An Exploratory Look at Supermarket Shopping Paths

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

July 2004

Abstract

We present exploratory analyses of an extraordinary new dataset that reveals the path taken by individual shoppers around an actual grocery store, as provided by RFID (radio frequency identification) tags located on their shopping carts. In order to explore the spatial patterns observed in the data, we present a multivariate clustering algorithm not yet seen in the marketing literature that is able to handle data sets with unique (and numerous) spatial constraints. The resulting output conveniently summarizes each cluster with an observed shopping path, thus enabling us to familiarize ourselves with the "canonical trip types" that are typical of grocery store travel.

*Jeffrey S. Larson is a doctoral student, Eric T. Bradlow is an Associate Professor of Marketing and Statistics, and Peter S. Fader is Frances and Pei-Yuan Chia Professor of Marketing at The Wharton School of the University of Pennsylvania. All correspondence on this manuscript should be sent to Jeffrey Larson, larsonj@wharton.upenn.edu, 215-898-2268, suite 700 JMHH, 3730 Walnut Street, Phila. PA 19104. The authors wish to thank Sorensen Associates for providing the data and, in particular, Herb Sorensen for his support and guidance.

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 trip 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 (i.e., 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 Figure 1, now allows us to examine these and other important behavioral questions.

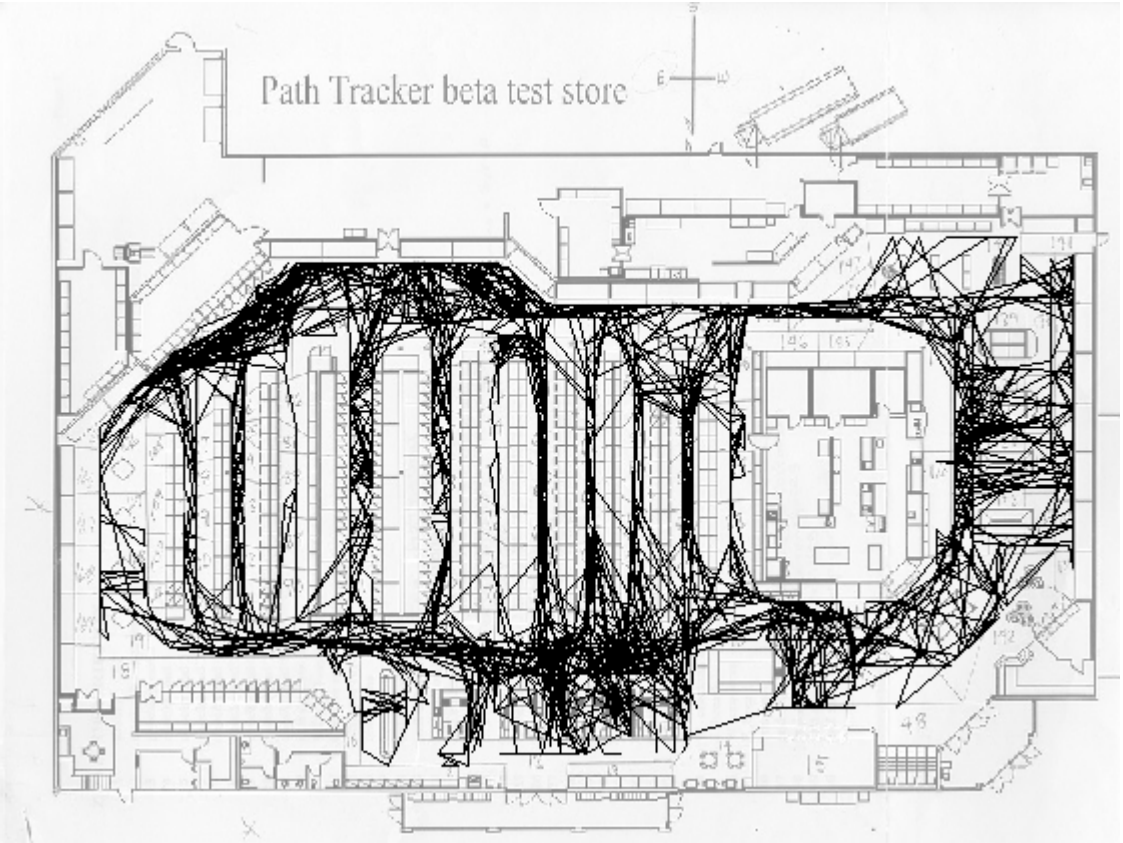


FIGURE 1: Layout of PathTracker® data from 20 customers

No, Figure 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 five seconds that is received by receptors installed at various locations throughout the store. The arrival latencies of the signals at the receptor locations are used to triangulate the position of the grocery cart. Thus, for every shopping trip, data are recorded regarding the cart's two-dimensional location coordinates, (x_{it}, y_{it}) for shopper i at his t^{th} observation (hereafter referred to as "blinks"), at five 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 trip 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 trip. 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

[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.

move through a supermarket. Similar research ideas, summarizing large sets of "behavioral" curves as in Figure 1 have been explored using principal components analysis methods (Bradlow 2002, Jones and 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 trip. 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. Finally, we conclude with directions for future research.

Overcoming Data Obstacles

The travel portion of our data consists of a "right-ragged" array of location coordinates, where every row is a shopping trip, 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 trips ranging in length from 25 blinks for a two-minute trip, to 1500 blinks for a two-hour trip. The mean trip consists of 205 blinks (just over 16 minutes), and the median has 166 blinks (just over 13 minutes)[2]. The trip 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 trip comparisons (differing lengths of trips)

(3) Spatial constraints (aisle layout and other physical obstructions in the store).

We next describe our solutions to these issues.

(1) Size of the dataset

With over 27,000 trips, and as many as 1500 pairs of coordinates (blinks) per trip, memory limitations for implementing a clustering algorithm posed significant problems. To make the analysis feasible, we drew a systematic sample of 9000 trips, drawing every 3rd trip

[2] When Sorensen Associates initially assembled the dataset, they observed a number of very long trips -- up to six hours in duration. These trips 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 trips lasting over two hours.

from a random starting point. Some of these trips were deleted due to data problems (the transponder stopped working for longer than a minute), leaving us with 8751 trips. With an additional three-way split of the 8751 trips 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 problem.

(2) Ragged array trip comparisons

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

(3) Spatial Constraints

With each trip standardized as a 100-blink (percentile) trip, 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 trips within a cluster, computed as a pointwise average coordinate-by-

coordinate (as in k-means algorithms) will not be a feasible store trip, 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, described next, that is able to handle the multitude of spatial constraints (Kaufmann and Rousseeuw 1990).

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. Whereas typical k-means clustering begins with a random clustering of all observations, 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 trips to serve as the initial centers. These observations (trips) are called medoids. Each of the remaining observations (trips) 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, 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

[3] In the 100-blink trips, the first and last blinks match the first and last blinks of the actual trips. So, technically, the

shopper paths, making the store densely covered by actual trips. 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.

A graphical illustration, below in Figure 2, explicates the procedure. Four actual standardized trips are 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 trips (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 trips, the bold trip 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].

---

second blink represents the trip 1/99th of the way through the trip; the third blink is the location 2/99th's of the way through, etc. We simplified the explanation above for clarity.

[4] At times the observed path appears to clip the corner of an infeasible area or appears to go through an aisle shelf. Since trip 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".

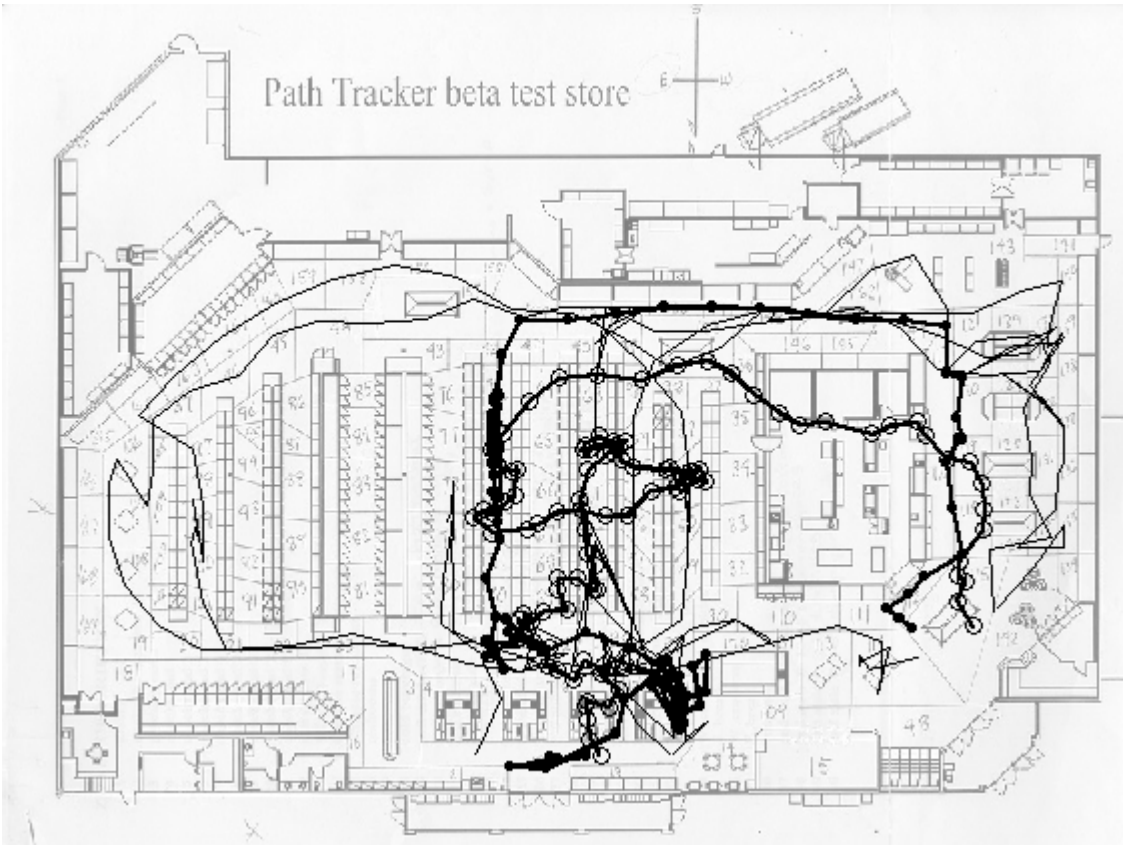


FIGURE 2: Illustration of our modified k-means clustering algorithm subject to spatial constraints.

The previously described algorithm has several desired properties: (1) it clusters shoppers according to similarity of travel behavior, and (2) yields a feasible trip (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 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 Figure 1. To show that these methods provide valuable information beyond what other possible techniques could provide, we present an alternative summary technique to which we can compare results.

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 trip paths. We accomplish this by taking each store trip and summarizing it by the amount of the trip 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 below in Figure 3.

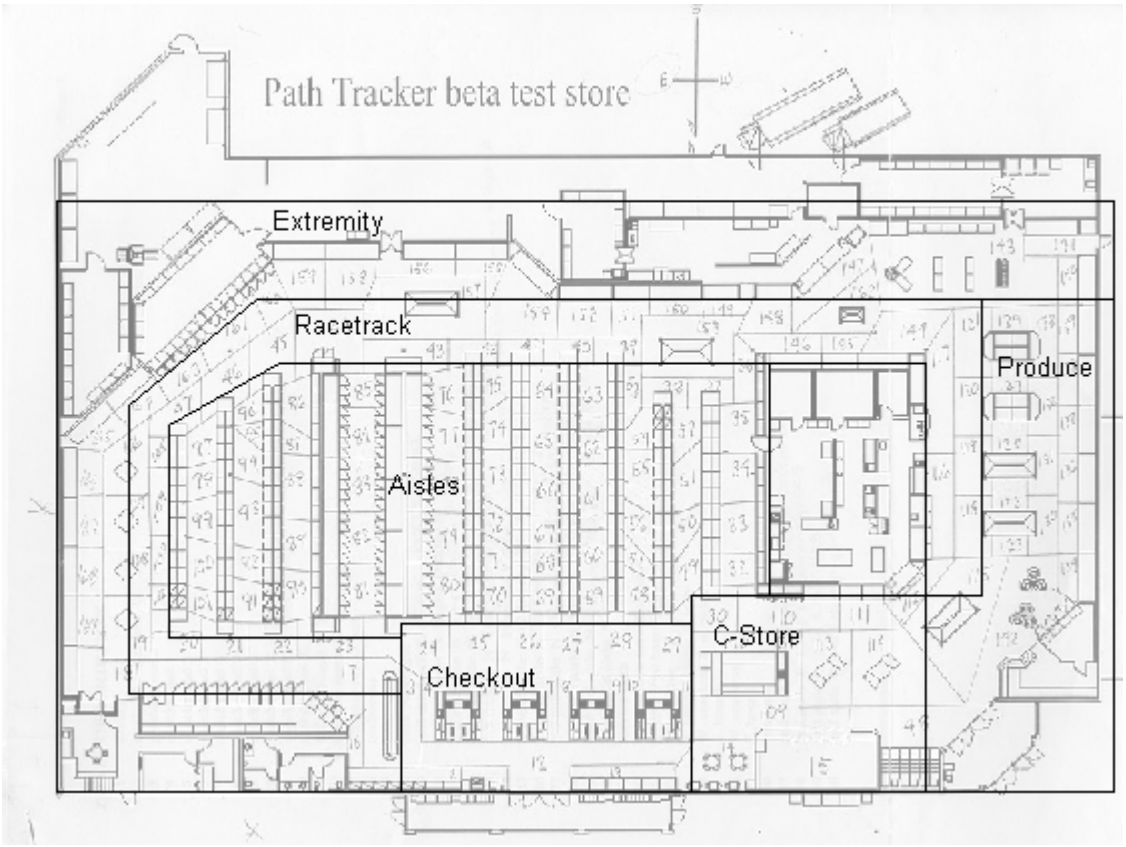


FIGURE 3: Illustrations of trip profiles

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 vs. 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. 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 trip. 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 trip, we record the percentage of the trip that occurred in each of these six mutually exclusive and exhaustive trips areas. Percentage of the trip was recorded as opposed to number of blinks as it allowed us to normalize out trip length from trip pattern. Figure 4 shows the distribution of each variable across our sample of 8751 trips. 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 trip.

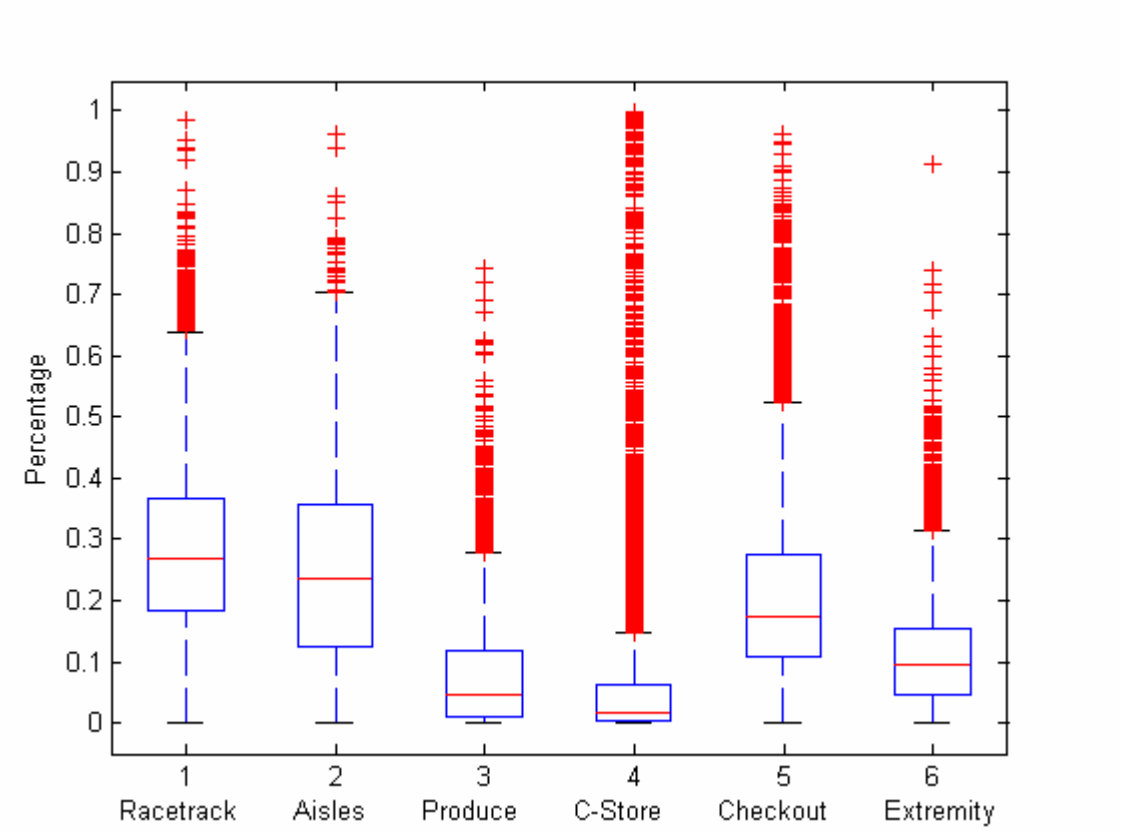


Figure 4: Percentage travel in each zone as distributed across the population.

A logical method to proceed based on these mutually exclusive variable profiles would be to use them in 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 (based on actual trips):

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

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

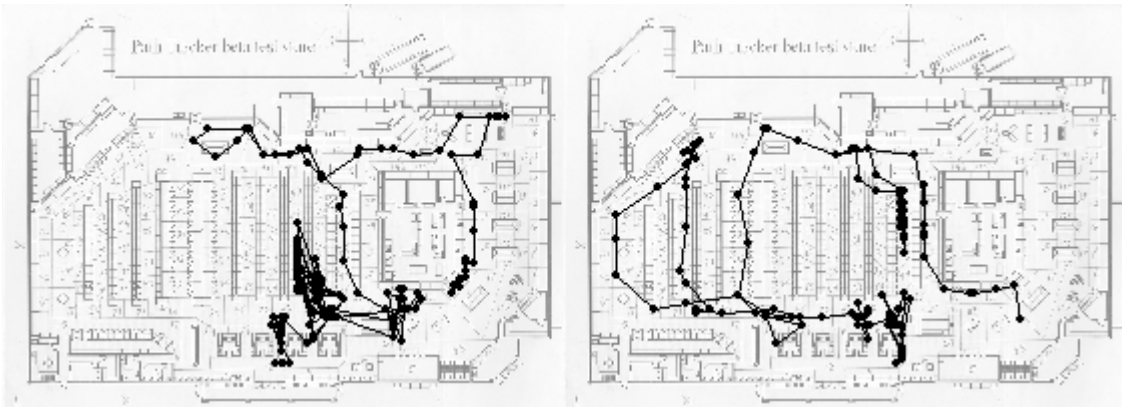


Figure 5: Trip A and Trip B

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 trips would no longer look the same. But the problem would still persist if, for instance, two trips 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 trips 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 trip lengths. This information can still be incorporated into our analysis by means we describe in the next section.

Clustering Results

The time dimension of shopping has been notably absent from most of the previous discussion. While we make all trips comparable to each other by creating standardized "percentile paths", we expect to see vastly different behaviors in a 5-minute versus a 30-minute

trip. For this reason, we chose to split the set of 8571 trips into three equally sized groups. The resulting splits yield a "low" group of 2917 trips with travel times ranging from 2 to 10 minutes, a "middle" group of 2916 trips ranging from 10 to 17 minutes, and a "high" group of 2918 trips lasting from 17 minutes to nearly two hours.

The splits by time have an intuitive appeal in that a longer "stock-up" trip is likely to be quite different from an intermediate "fill-in" trip. Similarly, trips under 10 minutes 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 trip that we had previously lost.

We now undertake a separate analysis for each of these three groups.

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. We ran the algorithm for several different numbers of clusters to find the "optimal" number. 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 is shown below in Figure 5.

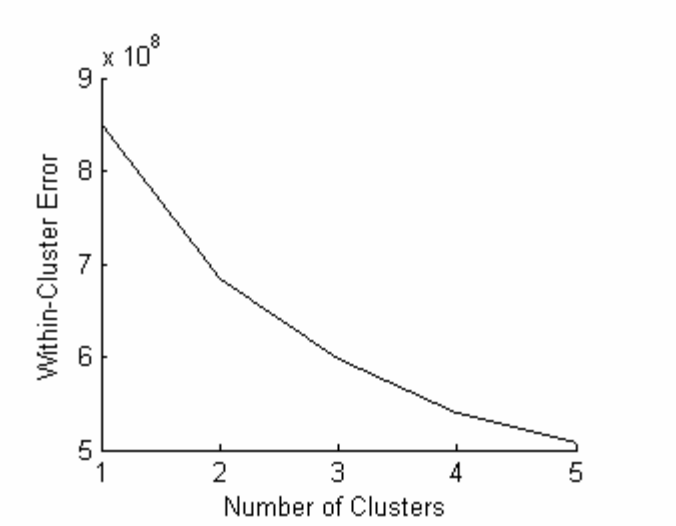


FIGURE 5: Scree plot for low group

We observe a major kink at 2 clusters and so we present the results of our algorithm for this particular configuration. Findings for alternative cluster solutions are available upon request. We present the resulting cluster centroids in Figure 6 below. The N associated with each centroid shows the number of trips assigned to that cluster.

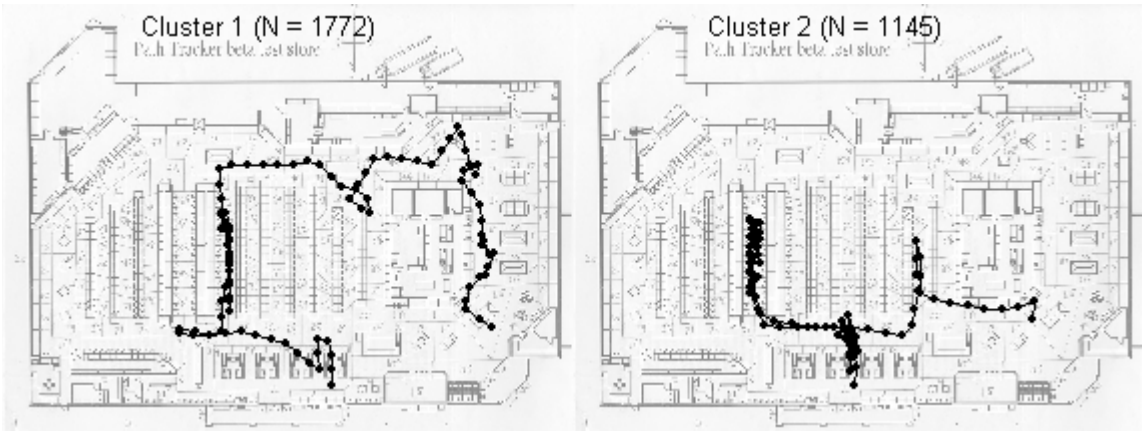
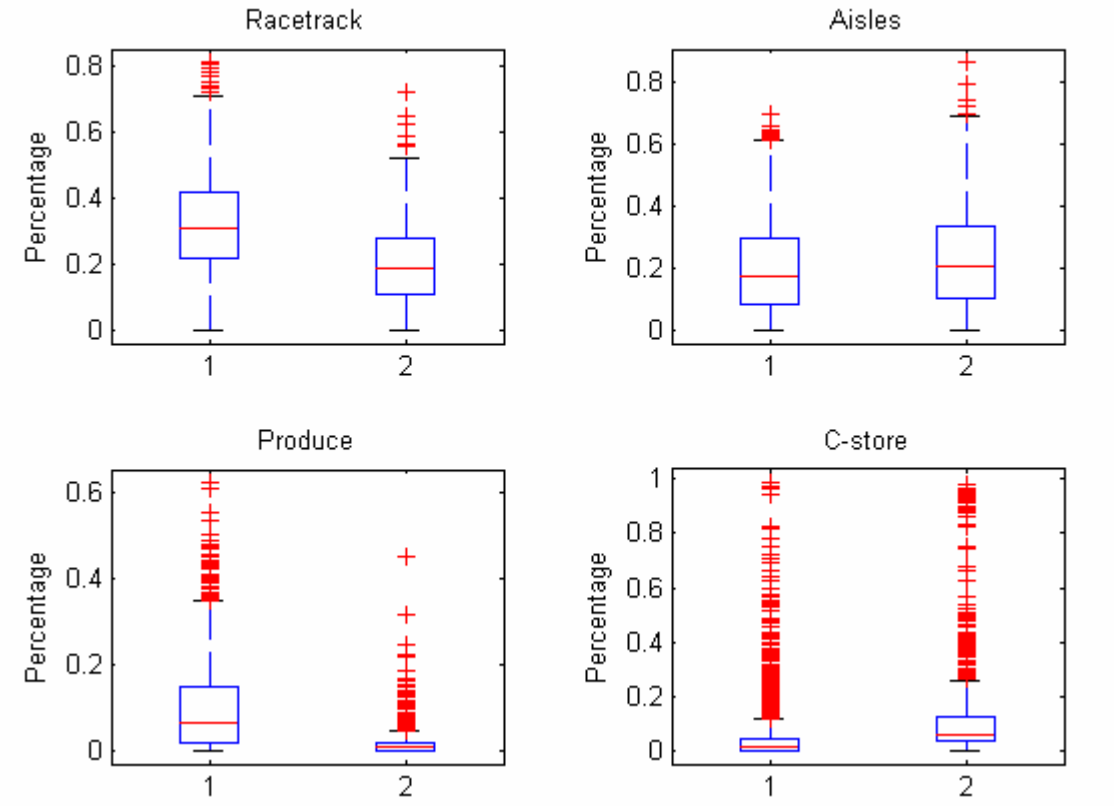


Figure 6: Low Group Medoids

For trips under ten minutes, 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 trips, whether or not they actually shop in the produce area, follow this default

path. However, a significant portion of short trips break the default pattern. This is likely due to time pressures—shoppers making shorter trips 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.

It is informative to examine how the results from our k-medoids clustering compares with the 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 trip. 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 trips within each cluster is represented in Figure 7.



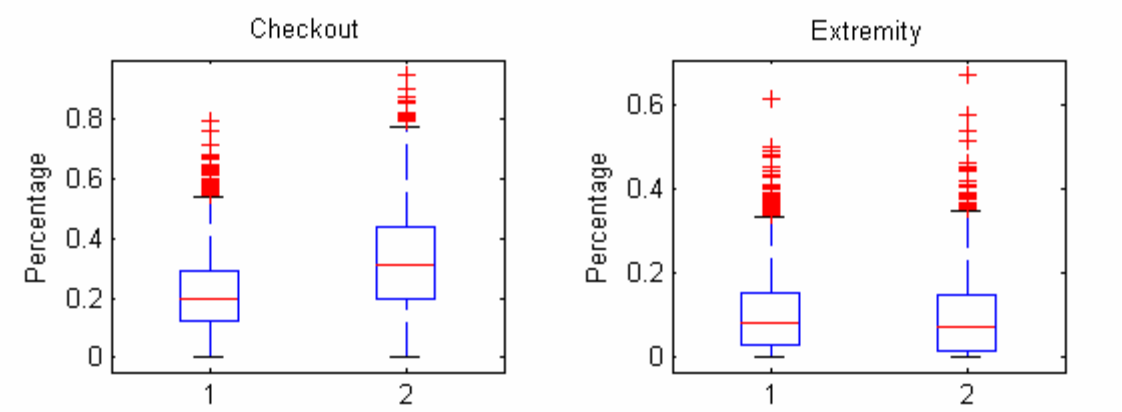


Figure 7: Differences in profiles between clusters for low group

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 trip length across these two clusters.

Medium Group

Since we expect more divergence in behavior with longer trips, we also expect to find more "canonical trip types". In other words, a cluster solution with more than two clusters will likely be appropriate for the medium length trips. We chose the four cluster solution for the medium group, as it adequately balances parsimony and fit. The scree plot appears below.

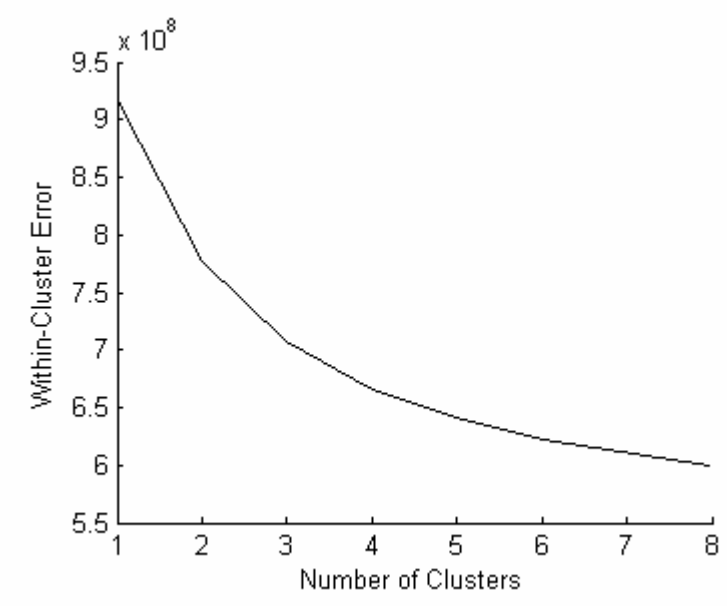


Figure 8: Scree Plot for Medium Group

Figure 9 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 trips at first glance appear to be more homogeneous than the two cluster medoids from the low group, as they all follow the 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.

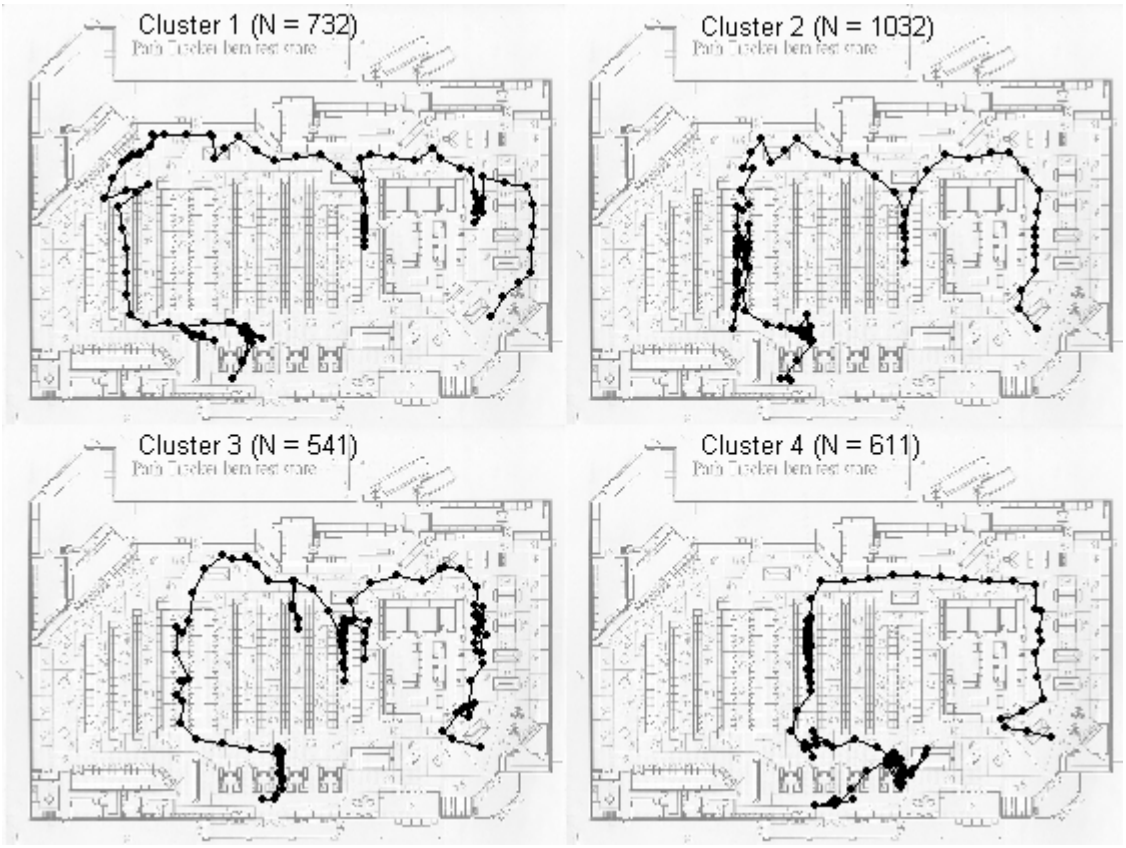


Figure 9: Medium cluster medoids

Again, the profiling variables are notably different across clusters, as seen by the boxplots in Figure 10. 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 trips, 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.

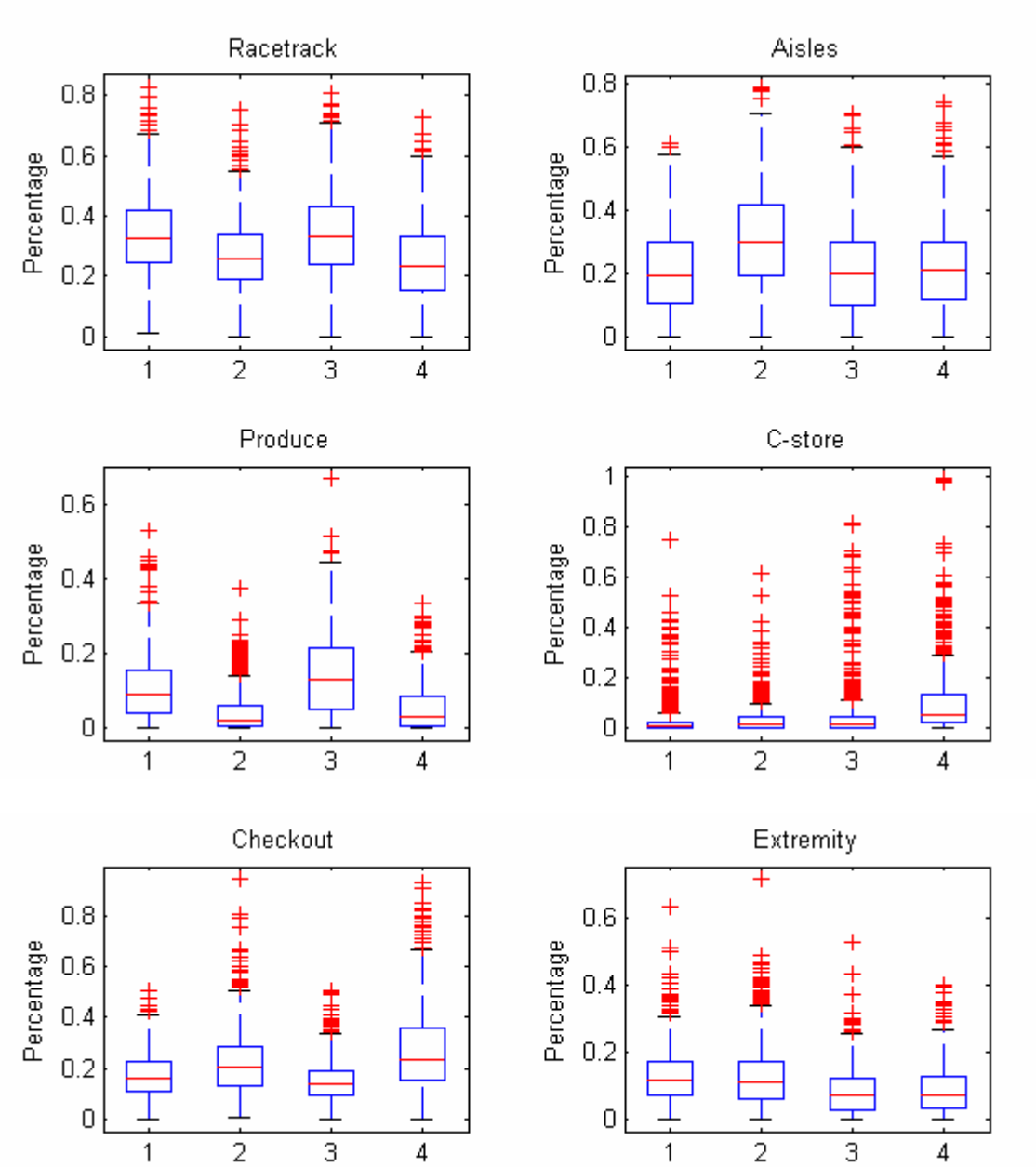


Figure 10: Differences in profiles across clusters for medium group

High Group

As the most variable group in trip length, we also expect to see a high degree of variability in the observed shopping patterns. After observing the scree plot and the solutions at several number of clusters, we chose eight clusters as a suitable balance of parsimony and fit. The scree plot appears below in Figure 11.

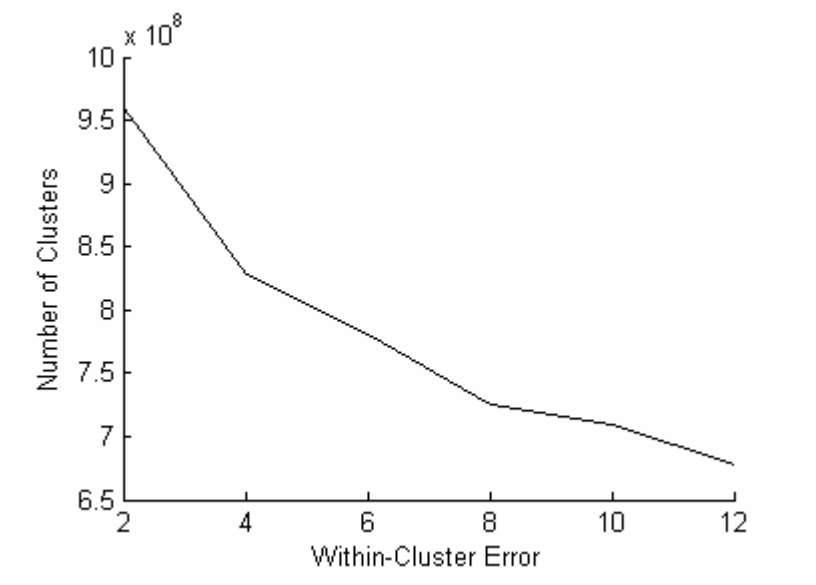


Figure 11: Scree plot for high group

Figure 12 presents the medoids for the eight clusters. As expected, we see a high degree of variability in trip type. Cluster 3 is the most unique trip. 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 minutes. Medoid 4 is also interesting: despite the absence of any self-imposed time constraint (as surmised by its length), these trips choose to break the default start path to go directly to the desired items in the aisles. Another trip dominated by aisle travel is trip 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. 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 trips, one of the most important distinguishing dimensions is not whether the trip 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 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 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 trips, 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 trip 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 trip, or they forget important purchases until later; perhaps product choices are themselves stochastic, influenced by store atmosphere. A better understanding the shopping process could lead to important discoveries for retailing.

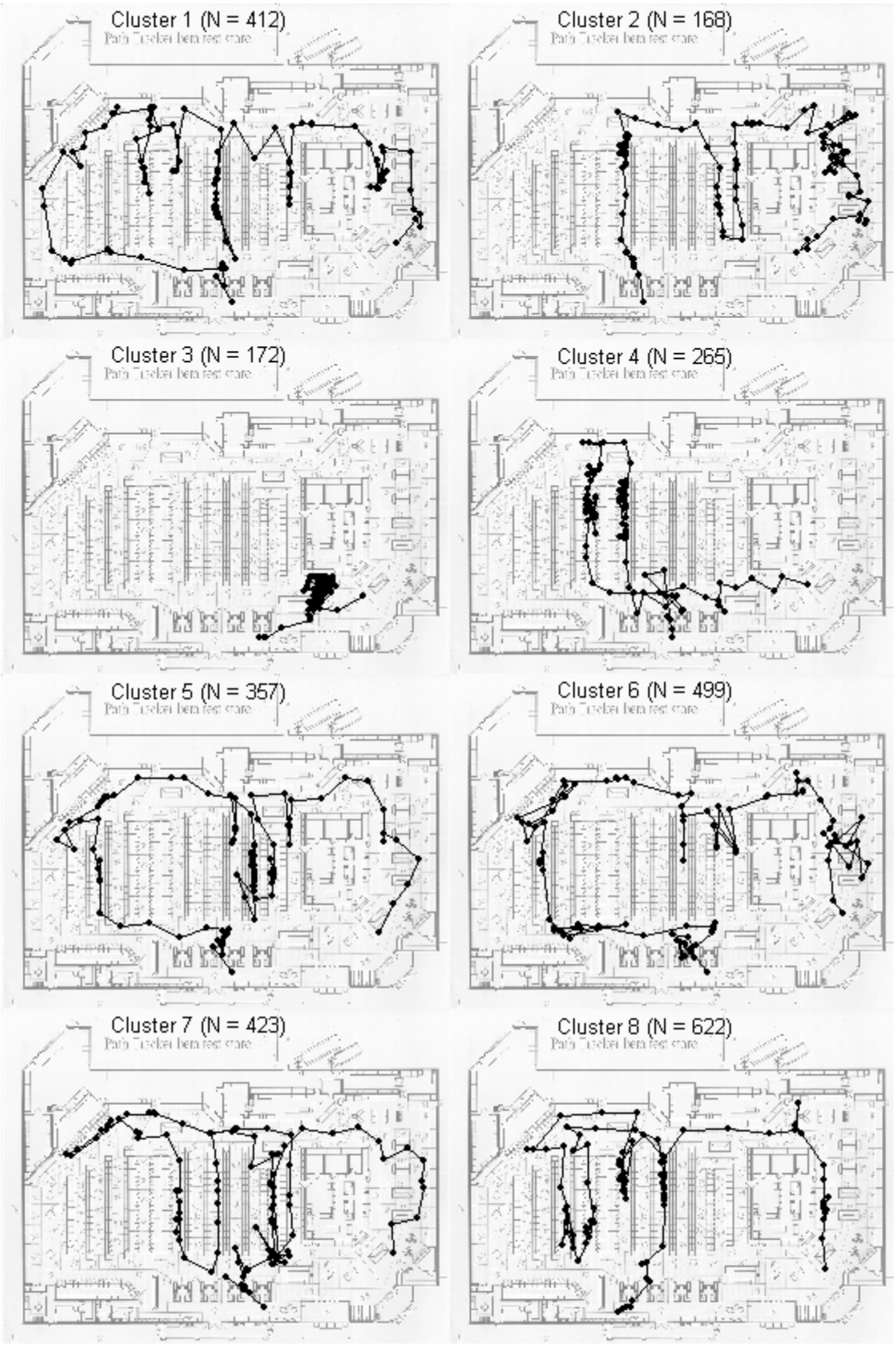


Figure 12: High group medoids

Again, we present the variation of the profiling variables across clusters in Figure 13. As expected, cluster 3 is high on 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 trips 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. 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 displays, again supporting the value of the k-medoid clustering method as opposed to a clustering algorithm based on summaries as described earlier.

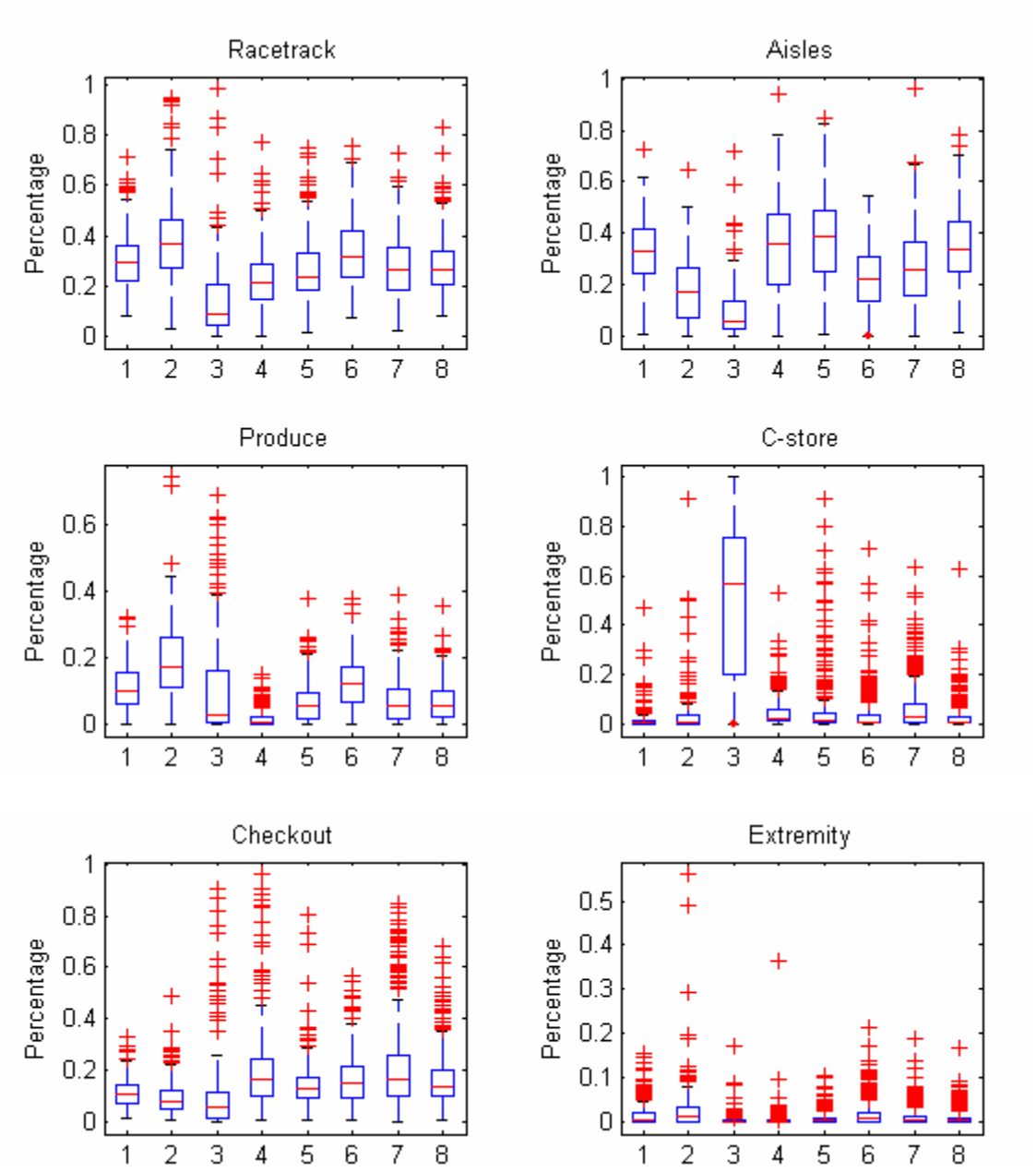


Figure 13: Differences in profiles across high clusters

Ties with Past and Future Research

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 et al (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.

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 complimentarity and substitutability of goods in ways that a more traditional "market basket" analysis can not 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 trip 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.

But 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.

References

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.

Jones, M.C., & Rice J.A. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *American Statistician,* 46(2), 140-145.

Farley, John U. and L. Winston Ring (1966), "A Stochastic Model of Supermarket Traffic Flow," *Operations Research,* 14(4), 555-567.

Kaufmann, L. and P.J. Rousseeuw. *Finding Groups in Data*. Wiley, 1990.

Mackay, David B. and Olshavsky, Richard W. (1975), "Cognitive Maps of Retail Locations: An Investigation of Some Basic Issues," *The Journal of Consumer Research,* 2, 197-205.

Park, C.Whan, Easwar S. Iyer and Daniel C. Smith (1989), "The Effects of Situational Factors on In-Store Grocery Shopping Behavior," *The Journal of Consumer Research,* 15 (4), 422-433.

Sorensen, Herb (2003), "The Science of Shopping," *Marketing Research*, 15 (3), 30-35.

Underhill, Paco (1999), *Why We Buy.* New York: Simon and Schuster.

Appendix

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 trips); 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 trip *i*

Initialization

1. C = Random draw of k numbers from Discrete Uniform (1, n) without replacement

2. Calculate D

${}_iD_j$ = distance of trip *i* from cluster centroid *j* = $\Sigma_t (\text{dist}(B_{it} - C_{jt}))^2$

3. Calculate CL

CL_i = cluster assignment for trip *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 trip *i* from cluster mean j = $\Sigma_t(\text{dist}(B_{it} - {}_jM_t))^2$

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.