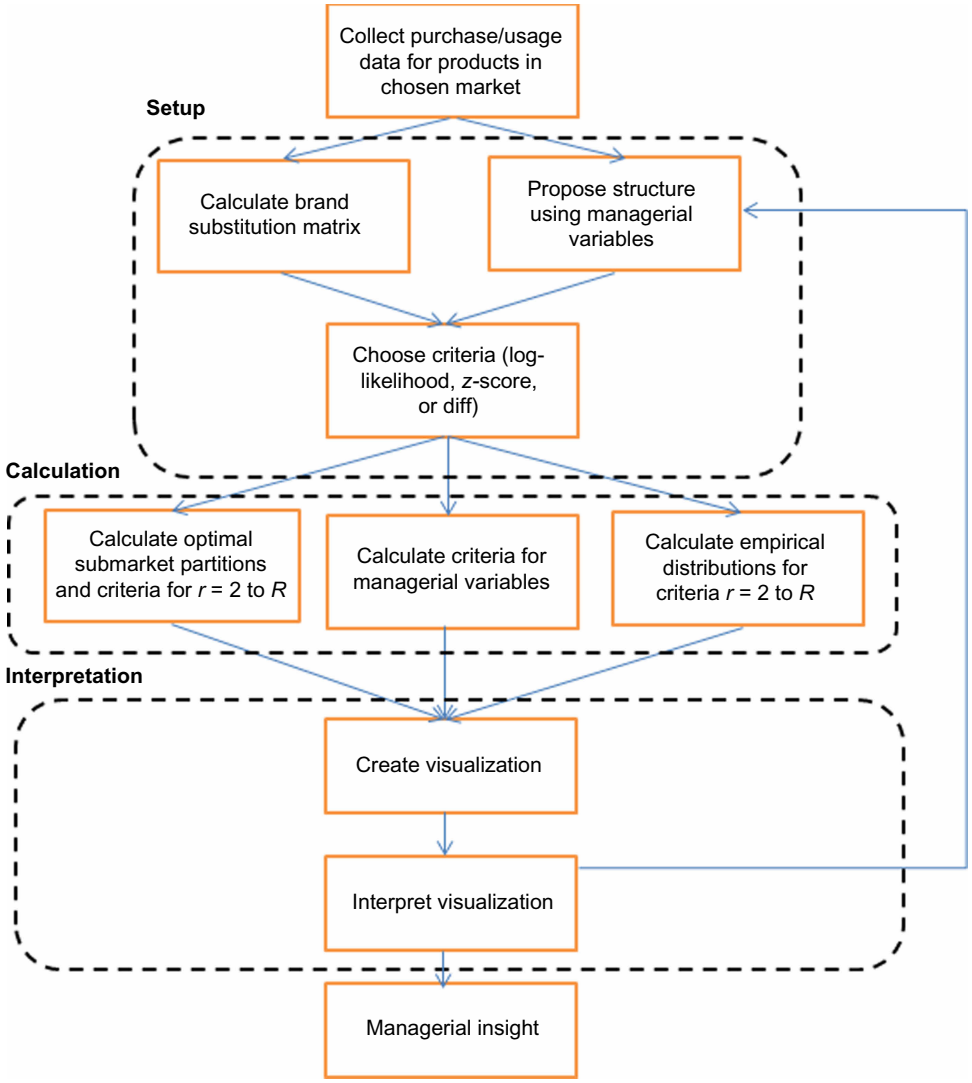
Brands can be characterized by categorical variables. In fact, market structure testing approaches often test submarket splits based on values of categorical variables. For the U.S. car market, UJH define submarkets using brand name and fuel type. For the U.S. cigarette market, Carter and Silverman (2004) define submarkets using size, brand, and color. To help analyze managerial hypotheses based on categorical features, we present the framework summarized in Figure 2.

In the setup stage, the brand substitution matrix $\mathbf{Q}$ is calculated from the product sales matrix $\mathbf{Y}$ using (15), several submarket partitions can be proposed based on managerial intuition and quantitative variables, and the evaluation criterion (log-likelihood, diff, or z-score) is chosen. In the calculation stage, the optimal submarket solution is calculated by optimizing the likelihood function given in (5), an empirical distribution of criterion values is created by calculating the criterion for random submarket configurations, and the criterion values are calculated for the managerial hypotheses. In the interpretation stage, the criterion values for the optimal solution, the managerial hypotheses, and the confidence intervals derived from the empirical distribution are plotted on one graph. Thus, one can characterize the managerial solutions with respect to both the optimal solution and the distribution significance levels. Additional insight can be gained by visualizing the brand substitution matrix using multidimensional scaling (MDS) and superimposing both the clustering solution and the values of the qualitative managerial variables.

Examples

In this section, we describe the analysis of several data sets using the tools introduced in the previous section. We analyzed data from four categories. For three of these categories, yogurt, detergent, and soup, data were taken from the Nielsen Scantrack data set. This data set includes four years' worth of purchases for consumers signed up to the Nielsen Scantrack panel. We chose this data set because the product names include tokens that can be easily converted to product attribute categories using simple text mining. The fourth category is restaurants and bars. This category is a little different from the others in that the data used to analyze the category are review instances from the Yelp.com website taken over a period of three years. Whereas panel

Figure 2 **(Color online) Process For Analyzing Managerial Data**



scanner data record purchases of consumers signed up for the panel, review instance data record reviewed visits from active online reviewers. For the Yelp.com data, when calculating the **Q** matrix, we assumed that each review was equal to one visit. The four data sets are summarized in Table 2, which gives the name, type, size, and inferred managerial submarket variables. The number of submarkets for each managerial submarket variable is given in parentheses. The actual submarkets are given for variables that have a small number of submarkets. We tested the fit of data against a range of distributions using the MATLAB distribution fitting tool. The log-likelihood distribution gave a strong fit for all four data sets. Graphs showing distribution fit are given in the Web appendix in Figures A1–A4.

We set the optimization criterion to be "log-likelihood" and ran the optimization algorithm 50 times for each number of submarkets from 2 to 40. We calculated all three criteria (LL, diff, and z-score) for the optimal solutions and for the solutions defined by the managerial variables. We also created managerial reference distributions with 1,000 random configurations for each number of submarkets. For analyzing the managerial variables, each criterion has its advantages. Diff is linear with respect to model error, the log-likelihood is closely tied to the underlying model, and the z-score is easily interpretable with respect to the material that a manager would learn in a basic business statistics class. A "criterion plot" using the z-score is given in Figure 3. Confidence levels from $p = 0.005$ to $p = 0.995$ are given for the empirical distribution, and were derived by calculating criterion values for 10,000 randomly generated configurations. Plots for the diff (Figure A5) and the log-likelihood (Figure A6) are given in the Web appendix.

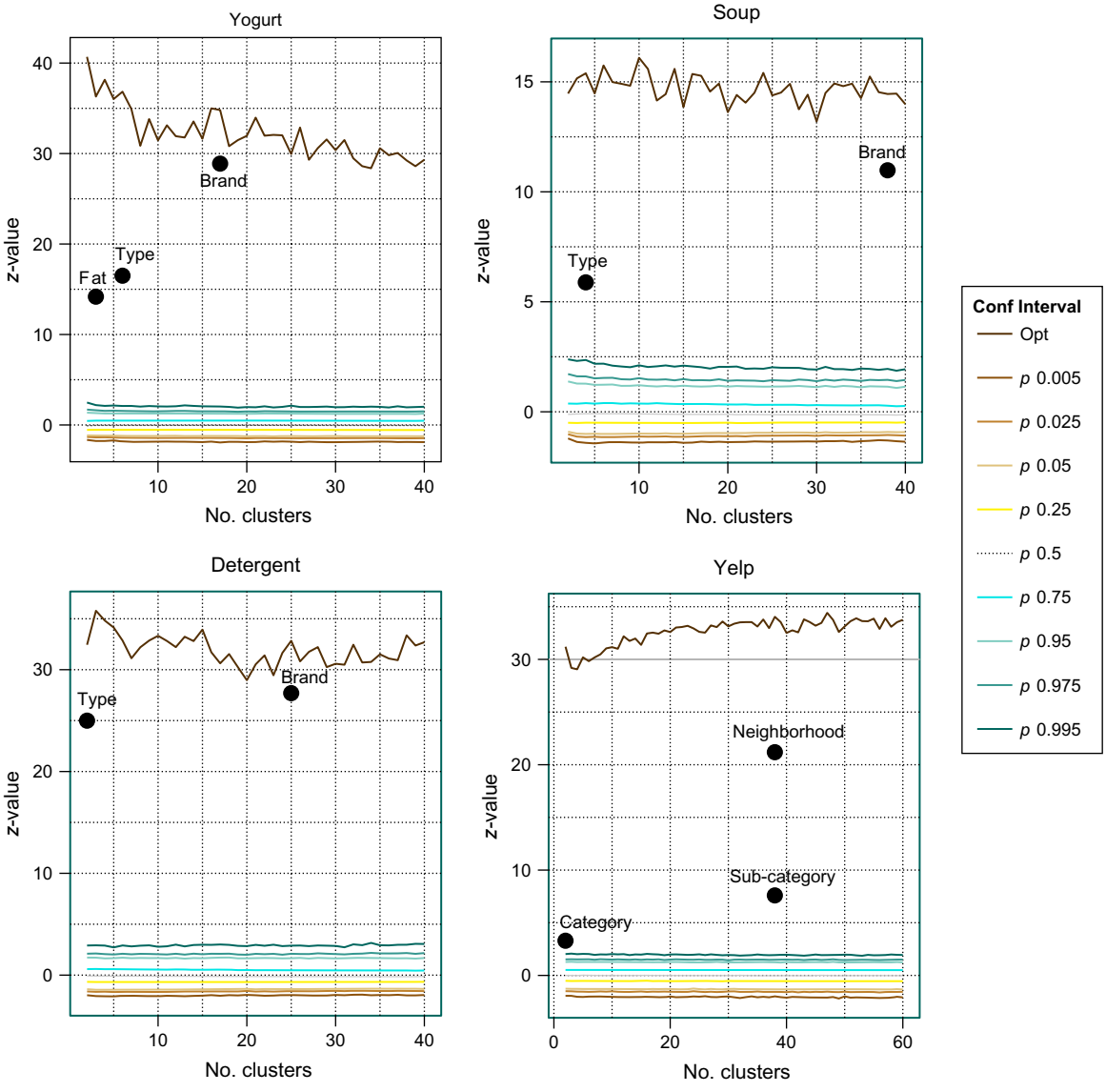
The first conclusion that one can make from the diagram in Figure 3 is that brand strongly defines market structure. This is especially true for detergent.

Table 2 Managerial Submarket Summary

Name	Type	No. cust.	No. prod.	Managerial
Yogurt	Scanner	2,599	370	Brand (18) Type (5) (fruit on bottom, soft whip, custard, frozen, standard) Fat (3) (non fat, low fat, standard)
Detergent	Scanner	3,091	561	Brand (25) Type (2) (liquid, dry)
Soup	Scanner	3,188	695	Brand (38) Type (4) (instant mix, ready mix, canned, ready serve)
Yelp	Review	18,720	1,909	Category, subcategory, neighborhood

Here, submarkets are strongly defined by both the brand and by the type (liquid/dry). The second is that all of the managerial variables chosen are highly significant. This suggests that even in cases where there is significant variety-seeking-based switching behavior, meaningful managerial variables are still significant and that a pure testing approach would always find such variables to be significant. Analyzing the relative level of significance with respect to both the empirical distribution and the optimal solution can

Figure 3 (Color online) Analysis of Managerial Variables Using the *z*-Score



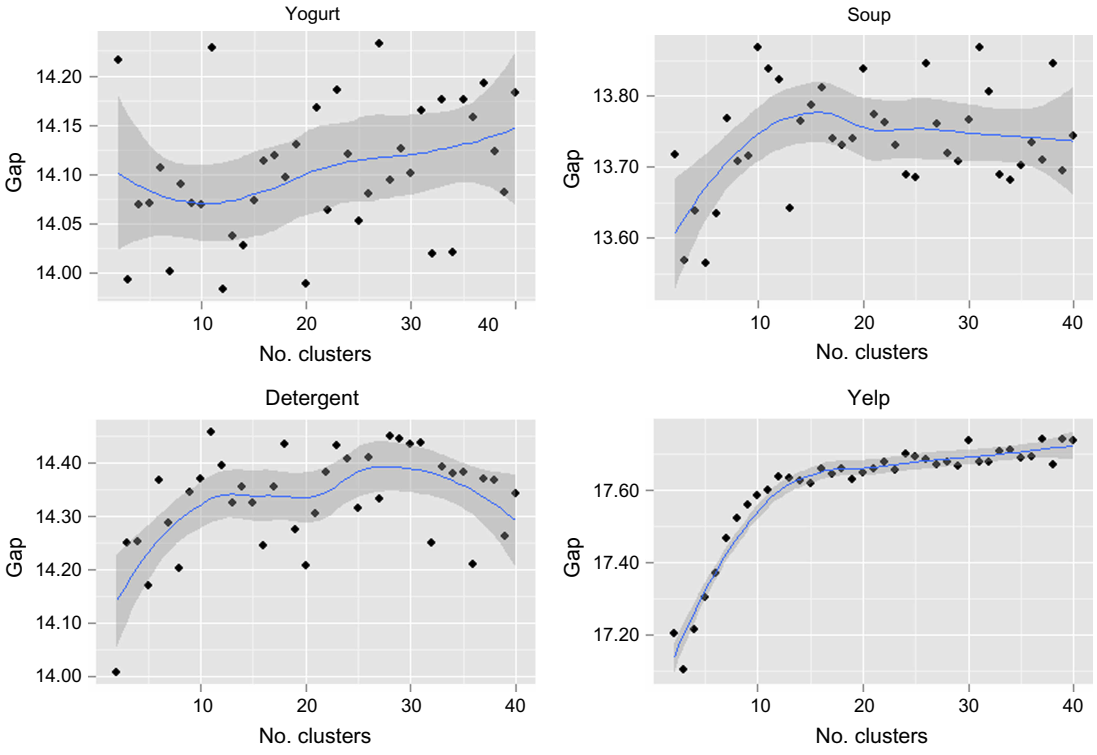
give more insight than a yes/no significance approach. For example, for the Yelp solution, both subcategory (the type of bar/restaurant) and neighborhood have a similar number of submarkets, but the neighborhood has much stronger significance. This is intuitive given the category. Variety switching is known to be high among categories such as restaurants, where products induce affective sensations in consumers (Van Trijp et al. 1996), whereas location and neighborhood are very important in physical retailing behavior (Bell 2014).

We performed both the best k validation and the LpOCV procedures described in the Complexity Control and Holdout Validation section. For the best k validation procedure, we used $k = 3$ and ran the procedure for $r = 2$ to 40 submarkets, generating a reference distribution of 10,000 random $k = 3$ agreements for each r. For all four data sets, the best three solutions were highly related, giving values of the Rand index between 0.8 and 1.0 and highly significant adjusted Rand index values. (In all cases, the best-three agreement value was above all 10,000 values in the empirical distribution.) A visual analysis of the adjusted Rand index results is given in Figure A7 in the Web appendix. For the LpOCV cross-validation procedure, we split the data set into 20, giving 20-fold cross-validation. Again, all of the results were significant with highly significant adjusted Rand index values ($p = 0.000$ for all combinations of cluster value and data set), showing that the solution is relatively robust to the addition/removal of brands. A visual analysis is given in Figure A8 in the Web appendix.

We ran the gap index procedure described in the previous section for each combination of data set and cluster size from $r = 2$ to 40. In Figure 4, the values of the index on the y-axis are plotted for the number of submarkets/clusters on the x-axis, along with a smoothed fitting curve. The overall pattern is quite definite for the Yelp data. The gap statistic increases rapidly up to 10 submarkets, which signifies that an r value of at least $r = 10$ may be required. The patterns for the gap statistics for the scanner product categories are less clear. For example, soup has the optimal gap statistic value at $r = 10$, but several r values for $r > 10$ give good values. The gap statistic values are in quite a narrow band, from 13.9 to 14.2 for yogurt, from 13.55 to 13.9 for soup, and from 14.0 to 14.5 for detergent. This suggests that the final choice for r could be left to interpretation and to the value of r for the managerial priors.

Figure 3 gives some insight into the relative importance and strength of the different managerial priors. However, additional insight can be gained by visualizing the submarket solutions and analyzing how different submarkets relate to the managerial priors. A visual analysis of the yogurt submarket solution for $r = 10$ is given in Figure 5. The brand substitution matrix was embedded into two dimensions using the R implementation of the SMACOF MDS algorithm (de Leeuw and Mair 2009). Using Tableau (Stolte et al. 2008), the submarket solutions were overlaid onto the overall submarket solution. Labels for the managerial

Figure 4 (Color online) Gap Index Across the Number of Submarkets



priors (brand, type, and fat) were added to the individual items. The item dots were made proportional to the item reliability, which was calculated using the submarket item reliability measure briefly described in the previous section. The submarket categories are relatively homogeneous with respect to the MDS visualization, which gives the submarket solution face validity.

The actual Tableau visualization is interactive and can be zoomed in and out to give more or less detail to allow for detailed analysis of submarkets, and in particular submarket boundaries. We demonstrate this in Figure 5 by showing the boundary between two submarkets, one associated with Nordica low-fat yogurts and the other associated with W-B-B and private label low-fat yogurts. In Figure 5, we also provide contingency tables, which can be used to show how the optimal submarket configuration relates to the underlying categorical managerial variables. The visualizations and the information can be utilized to help support product decisions in any of the scenarios described previously in the paper. Consider a scenario where a retailer wishes to introduce a set of premium private label products. Approximately 20% of U.S. consumer products are private label (Nielsen 2014). Although this is behind some European markets, for example, the United Kingdom at 40 + %, there has been steady growth in the market share of private label products, and retailers are increasingly positioning private label brands as "premium" brands, rather than as purely economy brands. The submarket results can be incorporated into a standard positioning framework of brand positioning; see for example, Green et al. (1988).

Like any segmentation exercise, this is a mixture of art and science. First, the significant submarket variables can be used to help explain the segments. For the yogurt category, we do this in Table 3. Our private label retailer could then develop "private label" products to compete in each submarket considered to be sufficiently profitable. For example, a range of premium "health" private label yogurts could be

Figure 5 (Color online) Yogurt Analysis

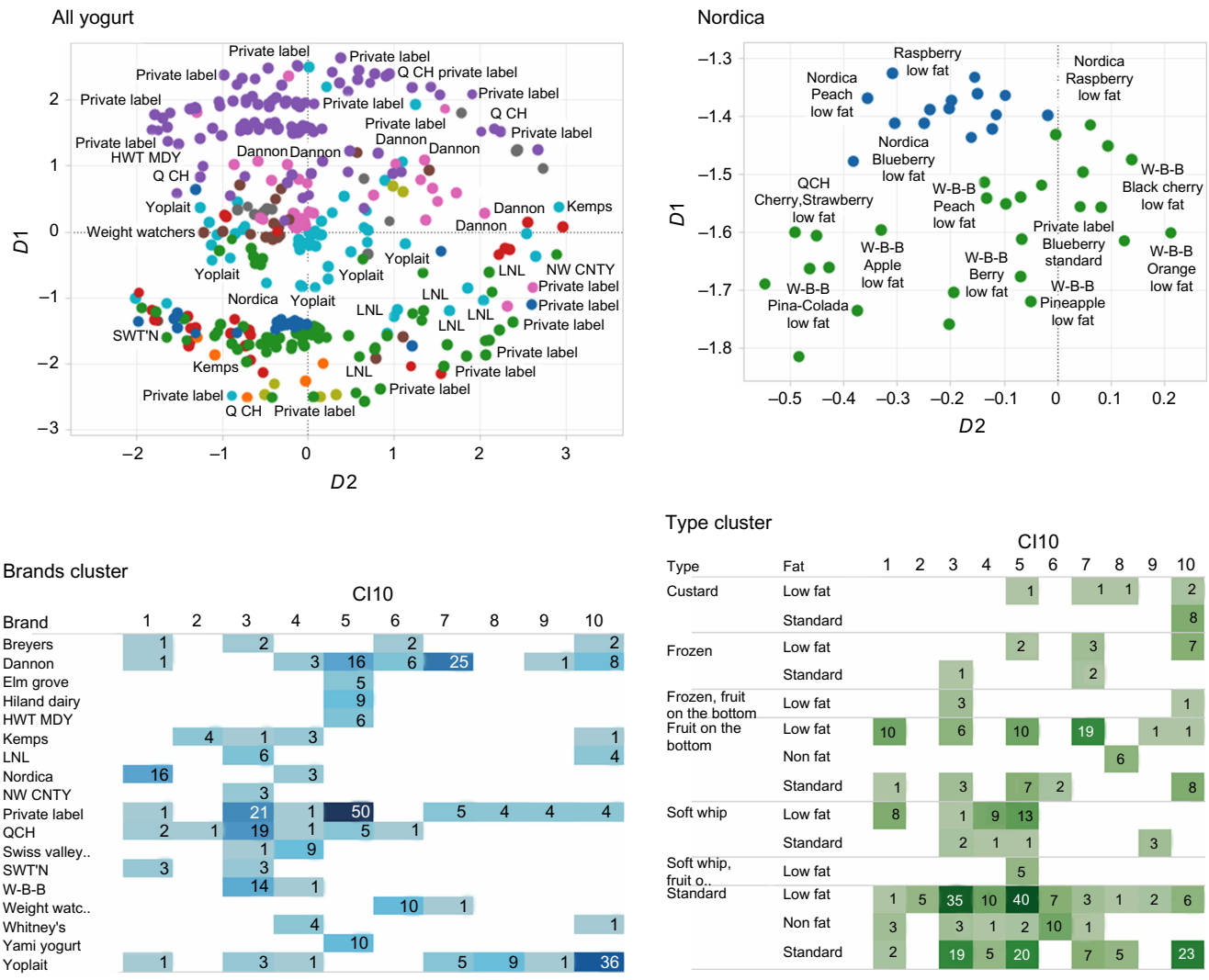


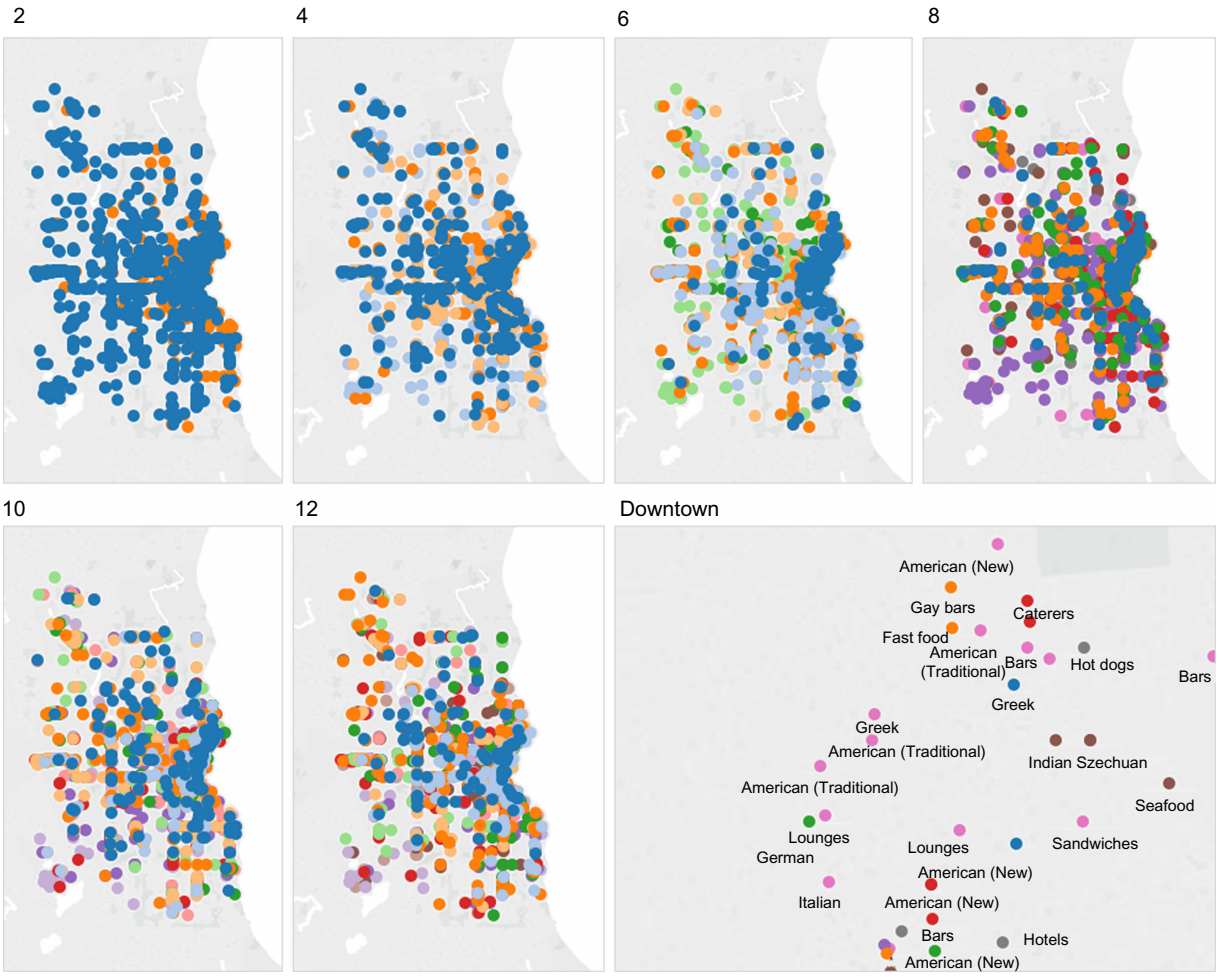
Table 3 Cluster Descriptions

Submarket 1: Dominated by Nordica yogurts, across all fat content levels, predominantly fruit on the bottom and soft whip
Submarket 2: Standard, low fat yogurts, from specialty producers Kemp farms and Q-CH
Submarket 3: A standard yogurt cluster, across all fat levels, with competition across multiple smart brands, with Q-CH, W-B-B, and private label brands dominating; submarket has the widest variety of flavors, indicating significant flavor variety seeking
Submarket 4: A cluster containing soft whip and standard yogurts across a range of brands; this cluster contains nearly all of the Swiss Valley brands; overrepresentation of plain yogurts
Submarket 5: The midmarket cluster; this submarket contains the vast majority of private label/generic brands and contains a range of yogurt types
Submarket 6: Predominantly nonfat and low-fat standard yogurts; this submarket contains all of the Weight Watchers yogurts and low fat yogurts from Dannon; this can be considered a "diet" segment
Submarket 7: The majority of the yogurts in this cluster are Dannon Fruit in the Bottom yogurts
Submarket 8: A predominantly Yoplait cluster, with most yogurts with some sort of berry flavor
Submarket 9: Generic/private label cluster
Submarket 10: A mostly Yoplait brand loyal cluster containing the vast majority of Yoplait yogurts

aimed at submarket 6. The success of the product introduction would be partially determined by the successful positioning of the range with respect to a submarket solution created from postintroduction sales data. If the new products were in the same submarket as the brands currently in submarket 6, then the positioning would be a success. If the new yogurts were colocated in a submarket with predominantly other private label/value brands, then the positioning would be a failure.

In the previous discussion, we noted that for the Yelp example, the "neighborhood" was a more important determinant of the submarket than the actual restaurant category. This provides an interesting spatial dimension to the analysis of submarket categories. In Figure 6, rather than overlay the cluster solutions onto

Figure 6 (Color online) Yelp Analysis



an MDS map, we overlaid the submarket solutions for $r \in \{2, 4, 6, 8, 10, 12\}$ onto the actual geographical locations of the restaurants and bars tested. Despite some heterogeneity due to product category, one can see clear spatial patterns. In the city of Milwaukee, most of the "cosmopolitan" restaurants are concentrated in the downtown and east side areas. There are several downtown/east side submarkets and several more suburban submarkets. There is more cuisine/category-based heterogeneity in the downtown submarkets than in the suburban submarkets. This is illustrated in the zoom-in analysis of the downtown area and may be because the greater variety of restaurants in this area leads to more category-based heterogeneity.

Overall, the idea of competitive submarkets and product substitution is a natural one, and as shown by the examples, the submarkets found with our method are strongly related to the submarkets defined by underlying managerial submarket variables. This lends strong face validity to the optimal submarket solutions. We have analyzed and provided visualizations for the yogurt and Yelp data. Similar visualizations are given for detergent and soup data in Figures A9 and A10 in the Web appendix.

Discussion and Future Work

The methodology described by Urban et al. (1984), provides a useful and theoretically grounded methodology for testing submarket structure. The examples given in Urban et al. (1984) show that by testing different submarket partitions, one can gain managerial insight into the structure of market competition. The methodology is general and can be used with either empirically derived brand substitution data or with directly observed perceptual product substitution preferences. Managerial applications for the methodology include analyzing positioning for new product entry, analyzing cannibalization for product portfolio optimization, and testing changes in product perceptions using perceptual product substitution data.

Our work contains several major contributions. We extend the methodology of UJH into a general maximum likelihood methodology for identifying and testing submarkets. We provide a set of tools to help analyze and interpret submarket solutions and control for the complexity of the solutions. These tools include LpOCV holdout analysis, *k*-best solutions analysis, a customized variant of the gap statistic, and a visualization methodology for testing submarket splits based on managerial intuition and/or managerial categorical variables. We show how the optimal submarket solutions can be visualized and interpreted by overlaying the solutions onto spatial representations of the products and by showing contingency tables of associations between the optimal submarket solutions and managerial categorical variables. Future work in the area of visualization could include the integration of our method with more complex multi-modal visualizations, such as those described by Ringel and Skiera (2015).

We tested our method using a series of experiments on generated data. The rationale behind the experiments was to test how well our method could identify underlying or latent submarket partitions for data with sales derived from a range of distributions and with random Gaussian noise added to the configurations. Overall, our method performed well. It outperformed a range of partitioning techniques carefully selected from the marketing literature in the task of identifying submarket partitions and obtained good results, even with high levels of noise.

The matrix used as input to both the original UJH technique and the extensions described in this paper is a market share substitutability matrix, which is summarized using relative market share information for individual consumers over time. The idea of "substitutability" is core to the idea of analyzing market structure. In fact, using substitutability data has several advantages over using pure brand-switching data. Weitz (1985) notes that complementary brands within a category can lead to significant variety switching, which is not a measure of brand substitutability. Day et al. (1979) note that brand-switching data can be inaccurate when multiple brands are purchased simultaneously. To create brand-switching data, purchases must be ordered, and the order of simultaneously purchased brand data can significantly affect brand-switching probabilities. This phenomenon does not adversely affect brand substitution data. However, there still may be some "history" effect that can also be present in substitutability data. The methods presented in this paper can easily be used to update market structure information. The frequency of updates and the length of the time period used to calculate the brand substitution data should be dependent on how quickly the market structure changes. There is scope for future research into these issues and on the development of statistical techniques to test the relative change in market structure over time.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2015.0958.

MATLAB software for the technique and associated R visualization routines can be found at https://sites.google.com/site/psychminegroup/software.

Acknowledgments

The authors would like to thank the editor, associate editor, and anonymous reviewers for their helpful and constructive feedback.

Appendix

We assume i.i.d. data with Gaussian errors. Assuming a Gaussian approximation to the binomial distribution and (3) gives (12), which is the likelihood function for a single submarket

$$L(\mu, \sigma \mid \hat{p}_i(\mathbf{S}_k)) = \frac{1}{\sqrt{\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))}\sqrt{2\pi}} \cdot \exp\left[-\frac{(\sum_{i\in\mathbf{S}_k} n_i\hat{p}_i(\mathbf{S}_k) - \sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k))^2}{2\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))}\right]. \quad (12)$$

This likelihood function, across all possible submarkets from $k = 1, \ldots, r$ is given in (13)

$$L(\mu, \sigma \mid \hat{p}_i(\mathbf{S}_1)\cdots\hat{p}_i(\mathbf{S}_r)) = \prod_{k=1}^{r} \frac{1}{\sqrt{\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))}\sqrt{2\pi}} \cdot \exp\left[-\frac{(\sum_{i\in\mathbf{S}_k} n_i\hat{p}_i(\mathbf{S}_k) - \sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k))^2}{2\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))}\right]. \quad (13)$$

The formulation consists of a product of the likelihood functions of the individual terms. The log-likelihood of (13) is given in (14)

$$LL(\mu, \sigma \mid \hat{p}_i(\mathbf{S}_1)\cdots\hat{p}_i(\mathbf{S}_r)) = \sum_{k=1}^{r} \log \frac{1}{\sqrt{\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))}\sqrt{2\pi}} - \sum_{k=1}^{r} \frac{(\sum_{i\in\mathbf{S}_k} n_i\hat{p}_i(\mathbf{S}_k) - \sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k))^2}{2\sum_{i\in\mathbf{S}_k} n_i p_i(\mathbf{S}_k)(1-p_i(\mathbf{S}_k))}. \quad (14)$$

The values of $p_i(\mathbf{S}_k)$ and $\hat{p}_i(\mathbf{S}_k)$ cannot be manipulated directly, but instead are dependent on the current assignment of products to submarkets. Without any loss of generality, the values of $\hat{p}_i(\mathbf{S}_k)$ and $p_i(\mathbf{S}_k)$ can be rewritten in terms of submarket assignments. Let there be N total products, and let ϕ_{ik} be the assignment of item i to submarket k. The assignments define a strict partition, so $\sum_{k=1}^{r} \phi_{ik} = 1$ for each i, and $\sum_{i=1}^{N} \phi_{ik} = 1$ for each k. The values of $p_i(\mathbf{S}_k)$ and $\hat{p}_i(\mathbf{S}_k)$ are given in (15) and (16), respectively

$$\begin{aligned} p_i(\mathbf{S}_k) &= \phi_{ik}\frac{\sum_{j=1}^{N}\phi_{jk}m_j - m_i}{1-m_i} \\ &= \phi_{ik}\frac{\sum_{j=1}^{N}\phi_{jk}(n_j/(\sum_{j=1}^{N}n_j)) - n_i/(\sum_{j=1}^{N}n_j)}{\sum_{j=1}^{N}n_j/(\sum_{j=1}^{N}n_j) - n_i/(\sum_{j=1}^{N}n_j)} \\ &= \phi_{ik}\frac{\sum_{j=1}^{N}\phi_{jk}n_j - n_i}{\sum_{j=1}^{N}n_j - n_i}; \end{aligned} \quad (15)$$

$$\hat{p}_i(\mathbf{S}_k) = \frac{1}{n_i}\phi_{ik}\left(\sum_{j=1}^{N}\phi_{jk}n_i(j)\right). \quad (16)$$

The reformulated log-likelihood function, given in terms of cluster assignments, is given in (5), with adjunct equations given in (6) and (7).

References

Ailawadi KL, Keller KL (2004) Understanding retail branding: Conceptual insights and research priorities. *J. Retailing* 80(4): 331–342.

Akaike H (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control* 19(6):716–723.

Bell DE, Keeney RL, Little JD (1975) A market share theorem. *J. Marketing Res.* 12(2):136–141.

Bell DR (2014) *Location is (Still) Everything: The Surprising Influence of the Real World on How we Search, Shop, and Sell in the Virtual One* (Houghton Mifflin Harcourt, Boston).

Boatwright P, Nunes JC (2001) Reducing assortment: An attribute-based approach. *J. Marketing* 65(3):50–63.

Borle S, Boatwright P, Kadane JB, Nunes JC, Shmueli G (2005) The effect of product assortment changes on customer retention. *Marketing Sci.* 24(4):616–622.

Carmone FJ Jr, Kara A, Maxwell S (1999) HINoV: A new model to improve market segment definition by identifying noisy variables. *J. Marketing Res.* 36(4):501–509.

Carter J, Silverman F (2004) An empirical approach to market partitioning: Application to the cigarette market. *J. Targeting, Measurement Anal. Marketing* 12(4):366–378.

Day GS, Shocker AD, Srivastava RK (1979) Customer-oriented approaches to identifying product-markets. *J. Marketing* 43(4): 8–19.

de Leeuw J, Mair P (2009) Multidimensional scaling using majorization: SMACOF in R. *J. Statist. Software* 31(3). http://jstatsoft.uibk.ac.at/article/view/v031i03.

DeSarbo W, De Soete G (1984) On the use of hierarchical clustering for the analysis of nonsymmetric proximities. *J. Consumer Res.* 11(1):601–610.

Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap* (Chapman and Hall, London).

Ehrenberg AS, Uncles MD, Goodhardt GJ (2004) Understanding brand performance measures: Using Dirichlet benchmarks. *J. Bus. Res.* 57(12):1307–1325.

Geyskens I, Gielens K, Gijsbrechts E (2010) Proliferating private-label portfolios: How introducing economy and premium private labels influences brand choice. *J. Marketing Res.* 47(10):791–807.

Glushko RJ, Maglio P, Matlock T, Barsalou LW (2008) Categorization in the wild. *Trends Cognitive Sci.* 12(4):129–135.

Gordon AD, Vichi M (1998) Partitions of partitions. *J. Classification* 15(2):265–285.

Green PE, Kim J, Carmone FJ (1990) A preliminary study of optimal variable weighting in *k*-means clustering. *J. Classification* 7(2): 271–285.

Green PE, Tull DS, Albaum G (1988) *Research for Marketing Decisions* (Prentice Hall, Upper Saddle River, NJ).

Grover R, Srinivasan V (1987) A simultaneous approach to market segmentation and market structure. *J. Marketing Res.* 24(2): 139–153.

Hartigan JA (1972) Direct clustering of a data matrix. *J. Amer. Statist. Assoc.* 67(337):123–129.

Herniter JD (1973) An entropy model of brand purchase behavior. *J. Marketing Res.* 10(4):361–375.

Hisano R, Mizuno T (2011) Sales distribution of consumer electronics. *Physica A: Statist. Mech. Appl.* 390(2):309–318.

Hruschka H, Natter M (1999) Comparing performance of feedforward neural nets and *K*-means for cluster-based market segmentation. *Eur. J. Oper. Res.* 114(2):346–353.

Hubert LJ, Arabie P (1985) Comparing partitions. *J. Classification* 2(1):193–218.

Iacobucci D, Arabie P, Bodapati A (2000) Recommendation agents on the Internet. *J. Interactive Marketing* 14(3):2–11.

Johnson SC (1967) Hierarchical clustering schemes. *Psychometrika* 32(3):241–254.

Kannan PK, Wright GP (1991) On "testing competitive market structures." *Marketing Sci.* 10(4):338–347.

Kohli R, Sah R (2006) Some empirical regularities in market shares. *Management Sci.* 52(11):1792–1798.

Krieger AM, Green PE (1999) A generalized Rand-index method for consensus clustering of separate partitions of the same data base. *J. Classification* 16(1):63–89.

Lattin JM, McAlister L (1985) Using a variety-seeking model to identify substitute and complementary relationships among competing products. *J. Marketing Res.* 22(3):330–339.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (John Wiley & Sons, New York).

MacKay DJC (2003) *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK).

Milligan GW, Cheng R (1996) Measuring the influence of individual data points in a cluster analysis. *J. Classification* 13(2):315–335.

Nielsen (2014) *The State of Private Label Around the World* (Nielsen, New York).

Novak TP, Stangor C (1987) Testing competitive market structures: An application of weighted least squares methodology to brand switching data. *Marketing Sci.* 6(1):82–97.

Punj G, Stewart DW (1983) Cluster analysis in marketing research: Review and suggestions for application. *J. Marketing Res.* 20(2): 134–148.

Qannari EM, Courcoux P, Faye P (2014) Significance test of the adjusted Rand index. Application to the free sorting task. *Food Quality Preference* 32(3):93–97.

Ringel DM, Skiera B (2015) Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Sci.* Forthcoming.

Schoenbachler DD, Gordon GL (2002) Multi-channel shopping: Understanding what drives channel choice. *J. Consumer Marketing* 19(1):42–53.

Shao J (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* 88(422):486–494.

Shocker AD, Stewart DW, Zahorik AJ (1990) Determining the competitive structure of product-markets: Practices, issues, and suggestions. *J. Managerial Issues* 2(2):127–159.

Shuweihdi F, Taylor CC (2014) Inference for similarity indices. Working paper, University of Leeds, Leeds, UK.

Srivastava RK, Leone RP, Shocker AD (1981) Market structure analysis: Hierarchical clustering of products based on substitution-in-use. *J. Marketing* 45(3):38–48.

Steenkamp J-BEM, van Heerde HJ, Geyskens I (2010) What makes consumers willing to pay a price premium for national brands over private labels. *J. Marketing Res.* 47(6):1011–1024.

Steinley D (2006) *K*-means clustering: A half-century synthesis. *British J. Math. Statist. Psych.* 59(1):1–34.

Stolte C, Tang D, Hanrahan P (2008) Polaris: A system for query, analysis, and visualization of multidimensional databases. *Comm. ACM* 51(11):75–84.

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc.: Ser. B Statist. Methodology* 63(2):411–423.

Tversky A, Sattath S (1979) Preference trees. *Psych. Rev.* 86(6):542–573.

Urban GL, Johnson PL, Hauser JR (1984) Testing competitive market structures. *Marketing Sci.* 3(2):83–112.

Van Trijp HCM, Hoyer WD, Inman JJ (1996) Why switch? Product category: Level explanations for true variety-seeking behavior. *J. Marketing Res.* 33(3):281–292.

Varian HR (1980) A model of sales. *Amer. Econom. Rev.* 70(4):651–659.

Weitz BA (1985) Introduction to special issue on competition in marketing. *J. Marketing Res.* 22(3):229–236.

Zahorik AJ (1994) A nonhierarchical brand switching model for inferring market structure. *Eur. J. Oper. Res.* 76(2):344–358.

Zhai Z, Liu B, Xu H, Jia P (2011) Clustering product features for opinion mining. *Proc. Fourth ACM Internat. Conf. Web Search and Data Mining* (ACM, New York), 347–354.