

## **An Integrated Model of Grocery Store Shopping Path and Purchase Behavior**

Sam K. Hui

Eric T. Bradlow

Peter S. Fader\*

This Version: January 31, 2007

---

\* Sam K. Hui is a doctoral candidate in Marketing, 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 and Peter S. Fader is the Frances and Pei-Yuan Chia Professor, Professor of Marketing, all at the Wharton School of the University of Pennsylvania. Corresponding author: Sam Hui. Email: [kchui@wharton.upenn.edu](mailto:kchui@wharton.upenn.edu). The authors are grateful for the data and assistance provided by Sorensen Associates and, in particular, the feedback and encouragement from Herb Sorensen.

## **An Integrated Model of Grocery Store Shopping Path and Purchase Behavior**

### **Abstract**

We develop an individual-level integrated Bayesian model for store shopping trips to capture the relationship between a consumer's shopping path through the store and her purchasing behavior. Using dynamic latent "attractions" of different zones and product categories in the store as fundamental constructs, our model jointly captures three key aspects of a consumer's within-store behavior: which zones she visits, how long she stays in each zone, and what purchases, if any, she makes within that zone. We calibrate our model using a novel dataset from the PathTracker<sup>®</sup> system, an electronic shopping cart monitoring system developed by Sorensen Associates.

Our study also contributes to in-store research by modeling the degree of forward-looking behavior that consumers may exhibit, an area of significant recent study in marketing and experimental economics. By incorporating the total process of in-store shopping (movement *and* purchases) in our model, retailers may obtain a more comprehensive understanding of consumers' grocery shopping behavior.

As an illustration of the model's potential usefulness, we conduct a hypothetical "policy experiment" in which two categories swap locations. We examine the impact of this change on the affected categories as well as other aspects of the shopping trip as a whole.

## 1. Introduction

Imagine the following scenario: two shoppers, Alfred and Bianca, are both standing in the middle of the frozen food aisle in their local grocery store. However, they followed different routes (as shown in Figure 1) to arrive at their current location. While Alfred (as shown by the dashed line) strolled up and down almost every aisle, Bianca (solid line) took a more direct approach. Who is more likely to purchase frozen food, Alfred or Bianca?

[Insert Figure 1 about here]

One may think that Bianca is more likely to make a purchase here, since she appears to be more “goal-directed”. Based on her behavior up to that point, it seems as if she had already decided to purchase frozen food before she entered the store, and hence took a relatively direct route to save time. On the other hand, one may argue that since Alfred has already visited more aisles, he may be in the mindset of a “stock up” shopping trip, and thus may have a higher baseline purchase probability. These contrasting perspectives would ordinarily be hard to test: such a task may require a large volume of data from an in-depth field experiment, and even then it would be difficult to identify the conditions that might lead to a specific prediction.

Rather than relying on a purely “data-driven” approach, however, we suggest the use of an integrated probability model of consumers’ paths and purchases calibrated on field data. Such a model will not only allow us to better understand the relationship between shopping paths and purchases, but will also enable us to assemble a more complete picture of each shopper’s in-store decisions and how she interacts with the product categories and the store environment. These understandings may have important implications for various retailing issues such as store layout (Vrechopoulos et al. 2004), shelf-space slotting fees (Bloom et al. 2000), and inventory management (Lummus and Vokurka 1999), to name just a few.

In the past, though, large datasets on shopping paths and purchases were difficult and costly to obtain. To collect such data, one would typically have to physically follow shoppers around the store or to rely on a large number of cameras (e.g., Farley and Ring 1966; Heller 1988). But today, advances in data-collection technology have helped overcome this hurdle. New technologies such as Radio Frequency Identification (RFID) enable researchers to track multiple shoppers' movements in real time. For instance, Sorensen Associates developed the PathTracker<sup>®</sup> system, a system that uses an RFID tag attached to each shopping cart to track shoppers' movements as they enter the store, until they reach the checkout counter. By combining individual-level movement data with their purchase records obtained from scanner data, the PathTracker<sup>®</sup> system generates datasets (as used in this paper) that include thousands of records of shoppers' paths and their corresponding purchases immediately and cost effectively (Larson et al. 2005; Sorensen 2003).

To extract managerially relevant information from a complex dataset of this type, our model jointly captures multiple aspects of each shopper's in-store behavior. Thus, this research is in the same spirit as papers such as Chintagunta (1993) and Park and Bradlow (2005), who developed integrated models that capture multiple behavioral aspects in the purchasing of packaged goods and internet auction bidding, respectively. To specify an integrated model of grocery shopping, three key elements of a grocery trip are considered: (1) which areas of the store a consumer chooses to visit, (2) whether she chooses to stay/shop at a given location, i.e., considers making a purchase in each of the areas, and (3) whether/what she actually purchases in each area. Central to these analyses is a set of latent, time-varying variables (called "attractions"), corresponding to each product category and location (zone) of the store. These latent variables evolve based on the shopper's path and purchases, and act in combination with other shopper-

specific characteristics to drive all of the above processes. In addition, shoppers are allowed to be forward looking, i.e., they plan ahead when deciding where they want to move next, a key insight drawn from the NEIO (e.g., Song and Chintagunta 2003; Sun et al. 2003) and empirical economics literature (Camerer and Ho 1999). Of course, shoppers may have different product preferences, movement tendencies, and “forward-lookingness”, thus a Hierarchical Bayes formulation (Rossi et al. 2006) is used to capture parameter heterogeneity.

Our integrated model gives us a novel perspective in understanding grocery shopping behavior. Traditional research in marketing typically models brand choice (e.g., Guadagni and Little 1983) or more recently the “market basket”, i.e., the entire portfolio of goods a shopper purchases as opposed to one category at a time (Bell and Lattin 1998; Manchanda et al. 1999). Our approach, in contrast, allows researchers to put themselves into the shoes of a consumer, and follow her footsteps as she moves around the store to understand the visit/shop/purchase decisions she makes along the way. Researchers can then study consumers’ grocery shopping behaviors by observing the entire shopping process, from entrance to checkout, rather than merely looking at limited “snapshots” (e.g., scanner data). From this perspective, we study how purchase patterns are related to shopping paths and provide a descriptive statistical model of shoppers’ movements and purchase tendencies.

The remainder of this paper is organized as follows. Section 2 gives an overview of the data collection and preparation procedures, and provides summary statistics of our PathTracker<sup>®</sup> dataset. In Section 3, we specify our model in detail and describe our estimation and computational approaches. In Section 4, we perform a simulation study to assess the identifiability of our model, which is crucial as a model of this kind has never been introduced into the marketing literature. In Section 5, we apply our model to field data, demonstrate its fit,

and interpret its parameter estimates. We then conduct a policy experiment in Section 6 to demonstrate some potential managerial applications of the model. Finally, Section 7 concludes with a discussion of model extensions and future research directions.

## **2. Data Overview**

In this section, we briefly describe the PathTracker<sup>®</sup> system and the format of our raw data. Then, we describe in detail how we transform the raw data into a format suitable for our modeling purposes. We spend a considerable amount of time discussing data preparation in order to highlight the general modeling and data challenges that marketing researchers are likely to face when they build similar models of path data (Hui et al. 2006a).

### 2.1 Data description

Our dataset contains paths and associated shopping basket data collected from March 14, 2004 to April 3, 2004 using the PathTracker<sup>®</sup> system, which was installed in a large supermarket in the Eastern United States. The system consists of a set of RFID tags and antennae: A small RFID tag is affixed under each shopping cart, and emits a uniquely coded signal every five seconds (“blinks”); this signal is then picked up by antennae around the perimeter of the store to locate the cart (Sorensen 2003). A PathTracker<sup>®</sup> dataset consists of shopping trips that include both the shopping path, represented by a list of (x,y) coordinates at five second intervals, and purchase records (in terms of product UPC’s) from scanner data. Each trip starts when a shopping cart is taken at the store entrance, and ends when it is pushed through the checkout line to the other side of the checkout counter.

Within the PathTracker<sup>®</sup> system, each product category’s shelf location is also represented by a pair of (x,y) coordinates. Together with the scanner data, this allows us to map each purchase back to the store location where it was made. Since we only have detailed

information about each product category's (but not each individual UPC's) location, we study purchase behavior at the product category level in this research; i.e., each purchased UPC is aggregated to its product category, and identified with the position of the product category in the store. Sometimes, a product category can be located in more than one area within the store, making it ambiguous where a shopper has purchased that item. In these cases, we explore and evaluate the likelihood function using several "assignment" heuristics, discussed in Section 2.3.3, to determine (or impute) which store location is associated with the purchase. In the future, as cart-level scanners (and/or item-level RFID tags) become available, this problem may be eliminated altogether. Another noteworthy caveat is that this dataset only include trips that use a shopping cart, which eliminates a substantial fraction of trips that contain a small number of purchases; thus, the basket sizes observed in our data tends to be slightly larger than commonly observed (Sorensen 2003).

## 2.2 Data cleaning

Since RFID is a relatively new technology, we take several steps to clean our data to exclude trips that may not correspond to valid shopping paths. First, we include only paths that start at the entrance and end at the checkout, indicating a completed trip. Second, we exclude paths that contain "inconsistent" purchase information, that is, if it contains a purchase of a product category that is not located at any position that the shopper visited during his entire trip. In this case, the shopping cart's location does not serve as a good proxy for the shopper's position, and thus the path is excluded.<sup>1</sup> After performing the above procedures, our dataset contains a total of 1051 paths and their corresponding purchase records. This dataset will be used for all of our subsequent analyses.

---

<sup>1</sup> We recognize that some shoppers "park" their carts at one place, then walk to other areas to obtain products. However, to keep our dataset "clean", we exclude such paths. As data collection technology further matures, such data cleaning needs will becomes less crucial.

## 2.3 Data preparation

Our model for consumers' in-store movement, as we will explain in Section 3, is by nature a discrete choice model (McFadden 1981). Thus, the raw data needs to be simplified to limit the number of possible locations (i.e., choice options).<sup>2</sup> This is a common procedure used by other researchers when building models to analyze eye-tracking data (e.g., Pieters et al. 1999) and pedestrian movements (e.g., Antonini et al. 2006). In the next subsection, we first discretize the grocery store into zones. Then we convert a shopping path into a series of sequential zone-visit decisions. Finally, we discuss the heuristics that assign purchases to zones when the location of the purchase is ambiguous.

### 2.3.1 Store discretization

We discretize our store by dividing it into distinct, non-overlapping zones. Each (x,y) coordinate pair on a shopping path is then mapped to a specific zone, and no further distinction is made, in this research, among (x,y) coordinates within the same zone.

Through a careful analysis of category locations and discussions with Sorensen Associates, we divided the grocery store into 96 zones of comparable sizes, as shown in Figure 2.<sup>3</sup> The location(s) of each product category across the 96 zones, along with its % penetration (fraction of the 1051 shopping baskets containing the category), are shown in Table 1.<sup>4</sup> Note, as

---

<sup>2</sup> In theory, we could model distance  $d$  and the angle of movement  $\theta$  within a blink (five seconds), perhaps using a correlated random walk which is common in models of animal movements (e.g., Bergman et al. 2000); however, the discrete choice model framework is more in line with the extant research in marketing as well as the level of data precision obtained here.

<sup>3</sup> In general, other methods of discretization are possible. The problem of finding an "optimal" discretization scheme is related to the edge detection/"wombling" problem in spatial statistics (e.g., Lu and Carlin 2005), which we will address in future research.

<sup>4</sup> Note that the scanner data portion of our dataset is noisier than typical scanner data. For instance, our data do not allow us to adequately tease apart the Skin Care and Eye Care category, and also the Baby Medical Needs/Diapers categories. Thus, these categories are lumped together in Table 1. In the model, however, their attractions are separately estimated.

mentioned before, a product category can appear in multiple zones. For example, Paper Towels are located both in zone 37 and zone 75.

[Insert Figure 2 about here]

[Insert Table 1 about here]

The procedure of discretization appears to “throw away” some of the data, since variations at a resolution finer than the zone are lost. However, this procedure leads to three main advantages. First, it simplifies modeling by limiting the number of possible locations in the store; we can then represent the shopping path as a series of “choose-1-out-of-n” problems, and hence model it using a discrete choice framework. Second, discretization brings the resolution of the model closer to the level of data precision and managerial decision making that typically occurs at the retail level. The location of the shopping cart is not a perfect proxy for the shopper’s location: the shopper will frequently move several steps away from the cart. It is therefore more reasonable to assume that the cart’s location gives us some indication to the general region where the shopper is located, rather than her exact location. Third, if data loss is indeed an issue, zones can be made smaller to refine the discretization and hence minimize data loss. Thus, the framework developed here is fully general.

One of the key challenges in modeling store movement data is the need to take into account the existence of physical barriers (e.g. aisles, walls) in the store. We do so by representing the store as a “graph”: a mathematical object defined by “nodes” that represent regions, and “edges” that depict the adjacency between different regions. A node is placed at the center of each zone. An edge is drawn between two nodes if they represent two adjacent zones, indicating that it is possible to move from one to the other without going through any other node. Figure 3 shows the grocery store represented by a graph of 96 nodes, referring to each of the 96

aforementioned zones. An assumption here is that adjacent nodes can be reached in one blink, while non-adjacent nodes cannot; this assumption has been empirically verified with the data.

[Insert Figure 3 about here]

By representing the grocery store as a graph, we implicitly take into account physical barriers within the store by the presence or absence of edges between nodes. For example (see Figure 3), although node A and node B are close together in Euclidean distance (they are in adjacent aisles), one would have to go through at least four intermediate nodes to go from A to B, due to the absence of an edge connecting them. 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 faithfully represents the distances between each zone in the grocery store by explicitly taking into account the multiple spatial constraints.

### 2.3.2 Path discretization

Having discretized the store into 96 zones, we convert each of the 1051 shopping paths by mapping each  $(x,y)$  coordinate on a path at each blink to its corresponding zone. If a shopper spends more than one blink in the same zone, we record the number of blinks that she spends in that zone. Thus, a path is converted into a sequence of zone visits, along with the number of blinks the person spent in each zone before moving to the next zone. From here on, we will refer to a zone transition as a “step”. An example of path discretization is shown in Figure 4. The top panel depicts the sequence of  $(x,y)$  coordinates in the raw data, while the bottom panel shows the corresponding discretized path.

[Insert Figure 4 about here]

### 2.3.3 Assignment of category purchases to zone visits

Several problems arise when allocating purchases to zone visits: (i) some product categories are located in more than one location in the store and thus there is some uncertainty about where these purchases are made; (ii) a shopper may enter a zone more than once, which makes it difficult to determine when the item was purchased; (iii) a shopper may buy multiple items from a product category, but we have no way of knowing which ones occurred together or in separate zone visits. To accommodate these possibilities, we assigned each purchase to a specific zone visit (henceforth referred to as a visit-occasion) based on a number of reasonable heuristics. In Section 5, we estimate our model using each of these heuristics and then compare them based on the marginal log-likelihood (Newton and Raftery 1994) of the data. We choose the assignment heuristic that results in the highest marginal log-likelihood, and is hence globally most likely; albeit for any given shopper, it is approximate. Four specific heuristics are considered and are listed as follows. In all cases, a “feasible visit-occasion” is defined as a visit to a location that contains the product category; we denote  $k$  as the number of items a shopper purchased from a certain product category.

(A) *LONGEST*: All  $k$  purchases are assigned to the *single* feasible visit-occasion where the shopper spent the *longest time*.

(B) *EVEN*: The  $k$  purchases are assigned to the  $k$  longest feasible visit-occasions, ordered by the amount of time the shopper spent in each feasible visit-occasion.

(C) *PROPORTIONAL*: The  $k$  purchases are assigned to the feasible visit-occasions proportional to the amount of time spent in each feasible visit-occasion.

(D) *LAST*: All  $k$  purchases are assigned to the last feasible visit-occasion.

While, of course, none of these is likely to be exact, they do span a reasonable set of allocations and allow us to assess the sensitivity of our findings to different imputation procedures (Little and Rubin 1987).

## 2.4 Summary statistics

Since our goal is to capture shoppers' in-store visit, stay, and purchase behaviors, we derive the following summary statistics that describe the data along those three key dimensions. These summary statistics are independent of the assignment heuristics discussed in Section 2.3.3. Thus, after we fit our model, they can be used to validate both the performance of our model and the chosen assignment heuristic. The summary statistics included for visit, stay, and purchase are discussed separately in Subsections 2.4.1-2.4.3 below.

### 2.4.1 Summary statistics for visit

We compute the total number of steps (i.e., zone transitions) that a shopper takes during the shopping trip, and we also compute the overall zone-to-zone transition probabilities. The histogram for the total number of steps is shown in Figure 5. In our dataset, the mean number of steps taken is 98.8 while the median is 90.0. The transitions that occur with highest frequency out of each zone are shown by the solid directed arrows in Figure 6, while the light shaded arrows indicate all possible movements.

[Insert Figure 5 about here]

[Insert Figure 6 about here]

Note from Figure 6 that there is a general tendency to “back-track” once a shopper enters an aisle; i.e., after a shopper enters an aisle, she is more likely to head out rather than traversing all the way through it. This interesting observation is consistent with the common “excursion” and lack of aisle-traverse behavior documented in Larson et al. (2005) and Sorensen (2003), and

can be valuable for determining shelf-slotting fees (i.e. mid-aisle shelf space may receive low traffic).

#### 2.4.2 Summary statistics for stay

To summarize shoppers' stay times during their trip, we compute (i) the total amount of time (in minutes) that a shopper spent in the grocery store, and (ii) the average amount of time that shoppers spent in each zone in the store. The histogram for total in-store time is shown in Figure 7. In our dataset, shoppers on average spend 48.6 minutes in store; the median in-store time is 43.8 minutes. The average amount of time shoppers spent in each zone (in minutes) is shown in Figure 8.

[Insert Figure 7 about here]

[Insert Figure 8 about here]

Figure 8 leads to several interesting insights about shopping behavior. First, shoppers on average spend a large amount of time in the area immediate to the entrance (zone 2 and 3), where produce products (fruits and vegetables) are located. Second, shoppers tend to move along aisles very quickly. Third, a shopper who follows the "typical" counter-clockwise movement through the store will in general tend to spend less and less time in a zone as her trip progresses, consistent with the observation in Sorensen (2003) that shoppers tend to speed up as they move towards checkout.

#### 2.4.3 Summary statistics for purchase

We compute (i) the total number of categories that a shopper purchased during his/her trip, and (ii) the % purchase incidence (penetration) for each product category. The histogram of the total number of categories purchased is shown in Figure 9. In our dataset, shoppers on average purchase from 6.7 categories.

[Insert Figure 9 about here]

### **3. Model Development**

We develop an integrated model to simultaneously describe each consumer's shopping path and purchase behavior. The intuition behind our model is based on an analogy between consumers and electric charges in physics. Imagine that  $X$ , a positive unit point charge, is placed in a space that contains a number of (fixed) point charges. If we know the exact positions and strengths of all the electric charges in the space, we can calculate  $X$ 's path using Coulomb's law (Halliday et al. 2004). Back to the grocery store setting, the "space" is the areas of the grocery store accessible to shoppers, the "other charges" are the product categories, and the "unit charge" is the shopper. Using the electric-field analogy, if we are given the exact "attraction" of each product category to the shopper, we can theoretically predict the future movement of the shopper, up to, of course, the stochastic nature of movements.

The strength of the above electric-field analogy is that it inherently takes into account the "forward-looking" nature of a shopper's within-store movement. In the electric field scenario,  $X$ 's movement is not only determined by charges in its immediate neighborhood, but also the ones that are distant from it. Its movement is determined by the vector sum of all the individual forces exerted by each charge. Likewise, in the supermarket setting, a shopper can be "forward looking", i.e., her movement depends not only on the attractions of product categories that are right next to her, but also on those that are further away. In other words, the shopper may try to move closer towards areas (zones) that she finds more attractive, even if she has to sacrifice her short-term utility by moving to a less attractive zone in the immediate next step.

Although this electric field analogy has some intuitive appeal, it does have several limitations, which provides insight into the richness/difficulty of our problem. First, while an

electric charge is a measurable physical quantity, the “attraction” of a product category to a shopper is an abstract psychological construct that cannot be directly measured. Thus, we treat attractions as latent constructs and infer their values from the observed shopping path and purchase information.

Second, while in physics we can construct scenarios in which charges are fixed and non-varying, it is unreasonable to assume, in the grocery store setting, that the attraction of a product category remains constant during the entire shopping trip. Instead, attractions may evolve depending on the consumer’s path and purchases. For instance, if a shopper had already purchased orange juice the first time she went past the juice section, she may be less inclined to stay and shop for orange juice the next time she reaches the same zone during a given trip. Thus, our model allows attractions to evolve based on a shopper’s visitation and purchase behavior, as we will explain in detail in Section 3.2.1.

Third, not only do shoppers move around in the grocery store, but they also stay at certain locations to look more closely at different products and make purchases. The “shop” and “buy” decisions go beyond the realm of the electric field analogy. To unify all three aspects of grocery shopping (visit, shop, and buy) into an integrated model, we have to explicitly link shoppers’ shop and buy decisions, not just their movements, to the attractions as well. To this end, we posit a discrete choice model that jointly captures shopper’s visit, shop, and buy decisions. We define two separate linear functions that relate attractions to latent quantities called “shop utility” and “buy utility”. A shopper will shop/buy, respectively, if her “shop utility”/“buy utility” exceeds a certain threshold. The intercept parameters in the linear functions capture the shopper’s baseline propensity to shop/buy, while the slope parameters capture the sensitivity of the shopper’s

shop/buy decisions to attractions, similar to the “discrimination” parameter commonly used in IRT-type models (Lord and Novick 1968).

In the remainder of this section, we will formalize the above discussion into an individual-level probability model. We present an overview of the shopper’s decision process in Section 3.1 and then describe each component of our model in detail in Section 3.2. For the sake of exposition, we focus first on a single shopper, and thus individual-level subscripts will be suppressed. In Section 3.3, we discuss how heterogeneity among shoppers’ purchase preferences, movement patterns, and forward-lookingness are captured using a Hierarchical Bayes framework. Finally, we specify prior distributions of the model parameters and outline the proposed estimation procedure in Section 3.4.

### 3.1 The shopper’s decision process

As discussed before, we discretize each path into a number of zone transitions, which we refer to as “steps”. A new step is initiated each time the shopper leaves one zone and goes to another zone, until she reaches checkout. At step  $t$ , we denote the zone that the shopper is located as  $x_t$ . At the first step ( $t=1$ ), the shopper is located at the entrance of the grocery store. From there, we model the shopper’s decision process at each zone as a sequence of three (nested) decisions: *visit*, *visit-to-shop*, and *shop-to-purchase*. Each of these decisions, as depicted in Figure 10, are driven by the latent attractions of product categories and zones, which we will define later.

[Insert Figure 10 about here]

First, the shopper makes a *visit* decision: she decides which zone she is going to visit next. If that zone is checkout, the trip ends. Otherwise, she makes a *visit-to-shop* decision: she decides whether she wants to shop at her current zone, or whether she is only passing through on her way to a different zone. We denote the shopper’s visit-to-shop decision by  $H_t$ , which takes the value 1

if a visit-to-shop conversion is made, and 0 otherwise. Note that we allow for the possibility that the shopper makes a visit-to-shop conversion ( $H_t = 1$ ), but decides not to buy anything.

Depending on whether she shops or not, she may stay at the zone for a different duration (presumably, the shopper will stay longer if she is shopping than passing through). We denote by  $S_t$  the number of blinks that the shopper stays at the current node in step  $t$ . Note that we are unable to directly observe whether someone is actually shopping or just passing through, and thus  $H_t$  is a latent construct that is central to our model (and of great relevance to managers as well); this is similar to the spirit of Hidden Markov models where a latent stochastic process drives the observed outcome (e.g., Montgomery et al. 2004).

Next, if she decides to shop, she needs to make a *shop-to-purchase* decision: she decides which product categories, if any, to purchase in that zone. We denote her category purchase incidence decision as a vector  $\vec{B}_t$ , where  $B_{jt} = 1$  if category  $j$  is purchased at step  $t$ , and 0 otherwise. If she does not make a *visit-to-shop* conversion, she does not make a purchase decision since she is only walking through the zone on her way to other zones.

Finally, the attractions are updated to take into account the behavior observed in the preceding zone(s). The shopper then decides which zones to visit next, and the decision process in Figure 10 is restarted.

### 3.2 The proposed model

In our model, each of the shopper's decisions (visit, visit-to-shop, and shop-to-purchase) are governed by latent constructs called category attractions and zone attractions. We define these constructs and their relationship to each other in Section 3.2.1. In Section 3.2.2 to 3.2.5, we describe how we model the shopper's three decisions as a function of category and zone attractions.

### 3.2.1 Category/zone attractions and baseline visit propensities

We define two sets of (related) latent variables to capture the “attractions” of product categories and of zones, respectively. A latent attraction is defined for each product category to model category purchase behavior; then, zone attractions are calculated based on the attraction of the product categories they contain to model zone visit behavior.

We define a vector of latent variables  $\bar{a}_t = (a_{1t}, a_{2t}, \dots, a_{Jt})'$ , where  $a_{jt}$  ( $j = 1, 2 \dots J$ ;  $t = 1, 2, \dots T$ ) denotes the “category attraction” of category  $j$  for the shopper at step  $t$ . These category attractions drive the model of purchase behavior—categories with higher attractions to the shopper are assumed to be more likely to be purchased. We then compute “zone attractions” based on the aggregation of “category attractions” of the product categories it contains. These “zone attractions” enter the model of shop and visit behavior, as we will discuss later. The zone attraction for zone  $i$  for the shopper at step  $t$  is defined as:

$$A_{it} = \log \left( \sum_{j \in C(i)} \exp(a_{jt}) \right) \quad (1)$$

where  $C(i)$  denotes the set of product categories available at zone  $i$ . This specification is similar to the “inclusive value” notion that is commonly used in nested-logit models (McFadden 1981). In our framework, the zone can be viewed as a “nest” that contains several product categories.<sup>5</sup>

As we have discussed earlier, category attractions may not be constant over time. Thus, we allow them (and hence the derived zone attractions) to evolve depending on the shopper’s visitation and purchase behavior up to step  $t$ . We use a parsimonious yet flexible specification to capture the basic evolution pattern of attractions, as follows:

$$a_{j,t+1} = a_{jt} + \Delta_b B_{jt} + \Delta_s I\{j \in C(x_t)\} \quad (2)$$

---

<sup>5</sup> Other specifications for Equation (1) are possible. For example, we may define zone attraction as the maximum of the attractions of the product categories it contains, i.e.,  $A_{it} = \max_{j \in C(i)} a_{jt}$ . We leave this for future research.

That is, we posit that after the shopper visited node  $x_t$ , the attraction of the categories contained in zone  $x_t$  will change by an amount indicated by  $\Delta_s$ . If  $\Delta_s$  is negative, the attraction of a product category decreases after a shopper visits the zone that contains it. If category  $j$  is purchased at step  $t$  ( $B_{jt} = 1$ ), then the attraction for category  $j$  will further change by an amount indicated by  $\Delta_b$ .

### 3.2.2 Model of visit

The shopper first decides which zone she wants to visit next. We denote the set of zones that are connected to zone  $x_t$ , under the graph structure we proposed earlier, by  $M(x_t)$ . The next zone  $x_{t+1}$  visited by the shopper must be a zone that is directly connected to  $x_t$ , i.e.,  $x_{t+1} \in M(x_t)$ . The shopper's choice of "next zone to visit" can thus be viewed as a choose 1-out-of- $n$  choice problem, with  $n$  being the number of zones in  $M(x_t)$ . Following a random utility framework, we define a latent visit utility  $u_{it}^v$  associated with the  $i$ -th zone. Latent utility  $u_{it}^v$  equals the sum of a zone-level baseline visit propensity  $Z_i$ , a "forward-looking" component  $G_{it}$  and a random, extreme-value distributed  $\varepsilon_{it}^v$ . The shopper will visit zone  $i$  in the next step if  $u_{it}^v$  is larger than the latent utility of any of the other zones in the current choice set  $M(x_t)$ .

The shopper is allowed to be "forward looking" (e.g., Sun et al. 2003) when deciding where to visit next. His choice involves a tradeoff between two aspects: (i) the intrinsic attraction of the new zone, and (ii) by going to the new zone, whether he will be closer to other zones of high attraction. We capture this tradeoff by defining  $G_{it}$  as the time-varying attraction of zone  $i$  ( $A_{it}$  as in Equation 1) plus a weighted sum of the attraction of all other zones. The weight associated with zone  $k$  is inversely proportional to the "distance" between zone  $k$  and the

focal zone  $i$ . Specifically, we define the “forward-looking” component of the latent utility of zone  $i$  as:

$$G_{it} = \kappa \left( A_{it} + \sum_{k \neq i} \frac{A_{kt}}{(1 + d_{ik})^\lambda} \right), \lambda \geq 0; \kappa \geq 0 \quad (3)$$

where  $d_{ik}$  denotes the length of the shortest path (on the graph) between zone  $i$  and zone  $k$ .  $\lambda$  is a parameter that governs how the shopper trades off immediate utility with the more forward-looking concern of reaching high attraction regions later on in his trip.  $\lambda = \infty$  means that the shopper is myopic, i.e., only concerned about the attractiveness of what is immediately ahead when making the visitation choice.  $\kappa$  is an individual-level parameter that measures the extent to which visit behavior can be explained by the zone attractions.

With this random utility framework, we can write down the likelihood regarding the shopper’s visit decision at step  $t+1$  (using Equation 3):

$$P(x_{t+1} = i) = P(u_{it}^v \geq u_{kt}^v \forall k \in M(x_t)) \quad (4)$$

$$= \left\{ \frac{\exp \left[ Z_i + \kappa \left( A_{it} + \sum_{l \neq i} \frac{A_{lt}}{(1 + d_{il})^\lambda} \right) \right]}{\sum_{k \in M(x_t)} \exp \left[ Z_k + \kappa \left( A_{kt} + \sum_{l \neq k} \frac{A_{lt}}{(1 + d_{kl})^\lambda} \right) \right]} \right\} \text{ if } i \in M(x_t), 0 \text{ otherwise.}$$

### 3.2.3 Model of visit-to-shop

At each step, the shopper may decide to slow down and shop in the current zone to contemplate a purchase. As we defined earlier,  $H_t$  equals 1 if a visit-to-shop conversion is made at step  $t$ , and 0 otherwise. We model visit-to-shop conversion using a different random utility framework than visitation. We posit that the shopper will perform a visit-to-shop conversion if her latent “shop utility” exceeds zero. Shop utility,  $u_t^s$ , is defined as a linear function of the current zone attraction,  $\alpha_s + \beta_s A_{it}$ , plus random error terms  $\eta_i$  (a zone-specific random effect),

and  $\varepsilon_t^s$ , which is assumed to follow an extreme value distribution.  $\alpha_s$  and  $\beta_s$  are person-specific parameters that capture the shopper's baseline shopping propensity and the extent to which his visit-to-shopping behavior is correlated with latent attractions, respectively. Thus, we have:

$$u_t^s = \alpha_s + \beta_s A_{it} + \eta_i + \varepsilon_t^s \quad (5)$$

$$P(H_t = 1 | \alpha_s, \beta_s, \bar{A}, \eta_i) = P(u_t^s > 0) = \frac{e^{\alpha_s + \beta_s A_{it} + \eta_i}}{1 + e^{\alpha_s + \beta_s A_{it} + \eta_i}}. \quad (6)$$

### 3.2.4 Model of stay time

We model the shopper's stay time and purchase behavior by two different behavioral processes depending on whether she makes a visit-to-shop conversion ( $H_t = 1$ ) or not ( $H_t = 0$ ). If the shopper has made a visit-to-shop conversion in the current zone, we model her stay time using a geometric distribution with parameter  $\tau_{x_t}^{shop}$  (a zone-specific parameter). On the other hand, if the shopper does not make a visit-to-shop conversion in the current zone, we model stay time as a geometric distribution with parameter  $\tau_{x_t}^{pass}$ . We assume that a shopper tends to spend longer in a zone if she is shopping than if she is only passing through. Thus, we assume that  $\tau_i^{pass} > \tau_i^{shop}$  for all  $i$  and parameterize the model by  $\text{logit}(\tau_i^{pass}) = \text{logit}(\tau_i^{shop}) + \delta_i$ ,  $\delta_i > 0$ . Formally,

$$[S_t | H_t = 1] \sim \text{geometric}(\tau_{x_t}^{shop}) \quad (7)$$

$$[S_t | H_t = 0] \sim \text{geometric}(\tau_{x_t}^{pass}) \quad (8)$$

$$\text{logit}(\tau_i^{pass}) = \text{logit}(\tau_i^{shop}) + \delta_i \quad \text{for all } i. \quad (9)$$

### 3.2.5 Model of purchase

As discussed earlier, we assume that purchase in a zone is possible only if a visit-to-shop conversion is made. Thus, if the shopper does not make a visit-to-shop conversion in the current zone ( $H_t = 0$ ), we assume  $B_{jt} = 0$  for all  $j$ .

When a visit-to-shop conversion is made ( $H_t = 1$ ), we model category purchase incidence using a random utility model. The shopper will buy from category  $j$  if it is available in her current zone and its “buy utility” is positive. “Buy utility” of category  $j$  is modeled as a linear function of the attraction of category  $j$ ,  $\alpha_b + \beta_b a_{jt}$ , plus a random error term  $\varepsilon_{jt}^b$ , which is assumed to follow an extreme value distribution. Similar to our model of stay,  $\alpha_b$  and  $\beta_b$  are person-specific parameters that capture the shopper’s baseline buying propensity and the extent to which shop-to-buy behavior is correlated with the latent attractions, respectively. This framework can accommodate impulse buying as well as planned purchase behavior. Our model is similar to the market basket model in Bell and Lattin (1998), where some or all of the categories in a zone may be purchased.

Formally, the random utility model for purchase is set up as follows:

$$u_{jt}^b = \alpha_b + \beta_b a_{jt} + \varepsilon_{jt}^b \quad (10)$$

$$P(B_{jt} = 1 | H_t = 1) = P(u_{jt}^b > 0) = \frac{e^{\alpha_b + \beta_b a_{jt}}}{1 + e^{\alpha_b + \beta_b a_{jt}}} \text{ if } j \in C(x_t), = 0 \text{ otherwise} \quad (11)$$

$$P(B_{jt} = 0 | H_t = 0) = 1 \text{ for all } j. \quad (12)$$

Finally, to obtain the likelihood of a path, we multiply together the likelihood of each of the processes in Figure 10, i.e., visit, stay, and buy, for each step. The overall likelihood of the data can then be calculated by multiplying the likelihoods across all paths. To summarize, through the use of latent attraction variables, our model implicitly links visit, shop, and purchase behaviors together. A graphical depiction of the integrated nature of our model and the relevant parameters is shown in Figure 11.

[Insert Figure 11 about here]

### 3.3 Hierarchical Bayes framework

Since consumers may have heterogeneous category preferences, shopping characteristics, and forward-looking tendencies, we allow the individual-level parameters to follow a Hierarchical Bayes specification. With this setup, each consumer has a different set of parameters, and are related through a common distribution; this allows us to borrow strength across customers to calibrate our model.

The parameter vector for the  $n$ -th consumer,  $(\vec{a}, \kappa, \alpha_s, \alpha_b, \beta_s, \beta_b, \Delta_s, \Delta_b, \lambda)_n$ , is assumed to be drawn from a set of common prior distributions. In the discussion below, we first specify the prior for the attraction vector  $\vec{a}$ , then the prior for the rest of the parameters.

For the attraction vector, we specify

$$\vec{a}_n \sim N(\vec{\mu}_A, \Sigma_A). \quad (12)$$

The variance-covariance matrix  $\Sigma_A$  allows us to borrow strength across categories by taking into account category complementarities. In particular, the  $(j, j')$ -th entry of  $\Sigma_A$  corresponds to the degree of complementarity between category  $j$  and category  $j'$ . For example, if category  $j$  and  $j'$  are complements, given that a person has purchased category  $j$ , we might expect that category  $j'$  is more likely to be purchased in the same trip as well. In this case, one may expect that the entry  $\Sigma_{A(j,j')}$  will be large and positive. In general,  $\Sigma_A$  could be an unrestricted  $N \times N$  matrix, with  $N$  being the number of categories. To reduce the number of parameters, we impose a 2-dimensional factor analytic structure on  $\Sigma_A$ .<sup>6</sup> Other studies that use a

---

<sup>6</sup> Our model can be generalized to include a D-dimensional map. In particular, we fit the model using D=2 and D=3; both model fits and parameter estimates are very similar. Thus, we restrict our attention to the D=2 case for ease of computation.

similar approach to capture dependence structures across categories include Hruschka et al. (1999). Formally, let  $z_j = (z_{j1}, z_{j2})$  be the “spatial position” of the  $j$ -th category. We model  $\Sigma_A$  as

$$\begin{aligned}\Sigma_{A[j,j]} &= \sigma^2 \\ \Sigma_{A[j,j']} (j \neq j') &= \sigma^2 \exp(-\|z_j - z_{j'}\|)\end{aligned}\quad (13)$$

where  $\|z_j - z_{j'}\| = \sqrt{(z_{j1} - z_{j'1})^2 + (z_{j2} - z_{j'2})^2}$ .

For model identification, the variance parameter  $\sigma^2$  is set equal to 1. The variance hyperparameters and the “positions”  $\vec{z} = (z_1, z_2, \dots, z_J)$  are given independent standard Gaussian diffuse priors  $N(0, 100^2)$  and are jointly estimated with other parameters in our model. For model identification, we set the first category at the origin, the second category on the x-axis, and the third category on the y-axis to control for shift, rotation around origin, and reflection about the x-axis respectively (Bradlow & Schmittlein 2000).

The other individual-level parameters (after suitable transformations) are assumed to follow standard multivariate Gaussian hyperpriors. Formally, we specify

$$(\log(\kappa), \alpha_s, \beta_s, \alpha_b, \beta_b, \Delta_s, \Delta_b, \log(\lambda))'_n \sim MVN(\bar{\mu}_I, \Sigma_I) \quad (14)$$

Similarly, zone-level parameters  $(Z_i, \tau_i^{pass}, \delta_i)$  for each zone are assumed to be drawn from a common multivariate Gaussian distribution. Formally,

$$\begin{pmatrix} Z_i \\ \text{logit}(\tau_i^{pass}) \\ \log(\delta_i) \end{pmatrix} \sim MVN(\mu_{ZONE}, \Sigma_{ZONE}). \quad (15)$$

For model identification, the mean hyperparameter associated with  $Z_i$  is set to 0.

### 3.4 Prior specification and estimation procedure

We specify a set of weakly informative priors for all hyperparameters in our model. The hyperparameter for means ( $\bar{\mu}_A, \bar{\mu}_I, \bar{\mu}_{ZONE}$ ) are given independent weakly informative  $N(0, 100^2 \mathbf{I})$  prior distributions, while the hyperparameter for variances ( $\Sigma_A, \Sigma_I, \Sigma_{ZONE}$ ) are given independent weakly informative Inv-Wishart(0.01,  $0.01 \mathbf{I}$ ) prior distributions. These weakly informative prior specifications are chosen to ensure that all posteriors will be proper, while at the same time allowing the data to dominate the posterior inference.

A MCMC procedure allows us to make inferences about our model parameters using samples from their posterior distributions. Conjugate hyperparameters are sampled using the Gibbs sampler (Gelfand and Smith 1990), while other parameters are sampled using a random-walk Metropolis-Hastings algorithm (Chib and Greenberg 1995). The scale of the Gaussian proposal distribution is set to allow for an acceptance rate of around 40%, as suggested by Gilks et al. (1995). To monitor and verify the convergence of the MCMC algorithm, three independent chains are run in parallel from over-dispersed starting points. Potential scale reduction factors (Gelman et al. 2003) are calculated to ensure convergence. Further details of the Gibbs sampler are available upon request.

#### **4. Simulation Study**

Since the proposed model is new to the literature, we perform a simulation study to ensure that our model and estimation procedure are able to produce accurate parameter estimates, and to assess whether the amount of data we have is adequate for model identification. To roughly replicate the size of our actual dataset, we simulate 1000 paths from a set of known parameters shown in Table 2. Then, the MCMC procedure is used to sample from the posterior distributions of our model parameters.

We choose the parameter values used for our simulation as follows. The zone-level parameters  $(Z_i, \tau_i^{pass}, \delta_i)$  are chosen so that the simulated data has similar stay and visit characteristics with the actual data. For the other parameters, the mean vector of category attractions  $\vec{\mu}_A$  is simulated from a  $N(0,1)$  distribution, while the coordinates of the position of each category are generated from a  $N(0,5)$  distribution. The mean vector for individual-level parameters  $\vec{\mu}_I$  is set to  $(0,0,1,0,1,-0.5,-0.2,0)'$ . Finally, the variance-covariance matrix  $\Sigma_I$  is set to  $0.01\mathbf{I}$  to allow shoppers to be heterogeneous in their individual-level parameters.

#### 4.1 Parameter estimation

Estimation results for the hyperparameters that govern the individual-level parameters (besides category attractions) are shown in Table 2. Plots of the true versus estimated parameters for category attractions and zone-level parameters are shown in Figure 12. In each of the panels of Figure 12, the true values of the parameters are plotted on the x-axis while the mean of each posterior distribution is plotted on the y-axis. As can be seen, the true parameter values for the mean category attractions vector  $\vec{\mu}_A$ , zone-level parameters  $Z_i, \tau_i^{pass}$ , and  $\delta_i$ , and the correlation between category attractions are accurately recovered by our estimation procedure.

[Insert Table 2 about here]

[Insert Figure 12 about here]

#### 4.2 Posterior predictive checks

In addition to demonstrating parameter recovery, we also investigate the fit of our model in terms of key summary statistics related to consumer's visit, stay, and buy behavior, as we discussed in Section 2.4. These posterior predictive checks (Gelman et al. 1996) on the simulated dataset allow us to understand how well our model can be expected to fit the summary statistics

of the actual data. To this end, we drew a sample of model parameters from the posterior distribution, and simulated 30 sets of data each with 1000 paths. Then, we calculated the summary statistics for each dataset and compare them against the values obtained from the original simulated dataset.

As can be seen in Figure 13, key summary statistics of the simulated dataset are adequately recovered. The mean absolute errors (MAE), as shown in the bottom three panels of the figure, are very small and indicate a good fit. In the upper three panels, we present histograms of aggregate statistics from the 30 simulated datasets and a vertical line for the original dataset. All three are in the middle of their histograms, again showing a good fit. The method here will allow us to assess the goodness-of-fit obtained when we apply our model to actual data, which we now discuss.

[Insert Figure 13 about here]

## **5. Empirical Application**

In this section, we apply our model to actual PathTracker<sup>®</sup> data. In order to assess the predictive validity of our model, we randomly divide our dataset of 1051 paths into a training sample of 851 paths, and a holdout sample of 200 paths. We calibrate our model on the training sample, and perform a holdout prediction task on the holdout dataset. In Section 5.1, we choose amongst the four aforementioned assignment heuristics and perform a posterior check to ensure that our model is capable of recovering key summary statistics. In Section 5.2, we assess the predictive performance of our model using holdout prediction, and compare our model against three important (nested) sub-models in term of both in-sample and holdout model fit. Finally, parameter estimates and substantive findings are presented in Section 5.3.

### 5.1 Model validation

As we discussed in Section 2.3.3, four different datasets, each of which has purchase locations imputed based on one of the four assignment heuristics we proposed, are used to fit our model. The marginal log-likelihoods for each dataset, calculated using Newton and Raftery’s method (1994), are listed in Table 3. The results in Table 3 suggest that the heuristic “*EVEN*” is most consistent with the actual data.<sup>7</sup> Thus, from here on, we will focus on the dataset with the category purchase locations imputed using this heuristic.

[Insert Table 3 about here]

We follow the same procedure as we did for the simulation study to assess how well our model recovers the set of summary statistics we considered in Section 4.2. We simulate 100 datasets from the posterior distribution of the model parameters, each with 851 paths (which replicates the size of our calibration dataset). Then, we calculate key summary statistics from each dataset, and compare them against those calculated from the actual data. The results are shown in Figure 14.

[Insert Figure 14 about here]

Figure 14 shows that our model recovers key summary statistics of the actual data fairly well. The top three panels show that data simulated from our posterior predictive distribution are able to replicate the key visit, stay, and purchase statistics of the dataset. The bottom three panels show that data simulated from our model have similar characteristics to the actual data in terms of average stay time (in minutes) in each zone, penetration of each product category, and zone-to-zone transition probabilities. Note that the fit of the transition probabilities has a MAE of 0.08, slightly higher than for the simulation results. This indicates that there are some aspects that affect movement that are not fully captured by our parsimonious model. Overall, however, we feel that the model fit is adequate with respect to these summary statistics.

---

<sup>7</sup> The posterior checks performed also showed that heuristic *EVEN* led to a better fit compared to the other heuristics.

## 5.2 Holdout prediction and model comparison

We perform a holdout prediction test on the 200 holdout paths to assess the out-of-sample predictive validity of the model. For each trip, we derive the posterior distribution of their individual-level parameters, using only the first half of each path to calibrate the model. The marginal log-likelihood of the holdout sample is computed, again using Newton and Raftery's (1994) importance sampling approach. Then, using the posterior distribution of their individual-level parameters, we draw 100 sample paths to complete each shopping trip. The summary statistics between the actual (holdout) dataset and the simulated paths are considered; the results are shown in Figure 15. Although (as expected) the model fit is worse than the in-sample fit, our model still provides a fairly good fit to the holdout data.

[Insert Figure 15 about here]

We also tested our model performance against benchmark models. To assess the extent to which the integrative nature of our model adds to its performance, we test the full model against nested sub-models that explicitly disables the linkage between purchase and visit/shop behavior. Both in-sample and holdout marginal log-likelihood are considered. The three submodels considered are as follows (see Figure 11):

Submodel I ( $\beta_s = 0$ ): By setting the parameter  $\beta_s$  to zero, the linkage between purchase and shopping/staying behavior is disabled.

Submodel II ( $\kappa = 0$ ): Setting  $\kappa$  to zero disables the linkage between purchase and visit behavior.

Submodel III ( $\lambda \rightarrow \infty$ ): Setting  $\lambda$  to infinity, as described in Section 3.2.2, will imply that consumers are myopic.

The results, as shown in Table 4, suggest that our full model provides a better description of the data (in terms of in-sample fit) and better holdout predictive performance (with respect to predictive log-likelihood) than any of the reduced submodels. This provides some evidence that our full integrated model is closer to actual behavior than the reduced models considered in Submodels I, II, and III.

[Insert Table 4 about here]

### 5.3 Parameter estimates and interpretation

The posterior distribution of the hyperparameters that govern the individual-level parameters are summarized in Table 5. These results offer a number of immediate insights. First, the reasonably large estimates of  $\kappa$  (mean of  $\log(\kappa)$  is -1.54) suggests that purchase behavior is indeed interrelated with visitation patterns. Second, the estimates for both  $\mu_{\beta_s}$  and  $\mu_{\beta_b}$  are positive, indicating that attractions are positively correlated with both visit-to-shop and shop-to-purchase decisions. Third, the estimates for both  $\mu_{\Delta_s}$  and  $\mu_{\Delta_b}$  are negative, suggesting that the attraction of a zone tends to decrease after a consumer visits the zone and/or purchases the product categories that it carries. This first finding regarding visitation is consistent with Soman and Shi (2003), who found that people in general tend to avoid (and dislike) backward-progression when deciding on a travel plan. Finally, the small estimates of  $\mu_\lambda$  suggests that consumers exhibit a certain (but not massive) degree of forward-looking behavior in their shopping paths. This is consistent with the finding in NEIO models, as well as the analysis in Hui et al. (2006b) where the researchers find evidence of forward-looking behavior for grocery shoppers.

[Insert Table 5 about here]

The posterior mean for the baseline attraction of each category is summarized in Table 6. Since purchase incidence is driven, in large part, by category attraction, we expect that category attractions should be positively correlated with simple purchase incidence statistics. Indeed, we find that the correlation between category attractions and purchase incidence is positive and highly significant ( $r = 0.58$ ;  $p < 0.001$ ). The product category that has the highest attraction is Fruit, with a posterior mean attraction of 2.70, which also has the highest observed purchase incidence (53.8%). In contrast, the second highest attraction category, Natural/Organic Food, has a very low observed purchase incidence (2.5%). This lack of purchase may be explained by the product's in-store location, and may suggest the possibility of relocating the Natural/Organic Food category. This shows the power and value of the model in its ability to sort out the inherent attractions of products per se from the regions of the store in which they reside. We explore this type of issue in more detail in the next section, by conducting a hypothetical policy experiment.

[Insert Table 6 about here]

Finally, we look at the different zone-level parameter estimates. The estimates for the parameter  $\tau^{shop}$  and  $Z_i$  for each zone are displayed in the form of a choropleth map (Banerjee et al. 2004) in Figures 16 and Figure 17 respectively. As expected, zones with low  $\tau^{shop}$  (and hence a long mean shopping time) generally correspond to zones where shoppers spend longer time. The correlation between  $\tau^{shop}$  and average observed time spent in the zone is negative and highly significant ( $r = -0.39$ ;  $p < 0.001$ ). On the other hand, the zones with high  $Z_i$  corresponds to zones which are visited more often. The correlation between  $Z_i$  and observed zone penetration is positive and highly significant ( $r = 0.43$ ;  $p < 0.001$ ).

[Insert Figure 16 about here]

[Insert Figure 17 about here]

## 6. Potential Managerial Applications

In this section, we demonstrate one set of potential managerial applications of our model by examining how the relocation of some product categories may affect consumers' shopping patterns. Before we begin, several important caveats are in order. First, our model is descriptive in nature. Since our dataset contains only one store, this policy experiment is an “extrapolation” outside of our calibration sample. Although we conducted a holdout task in Section 5.2 to provide some predictive validity, it was performed *conditional on a fixed store layout* and thus did not assess predictive performance when the store layout is changed. (A cross-store study, in future research, may alleviate this issue). Second, we assume that model parameters remain unchanged when the store layout is altered. This assumption may not hold in reality. This problem, also known as the Lucas Critique (Van Heerde et al. 2005), is a valid concern that can only be directly addressed by a field experiment, where the store layout is actually changed and its impact recorded. With these caveats, we stress that the following policy experiment serves as an illustration of the kinds of analyses that may be possible when our model is developed into a full-fledged decision tool for retailers.

In general, the penetration of a product category may be driven by both the product itself and by its in-store location. Taken together, these two factors determine the amount of traffic that a product category receives, and the conversion rate given that a shopper visits it. Thus, many retailers are interested in knowing how a category's penetration may change if it is relocated to another part of the store. In our experiment, we aim to obtain some insights into this issue. We use two different categories, Canned Vegetables and Pasta, to illustrate our approach. Both categories have moderate penetration (12.7% and 6.9%), and are located in different areas of the store; Canned Vegetables is located in zone 47 and 61, while Pasta is located in zone 30 (see

Table 1). We now switch the position of the two categories to examine how consumers' shopping patterns and purchase behavior will be hypothetically affected.

We simulated 100 datasets using the posterior distribution of parameters computed in Section 5.3, on the modified supermarket. Under the new store layout, the predicted penetration of Canned Vegetables increases from 12.5% to 16.2% ( $p < 0.001$ )<sup>8</sup>. This increase, however, is partly offset by the loss in predicted penetration of Pasta, which decreases from 7.7% to 5.5% ( $p < 0.001$ ). We also examine the impact of this purchase location switch on the predicted penetrations of other product categories. We find that the categories Coffee, Cereal, Canned Dried Fruit, Dried Beans/Peas, and Ethnic (TexMex) all show a slight (-0.8% on average) but significant decrease in predicted penetration ( $p < 0.05$ ). One interesting observation is that these categories are all located at the bottom-middle area of the store. This may indicate that the relocation of the two focal categories causes a certain change in consumers' paths through the store, and hence indirectly reduces the penetration of those categories. As a result of these and other changes, the overall average predicted basket size decreases slightly from 6.42 to 6.37; the change in basket size, however, is not statistically significant ( $p = 0.54$ ).

The above results lead to some interesting insights. It appears that switching the location of Canned Vegetable and Pasta categories affects not only their penetrations, but also those of other related categories as well. Thus, retailers need to carefully consider cross-category effects (i.e., the impact of a category's location/sales on other categories' sales) when making product placement decisions, and through our formal model of the underlying category and zone attractions one can capture those effects. Although the aforementioned results are preliminary in nature (and subject to the aforementioned caveats), they highlight the value of our integrated framework in modeling consumers' shopping path and purchase behavior.

---

<sup>8</sup> All p-values here are calculated with respect to the respective posterior predictive distributions.

## 7. Discussion and Conclusion

Many studies in marketing have focused on predicting brand choice or market baskets. In this paper, we take a step forward by incorporating the consumers' shopping paths in a joint modeling framework. To the best of our knowledge, this study is the first attempt in understanding consumers' in-store shopping paths and their purchases through an integrated statistical model. Using a set of latent variables that describe the "attraction" of each product category and zone, our model integrates three aspects of grocery shopping: (1) where shoppers visit and their zone-to-zone transitions, (2) whether (and for how long) they stay and shop in each zone, and (3) what product categories they purchase. Similar to NEIO models, we allow for forward-looking shoppers, and the data indeed suggest that they seem to exhibit some degree of "looking aheadness" as they move around the store.

We then applied our model to a sample of PathTracker<sup>®</sup> data provided by Sorensen Associates. Our model is able to replicate the data closely (in and out of sample) on various key summary statistics with respect to consumer visit, stay, and buy behavior. Although the primary goal of our model is descriptive, we demonstrate, through a policy experiment, its potential application to analyze the effect of store layout on purchase patterns. We acknowledge that the results are only preliminary in nature, but they provide a glimpse at the numerous managerial insights that may be possible from a broader use of this kind of model.

Finally, our model allows extensions in many directions. We conclude here by describing several of these extensions, in particular when additional information such as consumer characteristics and category/product-level profit margins become available.

(1) *Field experiment*: Given the profit margin of each product category and its capacity requirements, our model can be used to experiment with different store layouts through

simulations and thus offer recommendations on store layout and product placements. In collaboration with grocery retailers, these recommendations can be examined using a field experiment, by actually changing the location of product categories and comparing category sales/penetration before and after the change. As we pointed out earlier, carrying out these experiments is crucial if one wishes to use our model as a decision tool to improve store layout.

(2) *Cross-store study*: The PathTracker<sup>®</sup> system is being installed in an increasing number of supermarkets (and other types of retail stores) to track consumers' shopping patterns. Our model can easily be applied to the other stores to conduct a cross-store study, to study how store characteristics (e.g., square footage, number of aisles) are related to consumers' movement tendencies and shop/purchase behavior.

(3) *Data-driven zone definition*: To further extend our work to study paths and purchases across different stores, the division of a store into zones needs to be “automated”. Currently, the discretization of our store into regions (see Figure 2) is done manually based on discussions with Sorensen Associates. In general, it would be helpful to identify a set of zones that would “optimally” discretize the store and the associated shopping paths. This problem of identifying optimal zone boundaries is related to “wombling” (boundary analysis) techniques, a recent research area in spatial statistics (e.g., Lu and Carlin 2005).

(4) *Consumer characteristics*: The Hierarchical Bayes framework allows us to obtain individual-level parameters for each consumer. If consumer covariates (e.g., demographics/socioeconomics, attitudinal measures, and other behavioral data) were also available, for example, by bringing in data from a store loyalty card program, we can link these covariates to our model parameters.

Formally, we can extend Equation (14) as follows (Rossi et al. 2006):

$$(\log(\kappa), \alpha_s, \beta_s, \alpha_b, \beta_b, \Delta_s, \Delta_b, \log(\lambda))'_n \sim MVN(y'_n \gamma, \Sigma_I) \quad (16)$$

where  $y'_n$  denotes a vector of individual-level covariates for the  $n$ -th consumer. With this framework, our model may then offer empirical testing of different hypotheses that behavioral researchers are interested in. For instance, by studying the relationship between the coefficients  $\kappa$  and  $\lambda$  and consumer demographics, we may learn how forward-looking tendencies differ across shoppers of different gender and age (e.g., Otnes and McGrath 2001). Similarly, we can link individual-level preference for product categories, shopping characteristics, and movement patterns with individual-level demographics. One particularly interesting research direction is to study the “efficiency” of different types of shoppers (e.g., Hui et al. 2006b).

(5) *Dynamic promotions*: In the future, retailers may be able to offer promotions to shoppers in “real-time” during their grocery trip. Already, grocery stores in Europe have started to introduce portable scanning devices with which shoppers can scan their items as they put it into their carts; in the U.S., future plans are being made to provide information and generate offers dynamically (e.g., [www.mediacart.com](http://www.mediacart.com)). When such technology becomes more mature and widely available, retailers can use our model to identify consumers to whom they should offer coupons for certain products. For example, they may find that consumers with certain shopping patterns or “look-aheadness” are more price sensitive, and thus this line of research may help retailers target promotion efforts in a more efficient way. More generally, one may want to study how different components of the marketing mix affects model parameters; for example, how price affects category attractions, and how signage impacts baseline visit propensities, etc.

The study of paths and related behaviors extends well beyond the applications outlined in this paper. Path data, which includes the movement patterns of animals, traffic, and pedestrians, have been studied extensively in other fields. In marketing, path data arise naturally from eye-tracking applications, web clickstream data, or even Information Acceleration sessions (Hui et al.

2006a). As consumer-tracking technology (e.g., RFID) becomes more commonplace, we expect that path-related data will become more widely available and cost efficient in the near future.

The collection and analysis of paths to understand consumer behavior may one day become widespread in marketing, much like the analysis done on scanner data today. Thus, we believe that marketing researchers may benefit from a deeper study of path data, perhaps by borrowing analytical approaches from psychology, economics, and sociology, and the first step in marketing provided here.

## References

- Antonini, Gianluca, Michel Bierlaire, and Mats Weber (2006), "Discrete Choice Models of Pedestrian Walking Behavior," *Transportation Research B*, 40, 667-687.
- Banerjee, Sudipto, Bradley P. Carlin, and Alan E. Gelfand (2004), *Hierarchical Modeling and Analysis of Spatial Data*, Chapman and Hall.
- Bell, David R. and James M. Lattin (1998), "Shopping Behavior and Consumer Preference for Store Price Format: Why Large Basket' Shoppers Prefer EDLP," *Marketing Science*, 17(1), 66-88.
- Bergman, Carita M., James A. Schaefer, and S.N. Luttich (2000), "Caribou Movement as a Correlated Random Walk," *Oecologia*, 123, 364-374.
- Bloom, Paul N., Gregory T. Gundlach, and Joseph P. Cannon (2000), "Slotting Allowances and Fees: Schools of Thought and the Views of Practicing Managers," *Journal of Marketing*, 64(2), 92-108.
- Bradlow, Eric T., and David C. Schmittlein (2000), "The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science*, 19(1), 43-62.
- Camerer, Colin, and Teck-Hua Ho (1999), "Experience-weighted Attraction Learning in Normal Form Games," *Econometrica*, 67(4), 827-874.
- Chib, Siddhartha, and Edward Greenberg (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49(4), 327-335.
- Chintagunta, Pradeep K. (1993), "Investigating Purchase Incidence, Brand Choice and Purchase Quantity Decisions of Households," *Marketing Science*, 12(2), 184-208.
- Farley, John U., and L. Winston Ring (1966). A Stochastic Model of Supermarket Traffic Flow. *Operations Research*, 14(4), 555-567.
- Gelfand, A. E., and A. F. M. Simth (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.
- Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin (2003), *Bayesian Data Analysis, 2<sup>nd</sup> Ed.* Chapman & Hall.
- Gelman, Andrew, Xiao-Li Meng, and Hal Stern (1996), "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies," *Statistica Sinica*, 6, 733-807.
- Gilks, W. R., S. Richardson, D. J. Spiegelhalter (1995), *Markov Chain Monte Carlo in Practice.* Chapman & Hall.

- Guadagni P. M. & J. D. C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2(3), 203-238.
- Halliday, David, Robert Resnick, Jearl Walker (2004), *Fundamental of Physics, 7<sup>th</sup> Ed.* John Wiley & Sons.
- Heller, Walter (1988), "Tracking Shoppers Through the Combination Store," *Progressive Grocer*, 47-64.
- Hui, Sam K., Peter S. Fader, and Eric T. Bradlow (2006a), "Path Data in Marketing: An Integrative Framework and Prospectus for Model-Building," Working Paper.
- Hui, Sam K., Peter S. Fader, and Eric T. Bradlow (2006b), "The Traveling Salesman Goes Shopping: The Systematic Inefficiencies of Grocery Paths," Working Paper.
- Hruschka, Harald, Martin Lukanowicz, and Christian Buchta (1999), "Cross-Category Sales Promotion Effects," *Journal of Retailing and Consumer Services*, 6, 99-105.
- Larson, Jeffrey S., Eric T. Bradlow and Peter S. Fader (2005), "An Exploratory Look at Supermarket Shopping Paths," *International Journal of Research in Marketing*, 22, 395-414.
- Little, R.J.A., and Rubin D.B. (1987), *Statistical Analysis with Missing Data*, J. Wiley & Sons.
- Lord, F.M., and M.R. Novick (1968), *Statistical Theories of Mental Test Scores*, Addison-Wesley.
- Lu, H, and Carlin B.P. (2005), "Bayesian Areal Wombling for Geographical Boundary Analysis," *Geographical Analysis*, 37, 265-285.
- Lummus, Rhonda R., and Robert J. Vokurka (1999), "Defining Supply Chain Management: A Historical Perspective and Practical Guidelines," *Industrial Management and Data Systems*, 99(1), 11-17.
- Manchanda, P., A. Ansari and S. Gupta (1999), "The Shopping Basket: A Model for Multicategory Purchase Decisions," *Marketing Science*, 18 (2), 95-114.
- McFadden, D. L. (1981), *Structural Analysis of Discrete Data with Econometric Applications*. MIT press.
- Montgomery, Alan L., Shibo Li, Kannan Srinivasan, and John C. Liechty (2004), "Predicting Online Purchase Conversion Using Web Path Analysis," *Marketing Science*, 23(4), 579-595.
- Newton, Michael A., and Adrian E. Raftery (1994), "Approximating Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society B*, 56 (1), 3-48.

- Otnes, Cele, and Mary Ann McGrath (2001), "Perceptions and Realities of Male Shopping Behavior," *Journal of Retailing*, 77, 111-137.
- Park, Young-Hoon, and Eric T. Bradlow (2005), "An Integrated Model for Bidding Behavior in Internet Auctions: Whether, Who, When, and How Much," *Journal of Marketing Research*, 42(4), 470-482.
- Pieters, Rik, Edward Rosbergen, and Michel Wedel (1999), "Visual Attention to Repeated Print Advertising: A Test for Scanpath Theory," *Journal of Marketing Research*, 16, 424-438.
- Rossi, Peter E., Greg M. Allenby, and Robert McCulloch (2006), *Bayesian Statistics and Marketing*, Wiley.
- Soman, Dilip, and Mengze Shi (2003), "Virtual Progress: The Effect of Path Characteristics on Perceptions of Progress and Choice," *Management Science*, 49(9), 1229-50.
- Song, Inseong, and Pradeep K. Chintagunta (2003), "A Micromodel of New Product Adoption with Heterogeneous and Forward-Looking Consumers: Application to the Digital Camera Category," *Quantitative Marketing and Economics*, 1(4), 371-407.
- Sorensen, Herb (2003), "The Science of Shopping," *Marketing Research*, 15(3), 30-35.
- Sun, Baohong, Scott A. Neslin, and Kannan Srinivasan (2003), "Measuring the Impact of Promotions on Brand Switching when Consumers are Forward-Looking," *Journal of Marketing Research*, 40 (Nov), 389-405.
- Van Heerde, Harald J., Marnik G. Dekimpe, and William P. Putsis (2005), "Marketing Models and the Lucas Critique," 42 (1), 15-21.
- Vrechopoulos, Adam P., Robert M. O'Keefe, Georgios I. Doukidis, and George J. Siomkos (2004), "Virtual Store Layout: An Experimental Comparison in the Context of Grocery Retail," *Journal of Retailing*, 80, 13-22.

Category Name	Zones	%buy
Fruit	2,4	53.8%
Vegetables	3,4,5	50.4%
Butter/Cheese/Cream	38,39,82,83	38.0%
Carbonated Beverages	16,21,22,23	24.2%
Salty Snacks	62,63,64,92	23.2%
Cookies/Crackers	18,44,45,46,47,93	22.6%
Milk	38	22.6%
Ice Cream	57,58,59,60	19.6%
Bread	52,53,61,93	19.4%
Candy/Gum/Mints	60,91,92	17.3%
Cereal	49,50,94	17.1%
Eggs	36	14.7%
Canned Vegetables	47,61	12.7%
Baking Ingredients	18,24,25,26,27	12.2%
Frozen Prepared Dinners	55,56	11.9%
Drinks (others)	52,53,94	11.9%
Yogurt	81	11.5%
Pasta Sauce	14,30	11.2%
Fruit Juice	36	10.8%
Canned Dried Fruit	20,95	10.8%
Pet Care	60,65,66,67	10.7%
Meat/Poultry/Seafood Manufactured Prepack	31,35	10.3%
Canned Soup	44,61	9.7%
Frozen Pizza Snacks	55,56	9.1%
Bath Tissue	37,77	9.0%
Frozen Vegetables	54	8.6%
Peanut Butter/Jams	48,61	7.7%
Bottled Water	23,40	7.6%
Prepared Food/Dried Dinners	29,95	7.4%
Frozen Meat/Poultry/Seafood	54,56	7.0%
Pasta	30	6.9%
Frozen Drinks	57	6.1%
Pastry/Snack Cakes	51	5.8%
Granola Bars	19,94	5.3%
Bagels/Breadsticks	52,53,73	5.2%
Spices/Seasonings	16,26,46,95	4.9%
Magazines	77,91,92	4.9%
Condiments/Sauces	24,25,26	4.7%
Frozen Baked Goods	57,58	4.6%
Tobacco	90,91	4.6%
Household Cleaners	78,79	4.4%
Facial Tissue	76,84	4.4%
Paper Towels	37,75	4.4%
Coffee	50	4.3%
Frozen Potatoes/Onions	54	4.2%
Oral Care	74,91,92	4.2%
Prepackaged Deli Meat	34	4.2%
Frozen Dessert/Fruit	58,93	4.0%
Canned Seafood	40	3.7%
Non-Refrigerated Dressings	25	3.6%
Disposable Tableware	69,94	3.6%
Olives/Peppers/Pickles	24	3.5%
Dough Products	39	2.9%
OTC Medicines	74,88,91,92	2.9%
Beer	62,63,93	2.9%
Non-Carbonated Flavored Drinks	51	2.8%
Skin/eye care	84,85,86,87,88	2.6%

Category Name	Zones	%buy
Shampoo/Conditioner	81,82	2.5%
Laundry Supplies	78,79	2.5%
Natural/Organic Food	7	2.5%
Pudding/Dry Dessert	25	2.1%
Rice	42	2.1%
Shelf-Stable Milk	27	1.9%
Bakery Service	8,10	1.7%
Hot Beverage Add-Ins	49	1.7%
Canned RTE Meat Entrées	40	1.7%
Baby Food	71	1.6%
Stationery/School Supplies	69,70	1.6%
Wine	28,29	1.5%
Refrigerated Snacks	81	1.5%
Ethnic (Oriental)	41	1.5%
Ethnic (TexMex)	43	1.5%
Toaster Pastries	48	1.4%
Paper and Plastic Bags	68	1.4%
Special Diet Items	9	1.4%
Cooking Oil	27	1.3%
Salad Add-Ins	27	1.3%
Natural/Organic Snacks	11	1.3%
Canned Meat	40	1.2%
Toiletries	87,90,91,92	1.2%
Meat/Poultry/Seafood Fresh Prepack	32	1.2%
Ethnic (Hispanic)	43	1.1%
Rolls/Buns/Pitas	52,53	1.0%
Prepackaged Deli Prepared Lunch	14	1.0%
Prepared Food/Potatoes	45	1.0%
Tea	49	0.9%
Frozen Dough/Bread/Bagel	58	0.9%
Electronic Media	89	0.9%
Cosmetics/Deodorant	86	0.9%
Pancake/Syrup	26,48	0.9%
Deli Prepack	13,15	0.8%
Feminine Hygiene	72	0.7%
Dry Soup	45	0.7%
Hagazines	42,43	0.6%
Baby Medical Needs	71,72	0.6%
Baking Supplies	61	0.6%
Hair Color Accessories	83	0.6%
Batteries	80,84	0.5%
Light Bulbs	80	0.5%
Office Supplies	75	0.5%
Plastic Wrap	68	0.5%
Deli Service	12	0.4%
Dried Beans/Peas	43,47	0.4%
Natural/Organic Drinks	11	0.4%
Aluminum Foil	68	0.4%
Napkins	76	0.4%
Hot Chocolate Mix	49	0.3%
Deli Amenities	15	0.3%
Automotive Supply	67	0.1%
Apparel	73	0.1%
Meat/Poultry/Seafood Fresh Service	17,31	0.1%
Meat/Poultry/Seafood Fully/Partially Cooked	33	0.1%
Floral	2,6	0.0%
Natural/Organic (Others)	7	0.0%

Table 1. Locations of product categories.

	True value	Posterior Mean	Posterior Standard Deviation
$\mu_{\kappa}$	0.000	-0.001	0.009
$\mu_{\alpha_s}$	0.000	0.008	0.023
$\mu_{\beta_s}$	1.000	1.003	0.012
$\mu_{\alpha_b}$	0.000	0.007	0.009
$\mu_{\beta_b}$	1.000	1.022	0.016
$\mu_{\Delta_s}$	-0.500	-0.496	0.006
$\mu_{\Delta_b}$	-0.200	-0.203	0.010
$\mu_{\lambda}$	0.000	-0.010	0.010

Table 2. Estimation results for model hyperparameters in simulation study.

Heuristic	Log-likelihood
<i>EVEN</i>	-468673.0
<i>LONGEST</i>	-469721.3
<i>PROPORTIONAL</i>	-470808.3
<i>LAST</i>	-473603.4

Table 3. Comparison of different assignment heuristics.

	In-sample Marginal LL	Holdout marginal LL
Full Model	-468673.0	-112350.4
Submodel I ( $\beta_s = 0$ )	-470408.3	-112921.0
Submodel II ( $\kappa = 0$ )	-477039.0	-113078.1
Submodel III ( $\lambda \rightarrow \infty$ )	-470284.6	-112719.4

Table 4. Comparison between full model and Submodels I, II, and III.

	Posterior Mean	Posterior S.D.	95% Posterior Interval
$\mu_{\kappa}$	-1.364	0.018	(-1.399, -1.331)
$\mu_{\alpha_s}$	-1.608	0.075	(-1.711, -1.475)
$\mu_{\beta_s}$	0.466	0.023	( 0.431, 0.506)
$\mu_{\alpha_b}$	-2.544	0.041	(-2.621, -2.480)
$\mu_{\beta_b}$	1.189	0.031	( 1.135, 1.247)
$\mu_{\Delta_s}$	-0.341	0.010	(-0.360, -0.323)
$\mu_{\Delta_b}$	-0.201	0.012	(-0.223, -0.181)
$\mu_{\lambda}$	-0.751	0.017	(-0.782, -0.713)

Table 5. Estimation results for model hyperparameters in the actual data.

Description	Attraction	Description	Attraction
Fruit	2.70	Wine	0.03
Natural/Organic Food	2.24	Toaster Pastries	0.03
Special Diet Items	2.04	Bakery Service	0.02
Butter/Cheese/Cream	1.80	Olives/Peppers/Pickles	0.02
Salty Snacks	1.62	Natural/Organic Drinks	-0.03
Vegetables	1.59	Plastic Wrap	-0.06
Pastry/Snack Cakes	1.54	Cooking Oil	-0.09
Cereal	1.47	Eye Care	-0.09
Yogurt	1.27	Frozen Baked Goods	-0.10
Canned Vegetables	1.27	Deli Prepack	-0.10
Milk	1.22	Frozen Drinks	-0.12
Canned Dried Fruit	1.15	Oral Care	-0.20
Drinks (others)	1.05	Refrigerated Snacks	-0.23
Paper and Plastic Bags	1.04	Shelf-Stable Milk	-0.23
Cookies/Crackers	1.03	Frozen Meat/Poultry/Seafood	-0.24
Rice	1.00	Baby Food	-0.27
Facial Tissue	1.00	Magazines	-0.28
Meat/Poultry/Seafood Manufactured Prepack	0.99	Spices/Seasonings	-0.30
Frozen Vegetables	0.97	Light Bulbs	-0.31
Bath Tissue	0.95	Dried Beans/Peas	-0.32
Canned Seafood	0.90	Disposable Tableware	-0.32
Frozen Prepared Dinners	0.87	Frozen Dessert/Fruit	-0.33
Tobacco	0.86	Ethnic (TexMex)	-0.38
Baby Medical Needs	0.86	Deli Amenities	-0.39
Hot Beverage Add-Ins	0.82	Pancake/Syrup	-0.40
Aluminum Foil	0.77	Coffee	-0.44
Frozen Pizza Snacks	0.77	Hair Color Accessories	-0.44
Carbonated Beverages	0.72	Hard Liquor	-0.48
Prepared Food/Dried Dinners	0.66	Meat/Poultry/Seafood Fresh Service	-0.48
Pasta Sauce	0.64	Bagels/Breadsticks	-0.49
Cosmetics/Deodorant	0.64	Diapers	-0.51
Canned Soup	0.63	Salad Add-Ins	-0.51
Ice Cream	0.62	Stationery/School Supplies	-0.54
Non-Carbonated Flavored Drinks	0.60	Paper Towels	-0.68
Prepared Food/Potatoes	0.60	Frozen Potatoes/Onions	-0.69
Baking Ingredients	0.57	Toiletries	-0.75
Laundry Supplies	0.52	Beer	-0.78
Pudding/Dry Dessert	0.49	OTC Medicines	-0.78
Natural/Organic Snacks	0.48	Meat/Poultry/Seafood Fully/Partially Cooked	-0.80
Bread	0.48	Canned Meat	-0.83
Granola Bars	0.47	Automotive Supply	-0.84
Candy/Gum/Mints	0.46	Rolls/Buns/Pitas	-0.88
Pet Care	0.44	Skin care	-0.90
Dough Products	0.44	Dry Soup	-0.96
Non-Refrigerated Dressings	0.38	Meat/Poultry/Seafood Fresh Prepack	-1.01
Canned RTE Meat Entrées	0.36	Napkins	-1.02
Feminine Hygiene	0.30	Peanut Butter/Jams	-1.09
Prepackaged Deli Meat	0.30	Hot Chocolate Mix	-1.12
Ethnic (Oriental)	0.27	Batteries	-1.14
Pasta	0.26	Natural/Organic (Others)	-1.22
Condiments/Sauces	0.19	Shampoo/Conditioner	-1.42
Frozen Dough/Bread/Bagel	0.18	Office Supplies	-1.46
Electronic Media	0.18	Apparel	-1.60
Bottled Water	0.13	Baking Supplies	-1.63
Household Cleaners	0.10	Prepackaged Deli Prepared Lunch	-1.65
Eggs	0.08	Floral	-1.70
Ethnic (Hispanic)	0.05	Deli Service	-1.74
Fruit Juice	0.04	Tea	-1.96

Table 6. Posterior mean for category attractions, sorted in decreasing order.

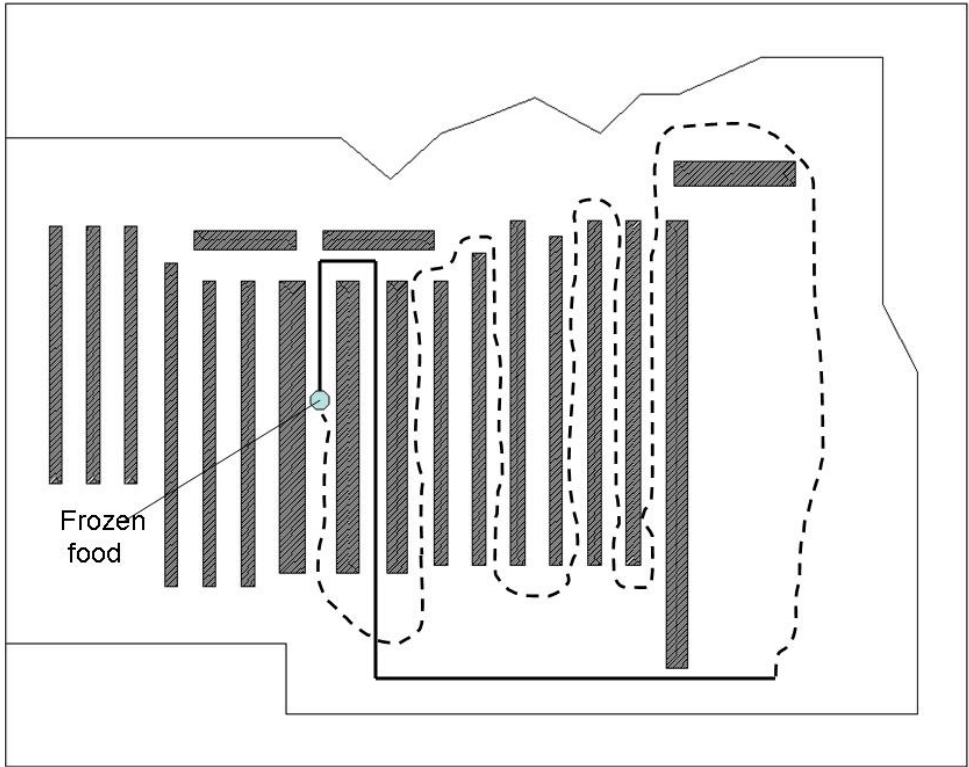


Figure 1. The paths of Alfred (dashed line) and Bianca (solid line).

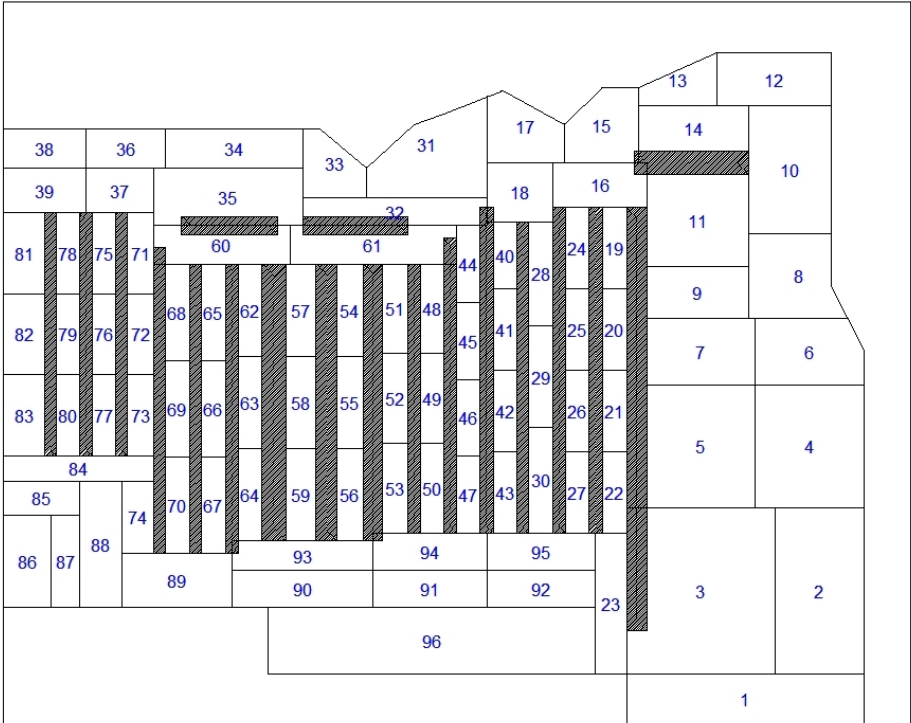


Figure 2. Grocery store divided into 96 zones.

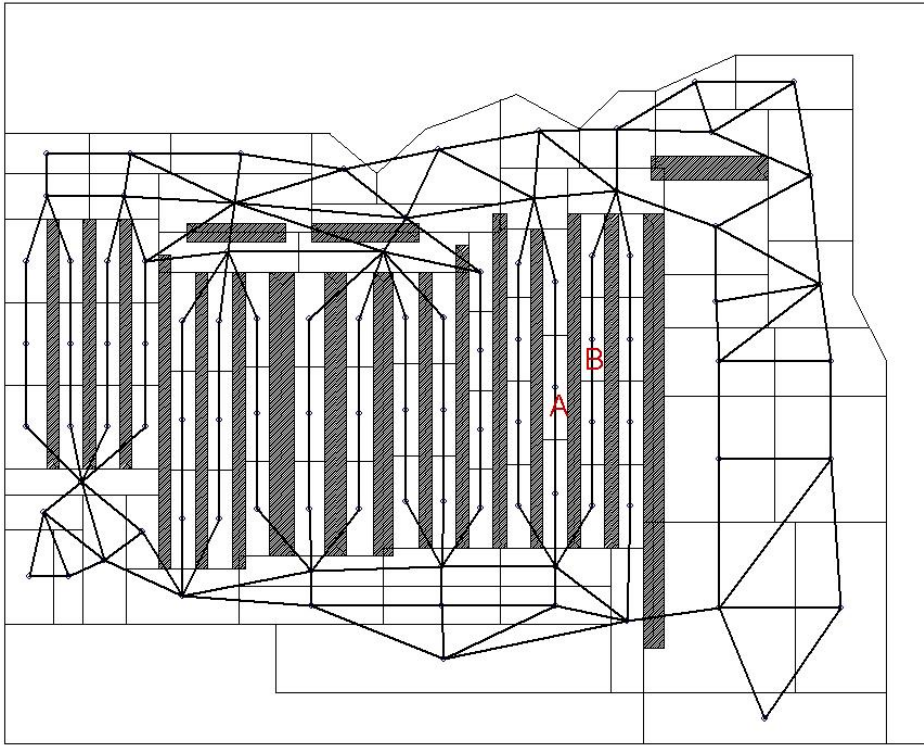
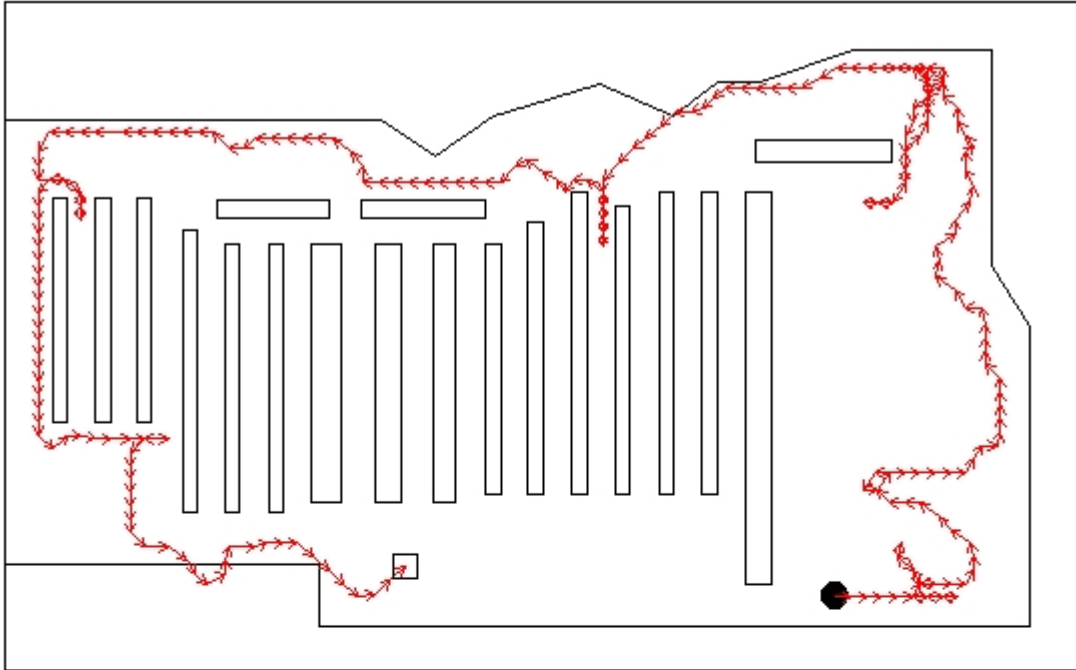


Figure 3. Grocery store represented by a graph of 96 nodes.

### Raw Path



### Path after discretization

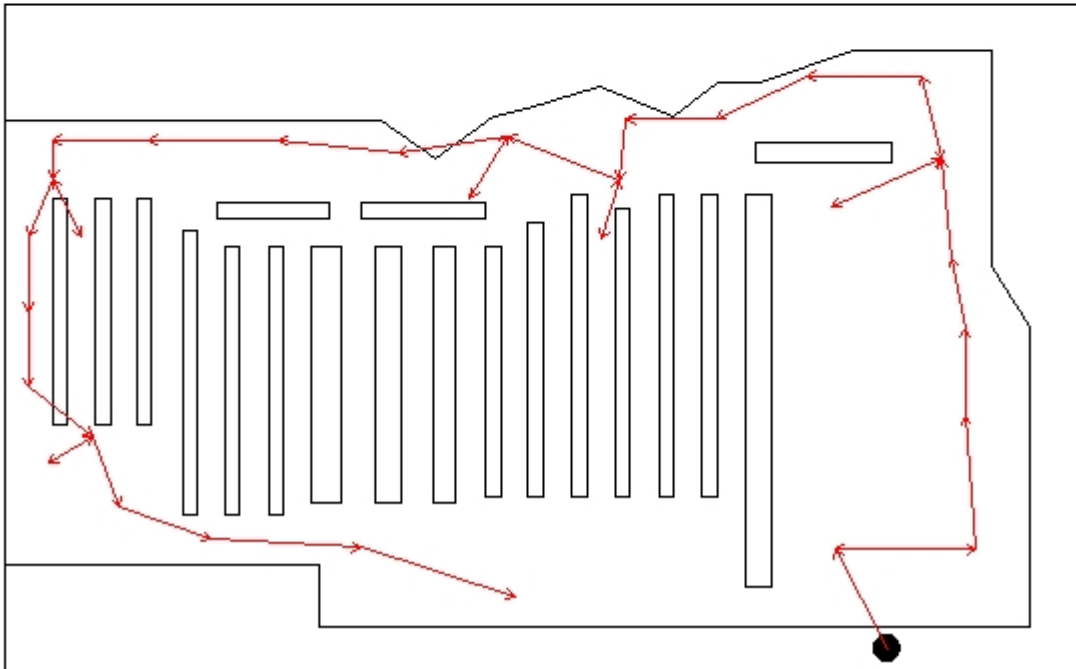


Figure 4. Example of path discretization.

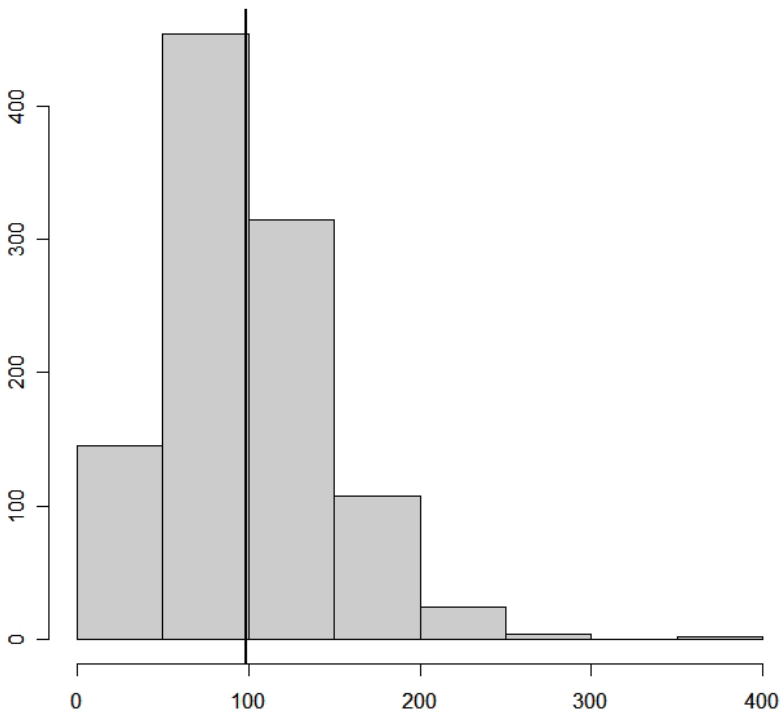


Figure 5. Histogram of number of steps (vertical line denotes the mean).

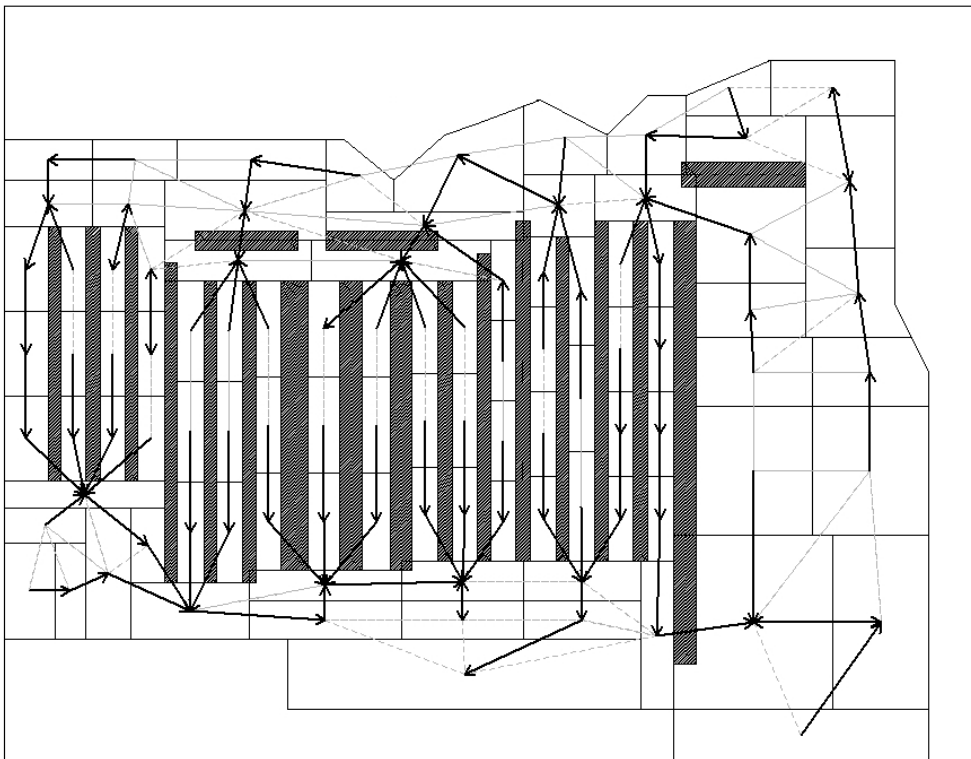


Figure 6. Most frequent transition out of each zone.

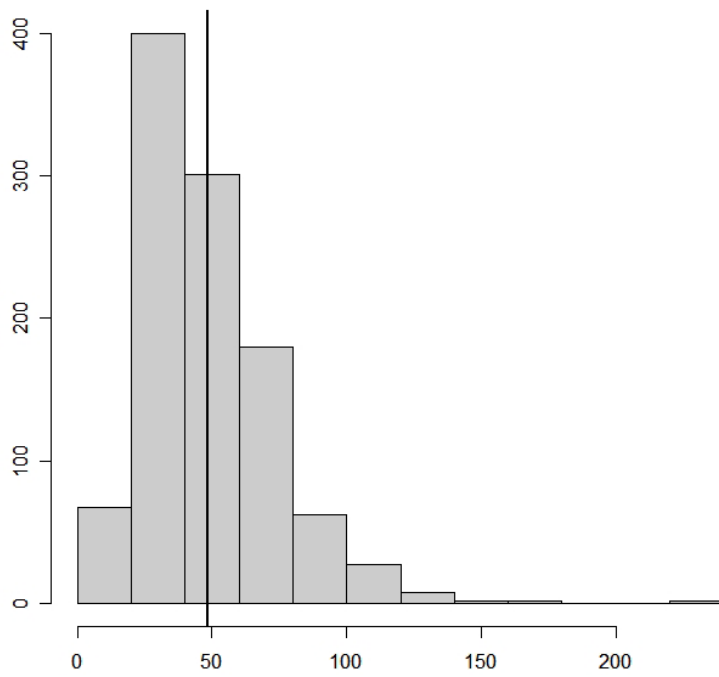


Figure 7. Histogram of total in-store time in minutes (vertical line denotes the mean).

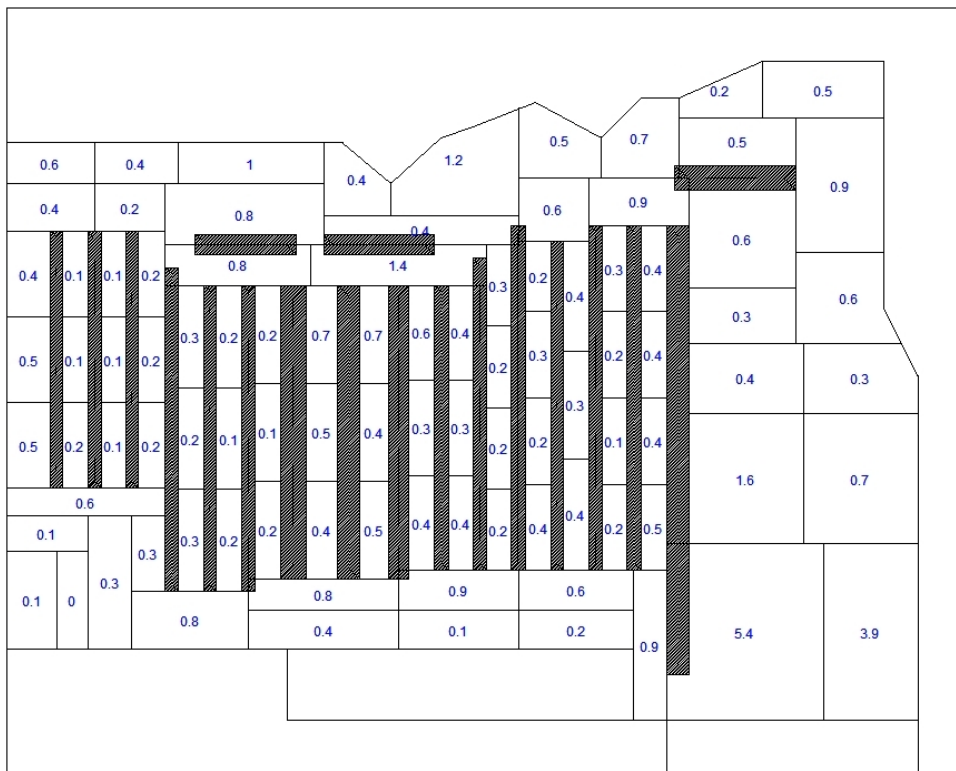


Figure 8. Average time a shopper spent (in minutes) in each zone.

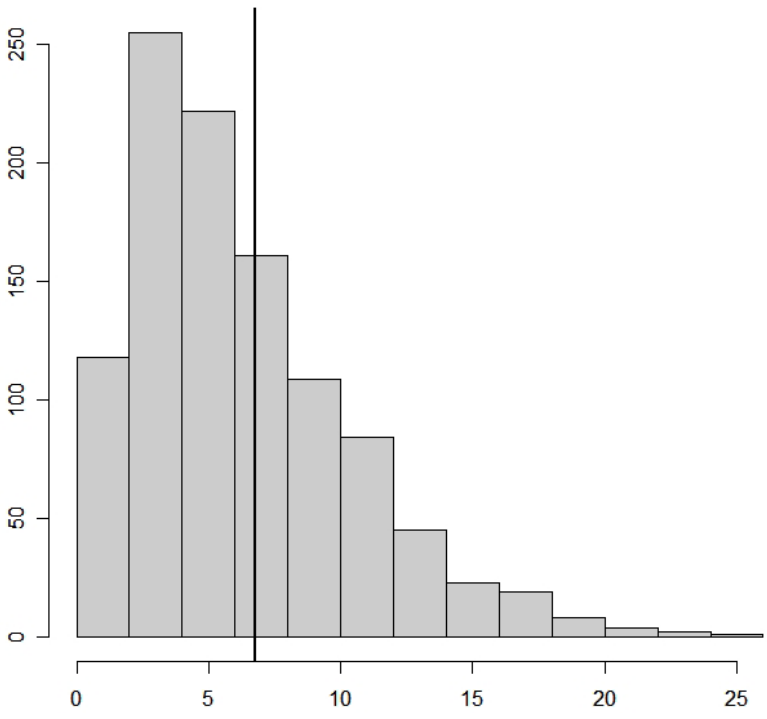


Figure 9. Histogram of the total number of product categories purchased (vertical line denotes the mean).

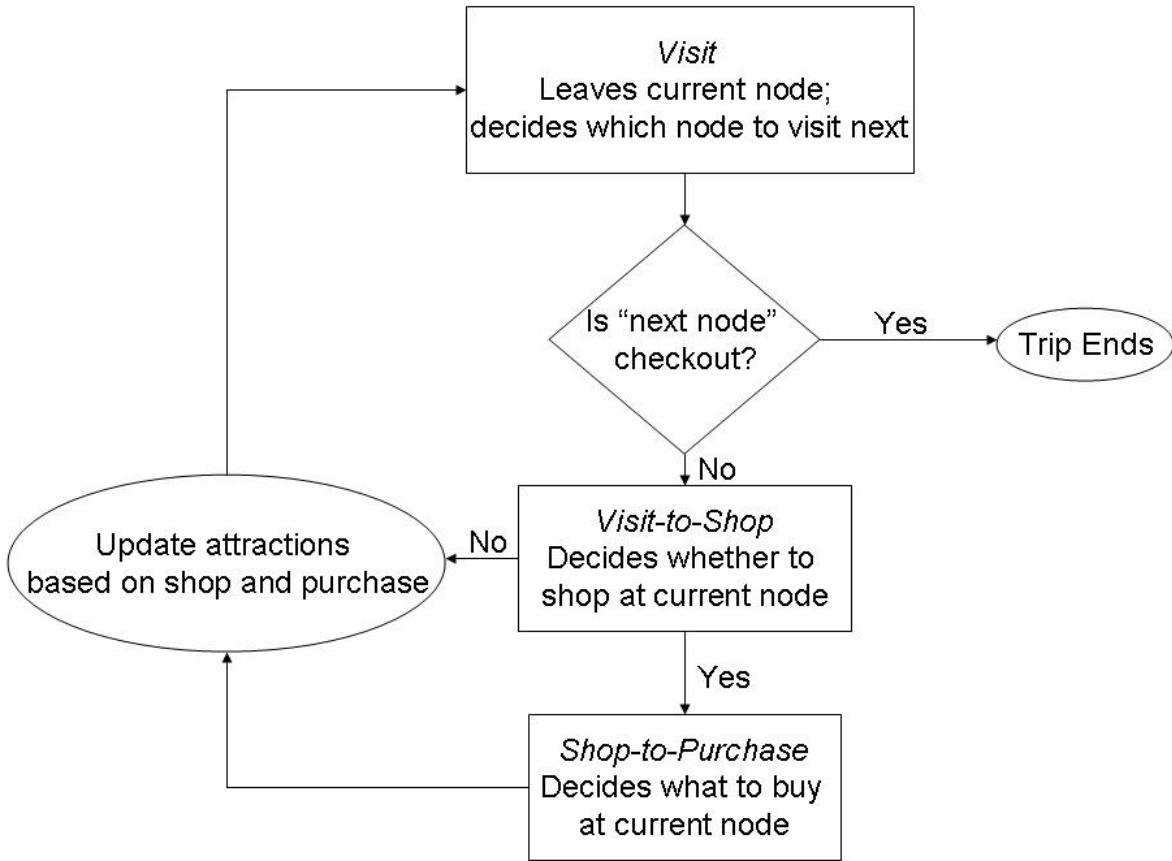


Figure 10. The shopper's in-store decision process.

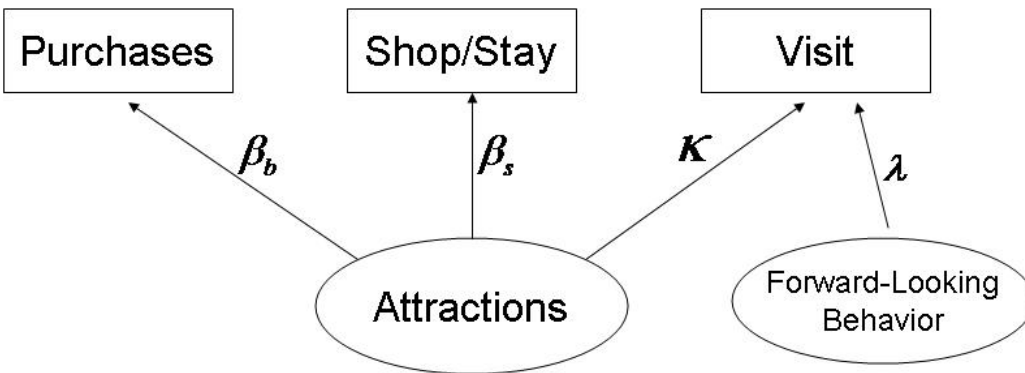


Figure 11. A schematic of the integrated model structure.

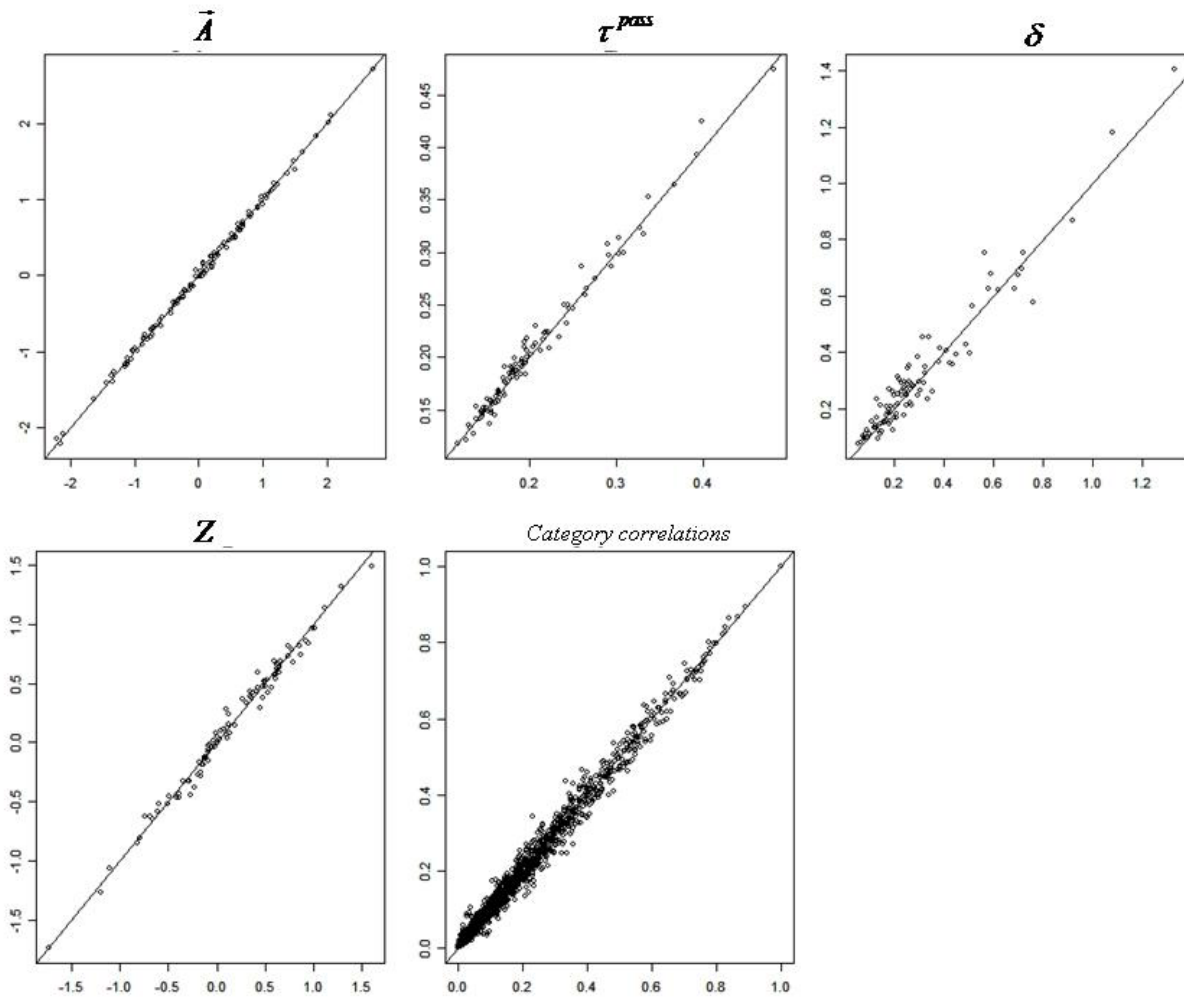


Figure 12. Estimation results for model parameters in simulation study not shown in Table 2. In each panel, the true values are plotted on the x-axis while the mean of the posterior distribution is plotted on the y-axis.

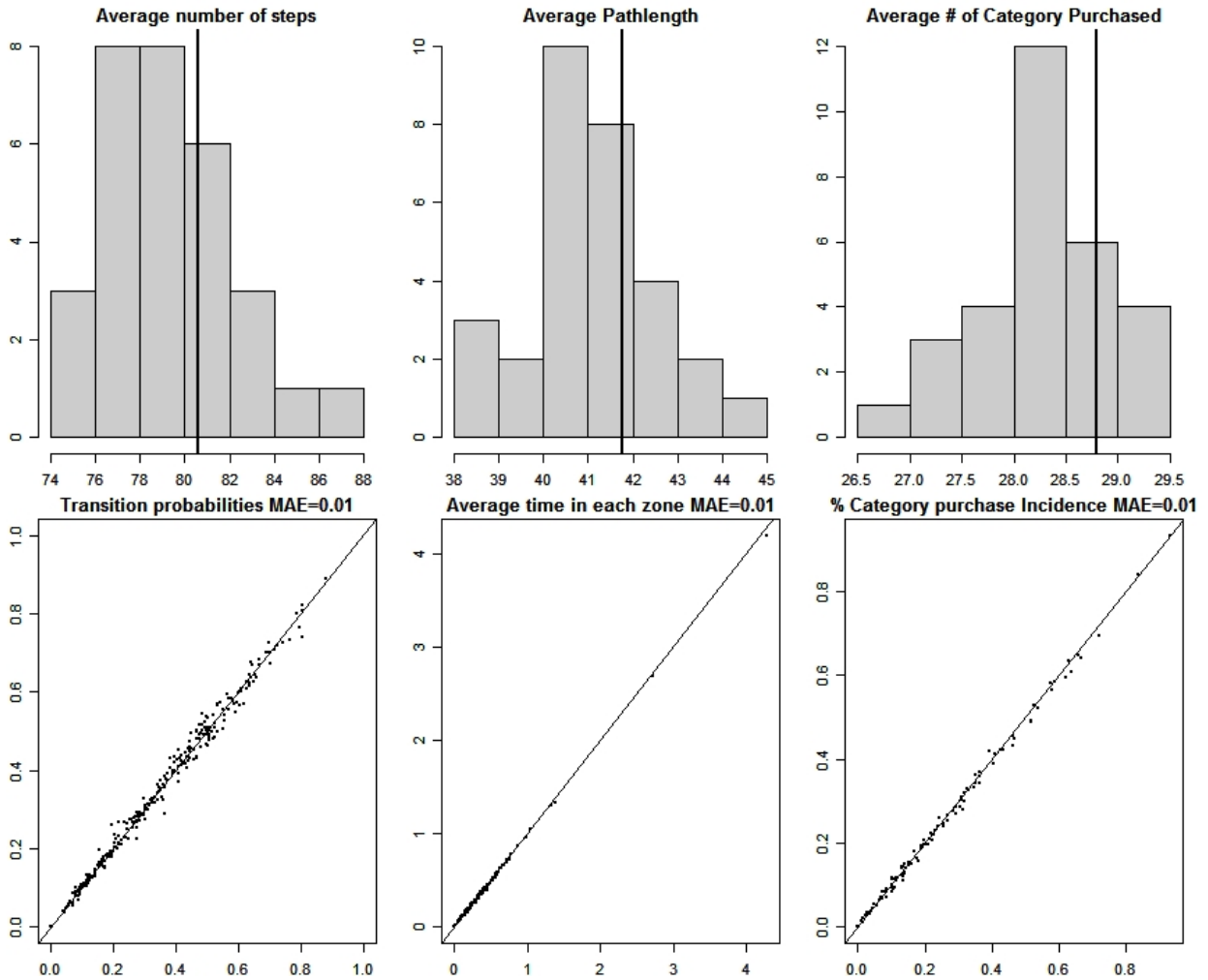


Figure 13. Posterior check for simulation study. In the upper three panels, histograms of summary statistics are drawn with the solid vertical line for the original simulated dataset. In the bottom three panels, the actual values of the summary statistics (calculated from the original simulated dataset) are plotted on the x-axis; the mean from the posterior sample is plotted on the y-axis.

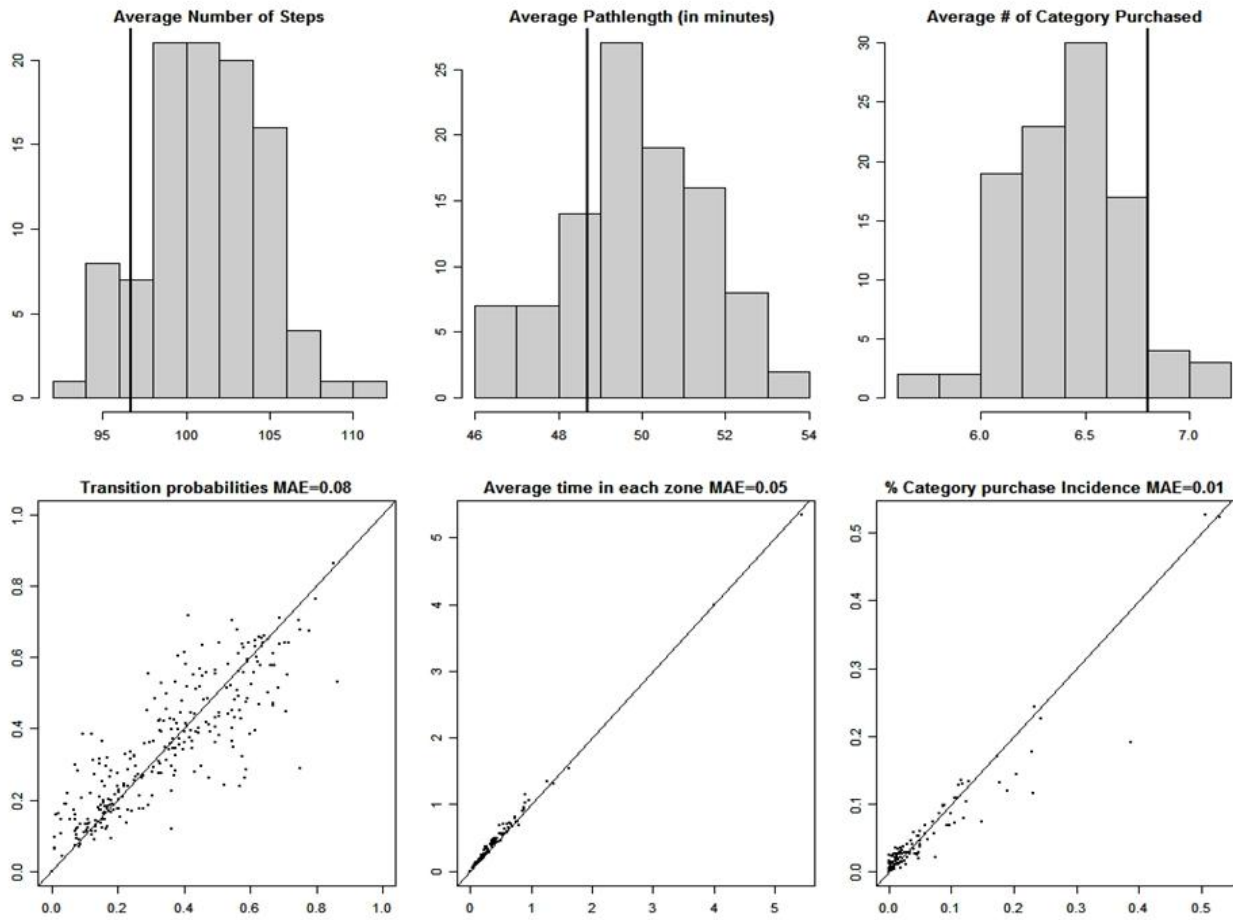


Figure 14. Posterior check for actual data. In the upper three panels, histograms of summary statistics are drawn with the solid vertical line for the actual (calibration) dataset. In the bottom three panels, the actual values of the summary statistics (calculated from the calibration dataset) are plotted on the x-axis; the mean from the posterior sample is plotted on the y-axis.

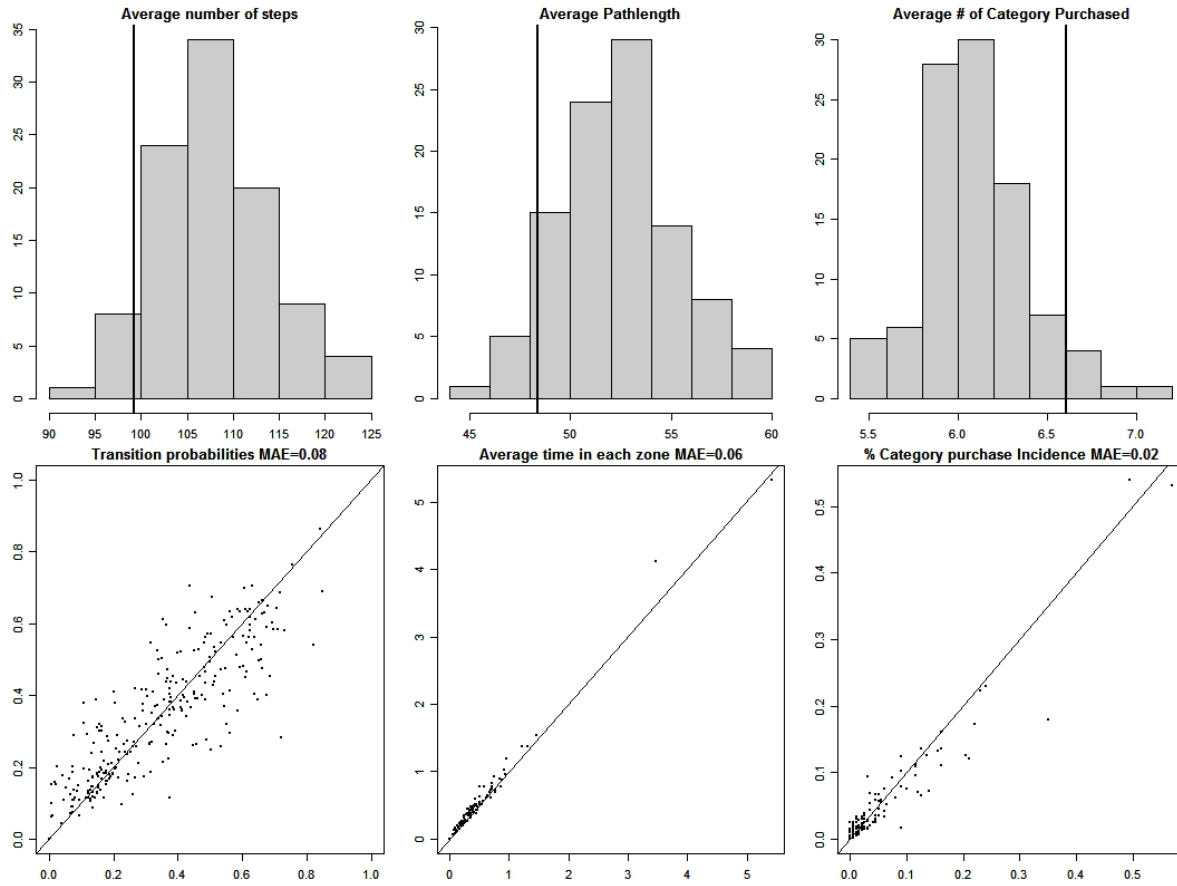


Figure 15. Holdout prediction posterior check. In the upper three panels, histograms of summary statistics are drawn with the solid vertical line for the holdout dataset. In the bottom three panels, the actual values of the summary statistics (calculated from the holdout dataset) are plotted on the x-axis; the mean from the posterior sample is plotted on the y-axis.

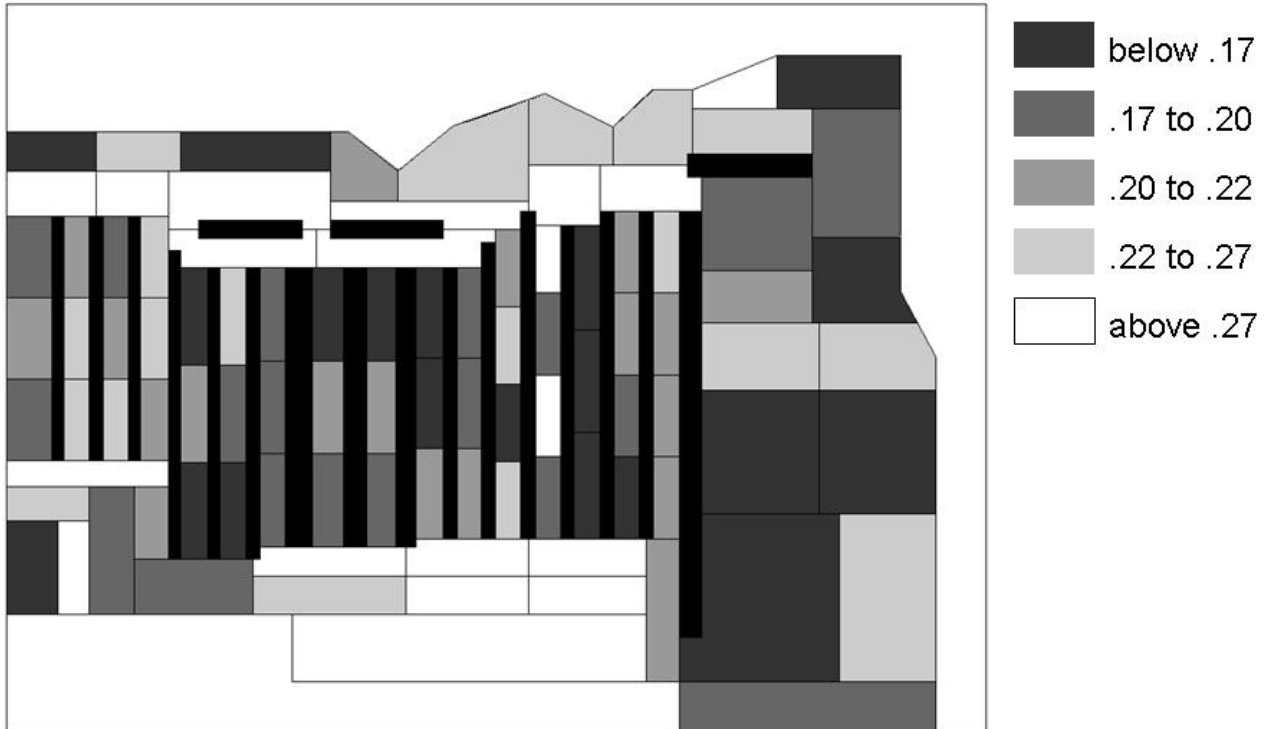


Figure 16.  $\tau^{shop}$  for each zone; zones with longer shopping time are shaded in darker gray.



Figure 17.  $Z_i$  for each zone; zones with higher  $Z_i$  are shaded in darker gray.