Mark Lett (2012) 23:505–516
DOI 10.1007/s11002-012-9181-6

Sequential market basket analysis

Wagner A. Kamakura

Published online: 22 May 2012

Abstract Market basket analysis (MBA) is a powerful and common practice in modern retailing that has some limitations stemming from the fact that it infers purchase sequence from joint-purchasing data. However, internet retailers automatically collect purchase-sequence data from their shoppers, and new technology is available for traditional (bricks and mortar) retailers to do the same, making it possible to analyze purchase sequences, rather than inferring them from joint purchases. This study first compares and contrasts traditional market basket analysis with a sequential extension, and then proposes a framework for purchase-sequence analysis, which is illustrated utilizing shopping trip data from one grocery store.

Keywords Market basket analysis · Shopping behavior

1 Introduction

Market basket analysis (MBA), also known as product affinity analysis, is already widely known and utilized by traditional (bricks and mortar) and Internet retailers. A Google search on the keyword set "market basket analysis" reveals over 40,000 hits, providing evidence of its prevalence in modern retailing.

The basic idea behind Market basket analysis is to find pairs or sets of products that are jointly observed in large samples of baskets, based on the assumption that purchase of one or more of the products within a set would lead to purchase of the remaining ones, thereby providing leads for cross-selling, bundling, product positioning, etc. The underlying assumption in market basket analysis is that joint occurrence of two or more products in most baskets imply that these products are complements in purchase (if not in consumption), and therefore, purchase of one will lead to purchase of others. As we will discuss later, market basket analysis makes

W. A. Kamakura (✉)
Fuqua School of Business, Duke University, One Towerview Road, Durham, NC 27708-0120, USA
e-mail: Kamakura@duke.edu

inferences about purchase sequence from data on joint purchases, which are potentially misleading.

In this study, we argue that with the advent of new technology, it is not necessary to make potentially misleading inferences about purchase sequences from joint purchases. Most internet retailers already collect data on the sequence items are added to shopping basket. Traditional retailers can also obtain the same sequence data using RFID technology (Sorensen 2003). Similarly, most services organizations know exactly the order by which their customers adopt new services. Given that data on purchase sequences are becoming more widely available, we propose an extension of market basket analysis into sequential market basket analysis, which takes advantage of these data, and develop a methodology for analyzing purchase sequences. The rest of this study proceeds as follows: first, we describe traditional market basket analysis based on joint purchases. We then present our proposed sequential market basket analysis, followed by en empirical comparison of the two approaches on the same purchase data, followed with a general discussion and directions for future research.

2 Traditional and sequential market basket analysis

The practice of market basket analysis has its origin in the data-mining literature, with the introduction of association-rule discovery (Anand et al. 1998). Despite the importance of this topic to retailing and e-commerce, there are surprisingly few published articles on market basket analysis in the marketing literature (Russell and Petersen 2000). One can find numerous articles on cross-category choice modeling (readers are referred to the excellent literature reviews by Russell et al. 1999 and Seetharaman et al. 2005 for details), but these models are rarely scalable beyond a few product categories and are therefore impractical for market basket analysis.

By far, the most common practice in market basket analysis is the identification of association-rules (Brijs et al. 2004). Each pair of products A and B is evaluated on three measures:

- *Support*—the joint probability of finding the pair AB across all baskets. A low support means that the pair is not relevant because it is not purchased frequently enough.
- *Confidence*—the conditional probability $p(B|A) = p(A \cap B)/p(A)$, which is often interpreted as the probability that purchase of A will lead to purchase of B.
- *Interest*—the ratio between the joint probability and the probability of joint occurrence under independence $\frac{p(A \cap B)}{p(A) \times p(B)}$. This measure discounts the joint probability by the "popularity" of the individual items in all baskets.

Product B is a good recommendation for shoppers who just added A to their basket if *interest* and *support* is high (i.e., they tend to occur jointly in most baskets), and *confidence*, the conditional probability of purchasing B given A, is also high. The three measures above are also combined to classify baskets based on their contents, leading to *trip classification*. Understanding the types of shopping trips a customer takes to a particular store at a particular time is critical for labor scheduling, product readiness and even changes in store layout.

Despite its popularity, market basket analysis has been criticized for the underlying assumption that joint occurrence (measured by support and interest) implies complementarity. A recent study by Vindevogel et al. (2005) looked at the price cross-elasticities for 2,700 pairs of products selected through association-rules, and demonstrated that more than half of them were in fact substitutes in terms of price response (positive cross-elasticities), while others were independent, so that only about 40 % of the selected pairs were true complements in price (negative cross-elasticities). We further argue that even when products are true complements, finding a strong conditional probability $p(B|A)$ based on joint occurrences does not necessary imply that purchase of A is followed by purchase of B. We later argue, using actual data from one retailer, that conditional probabilities based on joint occurrences are not necessarily consistent with the observed purchase sequences.

2.1 Data description

To illustrate the distinctive features of traditional market basket analysis and its sequential extension, we utilize shopping trip data provided by TNS-Sorensen for one grocery store, starting with measures from traditional market basket analysis and showing how they lead to different conclusions from sequential market basket analysis, based on actual sequences. The shopping trip data are collected by a RFID "tag" mounted under the shopping cart, which emits a uniquely coded signal every 5 s, collected by an array of antennae around the perimeter of the store. The signals collected by the antenna array are triangulated to produce the location of the shopper every five seconds, which is in turn integrated with purchase transaction (from the checkout), and store planogram data to produce a rich description of each shopping trip (Hui et al. 2009).

The data we consider includes only trips where more than five items were purchased. Fill-in trips with five items or less represent a substantial portion of all trips, but provide limited information about purchase sequences because there are few product categories being purchased and shoppers in a fill-in trip with few purchased items have a shorter travel path inside the store, which is determined by the few items they already planned to purchase prior to entering the store. This sampling scheme, following common practice in this literature, resulted into a final sample of 2,290 shopping trips including 15,206 observed purchases. Because actual purchases are only observed at the checkout counter, purchase sequences are inferred from the travel path, using the last observed path, in case of multiple paths between two purchased products.

2.2 Sequential market basket analysis

Traditional market basket analysis has been widely applied and quite useful, given the type of static basket data available in the past. However, as argued earlier, many firms have sequential data on how each basket was formed, so that purchase sequence does not have to be inferred from joint occurrences.

Table 1 lists 48 pairs of products with the strongest affinity, out of 56,508 observed pairs of purchased SKU's, based on the commonly-used measures of *Support*, *Relevance*, and *Confidence* that ignore the actual purchase sequence. We present in

Table 1 only the pairs with *Support* greater than 1 % for which we observed at least 1,000 joint incidences, sorted in decreasing order of *Interest*. The identities of the individual stock keeping units (SKU) are disguised behind 1500 product categories. *Support* and *Interest* were defined earlier, and *Confid* $(B|A)$ shows the estimated conditional probability of buying product B, given that product A was purchased, based on the premise that pairs with high *Interest* are complementary. We will next challenge this assumption by considering the sequence items were added to the basket, which will provide a better assessment of the affinities among the purchased items.

For sequential market basket analysis, given that we have the actual purchase sequences, we first compute $P(A \rightarrow B)$, the conditional probability of buying B after having purchased A, which we obtain dividing the number of purchases of the $A \rightarrow B$ sequence by the number of purchases of A alone plus the number of $A \rightarrow B$ purchases. In other words, $P(A \rightarrow B)$ is the probability of buying B *after* having bought A, which excludes all trips where B is purchased before A. Using this definition of sequential purchase probability, we then identify the following types of pairs:

- *Weak a→b sequence*—if $p(A \rightarrow B)p(B)$ and $p(B \rightarrow A) \sim p(A)$ [1]
  - This sequence is defined as "weak" because the first clause $\{p(A \rightarrow B)p(B)\}$ implies that prior purchase of A increases the likelihood of buying B, but the second clause $\{p(B - A) \sim p(A)\}$ implies that prior purchase of B does not substantially affect the likelihood of buying A. Therefore the evidence for an $A \rightarrow B$ sequence is only partial.
- *Strong A→B sequence*—if $p(A \rightarrow B)p(B)$ and $p(B \rightarrow A)p(A)$
  - This sequence is deemed "strong" because prior purchase of A increases the likelihood of buying B, while prior purchase of B reduces the likelihood of buying A. In this sequence, A is a clear antecedent of B.
- *Complements*— if $p(A \rightarrow B)p(B)$ and $p(B \rightarrow A)p(A)$
  - In these sequences, A and B are clear complements, because prior purchase of one increases the likelihood that the other is later bought.
- *Substitutes*—if $p(A \rightarrow B)p(B)$ and $p(B \rightarrow A)p(A)$
  - In these sequences A and B are clear substitutes because prior purchase of one reduces the likelihood that the other is later bought.
- *Independents*—otherwise
  - When the prior purchase of one product does not affect the probability of buying the other, and vice versa, there is no association between the two shopping decisions, and therefore we deem the pair independent.

Table 2 shows the sequential measures of affinity for the same pairs listed in Table 1. For easier interpretation, we report another statistic, $\mathrm{GAIN}[A \rightarrow B] = \frac{P[A-B]}{P[B]} - 1$, which indicates the percentage gain (or loss) in purchase probability for product B due to a previous purchase of product A, relative to the unconditional

[1] We use the mathematical symbol "~" to denote "approximately".

Table 1 Traditional market basket analysis statistics for selected pairs

PAIR	Product *A*	Product *B*	Traditional MBA			
			Support (%)	CONF (*B*|A) (%)	CONF (A|*B*) (%)	Interest
1	Entrees & Meal Components	Dinners & Skillet Meals	1.7	57.8	22.1	7.3
2	Pizza	Dinners & Skillet Meals	1.4	46.0	17.4	5.9
3	Grapes	Berries	1.5	22.6	24.0	3.7
4	Refrigerated	Potatoes	1.2	20.3	20.0	3.3
5	Refrigerated	Refrigerated	1.2	20.2	18.9	3.1
6	Grapes	Apples	1.1	17.3	16.1	2.5
7	Peaches/Nectarines	Bananas	1.7	41.5	9.9	2.5
8	Bananas	Apples	2.8	16.7	40.1	2.4
9	Refrigerated	Loaf Breads	1.7	25.2	14.9	2.3
10	Eggs	Loaf Breads	1.3	24.2	11.7	2.2
11	Grapes	Bananas	2.3	35.0	13.6	2.1
12	Berries	Bananas	2.1	34.1	12.5	2.0
13	Melons	Bananas	1.3	33.3	7.6	2.0
14	Citrus Fruits	Bananas	1.8	32.0	10.9	1.9
15	Bananas	Refr. Juices	1.7	9.9	31.6	1.9
16	Loaf Breads	Refr. Juices	1.1	9.5	20.3	1.8
17	Refrigerated	Bananas	1.8	29.3	10.7	1.7
18	Potatoes	Bananas	1.8	28.7	10.6	1.7
19	Bananas	Eggs	1.5	9.0	28.0	1.7
20	Regular Milk	Eggs	1.7	9.0	30.6	1.7
21	Regular Milk	Kids, Sweetened Cereals	1.3	6.9	30.5	1.7
22	Refrigerated	Bananas	2.3	27.7	13.4	1.7
23	Regular Milk	Refr. Juices	1.5	8.2	28.8	1.6
24	Refrigerated	Bananas	1.7	25.9	10.1	1.5
25	Regular Milk	Loaf Breads	3.0	16.3	26.8	1.5
26	Ice Cream & Premium	Regular Milk	1.5	25.8	8.1	1.4
27	Bananas	Regular Milk	4.3	25.6	23.3	1.4
28	Bananas	Loaf Breads	2.6	15.5	23.3	1.4
29	Refrigerated	Loaf Breads	1.2	15.3	11.1	1.4
30	Bananas	Yogurt	2.1	12.5	22.7	1.3
31	Yogurt	Regular Milk	2.3	24.5	12.4	1.3
32	Bananas	Refrigerated	1.6	9.5	22.2	1.3
33	Apples	Regular Milk	1.6	22.9	8.7	1.2
34	Grapes	Regular Milk	1.5	22.7	8.1	1.2
35	Refrigerated	Regular Milk	1.6	22.1	8.7	1.2
36	Refrigerated	Regular Milk	1.4	21.6	7.7	1.2
37	Potatoes	Regular Milk	1.3	21.1	7.1	1.1
38	Yogurt	Loaf Breads	1.2	12.8	10.6	1.1

Table 1 (continued)

PAIR	Product *A*	Product *B*	Traditional MBA			
			Support (%)	CONF (*B*|A) (%)	CONF (A|*B*) (%)	Interest
39	Regular Milk	Shelf Stable	1.2	6.3	20.7	1.1
40	Bananas	Refrigerated	1.2	7.4	18.9	1.1
41	Regular Milk	Drinking Water	1.1	5.8	20.2	1.1
42	Refrigerated	Regular Milk	1.3	19.5	6.9	1.1
43	Dinners & Skillet Meals	Regular Milk	1.5	19.2	7.9	1.0
44	Citrus Fruits	Regular Milk	1.1	19.1	6.0	1.0
45	Refrigerated	Regular Milk	1.1	17.4	5.8	0.9
46	Bananas	Wine	1.2	7.1	15.7	0.9
47	Regular Milk	Wine	1.3	6.9	16.8	0.9
48	Refrigerated	Regular Milk	1.4	16.7	7.4	0.9

probability ($P(B)$). This gain measure shows the increase in the odds of purchasing product B, among shoppers who *previously purchased* product A. Note that this measure of gain is distinct from the traditional measure of "lift" (or *Confidence*, as defined earlier), both conceptually and empirically. Our measure of gain is based on purchase sequence, which directly measures how the prior purchase of one product affects the likelihood that the other is subsequently purchased. The traditional measure only implies joint occurrence of the events.

At this point one could think about the asymptotic properties for these measures, to be used in statistical-significance tests. However, the sample sizes involved in market basket analysis are typically very large, and one is better off focusing on the product pairs with the largest joint incidences and concentrate on those pairs where the relevant gains are the largest.

The statistics listed in Table 2 show that the assumption of complementarity for the pairs with high support and interest is not necessarily confirmed by the purchase sequences. Complementarity implies that the probability of buying B after A and of buying A after B are higher than the respective unconditional probabilities (for B and A). This is only true for 12 of the 48 selected SKU pairs, out of all 56,508 observed pairs. Complementarity is confirmed only for the pairs with the highest *Interest*; this is what one would expect because *Interest* is the ratio between the joint probability and the probability expected under independence. For example, Table 2 shows that purchase of *Entrees & Meal Components* increases the probability of a subsequent purchase of *Dinners & Skillet Meals*, while purchase of *Dinners & Skillet Meals* increases the likelihood that *Entrees & Meal Components* are purchased later in the trip, thereby confirming these pairs as complements in terms of purchase sequence (purchase of one, increases the likelihood that the other is also purchased).

However complementarity is not confirmed for most of the other pairs where *Interest* is greater than 1 (thereby indicating complementarity). Take the sixth pair in Table 2 as an example; while *Interest* is still high (2.5) for this pair, our sequential analysis suggests that these two products are not complements; instead of

Table 2 Sequential Market basket analysis statistics for selected pairs

Pair	Product A	Product *B*	*P*[A→*B*] (%)	*P*[*B*] (%)	*P*[*B*→A] (%)	P[A] (%)	GAIN[A→*B*] (%)	GAIN[*B*→A] (%)	Relation
1	Entrees & Meal Components	Dinners & Skillet Meals	37.7	7.9	13.7	3.0	376	353	COMP
2	Pizza	Dinners & Skillet Meals	33.3	7.8	8.0	3.0	326	170	COMP
3	Grapes	Berries	14.9	6.2	11.2	6.6	142	72	COMP
4	Refrigerated	Potatoes	9.8	6.2	12.5	6.1	58	104	COMP
5	Refrigerated	Refrigerated	10.8	6.6	10.9	6.2	63	77	COMP
6	Grapes	Apples	12.1	7.0	6.2	6.5	73	−6	*A*→*B*
7	Peaches/Nectarines	Bananas	28.2	16.8	4.7	4.0	68	17	COMP
8	Bananas	Apples	7.1	7.0	29.4	16.8	1	74	COMP
9	Refrigerated	Loaf Breads	14.0	11.2	8.3	6.6	25	26	COMP
10	Eggs	Loaf Breads	12.7	11.2	6.7	5.4	13	25	COMP
11	Grapes	Bananas	24.7	16.8	5.8	6.5	47	−12	*A*→*B*
12	Berries	Bananas	19.9	16.8	6.9	6.2	18	12	COMP
13	Melons	Bananas	20.4	16.8	3.9	3.8	21	1	COMP
14	Citrus Fruits	Bananas	19.5	16.8	5.6	5.8	16	−2	*a*→*b*
15	Bananas	Refr. Juices	5.5	5.2	17.7	16.8	6	5	COMP
16	Loaf Breads	Refr. Juices	6.3	5.2	8.4	11.2	20	−24	*A*→*B*
17	Refrigerated	Bananas	17.4	16.9	5.6	6.2	3	−10	*a*→*b*
18	Potatoes	Bananas	18.6	16.8	4.9	6.2	11	−21	*A*→*B*
19	Bananas	Eggs	4.3	5.4	17.6	16.8	−21	5	*B*→*A*
20	Regular Milk	Eggs	4.2	5.4	19.6	18.4	−22	6	*B*→*A*
21	Regular Milk	Kids, Sweetened Cereals	3.7	4.2	17.6	18.4	−12	−4	SUBS
22	Refrigerated	Bananas	16.1	16.8	7.2	8.1	−4	−12	SUBS
23	Regular Milk	Refr. Juices	4.4	5.2	16.6	18.4	−17	−10	SUBS
24	Refrigerated	Bananas	16.8	16.8	4.5	6.5	0	−31	*a*→*b*

Table 2 (continued)

Pair	Product A	Product *B*	*P*[A→*B*] (%)	*P*[*B*] (%)	*P*[*B*→A] (%)	P[A] (%)	GAIN[A→*B*] (%)	GAIN[*B*→A] (%)	Relation
25	Regular Milk	Loaf Breads	8.1	11.2	16.6	18.4	−27	−10	SUBS
26	Ice Cream	Regular Milk	10.6	18.4	5.5	5.8	−43	−5	SUBS
27	Bananas	Regular Milk	14.1	18.4	13.8	16.8	−24	−18	SUBS
28	Bananas	Loaf Breads	5.9	11.2	16.7	16.8	−47	−1	SUBS
29	Refrigerated	Loaf Breads	7.9	11.2	6.2	8.1	−29	−24	SUBS
30	Bananas	Yogurt	6.6	9.3	13.0	16.8	−29	−23	SUBS
31	Yogurt	Regular Milk	12.5	18.4	7.3	9.3	−32	−21	SUBS
32	Bananas	Refrigerated	4.5	7.2	13.7	16.8	−38	−18	SUBS
33	Apples	Regular Milk	11.2	18.4	5.2	7.0	−39	−26	SUBS
34	Grapes	Regular Milk	15.0	18.4	3.4	6.5	−19	−48	SUBS
35	Refrigerated	Regular Milk	13.6	18.4	4.1	7.2	−26	−44	SUBS
36	Refrigerated	Regular Milk	9.8	18.4	4.8	6.6	−47	−27	SUBS
37	Potatoes	Regular Milk	11.8	18.4	3.7	6.2	−36	−41	SUBS
38	Yogurt	Loaf Breads	4.5	11.2	7.4	9.3	−59	−20	SUBS
39	Regular Milk	Shelf Stable	4.0	5.6	9.1	18.4	−29	−50	SUBS
40	Bananas	Refrigerated	4.2	6.6	9.5	16.8	−36	−43	SUBS
41	Regular Milk	Drinking Water	3.3	5.3	10.1	18.4	−37	−45	SUBS
42	Refrigerated	Regular Milk	11.9	18.4	3.2	6.5	−35	−51	SUBS
43	Dinners & Skillet Meals	Regular Milk	8.0	18.5	5.1	7.6	−56	−32	SUBS
44	Citrus Fruits	Regular Milk	9.8	18.4	3.3	5.8	−47	−42	SUBS
45	Refrigerated	Regular Milk	9.0	18.4	3.1	6.1	−51	−49	SUBS
46	Bananas	Wine	2.4	7.5	11.2	16.8	−68	−33	SUBS
47	Regular Milk	Wine	2.8	7.6	10.9	18.5	−62	−41	SUBS
48	Refrigerated	Regular Milk	10.5	18.4	3.2	8.2	−43	−61	SUBS

complementarity, Table 2 suggests a clear purchase sequence for these two items. For this pair, *Gain* ($A \rightarrow B$) is 73 % while *Gain*($B \rightarrow A$) is −6 %, which is strong indication that purchase of the SKU in *Grapes* leads to subsequent purchase of the SKU in *Apples*. Similar conclusions can be drawn for pairs #11, #16 and #18, while the reverse (purchase of B leads to subsequent purchase of A) for pairs #19 and #20, despite the fact that the traditional *Interest* measure is still greater than 1.5, indicating complementarity.

The most conflicting evidence between the conclusions drawn from traditional market basket analysis and the empirical evidence provided by actual purchase sequences is found in pairs such as #28 (*Bananas* and *Loaf Breads*) and # 29 (*Refrigerated* and *Loaf Breads*), where the previous purchase of one component of the pair decreases the likelihood that the other component will be purchased later in the same trip, suggesting that these pairs contain substitutes, rather than complements, despite the fact that *Interest* is relatively high (1.4) among pairs of high *Support* (2.6 % and 1.2 %, respectively).

The empirical comparisons between predictions drawn from traditional market basket analysis and actual purchase sequences discussed above suggest that there is valuable information in the actual sequence by which shoppers add items to their basket, and therefore models for basket analysis should take purchase sequence into account. We also compared the classifications based on the traditional and our sequential approach across a larger sample of 249 product pairs with *Support* greater than 0.5 %, defining pairs as complements, if the traditional approach showed *Interest* greater than 1.0 (i.e., joint probability greater than the product of the marginal probabilities), and as substitutes if *Interest* was smaller than 1.0. The results from this comparison are shown in Table 3, where one can see that all pairs deemed as substitutes by traditional Market basket analysis, are also classified as substitutes with our sequential approach. However, of the 218 pairs classified as complements by the traditional approach, only 27 % were classified as such by our sequential extension. In fact, almost half (107) of these "complement" pairs, according to the traditional model, were identified as strong substitutes by our sequential approach, even though the average *Interest* for these 107 pairs was 1.3, indicating complementarity based on the traditional method. Most importantly, because our proposed approach uses information about purchase sequence, it was also able to identify pairs with a clear sequence, where purchase one products increases the likelihood of buying the other ($A \rightarrow B$ or $B \rightarrow A$).

3 Discussion and conclusions

The main purpose of our study was to incorporate a longitudinal component into market basket analysis, looking at the sequential formation of the basket, rather than its final composition only. We demonstrated empirically that sequential market basket analysis provides valuable insights into how the purchase of one product leads to the purchase of another, which cannot always be properly inferred from the final basket compositions. Given that sequential data on market basket formation is now available with new data-gathering technology, both for traditional and internet retailers, there is no good reason for overlooking the additional insights embedded in purchase sequences.

Table 3 Classification comparisons of product pairs based on Market basket analysis and SMBA

			Classification based on MBA		
			Substitutes (INT<1)	Complements (INT>1)	
Classification based on SMBA	$a \rightarrow b$	Count	0	3	3
		% within MBA	0.0 %	1.4 %	1.2 %
	$A \rightarrow B$	Count	0	22	22
		% within MBA	0.0 %	10.1 %	8.8 %
	$b \rightarrow a$	Count	0	1	1
		% within MBA	0.0 %	0.5 %	0.4 %
	$B \rightarrow A$	Count	0	22	22
		% within MBA	0.0 %	10.1 %	8.8 %
	COMP	Count	0	59	59
		% within MBA	0.0 %	27.1 %	23.7 %
	subs	Count	0	4	4
		% within MBA	0.0 %	1.8 %	1.6 %
	SUBS	Count	31	107	138
		% within MBA	100.0 %	49.1 %	55.4 %
Total		Count	31	218	249
		% within MBA	100.0 %	100.0 %	100.0 %

Even though the empirical illustration presented here used baskets in a traditional grocery retailer, the concepts and methodology are applicable (perhaps even more so) to internet retailers, because the sequence of basket formation is automatically recorded, and these retailers can more easily re-configure their (website) layouts according to the trip types than traditional retailers.

One must be aware, however, that observed purchase sequences (as other forms of shopping behavior) are induced by store (or website) layout. However, our proposed measures are normalized by the number of times a product is purchased first, which corrects for the sample-size differences in first-purchases across products. Nevertheless, store layout will affect the sample sizes for the observed pair-wise sequences, potentially leading to small samples for some product pairs. Fortunately, market basket analysis typically involves very large samples of baskets, rendering concerns regarding statistical power less critical.

Most of the recent research on shopping trips focus on travel paths within the store and its impact on purchasing (Otnes and McGrath 2001; Larson et al. 2005; Hui et al. 2009). In contrast to this literature, which emphasize *travel* (browsing) paths and, in some cases (such as Hui et al. 2009), their impact on buying behavior, our main focus is on basket formation, or the sequence that items are added to the basket. In other words, we consider the path of basket formation, focusing only on the purchasing events, abstracting from the browsing path (i.e., non-purchasing events). We do this because our main interest is in sequential Market basket analysis and for two other more pragmatic reasons. First, by focusing only on the sequence of purchase events,

we can ignore time, thereby avoiding methodological issues related to differences in travel speed across shoppers. Second, by ignoring the non-purchase events, we avoid the causality between travel path and purchase events, which depends on whether the shopper has specific goals (such as a shopping list) in mind before starting the trip (Hui et al. 2009).

Because our main interest is in the *purchase* sequence within a shopping trip, our research might seem related to the literature on *activity pattern analysis* (Paas and Molenaar 2005), which infers the sequence by which consumers acquire products from the portfolio already owned by a sample of consumers. However, this literature is mainly focused on *inferring* the acquisition sequence from observed portfolios accumulated over several years and is therefore akin to traditional market basket analysis, because they are based solely on co-occurrence. In contrast, we focus on the *observed* acquisition sequences within a single shopping trip and the actual (rather than inferred from co-occurrence) sequence of these purchases.

Our study focused only on pair-wise purchase sequences, while the database contains information about the entire purchase sequence, along with the path followed by the shopper during each shopping trip. This type of data calls for an integrated model combining browsing with purchase data along each shopping trip, something that can be handled by a properly formulated hidden Markov model. A good example of hidden Markov model in Marketing, which almost serves as a template for the browsing/shopping data used in our study, is the household life-stages and lifecycle model proposed by Du and Kamakura (2006). In their study, Du and Kamakura (2006) track a sample of households over more than three decades, capturing their composition over time. Applying a hidden Markov model to these data, they identify a typology of American households across the three decades, and measure how the sampled households evolved through these different types over time. To see the parallel with our current problem, consider each shopping trip as the equivalent of a household in the Du and Kamakura study, tracked in equal intervals (e.g. 60 s), gathering data on the location coordinates of the shopping cart, along with the product categories being browsed and purchased. The same type of modeling approach would identify the most common "stations" in a store across shopping trips, both in terms of browsing and actual purchases, and measure how shoppers move among these "stations," as well as understanding how purchases are affected by these movement. We leave this extension for future research on the relationship between browsing and actual shopping.

References

Anand, S. S., Patrick, A. R., Hughes, J. G., & Bell, D. A. (1998). A data mining methodology for cross-sales. *Knowledge-Based Systems, 10*(7), 449–461.

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (2004). Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery, 8*(1), 7–23.

Du, R., & Kamakura, W. (2006). Household life cycles and lifestyles in the United States. *Journal of Marketing Research, 43*(1), 121–132.

Hui, S. K., Bradlow, E., & Fader, P. S. (2009). Traveling salesman goes shopping: the systematic deviations of grocery paths. *Marketing Science, 28*(3), 566–572.

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing, 22*(4), 395–414.

Otnes, C., & McGrath, M. A. (2001). Perceptions and realities of male shopping behavior. *Journal of Retailing, 77*, 111–137.

Paas, L. J., & Molenaar, I. W. (2005). Analysis of acquisition patterns: a theoretical and empirical evaluation of alternative methods. *International Journal of Research in Marketing, 22*, 87–100.

Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing, 76*(3), 367–392.

Russell, G. J., Ratneshwar, S., Shocker, A. D., Bell, D. R., Bodapati, A., Degeratu, A., Hildebrandt, L., Kim, N., Ramaswami, S., & Shankar, V. (1999). Multiple category decision making: review and synthesis. *Marketing Letters, 10*(3), 319–332.

Seetharaman, P. B., Chib, S., & Ainslie, A. (2005). Models of multi-category choice behavior. *Marketing Letters, 16*, 239–254.

Sorensen, H. (2003). The science of shopping. *Marketing Research, 15*(3), 30–35.

Vindevogel, B., Van den Poel, D., & Wets, G. (2005). Why promotion strategies on market basket analysis do not work. *Expert Systems with Applications, 28*(3), 583–590.