European Journal of Marketing

Market basket analysis insights to support category management
Andres Musalem, Luis Aburto, Maximo Bosch,

EJM
52,7/8

Market basket analysis insights to support category management

Andres Musalem and Luis Aburto
University of Chile, Santiago, Chile, and
Maximo Bosch
Universidad de Las Americas, Santiago, Chile

1550

Received 3 June 2017
Revised 23 November 2017
28 February 2018
Accepted 10 March 2018

Abstract

Purpose – This paper aims to present an approach to detect interrelations among product categories, which are then used to produce a partition of a retailer's business into subsets of categories. The methodology also yields a segmentation of shopping trips based on the composition of each shopping basket.

Design/methodology/approach – This work uses scanner data to uncover product category interdependencies. As the number of possible relationships among them can be very large, the authors introduce an approach that generates an intuitive graphical representation of these interrelationships by using data analysis techniques available in standard statistical packages, such as multidimensional scaling and clustering.

Findings – The methodology was validated using data from a supermarket store. The analysis for that particular store revealed four groups of products categories that are often jointly purchased. The study of each of these groups allowed us to conceive the retail store under study as a small set of sub-businesses. These conclusions reinforce the strategic need for proactive coordination of marketing activities across interrelated product categories.

Research limitations/implications – The approach is sufficiently general to be applied beyond the supermarket industry. However, the empirical findings are specific to the store under analysis. In addition, the proposed methodology identifies cross-category interrelations, but not their underlying sources (e.g. marketing or non-marketing interrelations).

Practical implications – The results suggest that retailers could potentially benefit if they transition from the traditional category management approach where retailers manage product categories in isolation into a customer management approach where retailers identify, acknowledge and leverage interrelations among product categories.

Originality/value – The authors present a fast and wide-range approach to study the shopping behavior of customers, detect cross-category interrelations and segment the retailer's business and customers based on information about their shopping baskets. Compared to existing approaches, its simplicity should facilitate its implementation by practitioners.

Keywords Customer segmentation, Marketing analytics, Category management, Market basket analysis, Multidimensional scaling, Retail Management

Paper type Research paper

Retailers face the problem of managing tens of thousands of products. For each of these products, retailers need to make marketing decisions concerning price, promotions,

This work was partially funded by the Complex Engineering Systems Institute (CONICYT – PIA – FB0816), by the Fondecyt Grant 1161333 (CONICYT) and by the National Doctorate Program (CONICYT).

European Journal of Marketing
Vol. 52 No. 7/8, 2018
pp. 1550-1573

0309-0566
DOI 10.1108/EJM-06-2017-0367

assortment, space and location for each stock keeping unit (SKU). The decisions for each SKU not only affect its own sales but also the sales of many other products (Ma *et al.*, 2012). In practice, it is too complex to consider all possible interactions in this marketing decision process (Gooner *et al.*, 2011). The category management approach is an attempt to tackle this challenge through the decomposition of this management problem into a set of sub-problems. Each of these is then considered as being practically independent from the rest. This is achieved by grouping highly interrelated products into categories in such a way that products contained in one category are almost independent from products of other categories. This property enables retailers to manage their categories as strategic business units (SBUs) with significant autonomy (Nielsen, 1992). In some cases, the management of each SBU is supported by partnerships between retailers and a specific manufacturer, which assumes the role of a category captain (Mouzas, 2006).

In category management, the definition of SBUs (categories) is usually done by grouping items assumed to exhibit high substitutability or sometimes high complementarity (Gooner *et al.*, 2011). For example, if the price of one product (e.g. Tide detergent) has a strong impact on the demand for other products (e.g. Gain detergent and Snuggle softener), these products may be grouped into the same SBU. Evidently, this approach significantly reduces the complexity of the retailer's problem, but it also ignores cross-category opportunities that arise because of the existence of other types of interrelations (Seetharaman *et al.*, 2005; Srinivasan *et al.*, 2008). For example, category management initiatives often involve the elimination of slow-moving SKUs. In this regard, Dupre and Gruen (2004) reported a 25 per cent reduction in the number of SKUs for a German retailer in the cleaning and household supply category. Assortment reductions may yield better performance metrics (e.g. contribution margin) for the focal category but may have negative externalities on the demand of other related categories, when interrelations among categories are ignored.

As proposed by Manchanda *et al.* (1999), there are different types of interrelations. Specifically, a distinction can be made between marketing interrelations (marketing cross-effects) and non-marketing interrelations (coincidence). The first type is associated with sets of products such that a marketing decision concerning one SKU (e.g. a promotion or price cut on Tide detergent) affects the sales of others (e.g. Snuggle softener). The detection of these interrelations can be made, for example, by estimating cross-price elasticities of demand. For instance, Manchanda *et al.* (1999) find significant (and asymmetric) cross-price elasticities of purchase incidence between detergent and softener and between cake mix and frosting. These cross-price elasticities are estimated to be smaller than the own-price elasticities. Furthermore, Ma *et al.* (2012) find that the magnitude of these cross-price elasticities varies greatly across brands, as certain brands within a product category (e.g. Betty Croker frosting) have a much larger impact on the sales and profits of another related product category (e.g. cake mix). In real-world situations, the task of estimating the full set of cross-elasticities across all possible pairs of products becomes very complex, and the risk of finding elasticities that just by chance are statistically significant becomes very relevant (Montgomery, 1997).

The second type of interrelations (coincidence) is associated with any non-marketing factors that drive the joint purchase of a pair of products. For instance, the fixed cost of visiting a store (e.g. traveling time and cost) may increase the size of the shopping basket (Ho *et al.*, 1998), potentially improving the chances that products from different categories (e.g. Tide detergent and Snuggle softener) might be purchased in the same shopping trip.

The category management approach typically takes into account only a subset (cross-price effects) of the full range of interrelations, and therefore, it could be enhanced by detecting and considering other interrelations that exist among different product categories, such as coincidence effects, as argued by Manchanda *et al.* (1999). The identification of cross-category interrelations can be a powerful piece of information in the process of understanding and managing the retailer's business. Specifically, it will become evident that a retail store can be envisioned as the sum of several sub-businesses. This understanding of the retailer business will provide a basis for the coordination of marketing decisions across different categories within the same store (Hruschka *et al.*, 1999).

In this paper, we will present an approach to detect these cross-category interrelations. In addition, we will show how starting from these interrelations, it is possible to produce a partition of the retailer's business and a segmentation of shopping trips. We are particularly interested in developing and introducing an approach that can be easily applied and replicated by practitioners. In this respect, the following properties should be desirable for an approach serving this purpose:

- wide range, i.e. suitable for analyzing several product categories simultaneously;
- low complexity and fast replication, i.e. being able to execute the analysis by using standard statistical software in a relatively short amount of time; and
- data availability, i.e. the necessary data should be easily obtained from the retailer's data sources and intuitive graphical display of the key managerial insights.

Most of the previous approaches reported in the marketing literature share some, but typically not all, of these properties. In particular, most of them show applications for no more than five product categories (Andrews and Currim, 2002; Erdem,1998; Heilman and Bowman, 2002; Ma *et al.*, 2012; Manchanda *et al.*, 1999; Russell and Kamakura, 1997; Russell and Petersen, 2000; Seetharaman *et al.*,1999); while others require the estimation of a considerable number of choice models (Bell and Lattin, 1998; Chib *et al.*, 2002). In contrast, Tanusondjaja *et al.* (2016) apply the duplication of purchase law (Goodhardt *et al.*, 1984 and Ehrenberg *et al.*, 2004) to study joint purchase patterns across 28 categories. For a given pair of categories, they compute a duplication factor that measures how the likelihood of buying one of the product categories increases if the other category is also purchased. Mild and Reutterer (2003) use collaborative filtering techniques to predict purchases across 29 categories.

Most of the papers mentioned above typically do not provide a friendly visualization of the interrelations among product categories. We will show that a graphical display of the categories that are more closely interrelated provides additional insights for supporting marketing decisions such as cross-category promotion, store layout and role definition. The following table provides a comparison of some of the methods used in the literature and our proposed approach along several dimensions. The proposed approach has advantages in terms of most (albeit not all) of these dimensions. In particular, the ability to consider a large number of categories and its simplicity are important advantages for this approach to benefit practitioners and analysts (Table I).

More specifically, in this article, we analyze scanner data by using techniques that are available in several statistical packages (e.g. multidimensional scaling [MDS] and cluster analysis) to study interrelations among 33 different product categories. Although MDS models have been widely used in perceptual analysis (Ghose, 1998; Chintagunta *et al.*, 2002; González-Benito *et al.*, 2009), choice analysis within a category (Andrews and Manrai, 1999; Elrod, 1988) and cross-citation analysis (Galvagno, 2011), we will describe a different application of this

Table I. Comparison of existing approaches

Reference/ approach	No. of categories	Complexity (model)	Choice data/ customer ID	Marketing info.	Visualization of cross-category interaction	Focus
Andrews and Currim (2002)	3	High (intercategory logit mixture)	Panel/Yes	Yes	No	Correlate sensitivity to marketing efforts across product categories
Erdem (1998)	2	High (Bayesian learning model)	Panel/Yes	Yes	No	Impact of marketing mix strategy in one product category on quality beliefs in a different category
Hruschka *et al.* (1998)	28	Medium (multivariate logit model)	Transactional/ No	Yes	Yes	Measure cross-category dependence and sales-promotion effects
Manchanda *et al.* (1999)	4	High (multivariate probit)	Panel/Yes	Yes	No	Measure complementarity and coincidence cross-category interactions
Russell and Kamakura (1998)	4	High (latent class Poisson-multinomial model)	Panel/Yes	No	No	Measure correlations for brand preferences across different categories
Russell and Petersen (2000)	4	Medium (multivariate logit model)	Panel/Yes	Yes	No	Measure cross-category complementarity and substitution
Seetharaman *et al*, (1999)	5	High (multinomial probit model with Bayesian variance decomposition)	Panel/Yes	Yes	No	Measure correlation in state-dependence and marketing mix sensitivity across product categories
Bell and Lattin (1998)	12	High (three-level nested logit)	Panel/Yes	Yes	No	Study store choice as a function of marketing mix decisions across multiple categories
Tanusondjaja *et al.* (2016)	28	Low (duplication of purchase law)	Transactional/ No	No	No	Study cross-category purchase patterns using the duplication coefficient
Chib *et al.* (2002)	12	High (multivariate probit)	Panel/Yes	Yes	No	Measure complementarity and coincidence cross-category interactions
Ma *et al.* (2012)	2	High (multivariate multinomial)	Panel/Yes	Yes	Yes	Studies complementarity and coincidence across categories and brands
Mild and Reutterer (2003)	29	Low (collaborative filtering and Jaccard proximity between customers)	Transactional/ No	No	No	Prediction of cross-category purchases
This Paper	3	Low (MDS and cluster analysis)	Transactional/ No	No	Yes	Measure dependence across large numbers of categories and identify representative basket types

technique to the analysis of interrelations among categories and the segmentation of shoppers according to their basket composition. This procedure can be executed for a wide range of products in a very short amount of time. As we mentioned before, these properties facilitate its implementation by a retailer, something that is very appealing from an applied and managerial point of view. Furthermore, this procedure does not require marketing mix information (e.g. price and feature) or identifiers for individual customers.

This last feature implies that the approach has some advantages, but also some disadvantages. An obvious advantage is that it is not necessary to collect marketing mix or household information. Even though individual information can be obtained by tracking a panel of customers, when using panel data, we run into the risk of using a sample of shoppers that might not properly represent the universe of shoppers. Furthermore, if the number of customers included in the sample is not big enough – as it is often the case in this type of micro-level analyses (Russell and Kamakura, 1997) – it will be also challenging to obtain a reliable segmentation. In addition, macro-level analyses, i.e. transaction data aggregated by day, week or month and by store, chain or market, do not contain enough information to analyze shopping basket composition. Given that the analysis presented in this paper is based on the universe of transactions, instead of a potentially biased sample (e.g. panel data from market research studies or loyalty card data sets), representativeness problems do not apply to our results. This is relevant if a large proportion of customers are not loyalty card holders or decline participating in market research studies. Using transactional data may yield large databases that could potentially be more difficult to handle. This issue can be easily addressed by selecting random samples of transactions. Also, the computational burden is not that large as the proposed approach uses the individual transactions only once to compute aggregate measures of joint purchase propensity between pairs of product categories.

It is also necessary to acknowledge that the procedure discussed in this paper cannot be used to disentangle the source of the different cross-category interrelations (e.g. marketing or non-marketing interrelations). But, despite this limitation, we believe that its practical advantages will help practitioners to make better use of their information resources to understand their customers.

Finally, it is important to note that value of the proposed approach does not stem from the specific insights we derive for the store under analysis in this paper, but from the features of the proposed methodology. In fact, the approach may be used beyond the supermarket industry. In particular, its application should be valuable for any business or activity where customers buy, hire or consume bundles of products or services from the same provider (e.g. department stores, financial institutions and online media).

1. Empirical setting

The data come from a mid-sized supermarket in Latin America. Its assortment approximately consists of 7,000 different products. The average basket includes products from 3.6 different categories. The data set contains information for each purchase (transaction) recorded in a single month (July 2000), specifically products sold, units sold, date and time. The data were organized by defining groups of products determining which products belong to each of the 33 product categories analyzed in this research (e.g. cereals and rice). These product categories were selected by first identifying the 25 categories most frequently purchased at this store. We then added eight categories considering products that, in spite of having a relatively low frequency of purchase, exhibited either a relatively high participation in the ticket in terms of dollar value (e.g. diapers) or belonged to a basket with high total dollar value (e.g. shampoo and conditioner).

Transactions not including products from at least one of these categories were not considered; this corresponds to 13.7 per cent of all transactions.

As given by Fader and Lodish (1990), to provide some descriptive information about our data set, we constructed a set of variables to describe each product category under study. The analysis of these variables will complement the conclusions we will obtain about the segmentation of shopping trips. These descriptive variables are as follows:

- LNT_j: This is the number of transactions, including category j. This variable measures how frequently each product category is purchased, and a natural logarithm transformation is used to facilitate scaling issues.
- $LCatSales_j$: This is the total sales of category j. This variable measures the size of each category in terms of dollar sales, and as before, a natural logarithm transformation is used.
- $LTotExp_j$: This is the total expenditure of shoppers of category j. This log-transformed variable measures the overall contribution of shopping trips, including each category to the revenues of the store.
- $AvgTSize_j$: This is the average transaction size ($) of purchases, including category j. This variable measures the average expenditure of a shopping trip including each category.
- $LRSCatExp_j$: This is the ratio between sales of category j and total expenditure of shoppers of category j. This log-transformed variable measures the contribution of each category to the total revenues generated by shopping occasions, including that category.
- $AvgNCat_j$: This is the average number of different categories included in purchases of category j. This variable measures the breadth of the average shopping basket, including product category j.

The values of these variables for each category are shown in Table II.

2. Definitions and methodology

We begin by quantifying the occurrence of purchases simultaneously, including products from different categories. To analyze their frequency, consider the following definitions:

- A = Set of purchases that include products from Category A.
- B = Set of purchases that include products from Category B.
- $A \cup B$ = Set of purchases that include either products from Category A, products from Category B or products from both categories.
- $A \cap B$ = Set of purchases that simultaneously include products from Category A and products from Category B.

According to Tan *et al.* (2002), several measures can be used for detecting association patterns. The most basic measure can be constructed by determining the fraction of all purchases that simultaneously include products from both categories (A ∩ B). This simple measure is referred to as support. This metric, however, is not sufficient to identify strong associations between product categories. For example, a high level of support may be observed when one of the categories is very frequently purchased, without necessarily implying any linkages between the categories. To address this issue, we need to normalize this observed probability of joint purchase (support) by the frequency of purchase (size) of

Category	LNT	LCatSales*	LTotExp*	AvgTSize*	LRSCatExp
Oil	9.09	9.57	12.01	11,150.38	−2.44
Baby food	7.94	8.35	10.95	12,063.01	−2.59
Cereals	8.61	8.86	11.54	11,286.27	−2.69
Rice	8.79	8.99	11.87	12,979.30	−2.88
White sugar	9.20	9.76	12.08	10,730.08	−2.33
Powdered juice	8.54	8.20	11.23	8,894.32	−3.04
Soft drink	9.57	10.01	11.91	6,195.56	−1.90
Candy	8.03	8.21	10.66	8,361.63	−2.46
Noodles	9.07	9.40	11.99	11,107.87	−2.60
Cookies	8.99	9.00	11.56	7,849.05	−2.56
Flour	8.03	8.34	11.20	14,192.00	−2.86
Yogurt	9.07	9.05	11.70	8,262.45	−2.64
Cheese	8.62	9.02	11.38	9,412.57	−2.36
Tomato sauce	8.69	8.41	11.72	12,433.36	−3.31
Tea	8.95	9.09	11.87	11,132.68	−2.78
Ice-cream and frozen dessert	6.87	7.47	9.63	9,417.04	−2.15
Cold cuts	9.31	9.62	11.87	7,743.62	−2.25
Margarine	8.86	8.84	11.68	10,089.59	−2.84
Shampoo and conditioner	7.64	8.44	10.76	13,532.03	−2.32
Diapers	7.61	9.19	10.19	7,906.73	−1.00
Bakery	10.12	9.96	12.08	4,245.76	−2.12
Produce	9.48	9.51	11.76	5,875.71	−2.26
China, glasses, pots and pans	7.02	8.36	9.02	4,446.02	−0.66
Beer	7.77	8.08	9.95	5,323.15	−1.87
Wine	8.83	9.71	11.19	6,360.40	−1.48
Coffee	8.59	9.39	11.55	11,621.84	−2.16
Mayonnaise	8.39	8.61	11.38	11,934.85	−2.78
Sanitary towel	7.66	7.83	10.61	11,448.17	−2.77
Meat	9.17	10.24	11.72	7,665.16	−1.48
Toilet paper	9.42	9.61	12.16	9,213.59	−2.55
Detergent	8.99	9.78	11.90	11,024.16	−2.13
Milk	9.34	9.68	11.97	8,278.12	−2.29
Powdered milk	8.30	9.62	11.26	11,519.85	−1.64

Note: The values of the variables followed by a * were multiplied by a constant for confidentiality reasons

Table II.
Descriptive statistics for each product category

the categories. This will be accomplished by relying on two alternative metrics: the Jaccard ratio and the duplication factor (or lift).

In terms of the Jaccard metric, we start by identifying the transactions that include products from category A, B or both. We then estimate the fraction of these transactions that include products from both categories (i.e., A and B). This metric, known as the Jaccard similarity measure (Tan *et al.*, 2002), can be determined as follows:

$$\text{Jaccard}(\text{A}, \text{B}) = \text{P}(A \wedge B | A \vee B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

where $\wedge$ and $\vee$ denote the AND and OR logical operators; and $|s^2|$ denotes the number of elements of a set (i.e. its cardinality). To illustrate this definition, consider the following example where 100 purchases included beer, but not soft drinks; 200 purchases included soft drinks, but not beer; and 150 purchases included both soft drinks and beer. Then the conditional probability of joint purchase can be estimated as follows:

$$\text{Jaccard}(\text{SoftDrinks}, \text{Beer}) = \text{p}(\text{SoftDrinks} \wedge \text{Beer} | \text{SoftDrinks} \vee \text{Beer}) = \frac{150}{100 + 200 + 150} = 33.3\%$$

The second approach is based on the duplication factor (Goodhardt *et al.*, 1984; Ehrenberg *et al.*, 2004; Tanusondjaja *et al.*, 2016), which normalizes the observed joint purchase probability [P(A ∩ B)] by the expected probability of purchase of both categories under independence, this is P(A)P(B). This metric is also referred to as lift in the association rules literature (Tan *et al.*, 2002) and can be estimated as follows:

$$\text{Duplication}(\text{A},\text{B}) = P(A \wedge B | \text{A}) / \text{P}(\text{B}) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Applying this definition to the same example described above and assuming that the total of number of transactions at the store equals 3,000, the duplication factor can be estimated as follows:

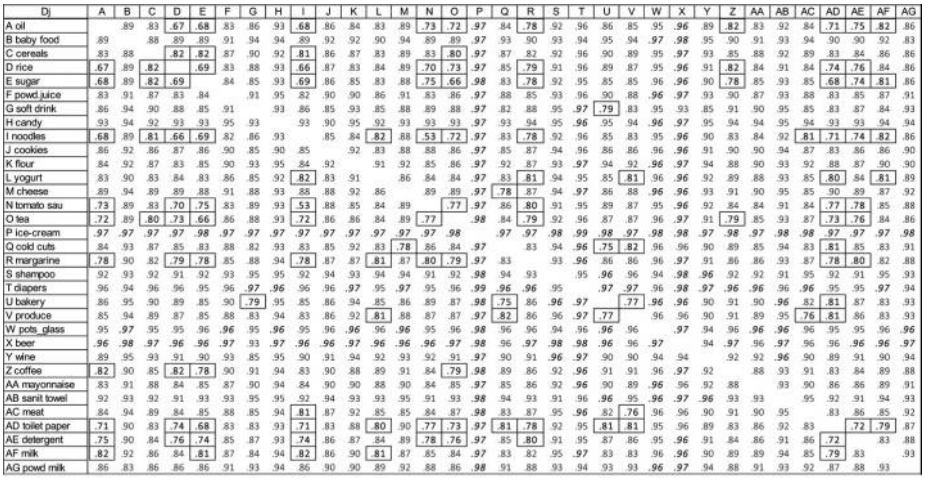
$$\text{Duplication}(\text{SoftDrinks}, \text{Beer}) = \frac{P(SoftDrinks \wedge Beer\text{eer})}{P(SoftDrinks)P(Beer)} = \frac{\frac{150}{3{,}000}}{\left(\frac{150 + 100}{3{,}000}\right)\left(\frac{150 + 200}{3{,}000}\right)} = 5.1$$

Both metrics (Jaccard and duplication) are symmetric, and we use them to construct a square matrix that describes the association between each pair of product categories. It is certainly not trivial to derive joint purchase insights by mere inspection of the matrix. Instead, it is more convenient to summarize this information by using multivariate analysis techniques and ideally generating a graphical representation of insights, which can be more easily understood and used by practitioners.

A useful technique to generate such an intuitive graphical representation is MDS (Urban and Hauser, 1993). This technique uses as an input the dissimilarity or similarity between different pairs of objects (i.e. product categories in our approach) and then yields a (multidimensional) map where more similar objects appear closer to each other. In our case, we will use the Jaccard and duplication metrics of joint purchase as similarity measures. Accordingly, pairs of products more frequently jointly purchased will be located closer to each other on this map. To implement this approach by using standard statistical packages, it may be convenient to rely on a dissimilarity instead of a similarity measure. For the Jaccard metric, this can be easily obtained by estimating the complement of the Jaccard ratio:

$$d_J(A,B) = 1 - \text{P}(A \wedge B | A \vee B) = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \text{Jaccard}(\text{A},\text{B}). \tag{3}$$

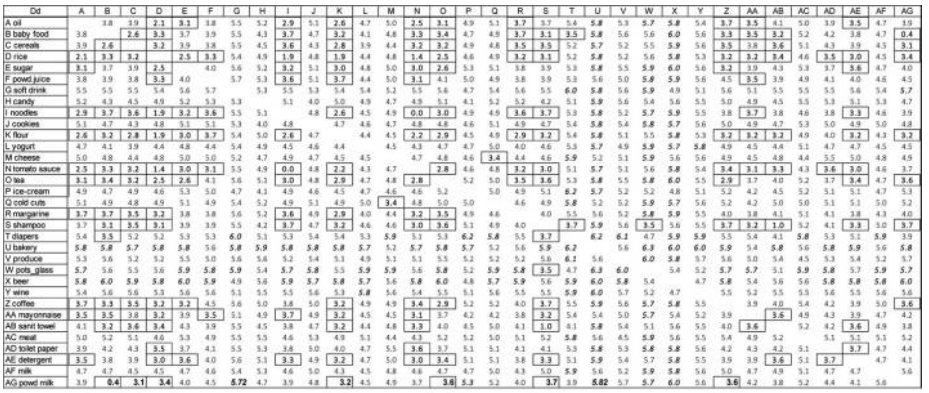
where $d_J(A,B)$ denotes the Jaccard dissimilarity (or distance) between product Categories A and B. For the duplication factor, we will compute the difference between the maximum



Dj	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
A oil		.89	.83	.67	.68	.83	.86	.93	.68	.86	.84	.83	.89	.73	.72	.97	.84	.78	.92	.96	.86	.85	.95	.96	.89	.82	.83	.92	.84	.71	.75	.82	.86
B baby food	.89		.88	.89	.89	.91	.94	.94	.89	.92	.92	.90	.94	.89	.89	.97	.93	.90	.93	.94	.95	.94	.97	.98	.95	.90	.91	.93	.94	.90	.90	.92	.83
C cereals	.83	.88		.82	.82	.87	.90	.92	.81	.86	.87	.83	.89	.83	.80	.97	.87	.82	.92	.96	.90	.89	.95	.97	.93	.85	.88	.92	.89	.83	.84	.86	.86
D rice	.67	.89	.82		.69	.83	.88	.93	.66	.87	.83	.84	.89	.70	.73	.97	.85	.79	.91	.96	.89	.87	.95	.96	.91	.82	.84	.91	.84	.74	.76	.84	.86
E sugar	.68	.89	.82	.69		.84	.85	.93	.69	.86	.85	.83	.88	.75	.66	.98	.83	.78	.92	.95	.85	.85	.96	.96	.90	.78	.85	.93	.85	.68	.74	.81	.86
F powd juice	.83	.91	.87	.83	.84		.91	.95	.82	.90	.90	.86	.91	.83	.86	.97	.88	.85	.93	.96	.90	.88	.96	.97	.93	.90	.87	.93	.88	.83	.85	.87	.91
G soft drink	.86	.94	.90	.88	.85	.91		.93	.86	.85	.93	.85	.88	.89	.88	.97	.82	.88	.95	.97	.79	.83	.95	.93	.85	.91	.90	.95	.85	.83	.87	.84	.93
H candy	.93	.94	.92	.93	.93	.95	.93		.93	.90	.95	.92	.93	.93	.93	.97	.93	.94	.95	.96	.95	.94	.96	.97	.95	.94	.94	.95	.94	.93	.93	.94	.94
I noodles	.68	.89	.81	.66	.69	.82	.86	.93		.85	.84	.82	.88	.53	.72	.97	.83	.78	.92	.96	.85	.83	.95	.96	.90	.83	.84	.92	.81	.71	.74	.82	.86
J cookies	.86	.92	.86	.87	.86	.90	.85	.90	.85		.92	.83	.88	.88	.86	.97	.85	.87	.94	.96	.86	.86	.96	.96	.90	.90	.90	.94	.87	.83	.86	.86	.90
K flour	.84	.92	.87	.83	.85	.90	.93	.95	.84	.92		.91	.92	.85	.86	.97	.92	.87	.93	.97	.94	.92	.96	.97	.94	.88	.90	.93	.92	.88	.87	.90	.90
L yogurt	.83	.90	.83	.84	.83	.86	.85	.92	.82	.83	.91		.86	.84	.84	.97	.83	.81	.94	.95	.85	.81	.96	.96	.92	.89	.88	.93	.85	.80	.84	.81	.89
M cheese	.89	.94	.89	.89	.88	.91	.88	.93	.88	.88	.92	.86		.89	.89	.97	.78	.87	.94	.97	.86	.88	.96	.96	.93	.91	.90	.95	.85	.90	.89	.87	.92
N tomato sau	.73	.89	.83	.70	.75	.83	.89	.93	.53	.88	.85	.84	.89		.77	.97	.86	.80	.91	.95	.89	.87	.95	.96	.92	.84	.84	.91	.84	.77	.78	.85	.88
O tea	.72	.89	.80	.73	.66	.86	.88	.93	.72	.86	.86	.84	.89	.77		.98	.84	.79	.92	.96	.87	.87	.96	.97	.91	.79	.85	.93	.87	.73	.76	.84	.86
P ice-cream	.97	.97	.97	.97	.98	.97	.97	.97	.97	.97	.97	.97	.97	.97	.98		.97	.97	.98	.99	.98	.97	.98	.98	.97	.98	.98	.97	.98	.97	.97	.97	.98
Q cold cuts	.84	.93	.87	.85	.83	.88	.82	.93	.83	.85	.92	.83	.78	.86	.84	.97		.83	.94	.96	.75	.82	.96	.96	.90	.89	.85	.94	.83	.81	.85	.83	.91
R margarine	.78	.90	.82	.79	.78	.85	.88	.94	.78	.87	.87	.81	.87	.80	.79	.97	.83		.93	.96	.86	.86	.96	.97	.91	.86	.86	.93	.87	.78	.80	.82	.88
S shampoo	.92	.93	.92	.91	.92	.93	.95	.95	.92	.94	.93	.94	.94	.91	.92	.98	.94	.93		.95	.96	.96	.94	.98	.96	.92	.92	.91	.95	.92	.91	.95	.93
T diapers	.96	.94	.96	.96	.95	.96	.97	.96	.96	.96	.97	.95	.97	.95	.96	.99	.96	.96	.95		.97	.97	.96	.98	.97	.96	.98	.96	.96	.95	.95	.97	.94
U bakery	.86	.95	.90	.89	.85	.90	.79	.95	.85	.86	.94	.85	.86	.89	.87	.98	.75	.86	.96	.97		.77	.96	.96	.90	.91	.90	.96	.82	.81	.87	.83	.93
V produce	.85	.94	.89	.87	.85	.88	.83	.94	.83	.86	.92	.81	.88	.87	.87	.97	.82	.86	.96	.97	.77		.96	.96	.90	.91	.89	.95	.76	.81	.86	.83	.93
W pots_glass	.95	.97	.95	.95	.96	.96	.95	.96	.96	.96	.96	.96	.96	.96	.98	.96	.96	.96	.94	.96	.96		.97	.94	.96	.96	.96	.96	.96	.95	.95	.96	.96
X beer	.96	.98	.97	.96	.96	.97	.93	.97	.96	.96	.97	.96	.96	.96	.97	.98	.96	.97	.98	.98	.96	.96	.97		.94	.97	.96	.97	.96	.96	.96	.96	.97
Y wine	.89	.95	.93	.91	.90	.93	.85	.95	.90	.91	.94	.92	.93	.92	.91	.97	.90	.91	.96	.97	.90	.90	.94	.94		.92	.92	.96	.90	.89	.91	.90	.94
Z coffee	.82	.90	.85	.82	.78	.90	.91	.94	.83	.90	.88	.89	.91	.84	.79	.98	.89	.86	.92	.96	.91	.91	.96	.97	.92		.88	.93	.91	.83	.84	.89	.88
AA mayonnaise	.83	.91	.88	.84	.85	.87	.90	.94	.84	.90	.90	.88	.90	.84	.85	.97	.85	.86	.92	.96	.90	.89	.96	.96	.92	.88		.93	.90	.86	.86	.89	.91
AB sanit towel	.92	.93	.92	.91	.93	.93	.95	.95	.92	.94	.93	.93	.95	.91	.93	.98	.94	.93	.91	.96	.96	.95	.96	.97	.96	.93	.93		.95	.92	.91	.94	.93
AC meat	.84	.94	.89	.84	.85	.88	.85	.94	.81	.87	.92	.85	.85	.84	.87	.98	.83	.87	.95	.96	.82	.76	.96	.96	.90	.91	.90	.95		.83	.86	.85	.92
AD toilet paper	.71	.90	.83	.74	.68	.83	.83	.93	.71	.83	.88	.80	.90	.77	.73	.97	.81	.78	.92	.95	.81	.81	.95	.96	.89	.83	.86	.92	.83		.72	.79	.87
AE detergent	.75	.90	.84	.76	.74	.85	.87	.93	.74	.86	.87	.84	.89	.78	.76	.97	.85	.80	.91	.95	.87	.86	.95	.96	.91	.84	.86	.91	.86	.72		.83	.88
AF milk	.82	.92	.86	.84	.81	.87	.84	.94	.82	.86	.90	.81	.87	.85	.84	.97	.83	.82	.95	.97	.83	.83	.96	.96	.90	.89	.94	.85	.79	.83		.93
AG powd milk	.86	.83	.86	.86	.86	.91	.93	.94	.86	.90	.90	.89	.92	.88	.86	.98	.91	.88	.93	.94	.93	.93	.96	.97	.94	.88	.91	.93	.92	.87	.88	.93	

Figure 1.
Joint purchase dissimilarity matrix using the DJ Jaccard ratio*

Notes: *Jaccard dissimilarities one standard deviation below the mean are shown in bold. Jaccard dissimilarities one standard deviation above the mean are shown in italics



Dd	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
A oil		3.8	3.9	2.1	3.1	3.8	5.5	5.2	2.9	5.1	2.6	4.7	5.0	2.5	3.1	4.9	5.1	3.7	3.7	5.6	5.8	5.3	5.7	5.8	5.4	3.7	3.5	4.1	5.0	3.9	3.5	4.7	3.9
B baby food	3.8		2.6	3.3	3.7	3.9	5.5	4.3	3.7	4.7	3.2	4.1	4.8	3.3	3.4	4.7	4.9	3.7	3.1	3.5	5.8	5.6	5.6	6.0	5.6	3.3	3.5	3.2	5.2	4.2	3.8	4.7	0.4
C cereals	3.9	2.6		3.2	3.9	3.8	5.5	4.5	3.6	4.3	2.8	3.9	4.4	3.2	3.2	4.9	4.8	3.5	3.5	5.2	5.7	5.2	5.5	5.9	5.6	3.5	3.8	3.6	5.1	4.3	3.9	4.5	3.1
D rice	2.1	3.3	3.2		2.5	3.3	5.4	4.5	1.9	4.8	1.9	4.4	4.8	1.4	2.5	4.6	4.9	3.2	3.1	5.2	5.8	5.2	5.6	5.8	5.3	3.2	3.2	3.4	4.6	3.5	3.0	4.5	3.4
E sugar	3.1	3.7	3.9	2.5		4.0	5.6	5.2	3.2	5.1	3.0	4.8	5.0	3.0	2.6	5.3	5.1	3.8	3.9	5.3	5.8	5.5	5.9	6.0	5.6	3.2	3.9	4.3	5.3	3.7	3.6	4.7	4.0
F powd juice	3.8	3.9	3.8	3.3	4.0		5.7	5.3	3.6	5.1	3.7	4.4	5.0	3.1	4.1	5.0	4.9	3.8	3.9	5.3	5.6	5.0	5.8	5.9	5.6	4.5	3.5	3.9	4.9	4.1	4.0	4.6	4.5
G soft drink	5.5	5.5	5.5	5.4	5.6	5.7		5.3	5.5	5.3	5.4	5.4	5.2	5.5	5.6	4.7	5.4	5.6	5.5	6.0	5.8	5.6	5.9	4.9	5.1	5.6	5.1	5.5	5.5	5.5	5.6	5.4	5.7
H candy	5.2	4.3	4.5	4.9	5.2	5.3	5.3		5.1	4.0	5.0	4.9	4.7	4.9	5.1	4.1	5.2	5.2	4.2	5.1	5.8	5.6	5.4	5.6	5.5	5.0	4.9	4.5	5.5	5.3	5.1	5.3	4.7
I noodles	2.9	3.7	3.6	1.9	3.2	3.6	5.5	5.1		4.8	2.6	4.5	4.9	0.0	3.0	4.9	4.9	3.6	3.7	5.3	5.8	5.2	5.7	5.9	5.5	3.8	3.7	3.8	4.6	3.8	3.3	4.6	3.9
J cookies	5.1	4.7	4.3	4.8	5.1	5.1	5.3	4.0	4.8		4.7	4.6	4.7	4.8	4.8	4.6	5.1	4.9	4.7	5.4	5.8	5.4	5.8	5.7	5.6	5.0	4.9	4.7	5.3	5.0	4.9	5.0	4.8
K flour	2.6	3.2	2.8	1.9	3.0	3.7	5.4	5.0	2.6	4.7		4.4	4.5	2.2	2.9	4.5	4.9	2.9	3.2	5.4	5.8	5.1	5.5	5.8	5.3	3.2	3.2	3.2	4.9	4.0	3.2	4.3	3.2
L yogurt	4.7	4.1	3.9	4.4	4.8	4.4	5.4	4.9	4.5	4.6	4.4		4.5	4.3	4.7	4.7	5.0	4.0	4.6	5.3	5.7	4.9	5.9	5.7	5.8	4.9	4.5	4.4	5.1	4.7	4.7	4.5	4.5
M cheese	5.0	4.8	4.4	4.8	5.0	5.0	5.2	4.7	4.9	4.7	4.5	4.5		4.7	4.8	4.6	3.4	4.4	4.6	5.9	5.2	5.1	5.9	5.6	5.6	4.9	4.5	4.8	4.4	5.5	5.0	4.8	4.5
N tomato sauce	2.5	3.3	3.2	1.4	3.0	3.1	5.5	4.9	0.0	4.8	2.2	4.3	4.7		2.8	4.6	4.8	3.2	3.0	5.1	5.7	5.1	5.6	5.8	5.4	3.4	3.1	3.3	4.3	3.6	3.0	4.6	3.7
O tea	3.1	3.4	3.2	2.5	2.6	4.1	5.6	5.3	3.0	4.8	2.9	4.7	4.8	2.8		5.2	5.0	3.5	3.6	5.3	5.8	5.5	5.8	6.0	5.5	2.9	3.7	4.0	5.2	3.7	3.4	4.7	3.6
P ice-cream	4.9	4.7	4.9	4.6	5.3	5.0	4.7	4.1	4.9	4.6	4.5	4.7	4.6	4.6	5.2		5.0	4.9	5.1	6.2	5.7	5.2	5.2	4.8	5.1	5.2	4.2	4.5	5.2	5.1	5.1	4.7	5.3
Q cold cuts	5.1	4.9	4.8	4.9	5.1	4.9	5.4	5.2	4.9	5.1	4.9	5.0	3.4	4.8	5.0			4.6	4.9	5.8	5.2	5.2	5.9	5.7	5.6	5.2	4.2	5.0	5.0	5.1	5.1	5.0	5.2
R margarine	3.7	3.7	3.5	3.2	3.8	3.8	5.6	5.2	3.6	4.9	2.9	4.0	4.4	3.2	3.5	4.9	4.6		4.0	5.5	5.6	5.2	5.8	5.9	5.5	4.0	3.8	4.1	5.1	4.1	3.8	4.3	4.0
S shampoo	3.7	3.1	3.5	3.1	3.9	3.9	5.5	4.2	3.7	4.7	3.2	4.6	4.6	3.0	3.6	5.1	4.9	4.0		3.7	5.9	5.6	3.5	5.6	5.5	3.7	3.2	1.0	5.2	4.1	3.3	5.0	3.7
T diapers	5.4	3.5	5.2	5.2	5.3	5.3	6.0	5.1	5.3	5.4	5.4	5.3	5.9	5.1	5.3	6.2	5.8	5.5	3.7		6.2	6.1	4.7	5.9	5.9	5.5	5.4	4.1	5.8	5.3	5.1	5.9	3.9
U bakery	5.8	5.8	5.7	5.8	5.8	5.6	5.8	5.9	5.8	5.8	5.8	5.7	5.2	5.7	5.8	5.7	5.2	5.6	5.9	6.2		5.6	6.3	6.0	6.0	5.9	5.4	5.8	5.6	5.8	5.9	5.6	5.8
V produce	5.3	5.6	5.2	5.2	5.5	5.0	5.6	5.8	5.2	5.4	5.1	4.9	5.1	5.1	5.5	5.2	5.2	5.2	5.6	6.1	5.6		6.0	5.8	5.7	5.6	5.0	5.4	4.5	5.3	5.4	5.2	5.7
W pots_glass	5.7	5.6	5.5	5.6	5.9	5.8	5.9	5.4	5.7	5.8	5.5	5.9	5.9	5.6	5.8	5.2	5.9	5.8	3.5	4.7	6.3	6.0		5.4	5.2	5.7	5.7	5.1	5.9	5.8	5.7	5.9	5.7
X beer	5.8	6.0	5.9	5.8	6.0	5.9	4.9	5.6	5.9	5.7	5.8	5.7	5.6	5.8	6.0	4.8	5.7	5.9	5.6	5.9	6.0	5.8	5.4		4.7	5.8	5.4	5.6	5.6	5.8	5.8	5.8	6.0
Y wine	5.4	5.6	5.6	5.3	5.6	5.6	5.1	5.5	5.5	5.6	5.3	5.8	5.6	5.4	5.5	5.1	5.6	5.5	5.5	5.9	6.0	5.7	5.2	4.7		5.5	5.2	5.5	5.5	5.6	5.5	5.4	5.6
Z coffee	3.7	3.3	3.5	3.2	3.2	3.2	4.5	5.6	3.0	5.0	3.2	4.9	4.9	2.4	2.9	5.2	5.2	4.0	3.7	5.5	5.9	5.6	5.7	5.8	5.5		3.9	4.0	5.4	4.2	3.9	5.0	3.6
AA mayonnaise	3.5	3.5	3.8	3.2	3.9	3.5	5.1	4.9	3.7	4.9	3.2	4.5	4.5	3.1	3.7	4.2	4.2	3.8	3.2	5.4	5.4	5.0	5.7	5.4	5.2	3.9		3.6	4.9	4.3	3.9	4.7	4.2
AB sanit towel	4.1	3.2	3.6	3.4	4.3	3.9	5.5	4.5	3.8	4.7	3.2	4.4	4.8	3.3	4.0	4.5	5.0	4.1	1.0	6.1	5.8	5.4	5.1	5.6	5.5	4.0	3.6		5.2	4.2	3.6	4.9	3.8
AC meat	5.0	5.2	5.1	4.6	5.3	4.9	5.5	5.5	4.6	5.3	4.9	5.1	4.4	4.3	5.2	5.2	5.0	5.1	5.2	5.8	5.6	4.5	5.9	5.6	5.5	5.4	4.9	5.2		5.1	5.1	5.1	5.2
AD toilet paper	3.9	4.2	4.3	3.5	3.7	4.1	5.5	5.3	3.8	5.0	4.0	4.7	5.5	3.6	3.7	5.1	5.1	4.1	4.1	5.3	5.8	5.3	5.8	5.8	5.6	4.2	4.3	4.2	5.1		3.7	4.7	4.4
AE detergent	3.5	3.8	3.9	3.0	3.6	4.0	5.6	5.3	3.3	4.9	3.2	4.7	5.0	3.0	3.4	5.1	5.1	3.8	3.3	5.1	5.9	5.4	5.7	5.8	5.5	3.9	3.9	3.6	5.1	3.7		4.7	4.1
AF milk	4.7	4.7	4.5	4.5	4.7	4.6	5.4	5.3	4.6	5.0	4.3	4.5	4.8	4.8	4.7	4.7	5.0	4.3	5.0	5.9	5.6	5.2	5.9	5.8	5.6	5.0	4.7	4.9	5.1	4.7	4.7		5.6
AG powd milk	3.9	0.4	3.1	3.4	4.0	4.5	5.72	4.7	3.9	4.8	3.2	4.5	4.9	3.7	3.6	5.3	5.2	4.0	3.7	3.9	5.82	5.7	5.7	6.0	5.6	3.6	4.2	3.8	5.2	4.4	4.1	5.6	

Figure 2.
Joint purchase dissimilarity matrix using the dD Duplication ratio*

Notes: *Duplication factor dissimilarities one standard deviation below the mean are shown in bold. Duplication factor dissimilarities one standard deviation above the mean are shown in italics

value of the duplication factor for all pairs of categories (D*) and the duplication for a given pair (A,B):

$$d_D(A,B) = \mathrm{D}^* - P(A \wedge B|\mathrm{A})/\mathrm{P}(\mathrm{B}) = \mathrm{D}^* - \frac{|A \cap B|}{|A \cup B|} = \mathrm{D}^* - \mathrm{Duplication}(\mathrm{A},\mathrm{B}). \tag{4}$$

where d_D (A,B) denotes the duplication dissimilarity between product Categories A and B. These dissimilarities [equations (3) and (4)] are shown in Figures 1 and 2, respectively. We

note that as a robustness check, we computed the Jaccard measures for the first and second halves of the data and obtained almost identical results (the correlation between both sets of Jaccard metrics is 0.996, $p < 0.001$). This suggests that similar findings will be obtained if the first, second or both halves of the data are used.

In our application, the dissimilarity matrices serve as the input for the ALSCAL algorithm in the MDS module of SPSS. The resulting map of product categories will facilitate the discovery of subsets of products that have a high joint purchase probability, as they will be located close to each other on this map. These subsets can be more formally identified using cluster analysis techniques. Categories belonging to each of these clusters will exhibit higher probabilities of being included simultaneously in purchases of products from categories belonging to the same cluster. Please note that the identification of these clusters is performed in an objective manner without relying on groupings of product categories that might be anticipated by the analyst.

Finally, when segmenting customers or their shopping baskets, it is typically important to not only identify such segments but also characterize them. This can be done by relying on additional information, which can be easily obtained from the transactional data, as shown in Table II. To facilitate the characterization of the shopping basket segments in terms of these variables, we will graphically represent these variables directly into the map. This can be easily accomplished by estimating a linear model for each category descriptor as a function of each category's coordinates on the map. More specifically, in each of these linear models, the dependent variable corresponds to one of the descriptive variables, while the independent variables correspond to the each of the dimensions of the product category map (Urban and Hauser, 1993):

$$y_{lj} = c_l + \sum_{k=1}^{K} D_{lk} x_{kj} + \varepsilon_{lj} \tag{5}$$

where y_{lj} is the value of descriptive variable l for category j; c_l is the constant of the model; x_{kj} is the position of category j on dimension k of the product category map; D_{lk} is the coefficient that relates the position of each category on dimension k to the value of the descriptive variable l; and ε_{lj} is the error term in the linear model. The coefficients D_{lk} will be the ones used to project each of the descriptive variables on the MDS map. Finally, to facilitate the implementation and replication of the methodology described in this section, a step-by-step guide can be found in the Appendix.

3. Measuring interrelationships among product categories

Using the MDS technique and relying first on the Jaccard dissimilarity matrix, we obtained the horizontal and vertical coordinates for each product category shown in Table III (please note that the results for the duplication dissimilarity will be shown at the end of this subsection). The MDS procedure yields fit measures (stress and square correlation [RSQ]) that describe the extent by which the product category map accurately displays the dissimilarities. In our application, the values of these measures are: stress = 0.364 and RSQ = 0.555, where smaller values of stress and larger values of RSQ are preferable. It is indeed possible to improve this fit by adding more dimensions to this graphical representation. We note that the conclusions may change with the number of dimensions used. Ideally, one should add dimensions until the quality of the MDS solution (e.g. stress) does not substantially improve. For example, if six dimensions were used, we would obtain a sizable improvement: stress = 0.155, RSQ = 0.772. However, as we aim to produce an intuitive graphical representation of the prevalence of joint

Category	Dim 1	Dim 2
Oil	1.00	0.25
Baby food	−0.64	1.65
Cereals	0.96	0.72
Rice	0.93	0.48
White sugar	0.98	0.21
Powdered juice	1.10	0.65
Soft drink	0.13	−1.49
Candy	−1.93	−0.29
Noodles	0.93	0.19
Cookies	0.21	−1.37
Flour	0.30	1.48
Yogurt	0.72	−0.92
Cheese	−0.03	−1.51
Tomato sauce	0.91	0.52
Tea	0.93	0.39
Ice-cream and frozen dessert	−2.08	0.45
Cold cuts	0.35	−1.19
Margarine	1.00	0.11
Shampoo and conditioner	−0.98	1.46
Diapers	−1.76	0.93
Bakery	0.15	−1.37
Produce	0.21	−1.29
China, glasses, pots and pans	−1.92	−0.33
Beer	−1.81	−0.87
Wine	−0.97	−1.30
Coffee	0.57	1.04
Mayonnaise	0.32	1.11
Sanitary towel	−1.35	1.06
Meat	0.17	−1.20
Toilet paper	0.80	−0.15
Detergent	0.81	0.31
Milk	0.43	−0.99
Powdered milk	−0.43	1.27

Table III.
MDS solution in two dimensions based on the Jaccard ratio

purchases, we will focus on the product category map with two dimensions (Figure 3), while the solution in six dimensions will be used in the next subsections (adding more dimensions did not substantially improve the stress measure).

This figure reveals an empty zone at the center of the map surrounded by categories forming an ellipse. This means that there is not a central category, implying that each category tends to be included with higher probability in purchases that already include products from a specific subset of categories. Therefore, it should be possible to segment the whole set of transactions into a reduced set of basket types. A visual inspection of the product category map in Figure 3 reveals several groups of categories across the different regions of the map. In particular, the top-right corner contains non-perishable goods, such as flour and coffee. Accordingly, these non-perishable products are more likely to be purchased together. Similarly, the bottom-right corner includes fresh products such as cold cuts and yogurt, which require faster consumption. In the bottom-left corner, we can identify categories typically related to hedonic motivations (i.e. "consumed for experiential pleasure", Khan *et al.*, 2005), such as beer and wine. The top-left region includes hygiene-related products such as diapers and shampoo. Nevertheless, rather than relying on a visual inspection of the map to identify these basket

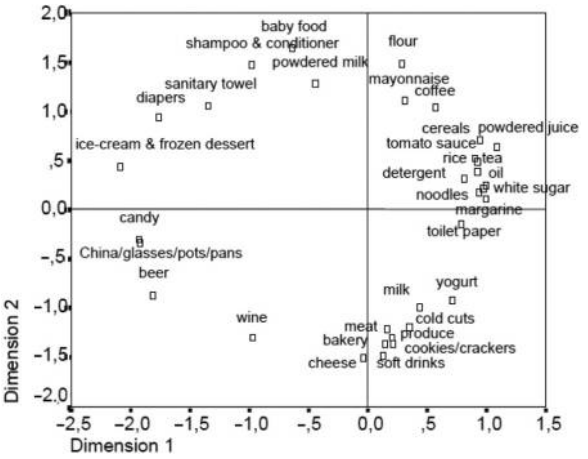


Figure 3.
MDS map of product categories based on the Jaccard metric

types, it is preferable to use an objective approach to identify sets of products that are more likely to be jointly purchased. This will be accomplished using a clustering technique that considers distance among the product categories in the map.

Considering the duplication dissimilarity matrix, we first explored the association between this metric and the Jaccard ratio by means of a scatterplot of the two metrics (Figure 4). This scatterplot shows a strong and positive association between the two measures: the correlation between these metrics is 0.562 ($p < 0.001$). Following a similar procedure as in the case of the Jaccard ratio, we obtained the map (Figure 5) for the duplication factor analysis. This map exhibits very similar features to those obtained using the Jaccard measure. In particular, once again, there is an empty zone at the center of the map surrounded by categories forming an ellipse. Through visual inspection, a similar, although not identical, grouping of categories emerges. In particular, several non-perishable product categories are located close to each other in the top-left quadrant; fresh product categories are located in the top-right quadrant; several hygiene product categories are located in the bottom-left quadrant; while hedonic product categories appear in the bottom-right quadrant. As both dissimilarity measures (Jaccard and duplication) yield similar insights, we will conduct the remaining analysis in this paper by using the Jaccard metric (more detailed results for the duplication dissimilarity measure are available from the authors upon request).

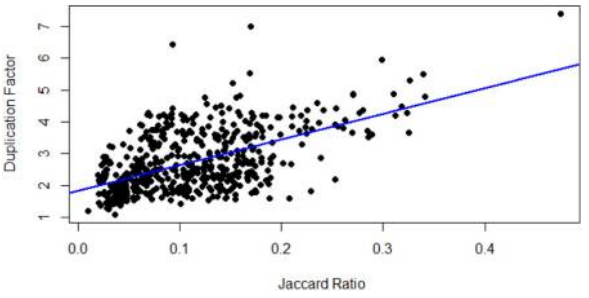


Figure 4.
Duplication factor vs Jaccard ratio

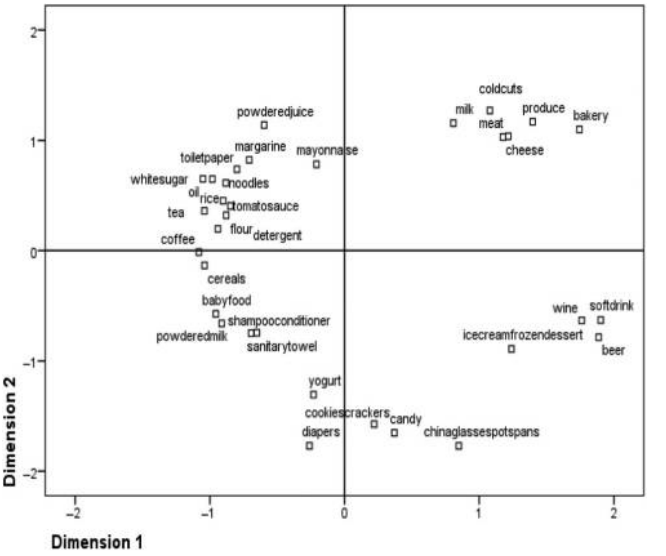


Figure 5.
MDS map of product categories based on the duplication metric

3.1 Identifying closely interrelated product categories

Groups of closely interrelated categories can be identified using the product category coordinates as inputs for a cluster analysis technique. In contrast with the last subsection where we described a product category map in two dimensions, in this subsection, we will rely on four additional dimensions to make better use of the information about joint purchases displayed in Table II. The corresponding coordinates are presented in Table IV. These six variables can then be used by the *k*-means procedure to classify product categories into basket types. The number of basket types to use can be determined by analyzing the quality of this classification. In this regard, the average silhouette is a measure of the cohesion and separation of a specific cluster solution (Rousseeuw, 1987; de Amorim and Hennig, 2015). We calculate the silhouette measure for different scenarios, ranging from grouping all categories into a single basket type and grouping them into six different basket types. The silhouette is maximized within this range by using four basket types.

This classification is shown in Figure 6. Four different types of baskets were identified, which match those that were obtained from visual inspection in the previous subsection. Specifically, we validate the four basket types previously described: non-perishable products (e.g. coffee, detergent and sugar); fresh and immediate consumption products; hygiene products; and hedonic products. Consequently, products from one of these types of baskets appear with higher probability in transactions that already include other products from the same basket.

These high probabilities of joint purchase might originate from different sources of complementarity. For example, products which are jointly consumed (i.e. which exhibit consumption complementarities) might be more likely to be included in the same shopping basket. However, there are pairs of products that are also jointly purchased with high probability, but are more difficult to anticipate, as there are no evident consumption complementarities. Consider for instance detergent and rice. Each of these does not enhance the consumption of the other. However, consumers often buy both products together. Therefore, the interrelationship among these two products does not arise from consumption

Category	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
Oil	1.43	0.54	0.79	0.25	0.30	0.22
Baby food	−0.61	1.31	−2.06	1.15	0.19	−0.21
Cereals	0.50	0.67	−1.54	0.53	−0.97	0.55
Rice	1.37	0.95	0.68	0.18	0.17	0.15
White sugar	1.44	0.48	0.41	0.43	0.35	0.53
Powdered juice	0.82	0.53	0.57	−1.23	−0.28	−1.69
Soft drink	−0.32	−2.02	0.32	1.05	0.56	0.46
Candy	−1.19	−0.08	−0.78	0.76	0.09	−2.57
Noodles	1.49	0.46	0.55	0.03	0.31	0.04
Cookies	0.07	−1.23	−1.65	0.71	−0.33	−0.69
Flour	−0.13	1.42	0.95	0.50	−1.86	0.10
Yogurt	0.60	−1.01	−1.40	−0.76	−0.59	0.15
Cheese	−0.60	−1.32	−1.02	−0.71	−1.59	0.54
Tomato sauce	1.38	0.97	0.74	−0.20	0.18	−0.31
Tea	1.37	0.74	0.18	0.69	0.23	0.62
Ice-cream and frozen dessert	−2.27	0.31	1.30	−0.03	−2.06	−0.90
Cold cuts	0.22	−1.77	−0.38	−0.61	−0.84	0.28
Margarine	0.95	0.24	−0.27	−0.72	−1.13	0.95
Shampoo and conditioner	−1.44	1.54	−0.54	−1.96	0.42	0.46
Diapers	−1.52	0.89	−1.06	−0.42	1.97	−1.63
Bakery	0.16	−2.32	−0.11	−0.23	0.21	0.11
Produce	0.42	−1.95	0.39	−0.66	0.45	−0.63
China, glasses, pots and pans	−2.31	0.43	0.33	−1.09	1.92	0.49
Beer	−2.46	−0.44	1.57	1.56	−0.16	0.18
Wine	−1.24	−0.98	1.23	1.60	1.05	0.49
Coffee	0.23	0.99	−0.08	1.61	−0.01	1.33
Mayonnaise	0.24	0.48	1.21	−0.53	−1.07	−1.52
Sanitary towel	−1.72	1.29	−0.24	−1.60	−0.19	1.31
Meat	0.47	−1.52	0.56	−0.94	0.77	−0.85
Toilet paper	1.33	−0.14	0.39	0.05	0.79	0.40
Detergent	1.19	0.63	0.57	−0.12	0.96	0.49
Milk	0.29	−1.20	0.09	−0.76	0.01	1.56
Powdered milk	−0.14	1.12	−1.72	1.47	0.13	−0.41

Table IV.
MDS solution in six dimensions

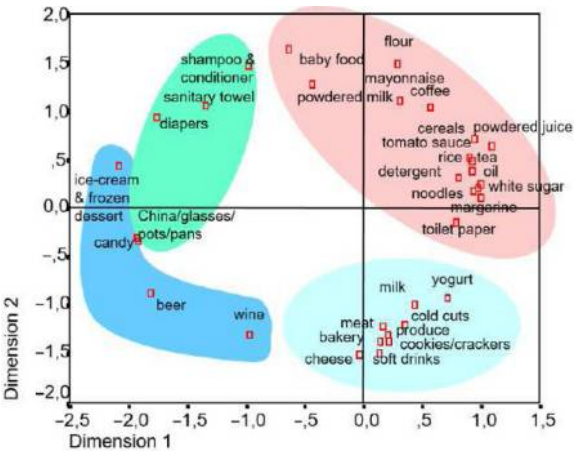


Figure 6.
Clusters of product categories

complementarities but from these two products being relevant for the same shopping occasion (e.g. weekly replenishment).

Furthermore, the shape of the four baskets is different, with some of them being more concentrated and compact than others. As we will later show, this is related to the basket size, i.e. the number of different categories included in each ticket. In particular, the immediate consumption basket (e.g. yogurt and milk) appears to be the most compact set. Focusing on transactions that include products from this set, the compactness of this basket implies that we should expect a larger number of these immediate consumption categories in the same transaction. In contrast, hedonic products (e.g. beer and wine) belong to the sparsest group. Therefore, transactions including hedonic products will include fewer categories from this basket within the same transaction. A similar conclusion applies to the hygiene basket. A more precise characterization of these baskets can be implemented by relying on the category descriptors listed in Table II, as shown in the next subsection.

3.2 Understanding differences among basket types

We now characterize and study the differences among the basket types identified in the previous section. This can be graphically accomplished by estimating linear models where the dependent variable corresponds to descriptive variables for each product category (Table II) and the independent variables correspond to the two dimensions of the MDS product category map (Table III). The results for each of these linear models are shown in Table V. Using the linear regression coefficients D_{lk} [equation (5)], it is possible to plot each of the category characteristics on the MDS map. These projections, which are shown in Figure 7, are useful to characterize each basket and identify differences among them.

This figure shows that categories from the non-perishable and immediate consumption baskets are those that account for the highest total expenditure and number of transactions. Furthermore, the non-perishable basket also exhibits a high average number of different categories in their transactions and a high average transaction size. This fact suggests that buyers of products from that basket are extremely important for the supermarket. They account for a large number of transactions which are sizable when measured in terms of dollar value and number of products from different categories.

The hygiene basket is located opposite to the immediate consumption basket. Product categories in this basket are associated with larger transactions, but they do not account for a large number of them. Finally, categories in the hedonic basket appear in shopping trips that include a small set of categories. This confirms the conclusions we drew in the previous subsection, where we discussed the sparseness of the hedonic and hygiene baskets, which implied a smaller number of categories from these baskets being jointly purchased.

Table V.
Linear model estimation of category descriptive variables as a function of MDS coordinates

	Coefficients			
Descriptor	Dimension 1	Dimension 2	R^2	Adjusted R^2
LNT	0.531**	−0.387**	0.82	0.81
LCatSales	0.321**	−0.315**	0.44	0.40
LTotExp	0.646**	−0.152*	0.81	0.80
AvgTSize	917.807**	2092.726**	0.76	0.74
LRSCatExp	−0.325**	−0.164*	0.42	0.38
AvgNCat	1.074**	1.062**	0.75	0.74

Notes: **$p < 0.01$; *$p < 0.05$

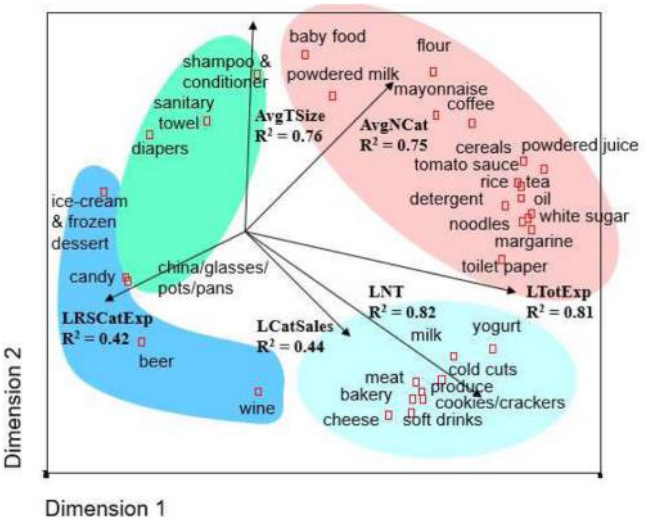


Figure 7.
MDS with projections of descriptive variables

So far, our discussion has been focused on categories. In the next section, we shift the focus of our analysis from product categories to shopping trips.

4. Segmentation of shopping trips

Relying on the four baskets previously identified, we now segment transactions based on the content of the corresponding shopping basket. Accordingly, we define the following segments: immediate consumption purchases (IC), non-perishable purchases (NP), hygiene purchases (H) and hedonic purchases (He). When more than half of the categories in a shopping trip belong to the immediate consumption basket, then we assign this shopping trip to the IC segment. The same definition applies for the non-perishable, hygiene segment and hedonic segment. Finally, we add a fifth mixed segment (M), which considers shopping trips for which half of the categories belong to one basket while the other half belongs to another. Given this segmentation of shopping occasions, it is possible to describe each segment in terms of:

- the fraction of purchase occasions that fall into each of the segments;
- the average dollar value; and
- the content of each transaction (Table VI).

This provides us with a quantitative characterization of the shopping behavior and the importance for the retailer of each of the segments previously defined.

First, we note that 84.9 per cent of the purchase occasions belong to one of these five segments. Therefore, by using these segments, we were able to classify most of the transactions of the supermarket. Furthermore, the average basket of each of the first four segments is highly concentrated (e.g. 90.9 per cent of the average basket of the IC segment corresponds to immediate consumption products). Given the results from the previous section, this is to be expected for the hygiene and hedonic segments, as these transactions include only one category most of the time. Interestingly, this high concentration suggests that the supermarket can be understood as four sub-businesses operating within the same store, each one with different service specifications. In fact, this business segmentation is consistent with some of the retail

Table VI.
Segmentation of shopping trips

	Shopping trip segment				
	Immediate consumption (IC)	Non perishable (NP)	Hygiene (H)	Hedonic (Ho)	Mixed segment (M)
Number of tickets	33,104	18,292	1,904	3,849	9,829
% from the total number of tickets	41.9	23.2	2.4	4.9	12.5
Average ticket size (TS) ($)*	4.32	11.21	5.07	3.01	6.80
Total expenditure of the segment ($)*	142,980.86	204,988.78	9,644.96	11,580.68	66,864.93
Average number of different categories	2.51	5.10	1.09	1.11	3.36
Basket composition					
% Immediate consumption products	90.9	14.4	0.1	0.9	44.7
% Non perishable products	6.9	83.2	0.1	0.2	36.8
% Hygiene products	0.5	1.1	99.4	0.0	4.0
% Social gathering products	1.7	1.1	0.1	98.8	12.6
First category most frequently purchased	Bakery (54.3%)	Toilet paper (44.40%)	China/glasses… (35.08%)	Beer (25.21%)	Bakery (35.09%)
Second category most frequently purchased	Soft drinks (29.9%)	Sugar (41.23%)	Diapers (31.32%)	Candy (16.17%)	Soft drinks (22.75%)

Note: *The figures were systematically modified (confidentiality)

formats that can be found in this industry in Latin America: bakeries selling fresh products, small grocery/convenience stores selling non-perishable products, pharmacies selling hygiene products and liquor stores, which in this region not only sell alcoholic drinks but also soft drinks. At the same time, it is important to acknowledge that if this methodology was applied to another retail store or format, the results will be different and may lead to a greater or a smaller number of sub-businesses depending on the degree of independence among the resulting clusters of product categories.

5. Managerial implications

The existence of cross-category interrelations allows us to produce a partition of the retailer's business and a segmentation of the shoppers according to the composition of their baskets. This raises important managerial implications as described next.

5.1 Role/function definition

The role/function assignment to each category is one of the most important steps in any category management process (Blattberg *et al.*, 1995). The consequences of this definition are directly related to assortment, price, display and promotion policies (Andersen Consulting and ECR Europe, 2000). In this respect, retailers select a limited number of categories for each function, such as traffic builder, transaction builder, profit contributor, cash generator and image generator.

As previously mentioned, each cluster of categories can be considered as a separate business exhibiting significant independence from the rest of the clusters (businesses) within the store. Our results imply that for each cluster (immediate consumption purchases, non-perishable purchases, hygiene purchases and hedonic purchases), a balance of functions should be designed to obtain an adequate combination of traffic, profits, cash and image. This also implies that each business should have an appropriate number of categories or subcategories to achieve different goals (e.g. traffic and profits). We note that this is markedly different from the traditional approach, where role assignment originates from the joint analysis of *all* product categories across *all* transactions identifying categories that are major contributors of traffic, profits, image, etc. without accounting for the differences among businesses within a store (Blattberg *et al.*, 1995). In particular, ignoring the underlying business structure could lead to some of these clusters (e.g. hygiene products) not having any categories assigned to an important role (e.g. traffic builder). In addition, with our methodology, we can identify product categories that, when considering all transactions, are not large contributors to specific goal (e.g. profits) but play a major role for a given business. A good example of this is the beer category, which might not be a large traffic contributor when considering all transactions and segments but is a very frequent and important category within the hedonic business. Consequently, the beer category should receive enough prominence and resources (e.g. shelf space and promotional display) to attract traffic from customers buying products belonging to the hedonic cluster.

5.2 Tactical decisions

As we previously mentioned, interrelations among products may originate from consumption compatibility, and these may give rise to opportunities for offering promotional bundles (e.g. coffee and cereal). However, these interrelations may also arise from shopping occasion compatibility (e.g. products such as noodles and detergent, which are not compatible in terms of consumption, but are often part of the same shopping occasion). These create new and hopefully untapped cross-category promotional opportunities. For example, a retailer may implement a promotional checklist, providing incentives (e.g. discounts) to customers if they buy at least one product from each of the categories belonging to a particular cluster. In addition, the retailer may offer promotional incentives to persuade a customer who is on a

certain shopping trip buying products that primarily belong to a certain cluster to return to the store to buy products from a different cluster.

5.3 Store layout

Our analysis also provides some guidelines for the design of the store layout. On the one hand, if the goal is to place products trying to minimize the total time that it takes a shopper to find all products in her basket, products should be located taking into account the probability of joint purchase of each set of products. More specifically, products appearing close to each other in Figure 3, such as those from the same business, should be placed at a shorter distance within the store. On the other hand, if the goal is to motivate the shopper to walk along more aisles of the store, categories with a high probability of joint purchase should be placed far away from each other. This may have negative consequences in terms of the shoppers' attitudes toward the retailer (e.g. service quality perceptions and customer satisfaction). As the analyzed store corresponds to a small supermarket, the retailer should aim at producing an efficient experience for the shopper, with categories belonging to the same business located close to each other.

5.4 Performance analysis

Traditional accounting does not consider the effects of promoting one category on the profits of other categories, as mentioned by Chen *et al.* (1999). The graphical representation of product categories based on joint purchase patterns enables managers to identify products that may be affected by a marketing decision concerning one specific category. In particular, categories with a higher probability of being included in the same purchase are obvious candidates for products that might be strongly impacted. Therefore, the evaluation of a marketing action concerning one category should consider not only the effects on the sales of that category but also the cross-effects on interrelated categories. Accordingly, it makes sense to set measures of the global performance for a cluster of product categories instead of metrics for a single product category. For instance, the ratio between gross margin and amount of space allocated to a cluster of product categories may be useful to analyze the efficiency of the allocation of space within the store. The same ratio computed for just one category rather than for a cluster of categories is not a good measure by itself, because it does not take into account the effects on other interrelated categories belonging to the same business. Through this myopic focus, a category not generating a significant margin would not be attractive to the retailer even if it attracts traffic to the store. In our view, the main challenge for the retailer is to evolve from category management, which is essentially a product-oriented approach to a business management approach, which should be driven by a comprehensive understanding of the customer shopping habits. The approach described in this article is, therefore, a useful step to achieve this transition.

6. Conclusions and future research

This paper presents a simple, fast and wide-range approach to study the shopping behavior of customers, detect cross-category interrelations and segment a retailer's business and customers based on information about their shopping baskets. This study allowed us to conceive the retail store under study as a set of four sub-businesses.

These conclusions then reinforced the need of a strategic coordination of marketing activities concerning interrelated categories within each business. Our results suggest that retailers could potentially benefit if they transition from the traditional category management approach, where retailers manage product categories in isolation, into a

customer management approach, where retailers identify, acknowledge and leverage interrelations among product categories.

It is important to note that our approach may be used beyond the particular industry under study in this paper (i.e. the supermarket industry). In this regard, this approach should be valuable for any business or activity where customers buy, hire or consume bundles of products or services from the same provider (e.g. department stores and financial institutions). This is also relevant for internet sites where visitors navigate through different pages of the same portal (e.g. espn.com) or purchase products from different categories (e.g. amazon.com). Moreover, this analysis is even more relevant in the case of dynamic or interactive websites that modify its design, advertising or content according to the information available about their visitors (Hauser *et al.*, 2009). Other areas of application include online music portals (e.g. Pandora and Spotify) that may use this analysis to study the musical preferences of its customers and social networks (e.g. Facebook) interested in studying which brands are more often followed or liked by the same users. In sum, we hope the approach presented in this paper might provide a powerful and practical approach to study customer behavior in multi-product settings.

Finally, in terms of future research, the proposed approach could be enhanced using additional information such as consumers' perceptions, relevance for different consumption or shopping occasions and management differences across product categories (e.g. perishability and inventory management). In addition, one could apply the methodology to the study of shopping behavior across multiple retail stores or formats (Hino, 2014). Another interesting avenue for further investigation is the study of the sequence of purchases within a shopping trip, i.e. the order in which products are added to the shopping cart. Our study does not contain this information, but the use of geolocation information (e.g. obtained through radio frequency identification technology) about customers and their trajectories inside the store could enrich our analysis (Hui *et al.*, 2009).

References

Andersen Consulting and ECR Europe (2000), *The Essential Guide to Day-to-Day Category Management*, ECR Europe, London.

Andrews, R.L. and Currim, I.S. (2002), "Identifying segments with identical choice behaviors across product categories: an intercategory logit mixture model", *International Journal of Research in Marketing*, Vol. 19 No. 1, pp. 65-79, available at: http://dx.doi.org/10.1016/s0167-8116(02)00048-4

Andrews, R.L. and Manrai, A.K. (1999), "MDS maps for product attributes and market response: an application to scanner panel data", *Marketing Science*, Vol. 18 No. 4, pp. 584-604, available at: http://dx.doi.org/10.1287/mksc.18.4.584

Bell, D.R. and Lattin, J.M. (1998), "Shopping behavior and consumer preference for store price format: why 'large basket' shoppers prefer EDLP", *Marketing Science*, Vol. 17 No. 1, pp. 66-88, available at: http://dx.doi.org/10.1287/mksc.17.1.66

Blattberg, R., Fox, E. and Purk, E. (1995), *Category Management-Complete Set*, Food Marketing Institute, Washington, DC.

Borg, I. and Groenen, P. (1997), "Modern multidimensional scaling", *Springer Series in Statistics*, Springer, New York, NY, available at: http://dx.doi.org/10.1007/978-1-4757-2711-1

Chen, Y., *et al.* (1999), "Accounting profits versus marketing profits: a relevant metric for category management", *Marketing Science*, Vol. 18 No. 3, pp. 208-229, available at: http://dx.doi.org/10.1287/mksc.18.3.208

Chib, S., Seetharaman, P.B. and Strijnev, A. (2002), "Analysis of multi-category purchase incidence decisions using IRI market basket data", *Advances in Econometrics*, pp. 57-92, available at: http://dx.doi.org/10.1016/s0731-9053(02)16004-x

Chintagunta, P., Dubé, J.-P. and Singh, V. (2002), "Market structure across stores: an application of a random coefficients logit model with store level data", *Advances in Econometrics*, pp. 191-221, available at: http://dx.doi.org/10.1016/s0731-9053(02)16009-9

De Amorim, R.C. and Hennig, C. (2015), "Recovering the number of clusters in data sets with noise features using feature rescaling factors", *Information Sciences*, Vol. 324, pp. 126-145, available at: http://dx.doi.org/10.1016/j.ins.2015.06.039

Dupre, K. and Gruen, T.W. (2004), "The use of category management practices to obtain a sustainable competitive advantage in the fast-moving-consumer-goods industry", *Journal of Business & Industrial Marketing*, Vol. 19 No. 7, pp. 444-459, available at: https://doi.org/10.1108/08858620410564391

Ehrenberg, A.S.C., Uncles, M.D. and Goodhardt, G.J. (2004), "Understanding brand performance measures: using dirichlet benchmarks", *Journal of Business Research*, Vol. 57 No. 12, pp. 1307-1325, available at: http://dx.doi.org/10.1016/j.jbusres.2002.11.001

Elrod, T. (1988), "Choice map: inferring a product-market map from panel data", *Marketing Science*, Vol. 7 No. 1, pp. 21-40, available at: http://dx.doi.org/10.1287/mksc.7.1.21

Erdem, T. (1998), "An empirical analysis of umbrella branding", *Journal of Marketing Research*, Vol. 35 No. 3, pp. 339-351, available at: http://dx.doi.org/10.2307/3152032.

Fader, P.S. and Lodish, L.M. (1990), "A cross-category analysis of category structure and promotional activity for grocery products", *Journal of Marketing*, Vol. 54 No. 4, p. 52. available at: http://dx.doi.org/10.2307/1251759

Galvagno, M. (2011), "The intellectual structure of the anti-consumption and consumer resistance field: an author co-citation analysis", *European Journal of Marketing*, Vol. 45 Nos 11/12, pp. 1688-1701, available at: http://dx.doi.org/10.1108/03090561111167441

Ghose, S. (1998), "Distance representations of consumer perceptions: evaluating appropriateness by using diagnostics", *Journal of Marketing Research*, Vol. 35 No. 2, p. 137, available at: http://dx.doi.org/10.2307/3151843

González-Benito, Ó., Martínez-Ruiz, M.P. and Mollá-Descals, A. (2009), "Using store level scanner data to improve category management decisions: developing positioning maps", *European Journal of Operational Research*, Vol. 198 No. 2, pp. 666-674, available at: http://dx.doi.org/10.1016/j.ejor.2008.10.015.

Goodhardt, G.J., Ehrenberg, A.S.C. and Chatfield, C. (1984), "The dirichlet: a comprehensive model of buying behaviour", *Journal of the Royal Statistical Society. Series A (General)*, Vol. 147 No. 5, pp. 621 available at: http://dx.doi.org/10.2307/2981696

Gooner, R.A., Morgan, N.A. and Perreault, W.D. Jr, (2011), "Is retail category management worth the effort (and does a category captain help or hinder)?", *Journal of Marketing*, Vol. 75 No. 5, pp. 18-33, available at: https://doi.org/10.1509/jmkg.75.5.18

Hauser, J.R., *et al.* (2009), "Website morphing", *Marketing Science*, Vol. 28 No. 2, pp. 202-223, available at: http://dx.doi.org/10.1287/mksc.1080.0459

Heilman, C.M. and Bowman, D. (2002), "Segmenting consumers using multiple-category purchase data", *International Journal of Research in Marketing*, Vol. 19 No. 3, pp. 225-252, available at: http://dx.doi.org/10.1016/s0167-8116(02)00077-0

Hino, H. (2014), "Shopping at different food retail formats: understanding cross-shopping behavior through retail format selective use patterns", *European Journal of Marketing*, Vol. 48 Nos 3/4, pp. 674-698, available at: http://dx.doi.org/10.1108/EJM-12-2011-0764

Ho, T.-H., Tang, C.S. and Bell, D.R. (1998), "Rational shopping behavior and the option value of variable pricing", *Management Science*, Vol. 44 No. 12- part-2, pp. S145-S160, available at: http://dx.doi.org/10.1287/mnsc.44.12.s145

Hruschka, H., Lukanowicz, M. and Buchta, C. (1999), "Cross-category sales promotion effects", *Journal of Retailing and Consumer Services*, Vol. 6 No. 2, pp. 99-105. available at: http://dx.doi.org/10.1016/s0969-6989(98)00026-5

Hui, S.K., Bradlow, E.T. and Fader, P.S. (2009), "Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior", *Journal of Consumer Research*, Vol. 36 No. 3, pp. 478-493, available at: http://dx.doi.org/10.1086/599046

Khan, U., Dhar, R. and Wertenbroch, K. (2005), "A behavioral decision theory perspective on hedonic and utilitarian choice", *Inside Consumption: Frontiers of Research on Consumer Motives, Goals, and Desires*, 1, Routledge, London, pp. 144-165.

Ma, Y., Seetharaman, P.B. and Narasimhan, C. (2012), "Modeling dependencies in brand choice outcomes across complementary categories", *Journal of Retailing*, Vol. 88 No. 1, pp. 47-62, available at: http://dx.doi.org/10.1016/j.jretai.2011.04.003

Manchanda, P., Ansari, A. and Gupta, S. (1999), "The "shopping basket": a model for multicategory purchase incidence decisions", *Marketing Science*, Vol. 18 No. 2, pp. 95-114, available at: http://dx.doi.org/10.1287/mksc.18.2.95

Mild, A. and Reutterer, T. (2003), "An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data", *Journal of Retailing and Consumer Services*, Vol. 10 No. 3, pp. 123-133, available at: http://dx.doi.org/10.1016/s0969-6989(03)00003-1

Montgomery, A.L. (1997), "Creating micro-marketing pricing strategies using supermarket scanner data", *Marketing Science*, Vol. 16 No. 4, pp. 315-337, available at: https://doi.org/10.1287/mksc.16.4.315

Mouzas, S. (2006), "Efficiency versus effectiveness in business networks", *Journal of Business Research*, Vol. 59 Nos 10/11, pp. 1124-1132, available at: http://dx.doi.org/10.1016/j.jbusres.2006.09.018

Nielsen (1992), *Category Management: Positioning Your Organization to Win*, NTC Business Books, Lincolnwood (Chicago), IL.

Rousseeuw, P.J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53-65, available at: http://dx.doi.org/10.1016/0377-0427(87)90125-7

Russell, G.J. and Kamakura, W.A. (1997), "Modeling multiple category brand preference with household basket data", *Journal of Retailing*, Vol. 73 No. 4, pp. 439-461, available at: http://dx.doi.org/10.1016/s0022-4359(97)90029-4

Russell, G.J. and Petersen, A. (2000), "Analysis of cross category dependence in market basket selection", *Journal of Retailing*, Vol. 76 No. 3, pp. 367-392, available at: http://dx.doi.org/10.1016/s0022-4359(00)00030-0

Seetharaman, P.B., *et al.* (2005), "Models of multi-category choice behavior", *Marketing Letters*, Vol. 16 Nos 3/4, pp. 239-254, available at: http://dx.doi.org/10.1007/s11002-005-5888-y

Seetharaman, P.B., Ainslie, A. and Chintagunta, P.K. (1999), "Investigating household state dependence effects across categories", *Journal of Marketing Research*, Vol. 36 No. 4, p. 488, available at: http://dx.doi.org/10.2307/3152002

Srinivasan, S., Pauwels, K. and Nijs, V. (2008), "Demand-based pricing versus past-price dependence: a cost–benefit analysis", *Journal of Marketing*, Vol. 72 No. 2, pp. 15-27, available at: http://dx.doi.org/10.1509/jmkg.72.2.15

Tan, P.-N., Kumar, V. and Srivastava, J. (2002), "Selecting the right interestingness measure for association patterns", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '02, ACM Press, Edmonton*, available at: http://dx.doi.org/10.1145/775052.775053

Tanusondjaja, A., Nenycz-Thiel, M. and Kennedy, R. (2016), "Understanding shopper transaction data: how to identify cross-category purchasing patterns using the duplication coefficient", *International Journal of Market Research*, Vol. 58 No. 3, p. 401, available at: http://dx.doi.org/10.2501/ijmr-2016-026

Urban, G.L. and Hauser, J.R. (1993), *Design and Marketing of New Products*, Prentice hall, Englewood Cliffs, NJ.

Appendix. Step-by-step guide to implement the methodology in "market basket analysis insights to support category management"

- Arrange the transactional data in a matrix with as many rows as transactions and as many columns as categories. The entries in the table are equal to 1 if a transaction includes the purchase of the corresponding category. For example, consider a data set with four transactions and four categories:(Table AI).

In this example, the first transaction involves the purchases of products from Categories 1 and 3, while the second transaction involves purchases of products from Categories 1 and 2. We will illustrate the methodology by using the Jaccard ratio. Similar steps are needed if the duplication factor is used instead of the Jaccard ratio:

- For each pair of categories, determine the number of transactions where both categories were purchased. In the example above, only one transaction involves the purchase of both Categories 1 and 2.
- For each pair of categories, determine the number of transactions where at least one of these categories was purchased. In the example above, four transactions involve the purchase of Category 1 and/or 2.
- Compute the ratio of the values determined in Steps 2 and 3 for every pair of categories. In the example above, this corresponds to 1/4 = 0.25 for the pair involving Categories 1 and 2. This value represents how often a pair of product categories are purchased together, conditional on at least one of them being purchased.
- Compute 1 minus the ratio from Step 4 = 0.75 for every pair of product categories.
- Use the ratios as distances in an MDS procedure. Start by using two dimensions. The MDS procedure will produce coordinates for each category on each of those two dimensions.
- Use the coordinates from the previous step to produce a two-dimensional map of product categories that represents how often products are jointly purchased, conditional on at least one of them being purchased.
- Improve the previous solution by adding an extra dimension and determine the change in the stress measure. Add more dimensions until the stress improvement becomes sufficiently small (Borg and Groenen, 1997, p. 38).
- As before, the MDS procedure will produce coordinates for each coordinate along the dimensions chosen in the previous step. These coordinates are then used to run a *k*-means cluster analysis to group products into basket types. The number of clusters is chosen by maximizing the silhouette coefficient.
- If additional information is available for the different product categories (e.g. number of tickets, average spending, assortment quality, rate of stockouts, etc.), these variables can be used to complement the conclusions from the MDS and cluster analysis procedures. More specifically, take each descriptor variable (e.g. average spending) and use it as a dependent variable in a linear regression where the independent variables are the

Table AI.
Illustrative data set

Category 1	Category 2	Category 3	Category 4
1	0	1	0
1	1	0	0
0	1	0	1
1	0	1	1

coordinates of each category for the MDS solution based on two dimensions. Estimate a separate regression for each descriptor (e.g. one regression for average spending, another for number of tickets). Use the coefficients to plot each descriptor within the MDS two-dimensional map. For example, suppose that using the average spending as a dependent variable produces coefficients equal to 1 and −0.5 for the x- and y-axes, respectively, and that the R^2 for this regression equals 0.6. Then, this variable (average spending) can be added (i.e. projected) to the MDS map by adding a vector that starts at ($x = 0$, $y = 0$) and ends at ($x = 1$, $y = -0.5$). Intuitively, this would suggest that categories with more positive x-coordinates and more negative y-coordinates display on average greater spending and that the x-axis is more strongly associated with greater spending than the y-axis.

- Note that the user may choose to adjust the length of the vector (without affecting its angle) to reflect the fit of the linear regression. In the previous example, the length may be chosen to be proportional to the 0.6 R^2.

Corresponding author
Andres Musalem can be contacted at: amusalem7@gmail.com