



University of Pennsylvania
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

2009


The Traveling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP-Optimality

Sam K. Hui
University of Pennsylvania

Peter S. Fader
University of Pennsylvania

Eric T. Bradlow
University of Pennsylvania

Follow this and additional works at: http://repository.upenn.edu/statistics_papers

 Part of the [Business Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Hui, S. K., Fader, P. S., & Bradlow, E. T. (2009). The Traveling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP-Optimality. *Marketing Science*, 28 (3), 566-572. <http://dx.doi.org/10.1287/mksc.1080.0402>

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/statistics_papers/10
For more information, please contact repository@pobox.upenn.edu.

The Traveling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP-Optimality

Abstract

We examine grocery shopping paths using the “Traveling Salesman Problem” (TSP) as a normative frame of reference. We define the “TSP-path” for each shopper as the shortest path that connects all of his purchases. We then decompose the length of each observed path into three components: the length of the TSP-path, the additional distance due to order deviation (i.e., not following the TSP-order of category purchases), and the additional distance due to travel deviation (i.e., not following the shortest point-to-point route). We explore the relationship between these deviations and different aspects of in-store shopping/purchase behavior. Among other things, our results suggest that (1) a large proportion of trip length is due to travel deviation; (2) paths that deviate substantially from the TSP solution are associated with larger shopping baskets; (3) order deviation is strongly associated with purchase behavior, while travel deviation is not; and (4) shoppers with paths closer to the TSP solution tend to buy more from frequently purchased product categories.

Keywords

Path models, Traveling Salesman Problem, Grocery Retailing

Disciplines

Business | Statistics and Probability

**The Traveling Salesman Goes Shopping:
The Systematic Deviations of Grocery Paths from TSP-Optimality**

Sam K. Hui

Peter S. Fader

Eric T. Bradlow*

January 2008

* Sam K. Hui is a doctoral candidate in Marketing, Peter S. Fader is The Frances and Pei-Yuan Chia Professor of Marketing, and Eric T. Bradlow is The K. P. Chao Professor, Professor of Marketing, Statistics, and Education, and Academic Director of The Wharton Small Business Development Center at the Wharton School of the University of Pennsylvania. Corresponding author: Sam K. Hui (kchui@wharton.upenn.edu). The authors are extremely grateful for the data provided by Sorensen Associates and, in particular, the valuable input and guidance from Herb Sorensen.

The Traveling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP-Optimality

Abstract

We examine grocery shopping paths using the “Traveling Salesman Problem” (TSP) as a normative frame of reference. We define the “TSP-path” for each shopper as the shortest path that connects all of his purchases. We then decompose the length of each observed path into three components: the length of the TSP-path, the additional distance due to *order deviation* (i.e., not following the TSP-order of category purchases), and the additional distance due to *travel deviation* (i.e., not following the shortest point-to-point route). We explore the relationship between these deviations and different aspects of in-store shopping/purchase behavior. Among other things, our results suggest that (1) a large proportion of trip length is due to travel deviation; (2) paths that deviate substantially from the TSP solution are associated with larger shopping baskets; (3) order deviation is strongly associated with purchase behavior, while travel deviation is not; and (4) shoppers with paths closer to the TSP solution tend to buy more from frequently purchased product categories.

1. Introduction

With the advent of new technologies, e.g., Radio Frequency Identification (RFID), researchers are equipped with better data to explore in-store shopping behavior, adding value to the ubiquitous scanner data analyses that have been pervasive over the past 25 years (Guadagni and Little 1983). For example, Burke (1996) studied consumers' grocery shopping patterns using a virtual (simulated) store; Sorensen (2003) tabulated purchase and time-of-stay statistics at different locations within an actual grocery store; Larson et al. (2005) categorized grocery paths using a clustering algorithm, and identified 14 different "canonical paths".

In contrast to these purely descriptive studies, we instead compare a large number of shopping paths and purchase baskets to the normative benchmark provided by the Traveling Salesman Problem (TSP). In the classic TSP, the salesman has to visit a number of cities before returning to his original starting point. The objective is to choose his order of visitation in order to minimize his travel distance while visiting all the required cities. By analogy, in the grocery setting, we define the *TSP-path* as the shortest route that connects the entrance, all the products that a shopper purchased, and the checkout counter.

We compare each shopper's observed behavior with his TSP-path and document the systematic departures that emerge. We focus on two types of deviations: First, the shopper may not follow the exact shopping order suggested by the TSP-path. We define this type of departure as *order deviation*. Second, given the actual order of purchases the shopper has chosen (TSP-optimal or otherwise), he may not follow the shortest point-to-point route. We define this source of departure as *travel deviation*. Thus, every observed path is decomposed into three parts: the travel distance of the TSP solution, the additional distance due to order deviation, and the additional distance due to travel deviation.

This decomposition leads to a number of empirical questions: How similar is each observed grocery path to its corresponding TSP solution? How will the contribution of each deviation vary across trips? Will one component dominate the others? Taking things a step further, we study the relationship between order/travel deviation and other more “classic” trip characteristics such as the number of items purchased on each trip, the number of aisles traversed, and total time in store. For instance, will longer trips be associated with higher or lower order deviations? On the one hand, longer trips may be more organized; yet on the other hand, there are more opportunities for choosing an order of visitation that is different from the TSP solution. Another interesting issue is whether/how category purchase incidence is in any way related to order/travel deviations. For example, what is the relationship between order/travel deviations and the number of items purchased? Do shoppers who travel routes closer to the TSP-path tend to shop disproportionately in certain categories? Our goal is to answer the above questions empirically in order to better understand shopping patterns as a whole as well as the nature of the deviations that we document here.

Our research is in the same spirit as other papers in marketing/economics that have compared observed behavior to a well-established normative paradigm. For example, Camerer et al. (2004) analyzed behavior in economic games, comparing it with the normative prescription of the Nash equilibrium. Likewise, Meyer and Assuncao (1990) analyzed consumers’ stockpiling strategies and documented the contexts in which consumers tend to underbuy or overbuy compared to their optimal solutions, calculated from sequential decision theory. In both cases, researchers took a logical optimality paradigm and carefully described how actual behavior departs from it. Other papers with similar goals include, Houser et al. (2004), MacGregor et al. (1999, 2000), Polivova (1974), Seale and Rapoport (2000) and Vickers et al. (2001).

The remainder of this paper is organized as follows. Section 2 discusses our analytical framework and how each observed path is decomposed into its TSP-path, order deviation, and travel deviation. Section 3 describes the data, and Section 4 discusses our empirical results. Finally, Section 5 concludes with a summary of our findings.

2. Analytic Framework

We define the TSP-path as the shortest path (in terms of total travel distance) that starts at the entrance, connects all of the observed purchases, and ends at checkout.¹ We obtain the TSP-path using two algorithms commonly employed to solve the TSP: exhaustive search (Lawler 1985) and simulated annealing (Goffe 1994), which are outlined in Technical Appendix A.

Once the TSP-path is derived, we carefully examine differences between it and its actual counterpart. Two types of aforementioned deviations are considered: order deviation and travel deviation. We illustrate these concepts using Figure 1.

[Insert Figure 1 about here]

In this simple example, the TSP-optimal order is $B \rightarrow A \rightarrow C$, with a total travel distance of $3 + 3 = 6$ units. The observed shopping order, $B \rightarrow C \rightarrow A$, results in a longer travel distance ($5 + 3 = 8$), assuming that the shortest point-to-point paths are taken when traveling between two locations. We define the difference between the travel distance of the optimal order and the observed order as order deviation, which in this case is $8 - 6 = 2$ units.

To measure travel deviation, we take the observed order ($B \rightarrow C \rightarrow A$) as given and look for excessively long routes when traveling from one location to the next. In Figure 1, although the shortest path from B to C requires 5 units of travel distance, the shopper took a more indirect

¹ Alternatively, one could consider the path that minimizes shopping time instead of distance. However, as discussed in Section 3, our data are not from a longitudinal panel, so it is impossible to tease apart individual speed differences among shoppers. Thus we cannot make any normative assessments about shopping time, per se.

path that required 7 units. Likewise, the shopper incurred $4 - 3 = 1$ units of travel deviation when traveling from C to A. Thus, there is a total of 3 units of travel deviation.

Thus, each observed path can be decomposed into three components: the TSP-path, order deviation, and travel deviation. Adding these three components together equals the total distance traveled. In our example above, the decomposition can be described by the following equation²:

$$\begin{array}{rcccccccc} \text{Observed Path} & = & \text{TSP-Path} & + & \text{Order Deviation} & + & \text{Travel Deviation} & & \\ 11 & = & 6 & + & 2 & + & 3 & & [1] \end{array}$$

3. Data

We apply our analytic framework to a dataset that contains consumers' shopping path data together with their purchases from a large supermarket in the eastern United States. We obtained our data from Sorensen Associates, an in-store research company that tracks shoppers' movement using its proprietary PathTracker[®] system based on RFID technology (Sorensen 2003). A small RFID tag is affixed under each shopping cart, and emits a uniquely coded signal every five seconds; this signal is then picked up by an array of antennae located throughout the store which can pinpoint the precise location of the shopping cart³ over time.

Our data preparation procedures, described more fully in Hui et al. (2007) and outlined in Technical Appendix B, yielded a total of 993 shopping paths and their corresponding purchase records. The procedure described in Hui et al. (2007) allows us to discretize the grocery store into a graph with 96 nodes, hence making each cart movement a selection among a finite set of edges. The division of the grocery store into zones is shown graphically in Figure 2 and Figure 3.

[Insert Figures 2 and 3 about here]

² To the best of our knowledge, this is the first time this decomposition has been derived and represents a contribution of this research that may aid researchers more broadly.

³ We recognize that the shopper's cart is a noisy proxy for his/her exact location; yet, it is a significant advance over having no tracking data. As per Sorensen (2003), more precise technologies are likely to be available soon.

For each path, we extract a number of key summary statistics (shown in Table 1) on shoppers' movement and purchases. These statistics include the total number of product categories purchased (out of a total of 116), total path distance traveled in the store, the number of unique zones visited (out of the aforementioned 96 zones), total time (in minutes) spent in the store, and the number of unique aisles that each shopper entered and traversed. Table 2 lists the top ten categories purchased based on the proportion of shoppers who made at least one purchase in each category. In Section 4, we relate these measures to our TSP-decomposition results.

[Insert Tables 1 and 2 about here]

4. Results

This section presents our empirical findings, which are summarized in Table 3. Section 4.1 describes the decomposition of paths into its TSP-path, order deviation, and travel deviation. Section 4.2 studies the relationship between order/travel deviations, basket size, and shopping path. Section 4.3 looks at the relationship between order/travel deviations and product categories purchased.

[Insert Table 3 about here]

4.1 TSP decomposition

The fractions of observed travel distance associated with the TSP-path, order deviation, and travel deviation, for each of the 993 paths, are shown graphically in the triangle plot in Figure 4, and the associated summary statistics are contained in Table 4. The triangle plot allows us to easily visualize the relationship among three variables that sum to 1. This figure yields several immediate insights. First, there is a great deal of variability in the decomposition of shoppers' paths, relative to the TSP-path, across the 993 trips. The percentage of total travel distance due to the TSP-path ranges from a low of approximately 5% to a high around 95%, with

an average of about 28%. In contrast, the extent of order deviation is quite limited – never exceeding 20%. This suggests that shoppers, in general, choose an order for their purchases that is fairly close or the same as the order suggested by the TSP solution.

Most of the trips lie in the lower right corner of Figure 4, indicating that travel deviation accounts for a large portion of the travel distance for the majority of grocery trips. So while the order of purchases is close to that of the TSP-path, shoppers spend a large portion of their in-store trip not following the shortest point-to-point routes. One potential reason (among others) for these large deviations is that shoppers may deliberately plan to visit some product categories to see whether promotions are available, but may not necessarily purchase from those categories. We investigate this issue through a sensitivity analysis in the Appendix.

[Insert Table 4 and Figure 4 about here]

4.2 Relationship between deviations, shopping basket, and trip characteristic

To explore the relationship between order/travel deviations and the characteristics of shopping paths mentioned earlier, we divide the 993 trips into four groups based on a median split along each deviation dimension. The summary statistics for each group are shown in Table 5, along with relevant visit and purchase characteristics, as described in aggregate in Table 1.

The first, and most obvious, contrast is between group 1 (low on both deviations) versus group 4 (high on both). It should come as no surprise that shoppers who exhibit the greatest deviations from the TSP solution tend to visit more zones, which means entering (and traversing) more aisles. It is not as obvious, a priori, that these shoppers will also buy more products, but the difference in basket size is large and highly significant ($p < .001$). Furthermore, we also note that the total time in the store is larger for shoppers in group 4 as compared to group 1 ($p < .001$).

A more illuminating contrast is between the two intermediate groups. In comparing group 2 to group 3, we see that order deviation tends to be more influential than travel deviation

in generating long trips with more aisles visits/traverses and larger baskets of purchased items. But a closer look at these two groups reveals some interesting differences that reflect the impact of order vs. travel deviation. For instance, while the average basket size is over 50% greater for group 2 vs. group 3, the mean number of zones visited is barely 10% larger. The latter difference is still statistically significant ($p=.003$), but it is indicative of the notion that the shoppers who exhibit a lot of travel deviation are visiting an “excessive number” of zones relative to the number of items that they purchase.

When we aggregate the data in Table 5 to look at each of the deviation dimensions by itself, we see another trend involving basket size. Specifically, mean basket size is far smaller for the groups with low order deviation, i.e., groups 1&3 (mean = 5.1) compared to those with high order deviation, i.e., groups 2&4 (mean = 9.1, $p<.001$). This observation is consistent with MacGregor and Ormerod (1996), who found that people’s performance in TSP problems generally worsens (i.e., more order deviation in our context) when they are given more locations to visit. But when we aggregate along travel deviation (groups 1&2 vs. 3&4) we see no difference in basket size (means of 7.0 and 7.2, respectively, $p=.35$). Thus, while travel deviation accounts for a large portion of most trips, order deviation has a much stronger association with purchasing behavior.

[Insert Table 5 about here]

4.3 Relationship between order/travel deviations and basket composition

Next, we study which product categories are most strongly associated with each of the four groups. To perform this analysis in a fair manner, we must normalize for the differences in basket size. To do so, we compute the number of purchases of each category for each group, and divide this by the total basket size of each group; these proportions are then compared across groups. Table 6 displays the product categories that are significantly (at $p<.05$) over-represented

in each group. We find that produce (e.g., fruits and vegetables), deli products (e.g., cheese/milk), and pre-packaged products tend to be associated with the groups that have low levels of order deviation; they seem to correspond to a well-organized shopping trip with a specific purpose, e.g., a shopper who brings a shopping list to shop for frequently purchased items. Along these lines, note that four of the ten most frequently purchased categories (in Table 2) are overrepresented in group 1. On the other hand, less frequently purchased household products are associated with higher order deviations. These purchases may correspond to a more impulsive shopping trip; on such trips, the shoppers may be shopping casually and choosing categories as they go along, without much concern about planning their trip. This results in a longer shopping path, and a seemingly haphazard path between purchases. Alternatively, it is also possible that these are “price shoppers” who are looking for promotions, an issue that we address using a sensitivity analysis in the Appendix. A further explanation is that the location of less frequently purchased items is more unknown, leading to greater travel deviation.

[Insert Table 6 about here]

5. Conclusion

In this research, we analyzed grocery shopping paths using the Traveling Salesman Problem (TSP) as a normative frame of reference. We decomposed the systematic deviations between the observed path and the corresponding solution of the TSP problem into two components: order deviation and travel deviation, and studied the relationship among these measures, purchase behavior, and shopping path characteristics.

Our results, as summarized in Table 3, offer a mixed answer to a question we raised in the introduction: “How similar are grocery trips to TSP-paths?” On the one hand, relatively few of them have a proportion of distance due to TSP-path that captures over 50% of their travel distance; but on the other hand, the degree of order deviation is very low in every case – never

exceeding 20% of the total distance. Thus shoppers tend to pick up their purchased products in an order close to that suggested by the TSP, but tend to depart from the shortest point-to-point path (i.e., travel deviations) as they move through the store.

Further, our analyses reveal consistent patterns about the interrelationship between order deviation and other characteristics of the trip. Specifically, trips with high order deviation tend to be longer trips with a greater number of product categories purchased and in-store time. Travel deviation is also associated with longer trips, but has no association with the overall basket size. We also find that trips with lower order deviation tend to be associated with frequently purchased categories.

These results have significant face validity; yet we believe that they were not obvious *a priori*. From a managerial standpoint, there is not a clear understanding of whether these deviations (and which type of deviation, to be more specific), are desirable or undesirable from a manager's perspective. On the one hand, deviations from the TSP solution give the shopper additional opportunities to see (and perhaps buy) more products – this might be a good outcome for the retailer. But on the other hand, part of the deviations may be due to confusing product placement and poor store layout – which could create dissatisfaction among shoppers. Our current dataset, by itself, cannot shed much light on this dichotomy, but it would not be hard to combine a similar analytic approach with attitudinal data covering aspects of shopper satisfaction to get a clear picture of the implications of deviation. We hope that our results will act as a springboard for future research in this area.

References

- Bucklin, Randolph E., and James M. Lattin (1991), "A Two-State Model of Purchase Incidence and Brand Choice," *Marketing Science*, 10 (Winter), 24-39.
- Burke, R. R. (1996), "Virtual Shopping: Breakthrough in Marketing Research," *Harvard Business Review*, Mar-Apr, 120-131.
- Camerer, C., F. T. Ho, and J. Chong (2004), "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 861-898.
- Goffe, W. L., G. D. Ferrier, and J. Rogers (1994), "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60, 65-99.
- Guadagni, P. M., and J. D.C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2(3), 203-238.
- Houser, D., M. Keane, and K. McCabe (2004), "Behavior in a Dynamic Decision Problem: An Analysis of Experimental Evidence Using a Bayesian Type Classification Algorithm," *Econometrica*, 72 (May), 781-822.
- Hui, S. K., E. T. Bradlow, and P. S. Fader (2007), "An Integrated Model of Shopping Paths and Purchase Behavior," *Working Paper*, available at <http://ssrn.com/abstract=960960>.
- Larson, J. S., E. T. Bradlow, and P. S. Fader (2005), "An Exploratory Look at Supermarket Shopping Paths," *International Journal of Research in Marketing*, 22 (4), 395-414.
- Lawler, E. L. (1985), *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley.
- MacGregor, J. N., and T. Ormerod (1996), "Human Performance on the Traveling Salesman Problem," *Perception and Psychophysics*, 58, 527-539.
- MacGregor, J.N., T.C. Ormerod, and E.P. Chronicle (1999), "Spatial and Contextual Factors in Human Performance on the Traveling Salesperson Problem," *Perception*, 28 (11), 1417-1427.
- MacGregor, J.N, T.C. Ormerod, and E.P. Chronicle (2000), "Model of Human Performance on the Traveling Salesperson Problem," *Memory and Cognition*, 28 (7), 1183-1190.
- Meyer, R. J., and J. Assuncao (1990), "The Optimality of Consumer Stockpiling Strategies," *Marketing Science*, 9(1), 18-41.
- Polivanova, N. I. (1974), "[Functional and Structural Aspects of the Visual Components of Intuition in Problem Solving," *Voprosy Psikhologii*, 4, 41-51.
- Seale, D. A., and A. Rapoport (2000), "Optimal Stopping Behavior with Relative Ranks: The Secretary Problem with Unknown Population Size," *Journal of Behavioral Decision Making*, 13(4), 391-411.
- Sorensen, H. (2003), "The Science of Shopping," *Marketing Research*, 15(3), 30-35.
- Vickers, D., M. Butavicius, M. Lee, and A. Medvedev (2001), "Human Performance on Visually Presented Traveling Salesman Problem," *Psychological Research*, 65, 34-45.

	Mean	S.D.	Min	Max
Number of product categories purchased	7.1	4.0	2.0	25.0
Total travel distance (in feet)	2513.0	1193.4	233.9	11234.4
Total in-store time (minutes)	49.8	25.2	7.9	238.3
Number of unique zones visited	49.9	14.2	5.0	83.0
Number of unique aisles entered	7.3	3.4	0.0	15.0
Number of unique aisles transversed	2.6	1.7	0.0	9.0

Table 1. Key summary statistics of the PathTracker[®] dataset.

Category	Proportion of shoppers who purchase the product category
Fruits	55.3%
Vegetables	52.2%
Butter/Cheese/Cream	40.0%
Carbonated Beverages	25.4%
Salty Snacks	24.4%
Cookies and Crackers	23.9%
Milk	23.7%
Ice cream	20.7%
Loaf Bread	20.5%
Cereal (Ready-to-eat)	18.1%

Table 2. Top 10 categories purchased.

TSP Decomposition (Section 4.1: Table 4 & Figure 5)
<ul style="list-style-type: none"> a. There is a great deal of variability in TSP-optimality across paths (5%-95%; average = 28%) b. Order deviation is small (always <20%; average = 3%) c. Travel deviation is large (average = 69%)
Relationship between deviations, basket size, and path characteristics (Section 4.2: Table 5)
<ul style="list-style-type: none"> a. Shoppers with paths that deviate more from TSP tends to (i) visit more zones, (ii) enter/traverse more aisles, (iii) spend longer time in store, and (iv) purchase more. b. Order deviation is strongly correlated to basket size. c. Travel deviation is uncorrelated to basket size.
Relationship between deviations and basket composition (Section 4.3: Table 6)
<ul style="list-style-type: none"> a. Paths closest to TSP (Group 1) tends to buy more frequently purchased categories. b. Paths with low order deviation (Group 1 & 3) tend to buy more produce, deli, and pre-packaged goods. c. Paths with high order deviation (Group 2 & 4) tend to buy more from categories that are less frequently purchased.

Table 3. Summary of our empirical findings.

	Mean	S.D.	Mean %	Min %	Max %
TSP-Path	612.0	189.4	27.5%	5.4%	94.7%
Order Deviation	89.5	107.9	3.1%	0.0%	17.1%
Travel Deviation	1811.5	1021.6	69.4%	5.3%	94.6%
Total Distance	2513.0	1193.4			

Table 4. Summary statistics from a TSP-decomposition analysis.

	Group 1	Group 2	Group 3	Group 4
Order Deviation (H/L)	L	H	L	H
Travel Deviation (H/L)	L	L	H	H
Number of shoppers	203	294	294	202
Mean % order deviation	0.4%	6.3%	0.6%	4.8%
Mean % travel deviation	59.5%	62.5%	78.6%	76.1%
Mean unique number of zones visited	38.2	52.1	48.9	59.7
Mean basket size (number of categories)	4.5	8.7	5.6	9.6
Mean unique number of aisles entered	4.7	7.7	7.1	9.6
Mean unique number of aisles traversed	1.4	2.8	2.5	3.7
Mean in-store time (in minutes)	28.8	47.9	50.5	72.2

Table 5. Summary statistics of clusters of shoppers (H: high; L: low).

Categories overrepresented in Group 1 (p<.05)					
Category	Group 1	Group 2	Group 3	Group 4	Overall
Butter-Cheese-Cream	7.4%	5.9%	5.1%	4.9%	5.6%
Milk	5.4%	3.5%	3.1%	2.4%	3.3%
Meat-Poultry-Seafood Manufactured Pre-pack	2.6%	1.3%	1.2%	1.7%	1.5%
Prepackaged Deli Prepared Lunch	0.4%	0.2%	0.0%	0.2%	0.2%
Fruits	10.9%	6.4%	9.7%	6.5%	7.8%
Vegetables	9.8%	5.8%	10.1%	6.0%	7.4%
Tobacco	1.3%	0.4%	0.8%	0.5%	0.7%
Categories overrepresented in Group 2 (p<.05)					
Category	Group 1	Group 2	Group 3	Group 4	Overall
Pudding/Dry Dessert	0.3%	0.5%	0.1%	0.3%	0.3%
Tea	0.0%	0.2%	0.0%	0.2%	0.1%
Canned Meat	0.2%	0.4%	0.0%	0.1%	0.2%
Pasta	0.5%	1.3%	0.8%	1.0%	1.0%
Paper Towels	0.5%	0.9%	0.7%	0.4%	0.7%
Prepared Food/Dry Dinner	0.3%	1.4%	0.9%	1.3%	1.1%
Categories overrepresented in Group 3 (p<.05)					
Category	Group 1	Group 2	Group 3	Group 4	Overall
Candy/Gum/Mint	2.6%	2.6%	3.3%	1.9%	2.6%
Fruits	10.9%	6.4%	9.7%	6.5%	7.8%
Vegetables	9.8%	5.8%	10.1%	6.0%	7.4%
Categories overrepresented in Group 4 (p<.05)					
Category	Group 1	Group 2	Group 3	Group 4	Overall
Baby Food	0.2%	0.0%	0.2%	0.5%	0.2%
Bagels/Breadsticks	0.3%	0.8%	0.6%	1.1%	0.8%
Bottled Water	0.9%	0.7%	1.0%	1.8%	1.1%
Coffee	0.2%	0.7%	0.3%	1.0%	0.6%
Cookies and Crackers	3.0%	3.4%	2.6%	4.1%	3.4%
Frozen Pizza/Snacks	1.0%	1.1%	1.2%	2.1%	1.4%
Household Cleaners	0.4%	0.6%	0.4%	1.0%	0.7%

Table 6. Comparison of product categories purchased for each group, controlling for basket size.

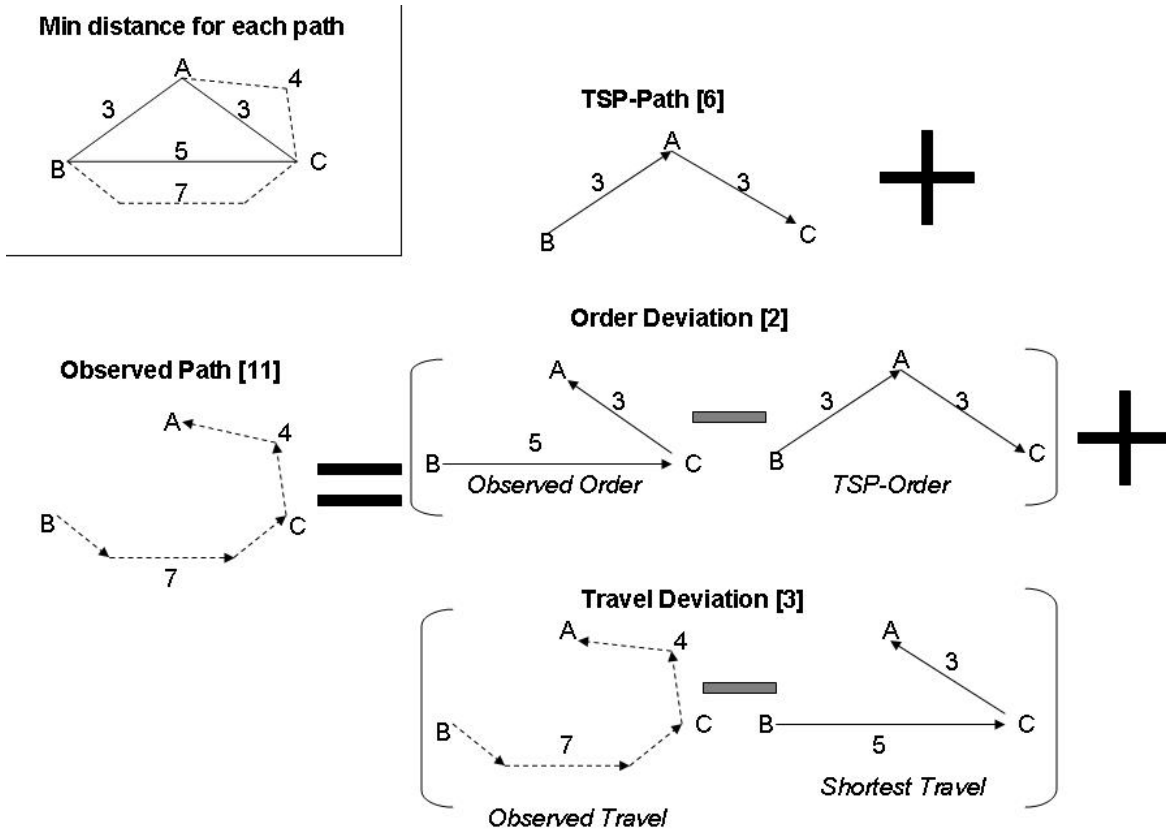


Figure 1. Deviation decomposition

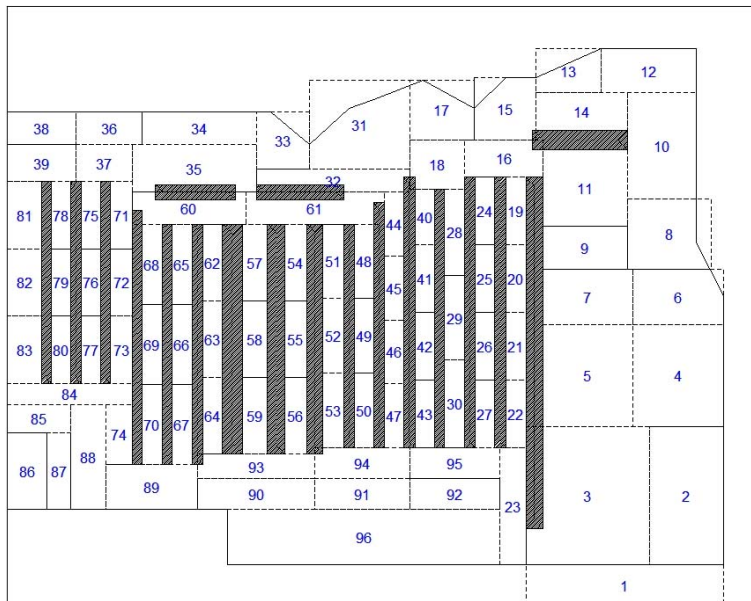


Figure 2. Grocery store divided into 96 zones

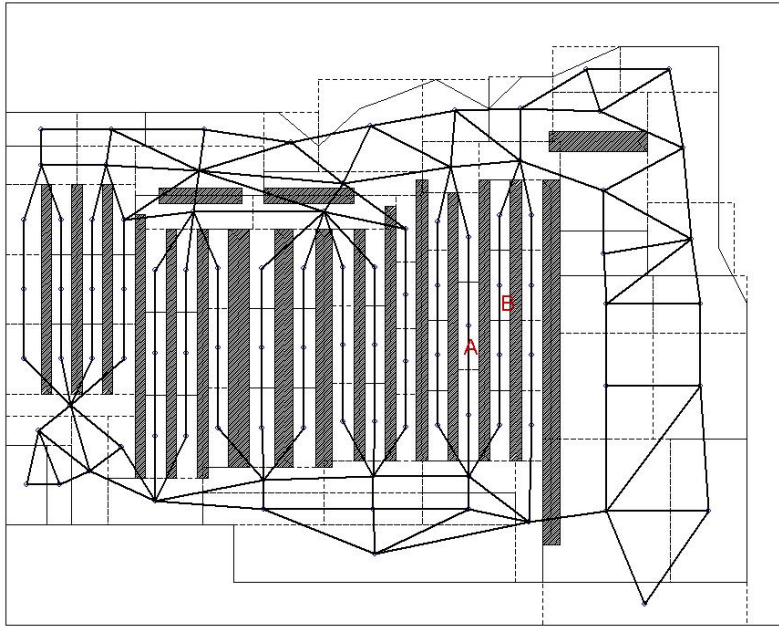


Figure 3. Grocery store represented by a graph of 96 nodes

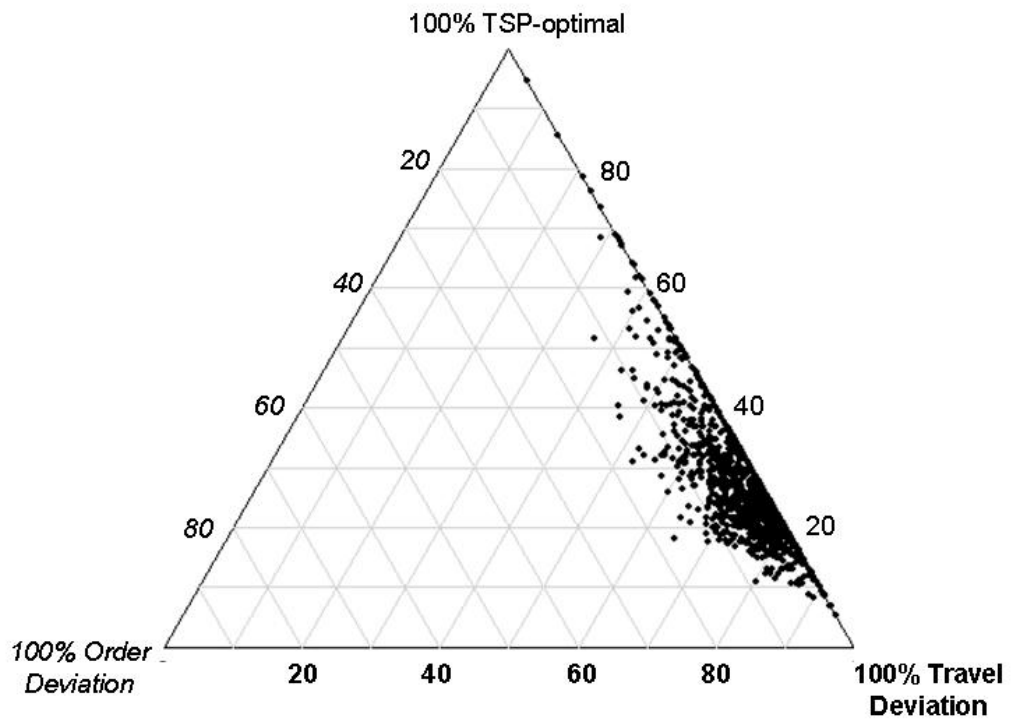


Figure 4. Triangle plot for optimal path, order deviation, and travel deviation. The different fonts and angled hashmarks indicate which scale corresponds to each dimension.

Regular Appendix: Sensitivity Analysis

The decomposition analysis in Section 4.1 is based on the assumption that shoppers come to the store with a fixed shopping list. But in reality, many categories purchases are unplanned (Bucklin and Lattin 1991), and a shopper may plan to visit some categories to check for promotions, but may not purchase from them if a suitable deal is not available. We conducted two sensitivity analyses to study how these violations of the “fixed shopping list” assumption affect our results.

In the first sensitivity analysis, we randomly assign, for each trip, some of the purchases to be “unplanned,” then we re-compute the TSP decomposition based on the reduced set of categories. The table below shows the results when 10%, 20%, 30%, 40%, and 50% of observed purchases are treated as unplanned.

Unplanned %	TSP-Path %	Order %	Travel %
0	27.5%	3.1%	69.4%
10	27.3%	3.0%	69.7%
20	26.4%	2.6%	71.0%
30	25.5%	2.3%	72.2%
40	24.0%	1.8%	74.2%
50	22.1%	1.3%	76.6%

By allowing some of the category purchases to be unplanned, the TSP-optimal portion of each path is reduced; this is expected because while the total observed distance is unchanged, the optimal distance (i.e., the minimum distance that the consumer needs to travel in order to complete his planned purchases) is reduced, thus reducing the extent of TSP-optimality (from 27.5% under the original no-unplanned purchase scenario to 22.1% when 50% of purchases are treated as unplanned). Because the movements towards unplanned purchases are treated as deviations from the main path, the fraction assigned to travel deviation therefore increases (from 69.4% to 76.6% as we go from 0% to 50% unplanned purchases). On the other hand, the extent

of order deviation decreases (from 3.1% to 1.3%), due to the removal of these unplanned purchases from the original shopping list.

In the second sensitivity analysis, we randomly add m categories to each consumer's shopping list to represent categories (store zones) that she chose to visit, but did not purchase from.. The results are shown in the table below.

m	TSP-Path %	Order %	Travel %
0	27.5%	3.1%	69.4%
1	29.8%	3.8%	66.4%
2	31.7%	4.3%	64.0%
3	33.2%	5.2%	61.6%
4	34.8%	5.8%	59.5%
5	36.1%	5.9%	58.0%

By allowing some category visits to be planned but not purchased, the fraction of the trip accounted for by travel deviation decreases from 69.4% to 58.0% as m rises from 0 to 5. This is because a portion of the travel deviation is now treated as planned *visitation* of certain categories. In addition, since the total observed trip length remains the same while the optimal path becomes longer, the fraction of distance due to the TSP-path goes up (from 27.5% to 36.1%). Finally, the fraction of order deviation goes up due to the inclusion of these additional product categories in the consumers' shopping list.

The above sensitivity analyses are valuable in three respects: (i) They show that our results are reasonably invariant with respect to violations of the TSP assumptions, namely unplanned purchases and category visits that are planned but do not result in purchases; (ii) they allow us to explore the directionality and magnitude of how our decomposition results are affected when the "fixed shopping list" assumption is violated; and (iii) they suggest that part of the travel deviation can be attributed to consumer search behavior.

Technical Appendix (Online Supplement)

A. Applying the TSP algorithm to grocery shopping paths

We use two algorithms to solve for the optimal solution: (i) exhaustive search, and (ii) simulated annealing. Exhaustive search is used when the number of purchase location is less than 11; otherwise, simulated annealing is used.

Algorithm I: Exhaustive Search

We generate every possible permutation of the order of purchase locations then compute the total distance associated with each order. The optimal solution is the order that results in the shortest total distance.

Algorithm II: Simulated Annealing

When the number of purchase locations is greater than 10, the number of possible permutations is too large to conduct an exhaustive search. In this case, we use simulated annealing to obtain the optimal solution. Our implementation here is based on the three-step algorithm of Goffe (1994). Readers are referred to Goffe (1994) for a full discussion of the theoretical foundations of the simulated annealing algorithm.

Step I: First, we define the “neighborhood” of an order. We define a “move” to be a pairwise switch of two locations in a sequence.

Step II: Then, we randomly choose a starting order. Each order has an equal probability of being chosen as the initial order.

Step III: We specify a “starting temperature” T_0 and a “cooling schedule” T_n (the temperature at step n). T_0 is defined so that initially, around 80% of the “uphill” moves (i.e., moves that lead to a large total travel distance) will be accepted, as recommended by Goffe (1994). We use an exponentially cooling schedule for T_n , which takes the following form:

$$T_n = T_0 k^n$$

where $k < 1$ is a tuning constant that controls how fast the temperature is cooled. In our case, k is chosen to be at least 0.99999 to ensure that the algorithm has a high probability of reaching the global optimal solution.

Step IV: We then start the annealing algorithm at the starting temperature, and loop until T_n falls below a small constant (0.001). At each iteration n , the following steps are performed:

- (i) A “proposal” order, denoted as C^* , is randomly generated from the neighborhood of the current order, denoted as C .
- (ii) The total distance associated with the proposal order, D^* , is calculated.

D^* is compared to the total distance associated with the current order, D , if $D^* < D$, then C is replaced by C^* . Otherwise, C is replaced by C^* with probability $\exp((D-D^*)/T_n)$.

B. PathTracker[®] data preparation and cleaning procedure (Hui et al. 2007)

A PathTracker[®] dataset consists of a large number of “trips” that include both the shopping path and purchase data for each one. Each trip starts when the shopping cart is taken at the store entrance and ends when the cart is pushed through the checkout line to the other side of the checkout counter. A shopping path is represented by a list of (x,y) coordinates at five second intervals (“blinks”) that indicate the current location of the shopping cart. Purchases are tracked via point-of-sale scanner data, indicating which products are purchased. Within the PathTracker[®] system, each location within the store is represented by a pair of (x,y) coordinates; together with the scanner data, we can map each purchase back to the store location where it was made. Sometimes, a product category can be located in more than one area within the store. In this case, we assign a purchase to the feasible location where the shopper spent the longest time shopping.

This assignment heuristic is based on the intuition that buying a product should take more time than just walking past an area. While we recognize that this assignment algorithm is imperfect, we believe that it provides a good starting point for our initial study on shopping behavior.

To prepare our data for analysis, we first screen out paths that contain purchases in only one category, because we are interested in determining whether a consumer's order of purchases is optimal. After eliminating these paths (roughly 25% of all paths), we have a total of 993 paths for analysis. Next, to facilitate our analysis, we discretize the store into 96 zones based on our discussion with Sorensen Associates. The discretization is shown in Figure 2.

At first sight, the procedure of discretization appears to “throw away” some of the data, since the variation of the data at a resolution finer than zones is lost. But this procedure leads to three main advantages: (1) It simplifies modeling by limiting the number of possible locations in the store; we can then represent the shopping path as a series of finite choice problems, which allows us to more efficiently solve for the TSP-optimal path; (2) the location of the shopping cart is not a perfect proxy for the shopper: the shopper can “park” his cart at a certain spot and then shop somewhere else. Thus, it is more reasonable to assume that the location of the cart gives us some indication to the general region where the shopper is located, rather than treating it as the shopper's exact location. (3) Similarly, we only know the general position of each product in the store, up to a certain degree of error. Thus, the procedure of discretization brings our analysis closer to the resolution of the measurement accuracy of our data.

After discretizing the store, we implicitly take into account the existence of physical barriers (e.g. aisles, walls) in the store by representing the store as a “graph”: a mathematical object defined by “nodes” that represent regions and “edges” that depicts the connectivity between different regions. A node is placed at the center of each zone. An edge is drawn between two nodes if they represent adjacent regions, indicating that it is possible to move from one to the

other without going through any other node. Figure 3 shows how the grocery store is represented as a graph of 96 nodes, referring to each of the 96 zones.

For example, although node A and node B are close to each other in Euclidean distance (they are in adjacent aisles), one would have to go through at least 6 intermediate nodes to go from A to B. The shortest travel distance between any pairs of locations in the store can be approximated by the distance of the shortest path connecting their respective nodes. Thus, the graph is a faithful representation of the distances between each zone in the grocery store, since it takes into account the multiple spatial constraints of the store.