ELSEVIER

Intern. J. of Research in Marketing 22 (2005) 395–414

International Journal of Research in Marketing

www.elsevier.com/locate/ijresmar

An exploratory look at supermarket shopping paths

Jeffrey S. Larson *, Eric T. Bradlow, Peter S. Fader

The Wharton School, The University of Pennsylvania, Suite 700 JMHH, 3730 Walnut Street, Philadelphia, PA 19104, USA

Abstract

We present analyses of an extraordinary new dataset that reveals the path taken by individual shoppers in an actual grocery store, as provided by RFID (radio frequency identification) tags located on their shopping carts. The analysis is performed using a multivariate clustering algorithm not yet seen in the marketing literature that is able to handle data sets with unique (and numerous) spatial constraints. This allows us to take into account physical impediments (such as the location of aisles and other inaccessible areas of the store) to ensure that we only report feasible centroid paths. We also recognize that time spent in the store plays an important role, leading to different cluster configurations for short, medium, and long trips. The resulting three sets of clusters identify a total of 14 "canonical path types" that are typical of grocery store travel, and we carefully describe (and cross-validate) each set of clusters. These results dispel certain myths about shopper travel behavior that common intuition perpetuates, including behavior related to aisles, end-cap displays, and the "racetrack." We briefly relate these results to previous research (using much more limited datasets) covering travel behavior in retail stores and other related settings.

Keywords: RFID (radio frequency identification); Clustering techniques; Exploratory analysis; Retailing; Shopping behavior; *k*-medoids clustering

1. Introduction

Most marketers have a well-established schema for shopper travel behavior within a supermarket—the typical customer is assumed to travel up and down the aisles of the store, stopping at various category locations, deliberating about her consideration set, choosing the best (utility maximizing) option, and then continuing in a similar manner until the path is complete. Despite the common presumption of this scenario, little research has been undertaken to understand *actual* travel patterns within a supermarket. How do shoppers really travel through the store? Do they go through every aisle, or do they skip from one area to another in a more direct manner? Do they spend much of their time moving around the outer ring of the store (a.k.a. the "racetrack"), or do they spend most of their time in certain store sections? Do most shoppers follow a single, dominant pattern, or are they rather heterogeneous? A rich new data source, as illustrated in Fig. 1, now allows us to examine these and other important behavioral questions.

No, Fig. 1 does not represent the random scribblings of a kindergartener. It is a subset of the PathTracker® data collected by Sorensen Associates, an in-store research firm, for the purpose of understanding shopper behavior in the supermarket. Specifically, Sorensen Associates affixed RFID (radio frequency identification) tags to the bottom of every grocery cart in an actual supermarket in the western U.S. These tags emit a signal every 5 seconds that is received by receptors installed at various locations throughout the store. The arrival laten-

* Corresponding author. Tel.: +1 215 898 2268.
E-mail address: larsonj@wharton.upenn.edu (J.S. Larson).

0167-8116/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijresmar.2005.09.005

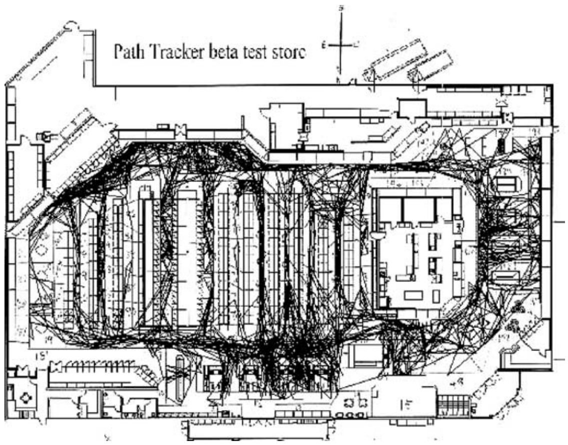

Fig. 1. PathTracker® data from 20 random customers.

cies of the signals at the receptor locations are used to triangulate the position of the grocery cart. Thus, for every shopping path, data are recorded regarding the cart's two-dimensional location coordinates, (x_{it}, y_{it}) for shopper i at his tth observation (hereafter referred to as "blinks"), at 5-second intervals, which can be used to determine each cart's route through the entire store[1]. While ideally, one might hope to obtain positioning data directly from the shoppers themselves, this is not currently available in an actual commercial setting. Therefore, we use customers' grocery carts as a proxy for their shopping path, since we know the exact shopper location when the grocery cart is moving and a good guess of the general vicinity of the shopper when the grocery cart is stationary. Regardless, the methodology developed in this paper will continue to be applicable as newer and better datasets become available. Finally, the time and location of the cart at the end of each path offers information about the checkout process; point-of-sale data can then be matched with the cart movement records to provide a complete picture of each shopping path. See Sorensen (2003) for more details about the PathTracker® system.

The goal of this research is to undertake exploratory analyses, useful for data summarization, inference, and intuition about shopper travel path data. Specifically, we want to identify typical in-store supermarket travel behaviors that will help us understand how shoppers move through a supermarket. Similar research ideas, summarizing large sets of "behavioral" curves as in Fig. 1 have been explored using principal components analysis methods (Bradlow, 2002; Jones & Rice, 1992); however, our goal here is not to explain the maximal variation across customers with principal curves, but instead to cluster respondents into "types" of shoppers and describe the prototypical path of a general cluster. Unfortunately, there are numerous challenges we face, since the application of standard clustering routines is not feasible due to the extremely large number of spatial constraints imposed by the physical supermarket layout (e.g. people can't walk through store shelves). For this reason, the contribution of this research is not limited to the empirical findings of the in-store path data, but also introduces to the marketing literature a multivariate clustering algorithm that can be applied to other settings with a large number of spatial constraints.

Although this new method represents a useful step forward in our ability to analyze multivariate data, we wish to emphasize our exploratory objectives: we want to use this procedure to help us identify predominant patterns that will catalyze future research. Given the newness of this area, we are not yet at the stage of being able to create (or test) formal theories of shopping behavior. In other words, in this paper we will raise more questions than provide answers, and we hope to motivate readers to pursue these research issues with complementary (and more conclusive) research methods.

The remainder of the paper is laid out as follows. First, we describe the data in more detail and explain various obstacles in undertaking exploratory analyses on this data (such as the numerous spatial constraints). Next, we detail the new-to-marketing clustering algorithm used to overcome these obstacles. We then present the results of the algorithm and the canonical shopper path profiles that emerge. The results are then displayed in relation to a set of variables that describe the travel areas of each path. We demonstrate that our methods enable us to cluster shopper paths along important dimensions that would be missed using simpler methods, lending support to the value of our techniques. We next perform a cross-validation of our results to assess the reliability of the findings. Finally, we conclude with a discussion section which summarizes the potential impact of the current findings and relates the current work to past and future research.

2. Overcoming data obstacles

The travel portion of our data consists of a "right-ragged" array of location coordinates, where every row

[1] The dataset originally came with some biases in the calculated locations due to electromagnetic variation in several areas of the store. For example, metal cans cause the signal from the tag to travel faster than in aisles with cardboard packages, so the location coordinates were biased. After extensive testing and calibration by Sorensen Associates, these biases have been corrected in the current dataset.

is a shopping path, and every pair of columns is what we term a "blink", or a coordinate point (x, y) in the store. In total, we have 27,000 shoppers' paths ranging in length from 25 blinks for a 2-min path, to 1500 blinks for a 2-h path. The mean path consists of 205 blinks (just over 16 min), and the median has 166 blinks (just over 13 min)[2]. The path is considered complete (and hence stops being tracked for our purposes) when the cart gets pushed through the checkout line and onto the other side of the checkout counter.

While for some datasets, performing exploratory data analyses may be straightforward, there are a number of significant challenges presented by this type of shopping path data. A proper analysis of such data must overcome the following obstacles:

(1) Memory limitations (size of data)
(2) Ragged array path comparisons (differing lengths of paths)
(3) Spatial constraints (aisle layout and other physical obstructions in the store).

We next describe our solutions to these issues.

2.1. Size of the dataset

With over 27,000 paths, and as many as 1500 pairs of coordinates (blinks) per path, memory limitations for implementing a clustering algorithm posed significant problems. To make the analysis feasible, we drew a systematic sample of 9000 paths, drawing every 3rd path from a random starting point. Some of these paths were deleted due to data problems (the transponder stopped working for longer than a minute), leaving us with 8751 paths. With an additional three-way split of the 8751 paths by total time in the store, which we explain later, this was sufficiently small to avoid computational problems but large enough to maintain the rich nature of the research question.

2.2. Ragged array path comparisons

The fundamental kernel of a cluster analysis is the ability to make distance comparisons among the units. The ragged-array nature of our in-store data (i.e. persons vary in their path times, or number of blinks) makes these distance calculations (computing a pairwise path distance) difficult. How can a path of 150 blinks be compared, in a reasonable fashion, to a path of 1200 blinks? To facilitate comparisons among paths of varying lengths, each path was recoded as a path of 100 percentile locations. The first blink represents the starting point; the second blink represents the store location 1% of the way through the path (measured in distance); the third blink is the location 2% of the way through, etc.[3]

2.3. Spatial constraints

With each path standardized as a 100-blink (percentile) path, they can be easily aligned for pairwise distance computation; however, the numerous store spatial constraints make the application of standard numerical clustering techniques still non-trivial. For example, although simple k-means clustering algorithms could be applied to the bivariate blink data, which would be equivalent to minimizing the pairwise squared distances at each percentile point within each cluster, it would almost certainly lead to infeasible cluster centroids that violate the store's spatial constraints by crossing through aisles and going to inaccessible areas of the store. That is, the average of multiple paths within a cluster, computed as a pointwise average coordinate-by-coordinate (as in k-means algorithms) will not be a feasible store path, and hence not a useful summary of store travel behavior for a given cluster. For this reason, we applied a new-to-marketing clustering algorithm, called k-medoids clustering (Kaufmann and Rousseeuw, 1990), described next, that is able to handle the multitude of spatial constraints.

3. Clustering algorithm

k-medoids clustering was developed primarily to make k-means clustering more robust to outliers. An additional advantage of k-medoids clustering is that its solution conforms to any spatial constraints that exist in the data. In particular, whereas typical k-means clustering begins with a random clustering of all observations,

[2] When Sorensen Associates initially assembled the dataset, they observed a number of very long paths — up to 6 h in duration. These paths did not seem to coincide with actual shopping behavior; for instance many appeared to be abandoned carts that stayed in one place for a long time before a store employee moved them away. Based upon our discussions with Sorensen Associates, we excluded all paths lasting over 2 h.

[3] In the 100-blink paths, the first and last blinks match the first and last blinks of the actual paths. So, technically, the second blink represents the path 1/99th of the way through the path; the third blink is the location 2/99th's of the way through, etc. We simplified the explanation above for clarity.

k-medoids clustering begins with a random selection of observations (four observations are selected for a four-cluster solution, two for a two-cluster solution, etc., to serve as cluster "centers"); in our case, we use a random selection of shopper paths to serve as the initial centers. These observations (paths) are called medoids. Each of the remaining observations (paths) is then assigned to the medoid that has the minimum (Euclidean) distance from it. The next step follows the usual k-means procedure in that the cluster centroid (cluster mean of each variable) is then computed. In our case, as discussed, this simple pointwise mean will yield a centroid path that is almost surely infeasible. So at this point, the k-medoids algorithm diverges from the k-means procedure by calculating the *observed* path (by definition a feasible path) within that cluster that is *closest* to the simple k-means centroid, yielding a new set of medoids. While these medoids may not necessarily be the closest feasible paths to the k-means centroid, they require little computation and lead to canonical paths that are actually observed in the data, a significant advantage when wanting to describe "typical" behavior among a set of shoppers. Also note that many in-store data sets are likely to have a very large number of shopper paths, making the store densely covered by actual paths. Thus, the difference between the theoretically closest feasible path to the simple k-means solution and the observed k-medoid path may be inconsequential. A detailed description of our algorithm is provided in the Appendix.

We make an interesting but justifiable choice in using Euclidean distance instead of travel distance to measure the distance between paths (percentile locations). For example, paths Q and T may be in aisles 2 and 3 respectively at percentile 40. Their Euclidean distance would only be some 10 ft, while the distance as measured by the required travel to arrive at the other location might be 30 ft. The justification for Euclidean distance is two-fold. First, studies on the difference between Euclidean and travel distance find correlations between the two measures as high as .99 and no lower than .90 (Apparicio, Shearmur, Brochu, & Dussault, 2003). In addition, the behaviors of interest in this application, such as the amount and pattern of aisle travel, do not depend on detailed differentiation of the travel distance between two aisles. In addition to justification on these grounds, Euclidean distance is much easier to apply and unambiguous, whereas the travel distance between two points depends on the specific route taken.

A graphical illustration, in Fig. 2, explicates the k-medoids procedure. Four actual standardized paths are

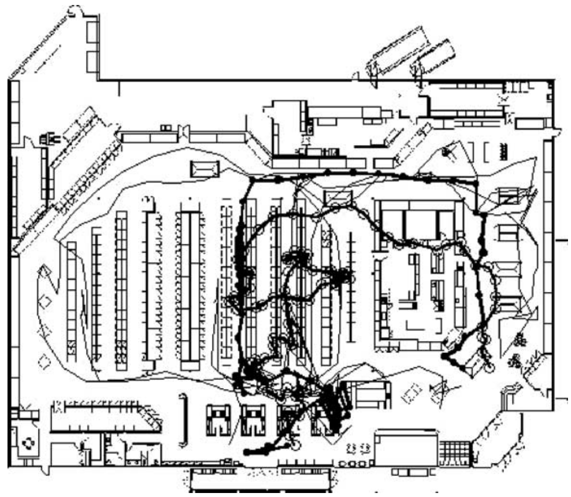

Fig. 2. Illustration of our modified k-means clustering algorithm subject to spatial constraints.

shown in the figure—three are represented with thin lines, the other with a bold line and dots at each blink. The other bold line, with circles denoting each blink, is the 100-blink pointwise mean of those paths (i.e. the naïve k-means cluster centroid). Note that it crosses through aisle shelving and travels through an inaccessible area of the store. Among the observed paths, the bold path with dots at each blink has the smallest sum of squared distance from the infeasible cluster mean, so it becomes the medoid for the given iteration.[4]

The algorithm we described has several desirable properties: (1) it clusters shoppers according to similarity of travel behavior, and (2) yields a feasible path (one that is actually observed) as a summary of the travel behavior manifested in each cluster. Thus, for K clusters, we end up with K canonical paths (medoids) providing a summary of the travel behavior in that store for each group of shoppers. This allows for a visual inspection of store travel behavior without the information overload shown in Fig. 1. To show that these methods provide valuable information beyond what other possible techniques could provide, we present an alternative

[4] At times the observed path appears to clip the corner of an infeasible area or appears to go through an aisle shelf. Since path locations are recorded at 5-second intervals, a shopper that rounds a corner during that time will appear to have traveled through the shelving. Since we know that these apparent discrepancies are artifacts of the nature of the data, we need not be concerned. Any other method to create a fictitious "closest feasible path" would run into worse problems trying to quantify the precise amount of shelf-crossing that is "feasible".

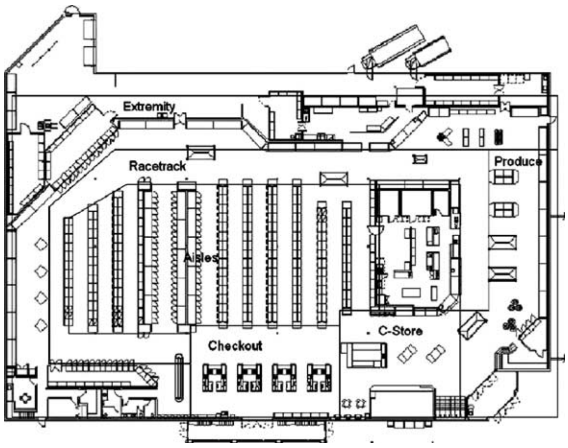

Fig. 3. Store subdivided into zones.

summary technique to which we can compare our results.

4. Profiling shopping paths by zones visited

Unlike standard cluster profiling, where the means of a set of variables can be computed for a cluster, our problem is more challenging in that we need to profile bivariate store paths. We accomplish this by taking each store path and summarizing it by the amount of the path spent in each of several strategically important "zones". We constructed the zones based upon discussions with Sorensen Associates (see Sorensen, 2003) and our own understanding of in-store shopping behavior. These zones are pictured in Fig. 3.

The Racetrack, the main thoroughfare on the outside edge of the aisles, is so named because travel in this section tends on average to be faster than travel in other zones. This is likely due to the higher amount of *travel* (but not necessarily *shopping*) that occurs here versus other areas. The Aisles are important because most people make the implicit assumption that the majority of shopping occurs there. The Produce section is of obvious importance to any grocery store, both in the minds of the consumer and in the financial statement of the store. The Convenience Store (C-Store in the figure) gets its name from the nature of the items in that section, many of which could be considered quick-stop items. The Checkout area is a necessary part of any shopping path. The Extremity consists of the shelving on the outside of the racetrack. In most stores, this includes, for example, the dairy section (often towards the end of the racetrack).

For each path, we record the percentage of the path that occurred in each of these six mutually exclusive and exhaustive paths areas. Percentage of the path was recorded as opposed to number of blinks as it allowed us to normalize out path length from path pattern. Fig. 4 displays via boxplots the distribution of each zone percentage variable across our sample of 8751 paths. The Convenience Store for the most part is not highly visited, but there are several outliers, indicating specialized behavior in this area. The existence of several outliers in every zone suggests that several shoppers are only shopping very select areas of the store in one shopping path.

A logical method to proceed based on these mutually exclusive variable profiles would be to use them in

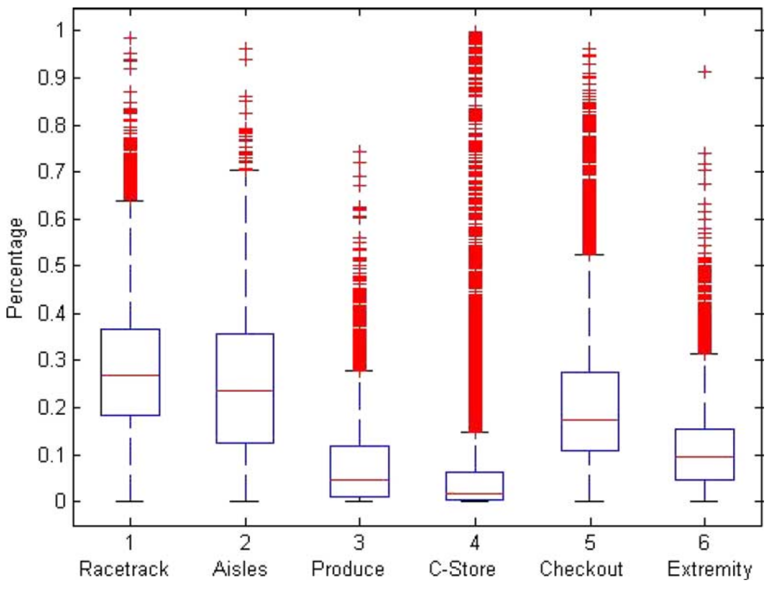

Fig. 4. Percentage travel in each zone as distributed across the population.

Table 1

	Racetrack	Aisles	Produce	C-Store	Checkout	Extremity
Path A	.2024	.3333	.0060	.1667	.2083	.0774
Path B	.2030	.3008	.0075	.1504	.2707	.0602

a straightforward k-means clustering algorithm. Perhaps this would allow one to find interesting patterns in the zone usage. While that might indeed lead to some interesting findings, it would come at a great loss of information. Consider the two following profiles, which are based on actual paths (Table 1). Under each store section is displayed the percentage of travel in that zone for the given path.

From the profiling variables, it appears that the two paths are nearly identical, aside from a slightly higher proportion of travel in the Checkout zone for Path B. However, their shopping paths actually show very different travel patterns (See Fig. 5).

One way to resolve this discrepancy would be to create more zones with less area. For example, if we made each aisle its own zone, the profiles of these two paths would no longer look the same. But the problem would still persist if, for instance, two paths traveled the same aisles in a different pattern, one going along the top of the aisles, the other going along the bottom. Even if the store were divided into hundreds of zones, two very different paths could have similar statistical profiles if they traveled in opposite directions. Zone divisions, no matter how well devised, will lead to a loss of information on two major accounts—order of the visits and precise locations visited. Our method, on the other hand, keeps both of these dimensions intact, while losing only time information in the standardization of the path lengths. This information can still be incorporated into our analysis by means we describe in the next section.

5. Clustering results by time

The time dimension of shopping has been notably absent from most of the previous discussion. While we make all paths comparable to each other by creating standardized "percentile paths", we expect to see vastly different behaviors in a 5-min versus a 30-min path. Consistent with this supposition, running k-medoids clustering indiscriminately on all paths leads to cluster solutions dominated mostly by path length. In fact, running hierarchical k-medoids clustering leads to high-order groupings of mostly uniform path length.[5]

For ease of exposition and description of the results of the algorithm, we chose to split the set of 8571 paths into three equally sized groups and run our k-medoids clustering on these groups. The resulting splits yield a "low" group of 2917 paths with travel times ranging from 2 to 10 min, a "middle" group of 2916 paths ranging from 10 to 17 min, and a "high" group of 2918 paths lasting from 17 min to nearly 2 h. This is akin to a hierarchical clustering where the three highest-order major groupings are based solely on time.

The splits by time also have an intuitive appeal in that a longer "stock-up" path is likely to be quite different from an intermediate "fill-in" path. Similarly, paths under 10 min don't leave time for the shopper to buy more than a few items. Obviously, shoppers from this group are looking to grab a few important items and leave. By splitting the analysis this way, we now incorporate the time dimension of the shopping path that we had previously lost.

We now undertake a separate analysis and discuss the results for each of these three groups.

5.1. Low group

To find the "optimal" number of clusters using our k-medoids procedure, we need to balance adequate fit (low within-cluster sum of squares) and parsimony. A number of techniques exist to help choose the "optimal" number of clusters. We employ here two different methods. A traditional approach is to find the cluster solution for various numbers of clusters and plot the within-cluster error by cluster number. This resulting "scree plot" should have an "elbow" at the correct number of clusters (Sugar, 1998). In addition to this more traditional method, we computed the KL statistic for each cluster solution (Krzanowsky & Lai, 1985). This statistic utilizes the error improvement from the cluster solution with k versus $k+1$ clusters to find the best solution. Specifically, it requires that we compute

$$\mathrm{DIFF}(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k$$

where k is the number of clusters, p is the number of variables used (200 for our application—both x and y coordinates for 100 locations), and W_k is a measure of within-cluster error for k clusters. The KL statistic is

$$\mathrm{KL}(k) = |\mathrm{DIFF}(k)/[\mathrm{DIFF}(k+1)]|.$$

The cluster solution with the highest KL statistic is the recommended solution. Note that KL(k) is not defined for $k=1$. The dual application of such different

[5] Available upon request.

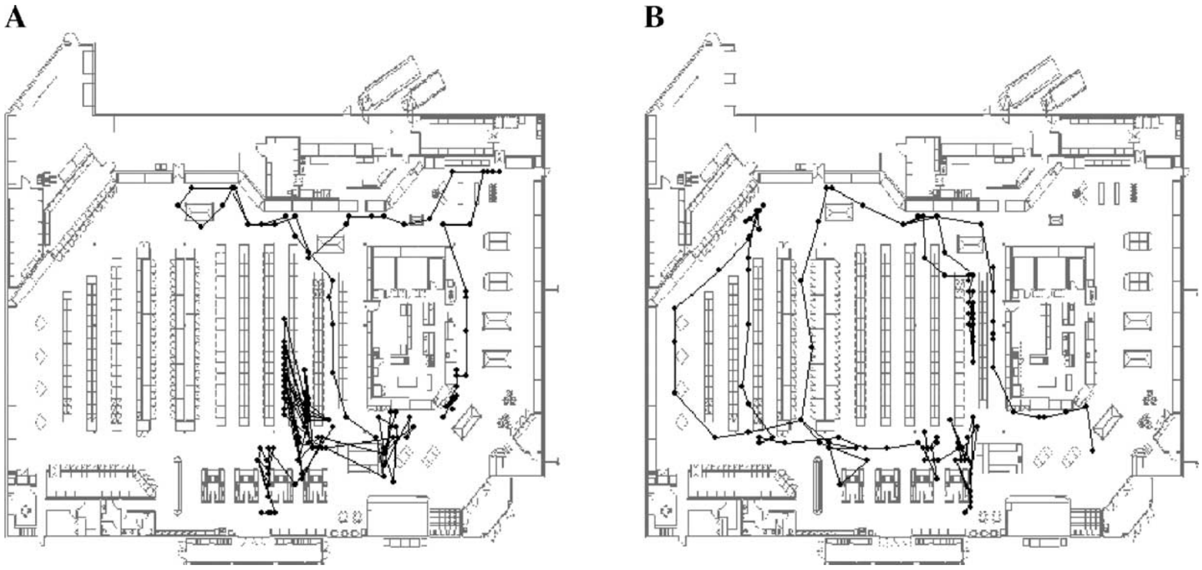

Fig. 5. Path A and path B.

techniques provides greater evidence for our choice of the number of clusters. As with *k*-means clustering techniques, our algorithm produces a local solution, so we ran the algorithm from 20 different (random) starting points for each number of clusters to ensure a suitable solution. The corresponding scree plot and display of the KL statistic by cluster number for the low time group are shown below in Fig. 6.

Though we observe no obvious "elbow" in the scree plot, there is an observable kink at 2 clusters. This choice is further justified by the KL statistic, which is highest at $k=2$. Findings for alternative cluster solutions are available upon request. We present the resulting cluster centroids in Fig. 7 below. The N associated with each centroid shows the number of paths assigned to that cluster.

For those under 10 min, there exist two distinguishing patterns. The store is laid out in such a way that most shoppers choose the "default" start path along the racetrack to the right of the infeasible zone (i.e., office/storage area between the aisles and the produce). Over half of the low group paths, whether or not they actually shop in the produce area, follow this default path. However, a significant portion of short paths (cluster 2) breaks the default pattern. This is likely due to time pressures—shoppers making shorter paths want to finish their few tasks as quickly as possible, and thus are less likely to follow the default traffic flow. We will see from the results of the longer groups that shoppers not faced with such self-imposed time constraints are more likely to follow the default path up the right-hand side of the store.

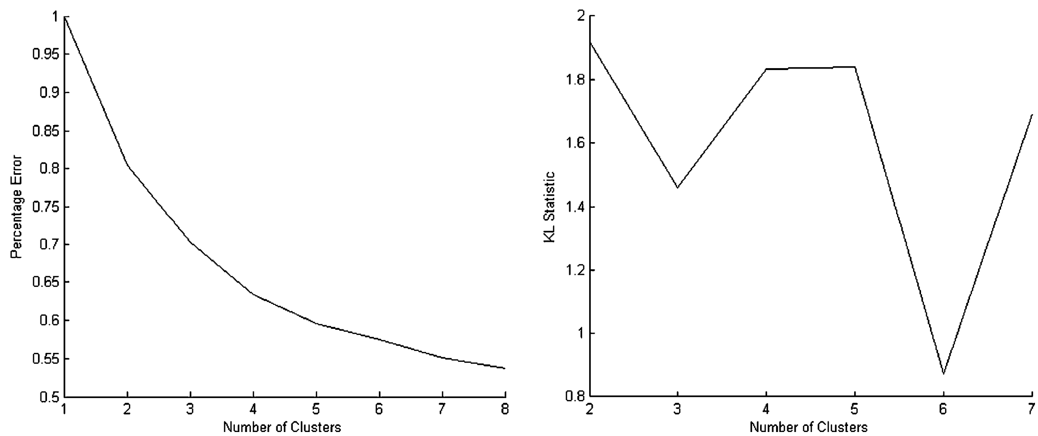

Fig. 6. Scree plot and KL statistic by cluster for low group.

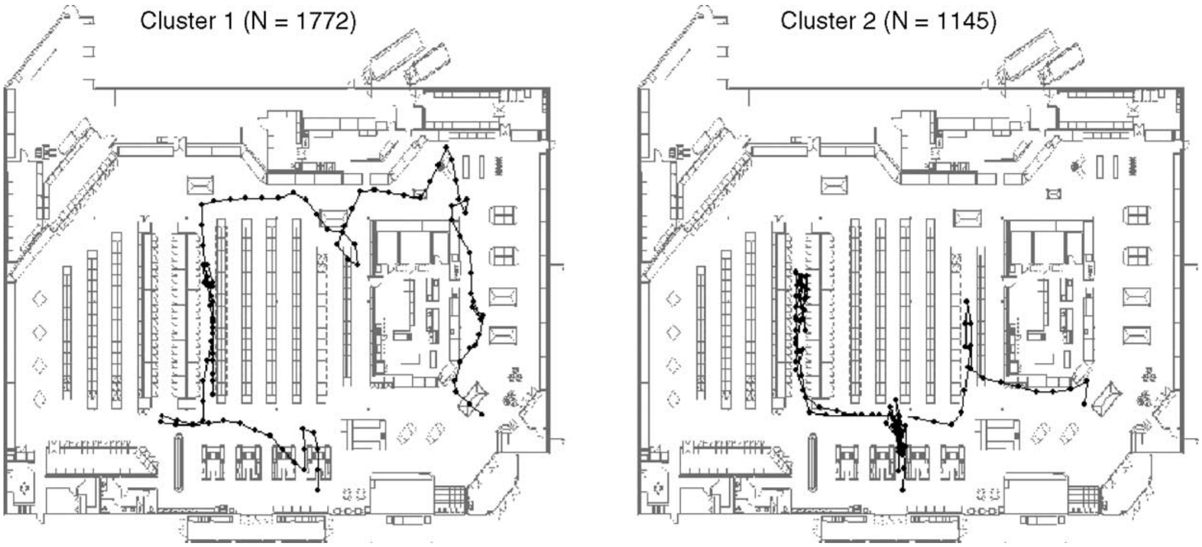

Fig. 7. Low group medoids.

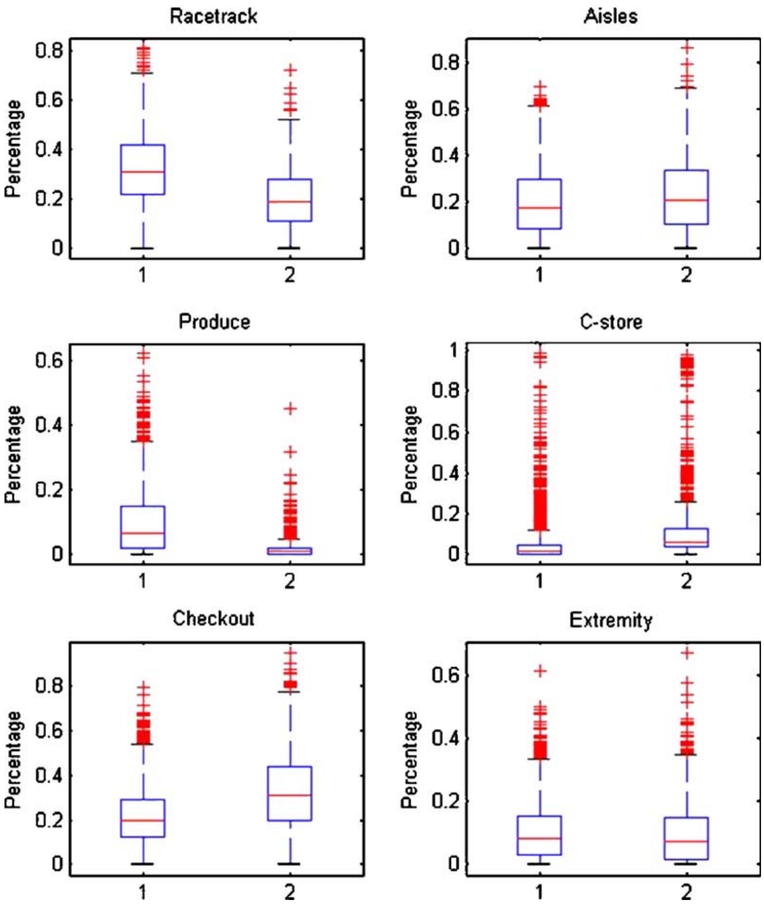

Fig. 8. Differences in profiles between clusters for low group.

It is informative to examine how the results from our k-medoids clustering compares with the store-zone profiling technique described earlier. Profiling, with its described weaknesses in information loss, still provides summary information that sheds valuable light on a certain type of shopping path. This is especially true when comparing "canonical" paths that emerge from the cluster analysis. The displayed medoids show the central travel tendency of each cluster, but a profile summary of the entire cluster provides further intuition as to the important clustering dimensions. The distribution of paths within each of the two low time clusters is represented in Fig. 8.

The clusters are visibly different on every dimension except perhaps Aisle and Extremity. The biggest differentiator appears to be the use of the default path, along which both Racetrack and Produce lie. No differences are observed in the total path length across these two clusters.

5.2. Medium group

Since we expect more divergence in behavior with longer paths, we also expect to find more "canonical path types". In other words, a cluster solution with more than two clusters will likely be appropriate for the medium length paths. Again, an observable kink at 4 clusters confirms the choice of four clusters as recommended by the KL statistic, seen below in Fig. 9. Fig. 10 presents the four resultant cluster medoids. Several interesting patterns emerge. Note that shoppers in this intermediate group appear to be less time constrained, as evidenced by a higher propensity to follow the default start path along the right-hand side of the store. All four paths at first glance appear to be more homogeneous than the two cluster medoids from the low group, as they all follow a similar start path and continue around the racetrack for some time. Upon further examination, however, we notice significant variation across the four groups. Clusters 1 and 3 are much more dominated by racetrack travel—cluster 1 because it follows the racetrack farther; cluster 3 because it spends more time in the smaller area of the racetrack that it covers. Clusters 2 and 4 follow the racetrack, but appear to be using the racetrack to travel to their next shopping destination, not to shop there. Finally, cluster 4 spends a long time in the checkout area. This could be due to a slow cashier, socializing, or actual shopping in that area. With the current data we are unable to answer that question, but it raises an issue that the retailer might want to examine. At first glance, clusters 2 and 3 appear to be extremely similar. Further inspection reveals the importance of the time dimension in classifying trips. Though the dominant pattern is similar, cluster 2 moves more quickly through the produce and into the second aisle. Thus, though they both go into aisle two, they do so at different times. In cluster 2, for instance, percentile 20 is in aisle 2 while the 20th percentile of cluster 3 is still in the produce area, a large difference.

Again, the zone profiling variables are notably different across clusters, as seen by the boxplots in Fig. 11. As observed from the medoids, clusters 1 and 3 display more racetrack travel, while cluster 2 is dominated by aisle travel. Clusters 2 and 4 exhibit almost no produce travel, consistent with the speed with which the medoid path went through that area. Indeed, cluster 4 as a whole spends more time in the checkout area. Note that only one of the four clusters displayed more aisle travel than racetrack travel. This may be evidence that the current store layout does a good job of accommodating medium

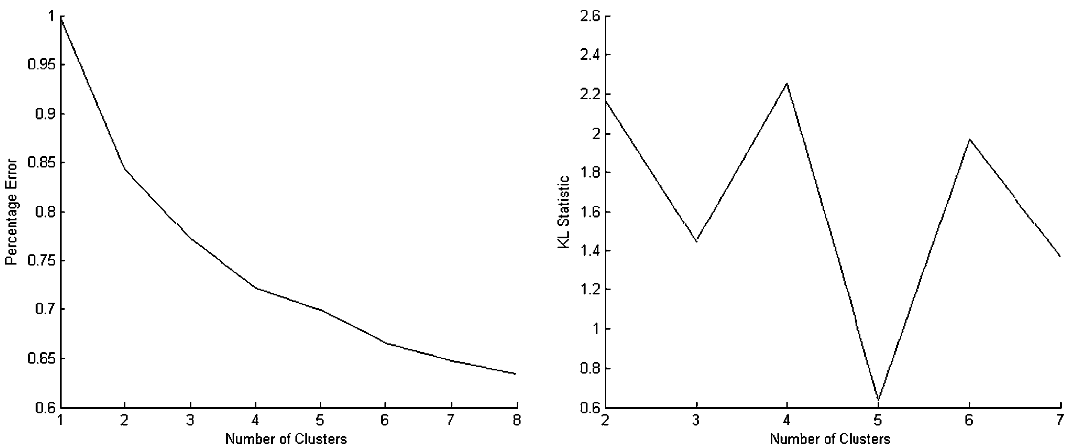

Fig. 9. Scree plot and KL statistic by cluster for medium group.

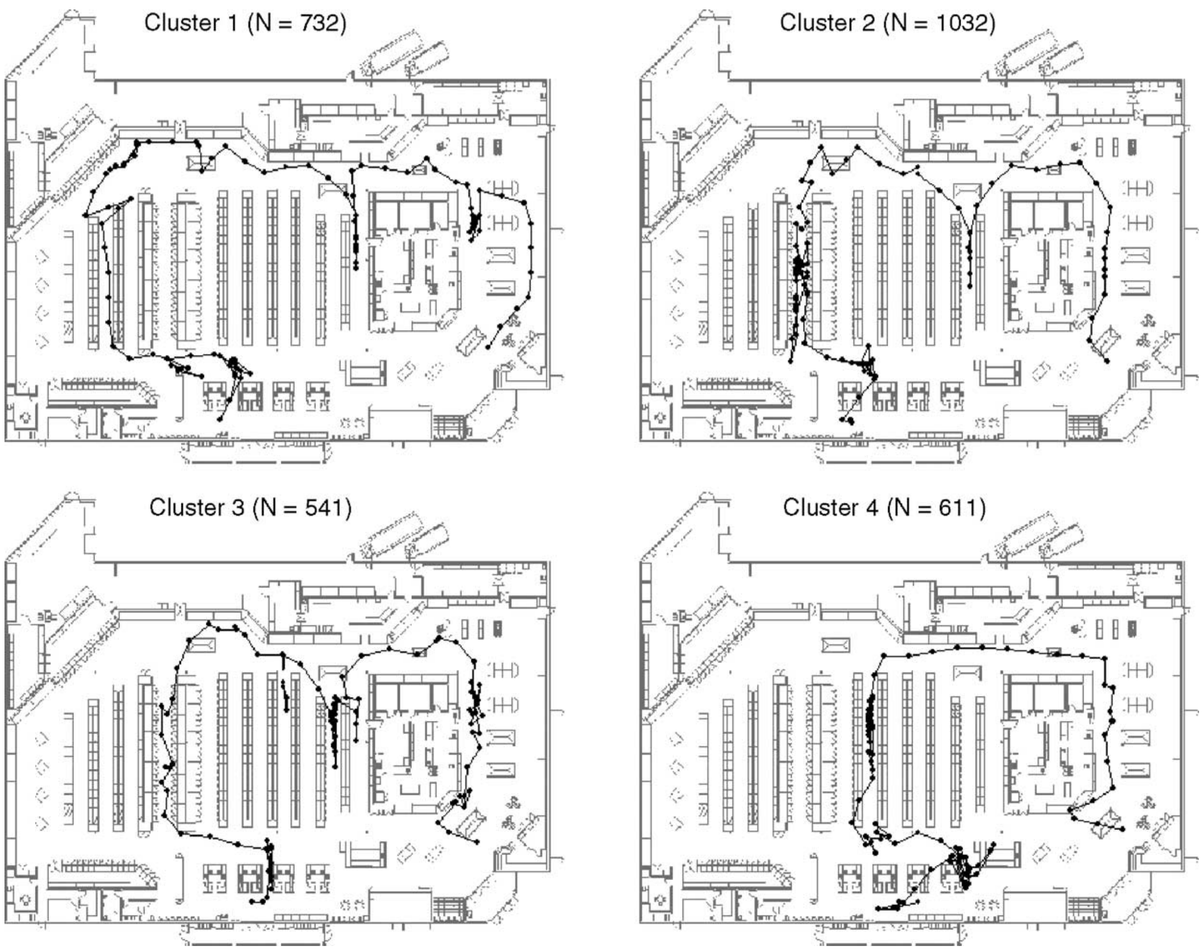

Fig. 10. Medium group medoids.

paths, which are likely for refilling key food items after a few days of depletion. Shoppers appear to be able to fill most of their basket by traveling the main thoroughfare and making quick excursions into the aisles.

5.3. High group

As the most variable group in path length, we also expect to see a high degree of variability in the observed shopping patterns. The kink at 8 clusters is the most obvious in the scree plot, and this observation is again confirmed by the KL statistic (Fig. 12).

Fig. 13 presents the medoids for the eight clusters. As expected, we see a high degree of variability in path type. Cluster 3 is the most unique path. The Convenience Store, with its quick stop items, also has a small Chinese food takeout counter, which likely kept many of this cluster's shoppers in the store for over 17 min. Medoid 4 is also interesting: despite the absence of any self-imposed time constraint (as surmised by its length), these paths choose to break the default start path to go directly to the desired items in the aisles. Another path dominated by aisle travel is path 5, which spends most of its time in a different set of aisles from those traveled by cluster 4. In no cluster do we see aisle travel that spreads across all twelve aisles. It appears that an important dimension that distinguishes aisle-traveling clusters is the choice of particular aisles in which to shop. Therefore, the commonly assumed travel pattern of complete aisle-by-aisle shopping is not supported by this analysis. The dominant travel pattern, if it includes any aisle travel at all, includes only select aisles.

As with the medium length paths, one of the most important distinguishing dimensions is not whether the path travels along the racetrack, for the vast majority do—it is their use of the racetrack, whether it be for shopping or travel. Cluster 1 seems to balance both, using the racetrack to travel to the important aisle purchases, but also spends extra moments there, likely for shopping purposes. Cluster 5, though it covers a

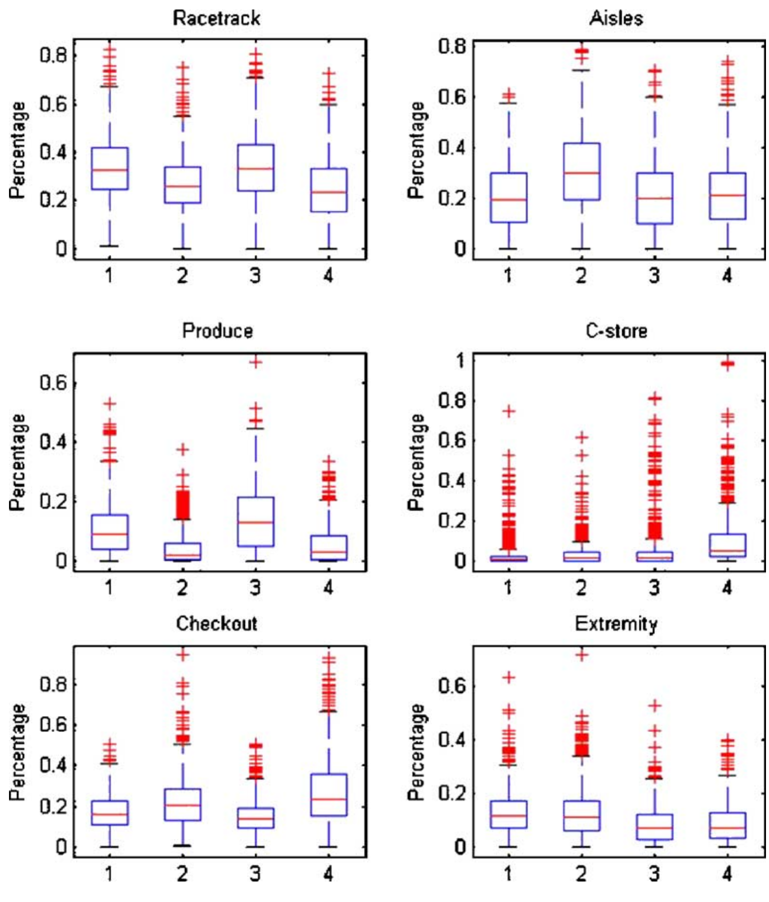

Fig. 11. Differences in profiles across clusters for medium group.

great deal of the racetrack, spends very little time there, moving on it only to arrive at more important destinations, specifically products located in select aisles and in the extremity. Cluster 2, though it does not appear to utilize the whole of the racetrack, spends a great deal of time in the racetrack sections it does travel, taking several major pauses on it. The sixth cluster exemplifies some of the same pattern seen in cluster 1; that is, shopping along the racetrack while taking quick excursions into the aisles for specific products (that is, entering and exiting the aisle from the same side). Though full-aisle traverses (entering one side of the aisle and

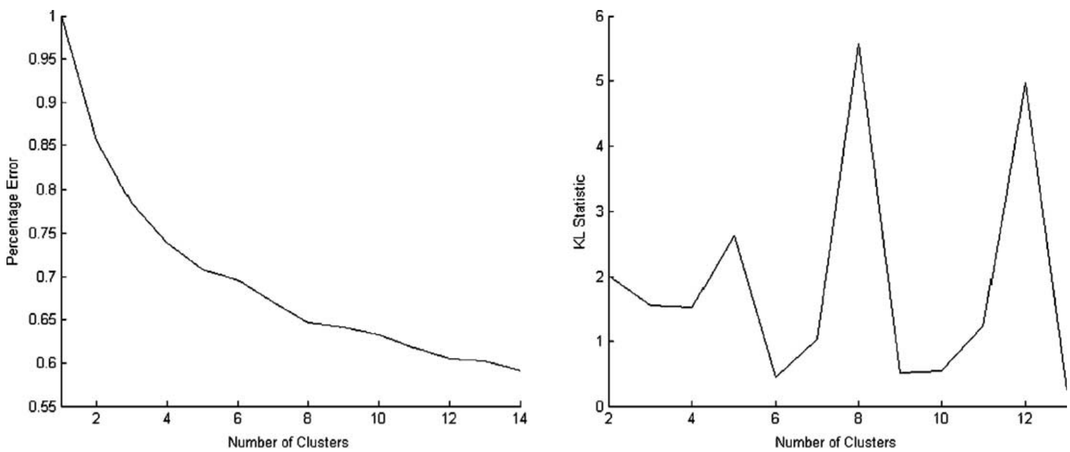

Fig. 12. Scree plot for high group.

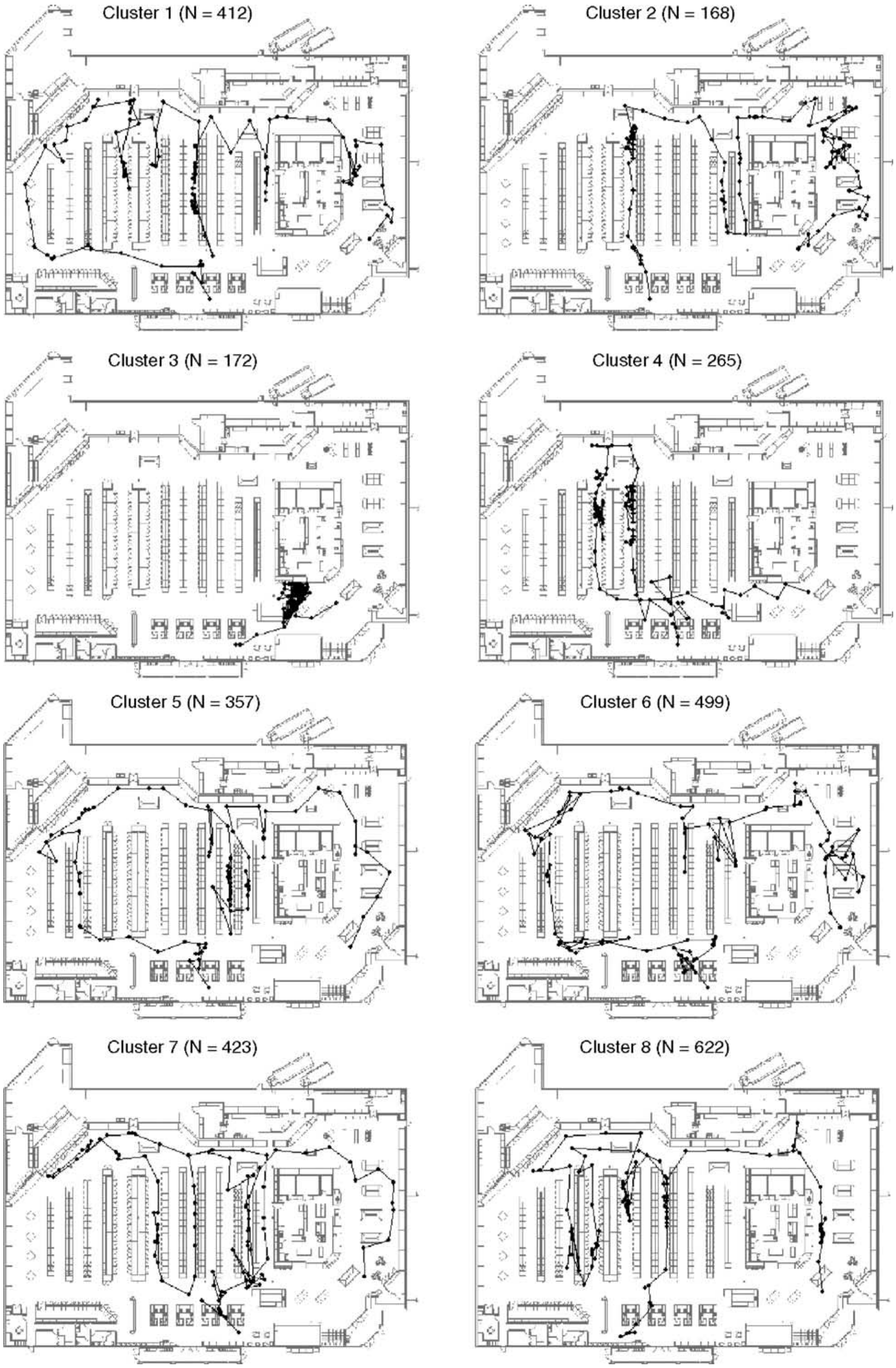

Fig. 13. High group medoids.

traveling all the way through it) are seen in several of the medoids, quick aisle excursions are far more common, attesting to the importance of good end-cap merchandising, since racetrack-with-excursions paths, as seen in clusters 1 and 6, will spend much of their time near these end-of-aisle displays.

Another "default shopping pattern" – forward progress shopping – is broken by clusters 7 and 8. These medoids, 7 and 8, display significant backtracking, shopping in aisles that were previously passed, whereas the other medoids tend to flow in a single direction towards the checkout, making necessary stops along the way. This can be viewed as evidence that most shoppers are looking to make their shopping path efficient, picking up the necessary products in an orderly, logical manner. There are many possible reasons why medoid paths 7 and 8 do not follow this logical flow. Perhaps they do not put forth the mental energy to organize their path, or they forget important purchases until later; perhaps product choices are themselves stochastic, influenced by store atmosphere. A better understanding of the shopping process could lead to important discoveries for retailing.

Again, we present the variation of the zone profiling variables across clusters in Fig. 14. As expected, cluster 3 is high on percent Convenience Store. Cluster 2, somewhat surprisingly, shows the highest racetrack travel, whereas the medoids seem to indicate a higher racetrack level from cluster 1 or cluster 6. The large clump of blinks at the top right of medoid 2 indicates that many of the paths in cluster 2 spend a long portion of their trip at the right of the store, thus inflating their racetrack score. This is further evidenced by the fact that cluster 2 has the highest produce travel. As mentioned, the profiling variables by themselves can be misleading, as the high racetrack statistics in cluster 2 may lead us to believe that members of that cluster tend to travel more of the racetrack. The results of the k-medoids analysis inform us that in reality, clusters 1 and 6 travel more of the racetrack. Clusters 4, 5 and 8, not surprisingly, are high on aisle travel, as is cluster 1, with its several excursions. The backwards pattern of clusters 7 and 8 are not at all evident from these boxplot displays, again supporting the value of the k-medoid clustering method as opposed to a clustering algorithm based on summaries.

6. Cross-validation

The results of clustering algorithms can sometimes be unstable, so we display the results of a cross-validation performed on an additional third of the data. This

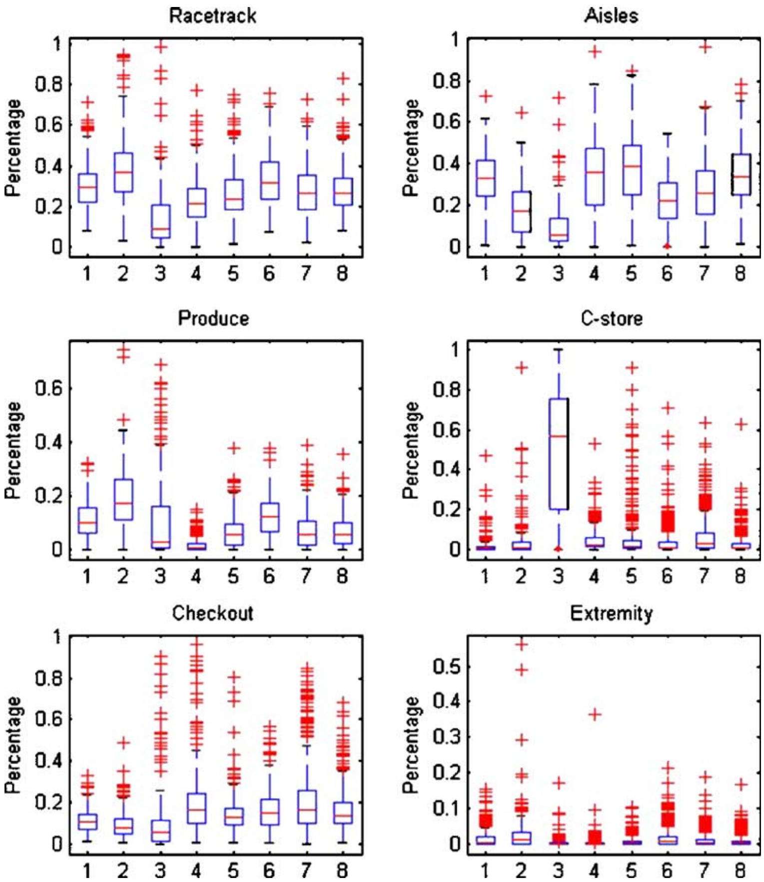

Fig. 14. Differences in profiles across clusters for high group.

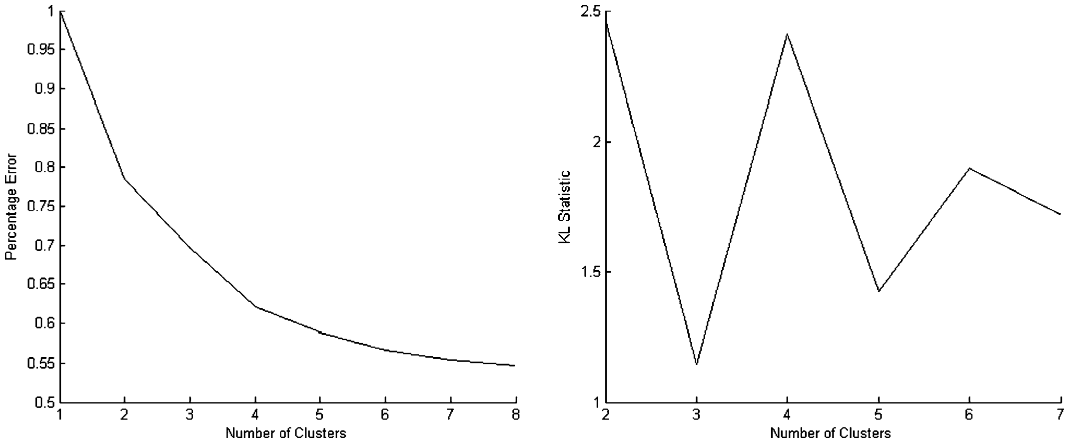

Fig. 15. Scree plot and KL statistic by cluster for low group.

cross-validation allows us to assess the stability of the results in the sense of whether a given store layout will consistently lead to same overarching patterns displayed in our particular sample.

6.1. Low group

Fig. 15 presents the scree plot and KL statistic by cluster for the low group in the cross-validation sample. The choice of the two-cluster solution is again justified.

Fig. 16 displays the cluster medoids for this two cluster solution. Similar to cluster 2 from our original low group, cluster 1a cuts away from the default path to go immediately into the second aisle. Also like the original cluster 2, cluster 1a here makes a lengthy excursion into aisle 8. Cluster 2a here also displays the same general pattern displayed by our original cluster 1.

The profiles of the two clusters provide further evidence of the stability of the resultant cluster solution. Comparison of these profiles with the original solution shows a nearly identical pattern of differentiation in travel behavior (Fig. 17).

6.2. Medium group

Consistent with our earlier results, the choice of four clusters is the best for the medium group, as can be seen from the same scree plot and KL statistics (Fig. 18).

Interestingly, each canonical path here (Fig. 19) corresponds to one of the canonical paths from our original results. The cluster 1b medoid displays nearly identical

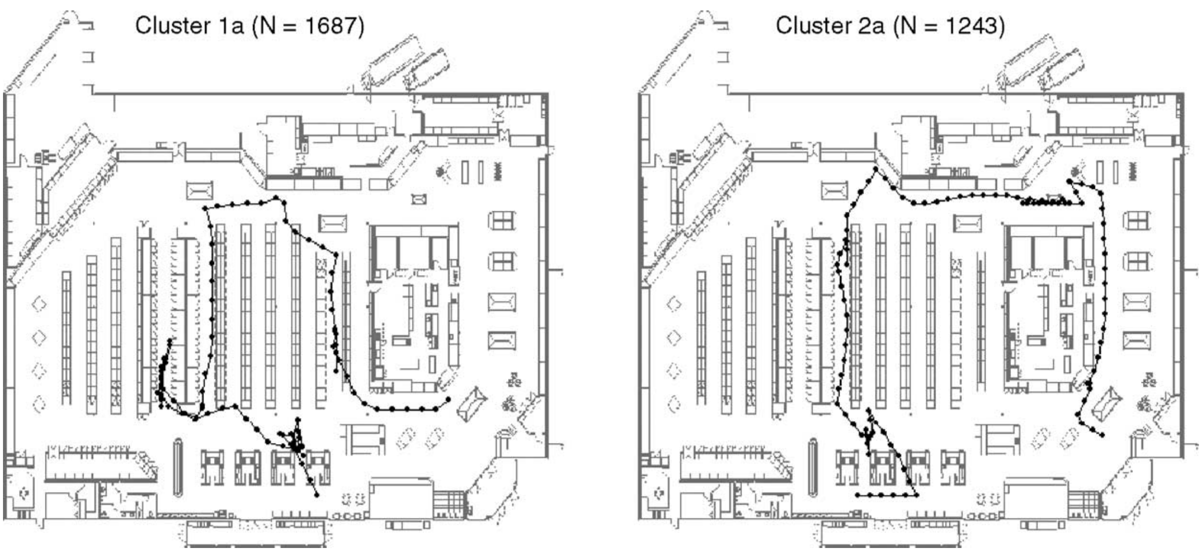

Fig. 16. Low group medoids.

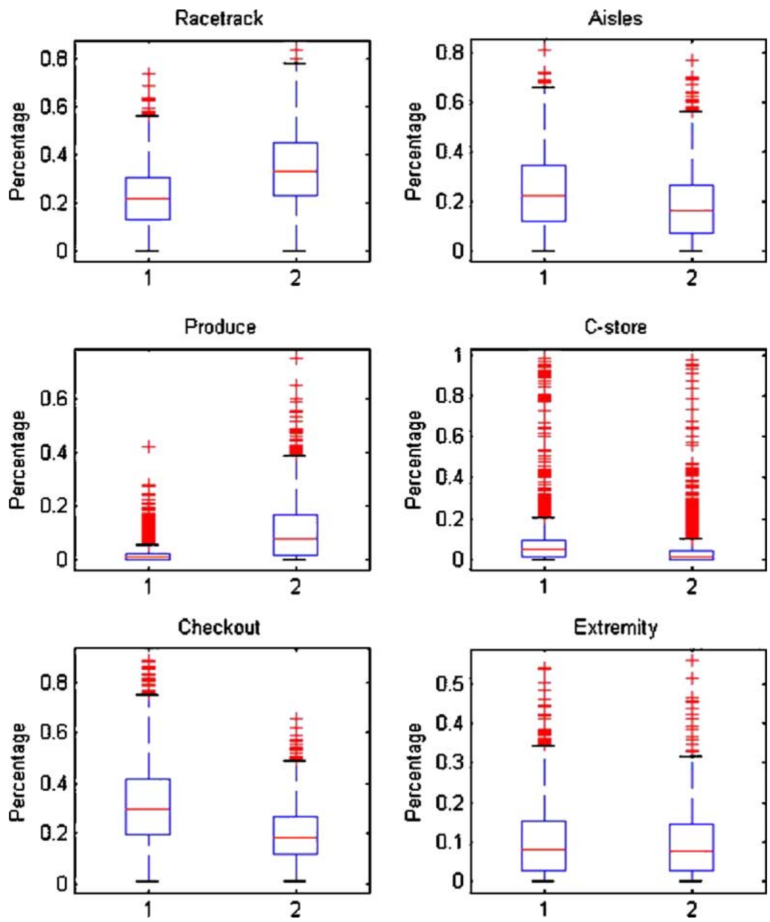

Fig. 17. Differences in profiles between clusters for low group.

behavior as the medoid for cluster 2 (Fig. 10) from our earlier results. This includes heavy traveling in the same aisle toward the left side of the store, and an excursion into an earlier aisle. Similarly, cluster 2b here shows behavior strikingly close to the original medium group's cluster 1. They both take the racetrack around nearly the whole store and travel down one of the last aisles. Cluster 3b displays very similar behavior to the original cluster

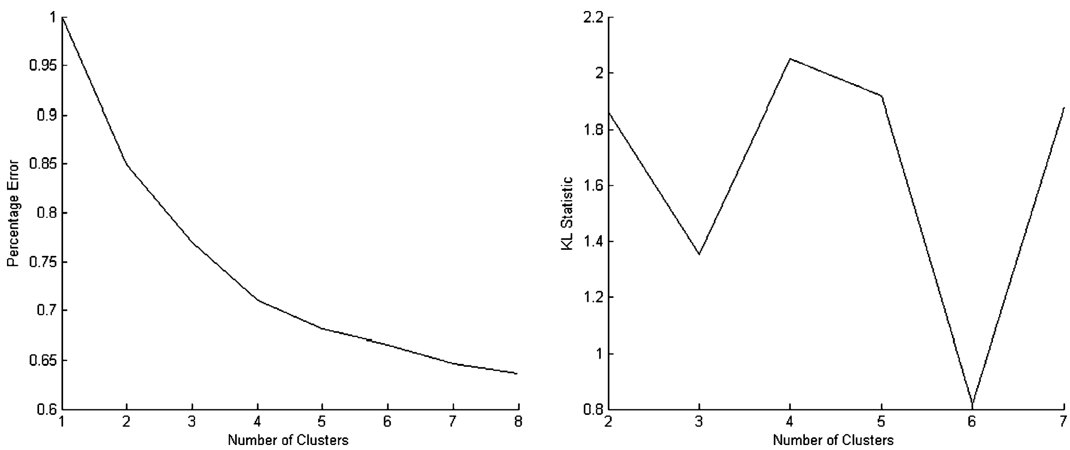

Fig. 18. Scree plot and KL statistic by cluster for medium group.

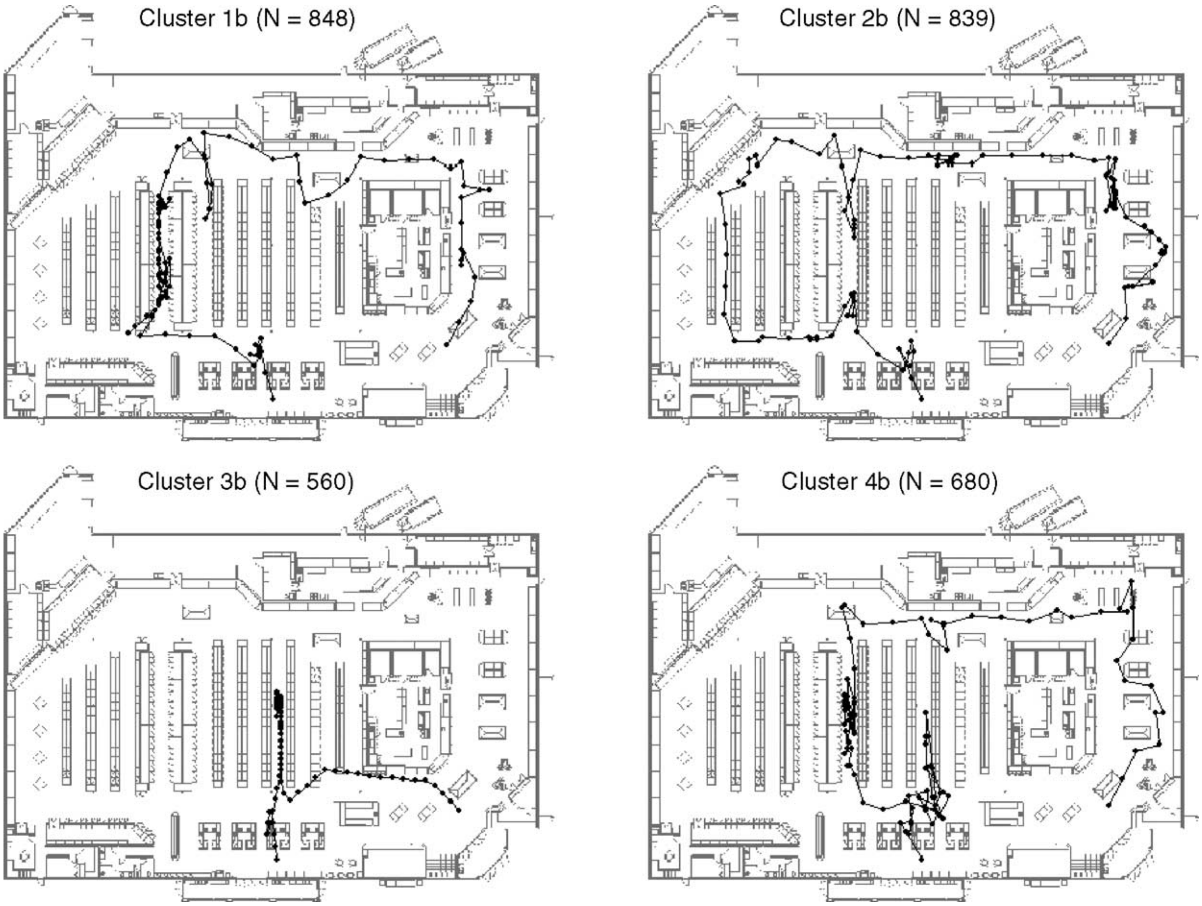

Fig. 19. Medium group medoids.

4, though it is not immediately obvious. Here, the medoid breaks the typical route to go directly into an aisle, where it spends the majority of the path. The original cluster 4 does not break the typical route, but it does travel expeditiously into an aisle, where it spends most of its time. Cluster 4b below does not correspond as cleanly to the first cluster 3, but the overall patterns displayed in this set of paths correspond rather well to the original results, giving evidence that the clustering results are relatively stable.

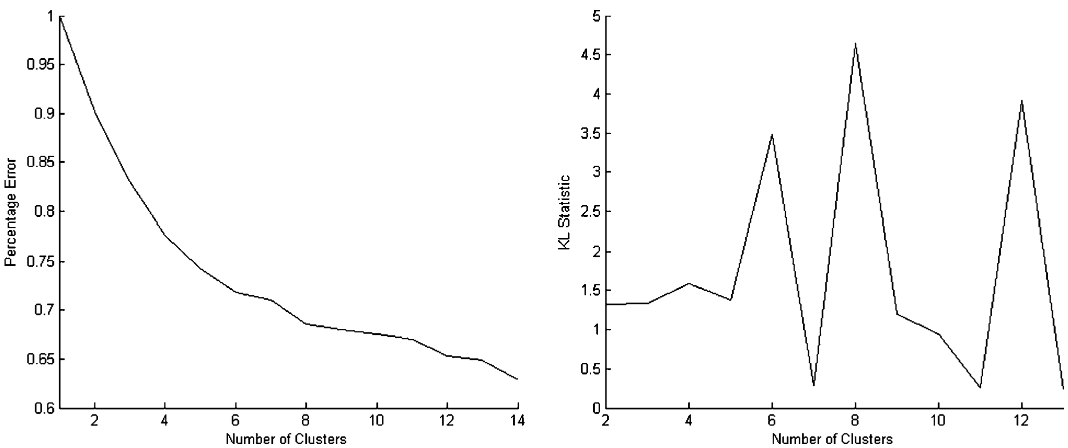

Fig. 20. Scree plot and KL statistic by cluster for high group.

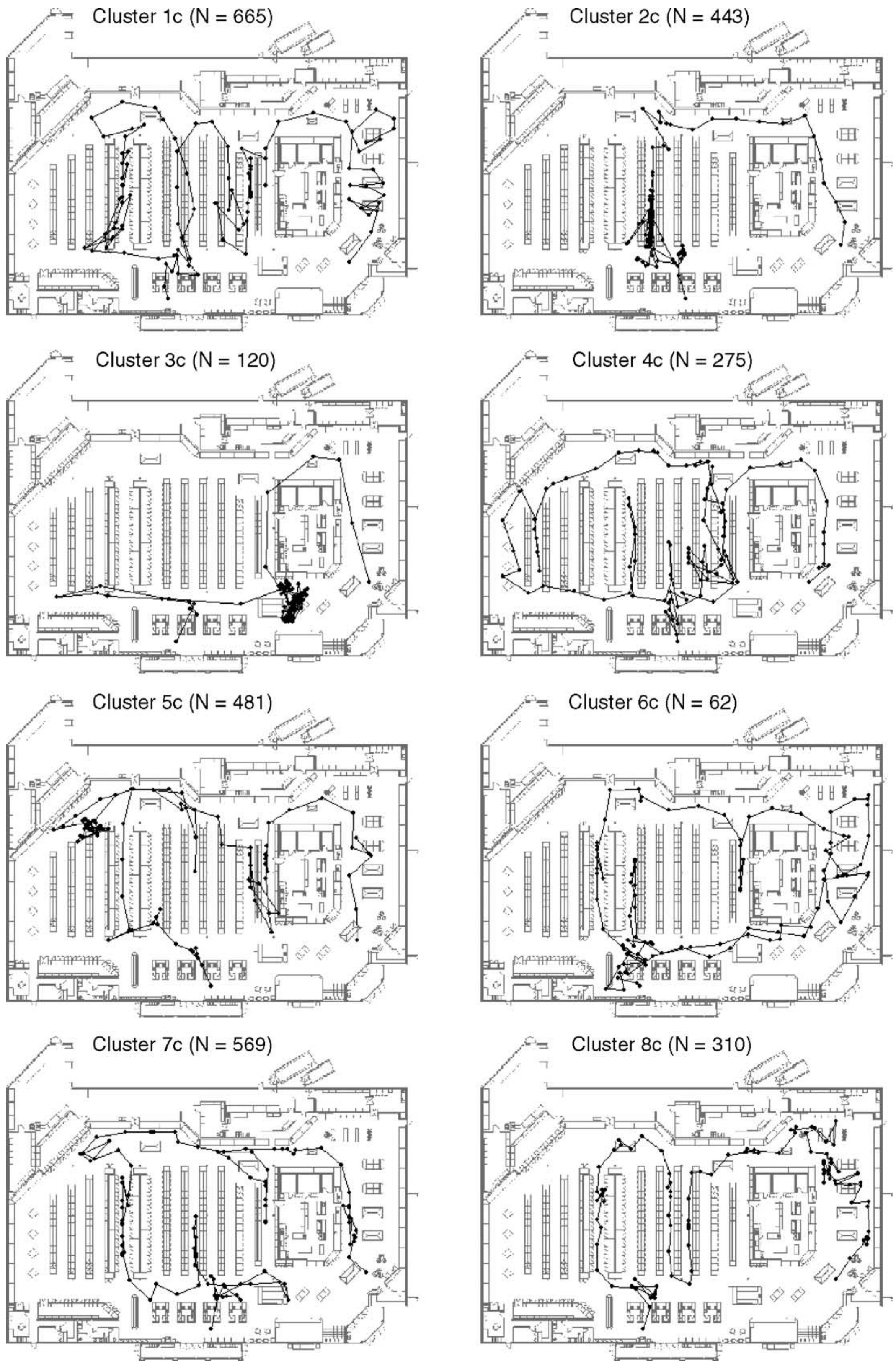

Fig. 21. High group medoids.

6.3. High group

With the high time group, the recommended number of clusters again is eight (Fig. 20), just as with the original data.

In summary, the correspondence between the paths shown in Fig. 21 with those in Fig. 13 is quite high, yet for brevity we omit the exact details. Suffice it to say that many of the dominant patterns seen in these results correspond to many of the same patterns from the original set of results. Inspection of the profiling variables (not shown) also demonstrates a similar pattern of spread across the eight clusters.

7. Discussion

Our main purpose in presenting these exploratory analyses was to familiarize other researchers with the existence of such data and stimulate ideas for future use. As such, we leave to future researchers the work of detailing the many potential managerial implications of work using this type of data. However, even the exploratory work we have presented here carries useful and actionable information for store managers.

A simple examination of the canonical paths resulting from the *k*-medoids clustering helps dispel a number of myths that our personal schemas about supermarket travel perpetuate. Of particular note is the extremely low occurrence of the pattern commonly thought to dominate store travel—weaving up and down all aisles. We note that most shoppers tend only to travel select aisles, and rarely in the systematic up and down pattern most tend to consider the dominant travel pattern. Those trips that do display extensive aisle travel tend to travel by short excursions into and out of the aisle rather than traversing the entire length of it. This simple observation has important implications for the placement of key products, the use of end-cap displays, etc. Products placed at the center of aisles will receive much less "face time" than those placed toward the ends. Of related interest is a practitioner study that found that placing familiar brands at the end of the aisles served as a "welcome mat" to those aisles, effectively increasing its traffic (Sorensen, 2005). Granted, the previous observations are specific to this particular store, and cannot without caution be applied to all grocery stores, but this template for identifying true store utilization patterns can be equally useful for any store layout. Informed decisions about store layout can only be made through direct observation of the current utilization of the store.

Once we observe that the aisles are utilized much less than common folklore leads us to believe, we turn our attention to the areas of the store that pick up this slack. Whereas previous folklore perpetuated the myth that the perimeter of the store was visited incidental to successive aisle traverses, we now know that it often serves as the main thoroughfare, effectively a "home base" from which shoppers take quick trips into the aisles. The relationship between the perimeter travel and aisle travel has sparked substantial practitioner interest. The data and techniques described in this paper form effective first steps at understanding this complex relationship. Shorter trips tend to stick predominantly to the perimeter and convenience store areas. This simple observation provides an important starting point for the targeting of particular shopper segments.

While the dataset featured in this paper is quite novel, we acknowledge that other researchers have addressed the general topic of in-store shopping patterns in the past. Every ten years or so, researchers seem to "rediscover" this topic, and have applied very different methods to capture it. One of the earliest such studies of shoppers was a paper by Farley and Ring (1966) who built a stochastic model to study zone-to-zone transitions within a store. Unfortunately, few researchers, to our knowledge, extended or applied their model. Coming from a psychological perspective, Mackay and Olshavsky (1975) examined consumer perceptions of store space, and Park, Iyer, and Smith (1989) sought to understand the impact that store knowledge and time constraints have on unplanned buying, failure to make planned purchases, and other purchase behaviors. Perhaps the most famous study, or series of studies, on in-store shopping behavior is *Why We Buy* (1999) by Paco Underhill. He uses anthropological methods to uncover a variety of behavioral patterns observed while tracking shoppers in different types of retail stores, but limits the depth of his research findings to basic suggestions about ways to enhance consumer convenience. Of all the facets of shopper behavior explored in previous research, none has focused on the complete shopper path as we have, thus making our research a useful step forward. A natural avenue of investigation would be an effort to tie the results and methods discussed in these earlier psychological and anthropological studies to the broader behaviors illustrated in the present study.

Another stream of related research deals not with the grocery store but with spatial movements in general. Some of the most directly related research in this line has examined individual pedestrian movements in museums and shopping malls (Batty, 2003). The chapter presents a number of useful models to describe individual movements that will prove useful to further research on shopper movements. The emphasis in that work is on pedestrian flow, rather than profit from a store. Earlier work in environmental psychology also studied pedestrian traffic flow inside an architectural space, with the hopes of improving architectural design (Winkel & Sasanoff, 1966). Though this work has obvious connections to the present work, its pure focus on traffic flows makes its application to grocery stores not entirely straightforward.

Other work in environmental psychology has an entirely different potential application to the current field of study. Anthropological studies about people's impressions of their surrounding neighborhood has postulated that people look for order on their environment out of an inherent need to organize it in their minds (Lynch & Rivkin, 1959). The way in which shoppers organize a store in their minds may have important implications for their subsequent movements. The current work provides a springboard from which this can be studied.

The exploratory analyses we have presented on this new realm of shopper behavior research are only a first step in understanding shopping behavior within the store. The present research focuses only on travel patterns without regard to purchase behavior or merchandising tactics. A study of the linkage between travel and purchase behavior seems a logical next step. Linking specific travel patterns to individual purchase decisions may lead to an improved understanding of consumer motivations for purchasing certain items, and can shed light on the complementarity and substitutability of goods in ways that a more traditional "market basket" analysis cannot capture.

Further exploration of travel behavior, independent of purchase, also seems another promising route for future research. In this paper, we have presented some exploratory techniques useful for knowledge building and intuition. A more formal model of travel behavior would lead to an increased understanding of shopper heterogeneity of travel and the underlying sources of said heterogeneity. Specifically, one could model travel as a series of "blink-to-blink" choices (with a careful focus on state dependence, since choices made earlier in the path probably have a great deal of influence on later choices). This would allow a more precise study of the key areas of the store—and perhaps merchandising activities—that may influence travel in a particular direction.

Before plunging deeply into such a complex model, we felt it was important to first understand this rich new dataset and the behavioral/computational issues it points to. We hope that this exploratory analysis serves as a useful catalyst for future research that will help us better understand the actual shopping patterns – as opposed to the widely accepted folklore – that take place in different types of retail environments.

Acknowledgements

The authors wish to thank Sorensen Associates for providing the data and, in particular, Herb Sorensen for his support and guidance.

Appendix A. *k*-means clustering algorithm to handle store spatial constraints

Notation

$_nD_k$ (n by k) matrix of distances between path n and cluster centroid k.

$_nd_k$ (n by k) matrix of distances between path n and cluster mean k.

$(CL_1, CL_2, \ldots CL_n)$ vector of cluster assignments; i.e. if $CL_{103}=12$, path 103 is assigned to cluster 12.

$(C_1, C_2, \ldots C_k)$ vector of cluster centroids (indexed to actual paths); i.e. if $C_{10}=1034$, the cluster centroid for cluster 10 is path 1034.

$_kM_{100}$ (k by 100) matrix of cluster means (mean position of cluster k at each of the 100 percentile locations).

B_{it} Blink t from path i.

Initialization

1. C=Random draw of k numbers from discrete uniform $(1, n)$ without replacement
2. Calculate D
   $_iD_j$=distance of path i from cluster centroid $j=\Sigma_t (\text{dist}(B_{it}-C_{jt}))^2$
3. Calculate CL
   CL_i=cluster assignment for path $i=\min_j ({}_iD_j)$
4. Calculate M
   $_jM_t=\text{mean}_i (B_{it})$; over all i such that $CL_i=j$
5. Calculate d
   $_id_j$=distance of path i from cluster mean $j=\sum_t(\text{dist}(B_{it}-{}_jM_t))^2$
6. Calculate C
   $C_j=\min_i ({}_id_j)$
7. Calculate D

Optimization

For $i=1$ to n

1. Calculate CL_i

2. If new $CL_i \neq$ old CL_i, then
   2a. Calculate M (update cluster means for new and old cluster)
   2b. Calculate d
   2c. Calculate C
   2d. Calculate D
3. Go to Step 1
4. Continue until new $CL_i =$ old CL_i for all i.

Note: Algorithm produces a local minimum. To find global minimum, the algorithm should be run from several starting points.

References

Apparicio, P., Shearmur, R., Brochu, M., & Dussault, G. (2003). The measure of distance in a social science policy context: Advantages and costs of using network distances in eight Canadian metropolitan areas. *Journal of Geographic and Information Decision Analysis*, *7*(2), 105–131.

Batty, M. (2003). Agent-based pedestrian modeling. In P. Longley, & M. Batty (Eds.), *Advanced spatial analysis: The CASA book of GIS* (pp. 81–105). Redlands, CA: ESRI Press.

Bradlow, E. T. (2002). Exploring repeated measures data sets for key features using principal components analysis. *International Journal of Research in Marketing*, *19*, 167–179.

Farley, J. U., & Ring, L. W. (1966). A stochastic model of supermarket traffic flow. *Operations Research*, *14*(4), 555–567.

Jones, M. C., & Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *American Statistician*, *46*(2), 140–145.

Kaufmann, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. Wiley.

Krzanowsky, W. J., & Lai, Y. T. (1985). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, *44*, 23–34.

Lynch, K., & Rivkin, M. (1959). A walk around the block. *Landscape*, *8*, 24–34.

Mackay, D. B., & Olshavsky, R. W. (1975). Cognitive maps of retail locations: An investigation of some basic issues. *The Journal of Consumer Research*, *2*, 197–205.

Park, C. W., Iyer, E. S., & Smith, D. C. (1989). The effects of situational factors on in-store grocery shopping behavior. *The Journal of Consumer Research*, *15*(4), 422–433.

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

Sorensen, H. (2005). In P. S. Fader (Ed.), *Management implications*. Philadelphia.

Sugar, C. (1998). *Techniques for clustering and classification with applications to medical problems*. Unpublished PhD dissertation, Stanford University, Palo Alto.

Underhill, P. (1999). *Why we buy*. New York: Simon and Schuster.

Winkel, G. H., & Sasanoff, R. (1966). An approach to an objective analysis of behavior in architectural space. *Architecture/development series*, *vol. 5*.