



University of Pennsylvania  
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

10-2009


# Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping Path and Purchase Behavior

Sam K. Hui

Eric T. Bradlow  
*University of Pennsylvania*

Peter S. Fader  
*University of Pennsylvania*

Follow this and additional works at: [https://repository.upenn.edu/statistics\\_papers](https://repository.upenn.edu/statistics_papers)

 Part of the [Applied Statistics Commons](#), [Behavioral Economics Commons](#), [Business Analytics Commons](#), [Econometrics Commons](#), [Management Information Systems Commons](#), [Management Sciences and Quantitative Methods Commons](#), [Marketing Commons](#), [Sales and Merchandising Commons](#), and the [Statistical Models Commons](#)

## Recommended Citation

Hui, S. K., Bradlow, E. T., & Fader, P. S. (2009). Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping Path and Purchase Behavior. *Journal of Consumer Research*, 36 (3), 478-493. <http://dx.doi.org/10.1086/599046>

This paper is posted at ScholarlyCommons. [https://repository.upenn.edu/statistics\\_papers/656](https://repository.upenn.edu/statistics_papers/656)  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping Path and Purchase Behavior

## **Abstract**

We examine three sets of established behavioral hypotheses about consumers' in-store behavior using field data on grocery store shopping paths and purchases. Our results provide field evidence for the following empirical regularities. First, as consumers spend more time in the store, they become more purposeful—they are less likely to spend time on exploration and more likely to shop/buy. Second, consistent with “licensing” behavior, after purchasing virtue categories, consumers are more likely to shop at locations that carry vice categories. Third, the presence of other shoppers attracts consumers toward a store zone but reduces consumers' tendency to shop there.

## **Keywords**

shopping behavior, Bayesian inference, econometrics

## **Disciplines**

Applied Statistics | Behavioral Economics | Business | Business Analytics | Econometrics | Management Information Systems | Management Sciences and Quantitative Methods | Marketing | Sales and Merchandising | Statistical Models | Statistics and Probability

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

Sam K. Hui

Eric T. Bradlow

Peter S. Fader\*

This Version: October 30, 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**

As behavioral hypotheses about in-store decision making become more common in the marketing literature, there is a growing need for richer, more complete datasets in order to test them more carefully. We introduce a novel PathTracker<sup>®</sup> dataset that captures consumers' in-store shopping processes, thus allowing researchers to start thinking about how to run such tests using actual field data. We propose an individual-level probability model that 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 she makes within that zone. After showing that our model offers an adequate description of the PathTracker<sup>®</sup> data, we discuss the issues involved in testing several behavioral hypotheses using the proposed framework.

## **1. Introduction**

Marketing researchers have recently shown increased interest in studying consumers' in-store decision making behaviors and the associated implications for store design and merchandising strategies. For instance, Dhar et al. (2007) studied consumers' "shopping momentum," where an initial purchase leads to a higher tendency of making more purchases; Inman et al. (2007) studied how category factors and customer characteristics drive unplanned purchases; Lee and Ariely (2006) studied how consumers' goals evolve during a shopping trip; finally, Argo et al. (2005) and Dahl et al. (2001) investigated how the social presence of other shoppers can affect consumers in a retail setting.

Testing the above behavioral hypotheses with field data would not only enhance the external validity of the proposed theories, but would also provide a bridge between laboratory studies and practical applications. A full-scale field study, however, would require data that are far richer than commonly used scanner data (e.g., Guadagni and Little 1983). Even more complete "market basket" data (e.g., Bell and Lattin 1998) are not nearly complete enough to capture any of the behaviors described above. Ideally, researchers need a dataset that covers the entire process of in-store shopping, i.e., the path that the shopper follows and the purchases she makes from the moment she enters the store to the moment she reaches the checkout counter. Equipped with this richer path data (Hui et al. 2007a), researchers can begin to study the complete set of in-store decisions rather than merely looking at limited "snapshots" (e.g., scanner data).

Until recently, however, suitable datasets on shopping paths and purchases were very difficult to obtain. To collect such data, one would typically have to physically follow shoppers around the store (e.g., Farley and Ring 1966) or rely on a large number of cameras (e.g., Heller

1988). These methods, as well as other related technologies (Burke 2005) can be costly and quite labor-intensive to collect and prepare large-scale datasets for analysis. But today, advances in data-collection technology have helped overcome these hurdles. New technologies such as Radio Frequency Identification (RFID) enable researchers to track multiple shoppers' movements in real time. For instance, Sorensen Associates (now known as TNS Sorensen) developed the PathTracker<sup>®</sup> system, which 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 ordinary point-of-sale 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). This dataset allows researchers to examine shoppers' behavior at a finer level than before.

In this paper, we introduce the PathTracker<sup>®</sup> dataset in detail, propose an integrated probability model to describe such data, and discuss how to test behavioral hypotheses using our model. We jointly captures three aspects of each shopper's in-store behavior: (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 (e.g., planning-ahead propensity) to drive all of the above processes. Finally, a Hierarchical Bayesian formulation (Rossi et al. 2006) is used to capture parameter heterogeneity across shoppers.

While the main focus of this paper is methodological, the contribution of this paper is threefold. First, we introduce a novel PathTracker<sup>®</sup> dataset to other researchers and outline research opportunities with this rich dataset. Second, we propose a new integrated probability model that adequately describes these data; such a model has never been proposed in the literature before. Third, on a substantive level, we discuss how behavioral hypotheses (e.g., shopping momentum, crowding/herding, goal gradient, and category assortment effects) can be built into our model framework in the near future and hence tested using field data.

The remainder of this paper is organized as follows. Section 2 describes the PathTracker system and provides summary statistics for our dataset. In Section 3, we specify our model and estimation approaches in detail. In Section 4, we apply our model to field data, demonstrate its fit, and interpret its parameter estimates. Section 5 discusses several behavioral hypotheses that can be tested under our general paradigm. Finally, Section 6 concludes with a discussion of future research directions.

## **2. PathTracker<sup>®</sup> Data**

In this section, we describe the PathTracker<sup>®</sup> dataset and data preparation procedures.

### **2.1 Data description**

Our dataset contains 1051 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 (but typical) 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 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.<sup>1</sup>

## 2.2 Data preparation

Our model for consumers' in-store movement, as we discuss in Section 3, is a discrete choice model (McFadden 1981). Thus, the raw data needs to be "discretized" to limit the number of possible locations (i.e., choice options). This is a common procedure used by other researchers when building models to analyze complex behaviors such as eye-tracking (e.g., Pieters and Wedel 2007) and pedestrian movements (e.g., Antonini et al. 2006).

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 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 1. The location(s) of each product category across the 96

---

<sup>1</sup> Sometimes a product category can be located in more than one area within the store. In these cases, we use statistical heuristics to impute where a purchase is made. Details are available upon request. When cart-level scanners becomes available in the future, this problem can be eliminated altogether as the time when an object is scanned and put into one's shopping cart will be known.



zones, along with its % penetration (fraction of the 1051 shopping baskets containing the category), are shown in Table 1.<sup>2</sup> Note, as mentioned before, that some product categories can appear in multiple zones; e.g., Paper Towels are located both in zone 37 and zone 75.

[Insert Figure 1 about here]

[Insert Table 1 about here]

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 zone. Figure 2 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 our data.

[Insert Figure 2 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, in Figure 2, although node A and node B, which are in adjacent aisles, are close together in Euclidean distance, 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

---

<sup>2</sup> Note that the scanner data portion of our dataset is less refined than datasets used in typical academic studies. For instance, our data do not allow us to adequately tease apart the Skin Care and Eye Care categories, 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.

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.

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 refer to a zone transition as a “step”. Figure 3 shows an example of path discretization; 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 3 about here]

## 2.3 Summary statistics

Since our goal is to capture shoppers’ in-store visit, stay, and purchase behaviors, hence a greater understanding of consumers’ underlying decision processes, we derive the following summary statistics that describe the data along those three key dimensions. The summary statistics included for visit, stay, and purchase are discussed separately in Subsections 2.3.1-2.3.3 below.

### 2.3.1 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 4. 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 5, while the light shaded arrows indicate all possible movements.

[Insert Figure 4 about here]

[Insert Figure 5 about here]

Note from Figure 5 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 through it. This interesting observation is consistent with the common “excursion” and lack of aisle-traversal 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).

### 2.32 Summary statistics for stay

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 6. 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 7.

[Insert Figure 6 about here]

[Insert Figure 7 about here]

Figure 7 leads to several interesting behavioral 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; possibly due to the “personal nature” of the goods. 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. This appears to provide field validation for the goal gradient hypothesis (Kivetz et al 2006, Nunes and Dreze 2006a).

### 2.3.3 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 8. In our dataset, shoppers purchase, on average, from 6.7 categories.

[Insert Figure 8 about here]

A key issue for retailers is to determine how much of these purchase decisions are driven by the attraction for each category, per se, versus the inherent attraction for the area(s) in the store in which they are located. It is critically important for a retailer to isolate these two effects, but it is nearly impossible to perform this task with raw data alone. It requires a formal model that can sort out underlying propensities for each of these decision factors. As we go on to develop the model, it will become clear how we accomplish this task, and we will demonstrate its implications with some real examples subsequently.

## 3. Model Development

We develop an integrated model to describe each consumer's shopping path and purchase behavior. 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. Finally, in Section 3.3, we embed our model within a Hierarchical Bayesian framework to allow for

heterogeneity among shoppers' purchase preferences, movement patterns, and planning-ahead tendencies.

### 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 9, are driven by the latent attractions of product categories and zones, which we will define later.

[Insert Figure 9 about here]

First, the shopper makes a *visit* decision: she decides which zone she is going to visit next. If that zone is the 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 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; 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). Further, 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 stays 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  $t$ .

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 9 is restarted there. We note that it is this sequential updating, either via movement to a new zone (a step), an additional blink in the given zone (staytime), or the purchasing of a product category (all possible consumer actions) that leads to the *dynamic modeling* of consumer behavior that is captured by our model.

### 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 a 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 inter-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.

We define a vector of latent variables  $\vec{a}_t = (a_{1t}, a_{2t}, \dots, a_{Jt})'$ , 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 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>3</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 specification to capture the basic evolution pattern of attractions, as follows:

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

That is, we posit that after the shopper visits 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

---

<sup>3</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.

purchased at step  $t$  ( $B_{jt} = 1$ ), then the attraction for category  $j$  will further change by an amount indicated by  $\Delta_b$ . While, a priori, we strongly hypothesize that  $\Delta_s$  and  $\Delta_b$  are likely to be negative, our model does not impose this as a formal constraint; we let the field data (and the model) identify the magnitude and direction of these updating parameters.

### 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  $M(x_t)$ . To capture this zone-choice decision, 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 “planning-ahead” 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 may plan ahead when deciding where to visit next. Her choice involves a tradeoff between two aspects: (i) the intrinsic attraction of the adjoining zone, and (ii) by going to the adjoining zone, whether she 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 “planning ahead” 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 graph-theoretical distance (i.e., the length of the shortest path) between zone  $i$  and zone  $k$ .  $\lambda$  is a parameter that governs how the shopper trades off immediate utility with the more planning-ahead concern of reaching high attraction regions later on in his trip. For instance,  $\lambda = \infty$  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 (above and beyond the zone-level baseline visit propensity  $Z_i$ ).

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 stay 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. To capture this decision, 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,

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

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

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

Note that through the specification in Equation (5)—(9), we assume that stay time and purchase are conditionally independent given the latent visit-to-shop state. Marginally, stay time and purchase are allowed to be positively correlated, as intuition would suggest.

### 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 as follows. 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 visit-to-shop,  $\alpha_b$  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 (Rook 1987) as well as planned purchase behavior (e.g., Block and Morwitz 1999). 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 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 9, 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 10.

[Insert Figure 10 about here]

### 3.3 Hierarchical Bayesian framework

Since consumers may have heterogeneous category preferences, shopping characteristics, and planning-ahead propensities, we embed our individual-model within a Hierarchical Bayesian framework. This leads to two key statistical advantages. First, with this setup, each consumer has a different set of parameters that are assumed to be drawn from a common distribution; this allows us to borrow strength across customers to calibrate our model. Second, by specifying a covariance matrix on the individual attraction parameters (see Appendix I), we can borrow strength across categories while taking into account category complementarities. The full details are discussed in Appendix I.

To confirm that our model is properly identified and our estimation procedure can recover the true parameter values, we conducted a simulation experiment; the details of which are described in Appendix II.

## **4. 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 4.1, we perform a set of posterior checks to ensure that our model is capable of recovering key summary statistics. In Section 4.2, we assess the predictive performance of our model using holdout prediction, and compare our model against three logical (nested) sub-models in term of both in-sample and holdout model fit. Section 4.3 presents parameter estimates and substantive behavioral findings.

## 4.1 Model validation

Though our proposed model is conceptually interesting, it is important to check our model fit against the actual data to see whether our model can recover important summary statistics of the dataset (which we described in Section 2.3). Towards this end, we use posterior predictive checks proposed by Gelman et al. (1996) to assess the adequacy of our model. Specifically, 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. That is, if the model fits the data well simulated data under the fitted model should look like the actual data along key summary dimensions. The results are shown in Figure 11.

[Insert Figure 11 about here]

Figure 11 shows that our model recovers key summary statistics of the actual data fairly well. The top three panels show that data simulated from the posterior predictive distribution is 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.

## 4.2 Holdout prediction and model comparison

We also 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 each customer's 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 (using Newton and Raftery's (1994) importance sampling approach) and then we draw 100 sample paths to

complete each shopping trip. The summary statistics that compare the actual (holdout) dataset and the simulated paths are considered; the results are shown in Figure 12. Although (as expected) the model fit is worse than the in-sample fit, our model still provides a good fit to the holdout data.

[Insert Figure 12 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 disable 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 10):

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 2, 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 2 about here]

#### 4.3 Parameter estimates and interpretation

The posterior distribution of the hyperparameters that govern the individual-level parameters are summarized in Table 3. These results offer a number of immediate insights about consumer behavior. 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 degree of planning-ahead behavior in their shopping paths. This is consistent with the finding in Hui et al. (2007b) where the researchers find evidence of forward-looking behavior for grocery shoppers.

[Insert Table 3 about here]

The posterior means for the baseline attractions of the 10 highest-attractiveness categories are summarized in Table 4. 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. This is well-aligned with the observation that Fruit 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 purchasing 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 return to this issue in more detail in the Section 5.1.

[Insert Table 4 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 13 and Figure 14 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  $Z_i$  correspond to zones that 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 13 about here]

[Insert Figure 14 about here]

## 5. Testing behavioral hypotheses

Our model specification and testing have focused primarily on three fundamental behaviors (visit, visit-to-shop, and shop-to-purchase), but the modeling framework as a whole is more flexible. As we have noted at several points throughout the text, we can introduce additional “behavioral tendencies” into the model and examine some associated hypotheses – many of which have already been tested in a laboratory setting – in a relatively straightforward manner. Proper field testing of these hypotheses will require additional data (especially from a cross-store setting where there is more variability in behavior and store layout); nevertheless, we



believe it is valuable to lay out a roadmap to show how these behavioral generalizations can be incorporated into our model when more suitable data become available in the near future (and we discuss some of these data-related issues in the next section of the paper).

(i) *Category assortment*: Researchers have been interested in consumers' perception of assortment in a category (Broniarczyk et al. 1998) and the effect of product assortment on buyer preferences (e.g., Chernev 2003; Simonson 1999), purchase probability (Chernev 2005, 2006), and consumption quantities (Kahn and Wansink 2004). The effect of category assortment can be incorporated into our model by allowing latent attractions to be a function of the variety of the assortment, e.g., using the definition proposed by Hoch et al. (1999).

(ii) *Shopping momentum*: Dhar et al. (2007) defined a general phenomenon which they termed “shopping momentum,” which refers to the tendency to purchase more items once an initial purchase is made. The researchers also found that the nature of the first purchase affects the strength of the shopping momentum effect. Within our model, we can incorporate “shopping momentum” by extending Equation (2) to (2\*) by including an indicator variable that represents whether an initial purchase has already been made:

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

where  $P_t$  takes value 1 if an initial purchase is made, and 0 otherwise. The coefficient  $\Delta_{Momentum}$  captures the magnitude of the shopping momentum effect. Presumably,  $\Delta_{Momentum}$  should be positive; i.e., all other categories in the store become more attractive (and hence more likely to be purchased) given that an initial purchase has already been made. One can also generalize (2\*) even further to allow strength of the shopping momentum effect to depend on whether the initial purchase is a “hedonic” or “utilitarian” item (Dhar et al. 2007).

(iii) *Licensing and spillover effects*: Since our data contain information about the sequence of purchases, we can explore how promotions obtained on items purchased earlier during the trip affect shoppers' behavior afterwards, i.e., how they “spillover” to consumer behavior for other categories (Janakiraman et al. 2006). In particular, one can investigate the idea of “mental accounting/budgeting” (e.g., Heath and Soll 1996; Thaler 1985) or “licensing” (e.g., Khan and Dhar 2006). For example, does getting a deal early in the trip increase consumers' probability to buy more items, given that they now think that they have “saved” early on? Coupled with data on promotions, we can extend our model to allow attractions to update based on the amount of promotions and savings that the shopper obtains up to any time. Similarly, does buying “virtue” items (e.g., vegetables, organic food) early on during the trip increase the propensity to buy “vice” items later on given that the consumer now feel that they have the “license” to do so (e.g., Kivetz and Keinan 2006; Wertenbroch 1998)? Intriguing issues such as these can be addressed by incorporating such effects into the attraction model in Equation (2).

(iv) *Crowding/herding*: Researchers have studied how grocery shoppers react to the presence of other shoppers in the store. Harrell and Anderson (1980) suggested that shoppers generally avoid crowded areas, and may reduce their shopping time in crowded conditions. Argo et al. (2005) and Dhal et al. (2001) stated that the “mere social presence” of other shoppers affects shopping behavior. On the other hand, the literature on “herding” effects (e.g., Banerjee 1992; Becker 1991) suggests that shoppers may prefer areas where other shoppers are located. Taken together, the effect of the number of other shoppers may be an inverted-U shape: a moderate shopper density may encourage visitation/shopping, while too much crowding may cause shoppers to avoid such areas. Given complete data on the position of each cart, our model can be extended and used to test these hypotheses. More specifically, one can incorporate “shopper density”, i.e.,

the number of shoppers in a certain zone divided by the area of the zone, as a covariate in the model of visit (Equation (4)) and the model of visit-to-shop (Equation (6)). The effects of crowding/herding can then be assessed quantitatively.

(v) *Non-stationary behaviors*: Shoppers' behavior may change in many ways over the course of their trip. As noted earlier, Sorensen (2003) observed that shoppers tend to speed up as they move towards the checkout counter. This observation may be related to the research on "goal gradient" (e.g., Kivetz et al. 2006), such that consumers tend to expend more effort and "accelerate" as they get closer to their goals (see also Nunes and Dreze 2006a). To incorporate this effect, we can generalize Equations (7) and (8) to allow shopping time to be a function of the distance to checkout. A similar issue is whether consumers tend to buy more unplanned items the longer they spend in the store, perhaps due to fatigue and thus a reduction of self-regulatory resources (Vohs and Faber 2007). This issue can be fruitfully explored by incorporating in-store time as a covariate for the variance term in the model of purchase, i.e., Equation (10).

## **6. Conclusions and Directions for Future Research**

In this paper, we have introduced an integrated modeling framework to capture, describe, and predict a consumer's shopping path and purchase behavior in a grocery store. 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.

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. A number of

academic and practical insights emerge from the model – chief among them is the ability to sort out how much of a product’s purchasing is due to its inherent attractiveness, per se, versus the propensity of shoppers to visit certain regions of the store (regardless of what products are carried there).

We discuss how the model can be extended so that various well-established behavioral hypotheses can be incorporated into our model framework, thus allowing us to validate existing behavioral theories using field data. Beyond these conceptual extensions, our model can also be extended through the incorporation of additional data structures that we do not have in our initial dataset. All of these are “close-in” extensions, which should be available to managers and academic researchers in the very near future.

(1) *Cross-store study*: The PathTracker<sup>®</sup> system is being installed in an increasing number of supermarkets (and other types of retail stores) around the world to track consumers’ shopping patterns. Our model can easily be applied to the other stores to conduct a cross-store study, to help us understand how store characteristics (e.g., square footage, number of aisles) are related to consumers’ movement tendencies and shop/purchase behavior. For instance, Meyers-Levy and Zhu (2007) demonstrated how ceiling height affect consumers’ information processing and with store varying layout information, this will be easily testable.

(2) *Consumer characteristics*: The Hierarchical Bayesian 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 (Nunes and Dreze 2006b), 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 planning-ahead tendencies differ across shoppers of different gender and age (e.g., Otnes and McGrath 2001; Yoon 1997). 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. 2007b).

(3) *Survey data*: By combining shopping path data with surveys, a lot of interesting behavioral questions can be addressed. For instance, one can ask consumers to state their shopping goals (Lee and Ariely 2007) before entering the store, and study how the propensity of unplanned purchase (Inman et al. 2007) is related to their path behavior. Further, one can also study how consumers’ path and purchase behavior changes under time pressure, a topic of recent interest for many behavioral researchers (e.g., Dhar and Nowlis 1999; Suri and Monroe 2003).

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. 2007a). As better consumer-tracking technologies (beyond RFID) become 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 routine analyses done with scanner data today.

Thus, we believe that marketing researchers will benefit from a deeper study of path data, particularly as they utilize theories and analytical approaches from psychology, economics, and sociology, as we have outlined here.

## References

- Anderson, Eric, and Duncan Simester (2004), "Long Run Effects of Promotion Depth on New Versus Established Customers: Three Field Studies," *Marketing Science*, 23(1), 4-20.
- Antonini, Gianluca, Michel Bierlaire, and Mats Weber (2006), "Discrete Choice Models of Pedestrian Walking Behavior," *Transportation Research B*, 40, 667-687.
- Argo, Jennifer J., Darren W. Dahl, and Rajesh V. Manchanda (2005), "The Influence of a Mere Social Presence in a Retail Context," *Journal of Consumer Research*, 32, 207-212.
- Banerjee, Sudipto, Bradley P. Carlin, and Alan E. Gelfand (2004), *Hierarchical Modeling and Analysis of Spatial Data*, Chapman and Hall.
- Becker (1991), "A Note on Restaurant Pricing and Other Examples of Social Influence on Price," *Journal of Political Economy*, 99, 1109-1116.
- 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.
- Block, Lauren G., and Vicki G. Morwitz (1999), "Shopping Lists as an External Memory Aid for Grocery Shopping: Influences on List Writing and List Fulfillment," *Journal of Consumer Psychology*, 8(4), 343-375.
- 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.
- Broniarczyk, S. M., W. D. Hoyer, L. McAlister (1998), "Consumers' Perception of the Assortment offered in a Grocery Category: The Impact of Item Reduction," *Journal of Marketing Research*, 35, 166-176.
- Burke, Raymond R. (2005), "The Third Wave of Marketing Intelligence," in Manfred Drafft and Murali Mantrala (Eds.): *Retailing in the 21<sup>st</sup> Century: Current and Future Trends*, Springer, 113-125.
- Chernev, Alexander (2003), "When More is Less and Less is More: The Role of Ideal Point Availability and Assortment in Consumer Choice," *Journal of Consumer Research*, 30 (2), 170-183.
- Chernev, Alexander (2005), "Feature Complementarity and Assortment in Choice," *Journal of Consumer Research*, 31, 748-759.
- Chernev, Alexander (2006), "Differentiation and Parity in Assortment Pricing," *Journal of Consumer Research*, 33 (September), 199-210.

- Dahl, Darren W., Rajesh V. Manchanda, and Jennifer J. Argo (2001), "Embarrassment in Consumer Purchase: The Roles of Social Presence and Purchase Familiarity," *Journal of Consumer Research*, 28(3), 473-81.
- Dhar, Ravi, Joel Huber, and Uzma Khan (2007), "The Shopping Momentum Effect," *Journal of Marketing Research*, 44(3), 370-378.
- Dhar, Ravi, and Stephen M. Nowlis (1999), "The Effect of Time Pressure on Consumer Choice Deferral," *Journal of Consumer Research*, 25(4), 369-384.
- Farley, John U., and L. Winston Ring (1966). A Stochastic Model of Supermarket Traffic Flow. *Operations Research*, 14(4), 555-567.
- Gelman, Andrew, Xiao-Li Meng, and Hal Stern (1996), "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies," *Statistica Sinica*, 6, 733-807.
- Guadagni P. M. & J. D. C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2(3), 203-238.
- Harrell, Gilbert and James C. Anderson (1980), "Path Analysis of Buyer Behavior Under Conditions of Crowding," *Journal of Marketing Research*, 17, 45-51.
- Heath, Chip, and Jack B. Soll (1996), "Mental Budgeting and Consumer Decision," *Journal of Consumer Research*, 23(1), 40-52.
- Heller, Walter (1988), "Tracking Shoppers Through the Combination Store," *Progressive Grocer*, 47-64.
- Hoch, S. J. E. T. Bradlow, and B. Wansink (1999), "The Variety of an Assortment," *Marketing Science*, 18(4), 527-546.
- Hui, Sam K., Peter S. Fader, and Eric T. Bradlow (2007a), "Path Data in Marketing: An Integrative Framework and Prospectus for Model-Building," *Working Paper*.
- Hui, Sam K., Peter S. Fader, and Eric T. Bradlow (2007b), "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.
- Inman, J. Jeffrey, Russell S. Winer, and Rosellina Ferraro (2007), "The Interplay between Category Factors, Customer Characteristics, and Customer Activities on In-Store Decision Making," *Working Paper*.



- Janakiraman, Narayan, Robert J. Meyer, and Andrea C. Morales (2006), "Spillover Effects: How Consumers Respond to Unexpected Changes in Price and Quantity," *Journal of Consumer Research*, 33 (December), 361-369.
- Kahn, Barbara E. and Brian Wansink (2004), "The Influence of Assortment Structure on Perceived Variety and Consumption Quantities," *Journal of Consumer Research*, 30(4), 519-533.
- Khan, Uzma, and Ravi Dhar (2006), "Licensing Effect in Consumer Choice," *Journal of Marketing Research*, 43, 259-266.
- Kivetz, Ran, and Anat Keinan (2006), "Repenting Hyperopia: An Analysis of Self-Control Regret," *Journal of Consumer Research*, 33 (September), 273-282.
- Kivetz, Ran, Oleg Urminsky, and Yuhuang Zheng (2006), "The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention," *Journal of Marketing Research*, 43, 39-58.
- 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.
- Lee, Leonard, and Dan Ariely (2006), "Shopping Goals, Goal Concreteness, and Conditional Promotions," *Journal of Consumer Research*, 33, 60-70.
- McFadden, D. L. (1981), *Structural Analysis of Discrete Data with Econometric Applications*. MIT press.
- Meyers-Levy, Joan, and Rui (Juliet) Zhu (2007), "The Influence of Ceiling Height: The Effect of Priming on the Type of Processing that People Use," *Journal of Consumer Research*, 34 (August), 174-186.
- 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.
- Nunes, Joseph S., and Xavier Dreze (2006a), "The Endowed Progress Effect: How Artificial Advancement Increases Effort," *Journal of Consumer Research*, 32 (March), 504-512.
- Nunes, Joseph S., and Xavier Dreze (2006b), "Your Loyalty Program is Betraying You," *Harvard Business Review*, 124-131.

- Otnes, Cele, and Mary Ann McGrath (2001), "Perceptions and Realities of Male Shopping Behavior," *Journal of Retailing*, 77, 111-137.
- Pieters, Rik, and Michel Wedel (2007), "Goal Control of Attention to Advertising: The Yarus Implication," *Journal of Consumer Research*, 34 (August), 224-233.
- Rook, Dennis W. (1987), "The Buying Impulse," *Journal of Consumer Research*, 14(2), 189-199.
- Rossi, Peter E., Greg M. Allenby, and Robert McCulloch (2006), *Bayesian Statistics and Marketing*, Wiley.
- Simonson, I. (1999), "The Effect of Product Assortment on Buyer Preferences," *Journal of Retailing*, 75, 347-370.
- 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.
- Sorensen, Herb (2003), "The Science of Shopping," *Marketing Research*, 15(3), 30-35.
- Suri, Rajneesh, and Kent B. Monroe (2003), "The Effects of Time Constraints on Consumers' Judgements of Price and Products," *Journal of Consumer Research*, 30(1), 92-104.
- Thaler, Richard (1985), "Mental Accounting and Consumer Choice," *Marketing Science*, 4(3), 199-214.
- Vohs, Kathleen D., and Ronald J. Faber (2007), "Spent Resources: Self-Regulatory Resource Availability Affects Impulse Buying," *Journal of Consumer Research*, 33 (March), 537-547.
- Wertenbroch, Klaus (1998), "Consumption Self-Control by Rationing Purchase Quantities of Virtue and Vice," *Marketing Science*, 17(4), 317-337.
- Yoon, Carolyn (1997), "Age Differences in Consumers' Processing Strategies: An Investigation of Moderating Influences," *Journal of Consumer Research*, 24(3), 329-342.

## Appendix I: Hierarchical Bayesian Specification

The parameter vector for the  $n$ -th consumer,  $(\bar{a}, \kappa, \alpha_s, \alpha_b, \beta_s, \beta_b, \Delta_s, \Delta_b, \lambda)_n$ , 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  $\bar{a}$ , then the prior for the rest of the parameters.

For the attraction vector, we specify

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

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>4</sup> Other studies that use a 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  $\|z_j - z_{j'}\| = \sqrt{(z_{j1} - z_{j'1})^2 + (z_{j2} - z_{j'2})^2}$ .

---

<sup>4</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.

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:

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

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:

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

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

To complete our Hierarchical Bayesian model specification, we specify a set of weakly informative, conjugate priors for all hyperparameters in our model. A MCMC procedure allows us to make inferences about our model parameters using samples from their posterior distributions (details available upon request).

## Appendix II: 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  $\bar{\mu}_A$  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  $\bar{\mu}_I$  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.

Estimation results for the hyperparameters that govern the individual-level parameters (besides category attractions) are shown in Table A1. Plots of the true versus estimated parameters for category attractions and zone-level parameters are shown in Figure A1. 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  $\bar{\mu}_A$ , zone-level parameters  $Z_i, \tau_i^{pass}$ , and  $\delta_i$ , and the correlation between category attractions are accurately recovered by our estimation procedure.

	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 A1. Estimation results for model hyperparameters in simulation study.

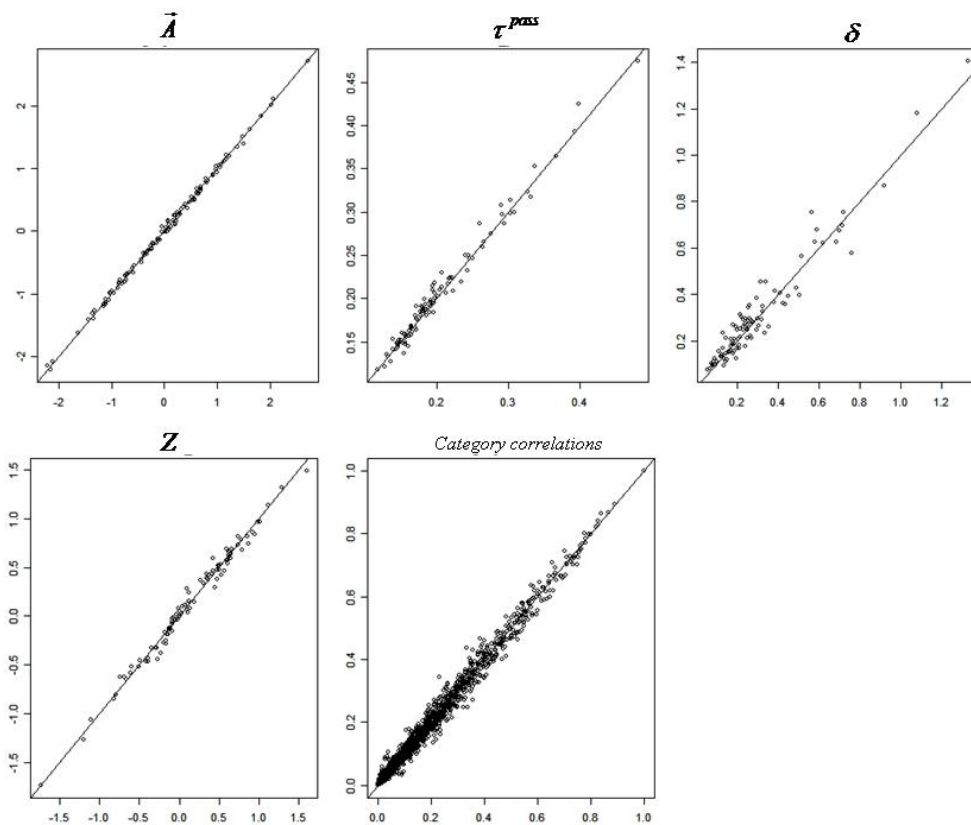


Figure A1. Estimation results for model parameters in simulation study not shown in Table 2. In each panel, the true values are plotted on the x-axis while the mean of the posterior distribution is plotted on the y-axis.

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%
Hag Liquor	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.

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

Table 2. 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 3. Estimation results for model hyperparameters in the actual data.

Product Category	Attraction
Fruit	2.70
Natural/Organic Food	2.24
Special Diet Items	2.04
Butter/Cheese/Cream	1.80
Salty Snacks	1.62
Vegetables	1.59
Pastry/Snack Cakes	1.54
Cereal	1.47
Yogurt	1.27
Canned Vegetables	1.27

Table 4. Posterior mean for category attractions for the 10 categories with the highest attraction, sorted in decreasing order.





Figure 1. Grocery store divided into 96 zones.

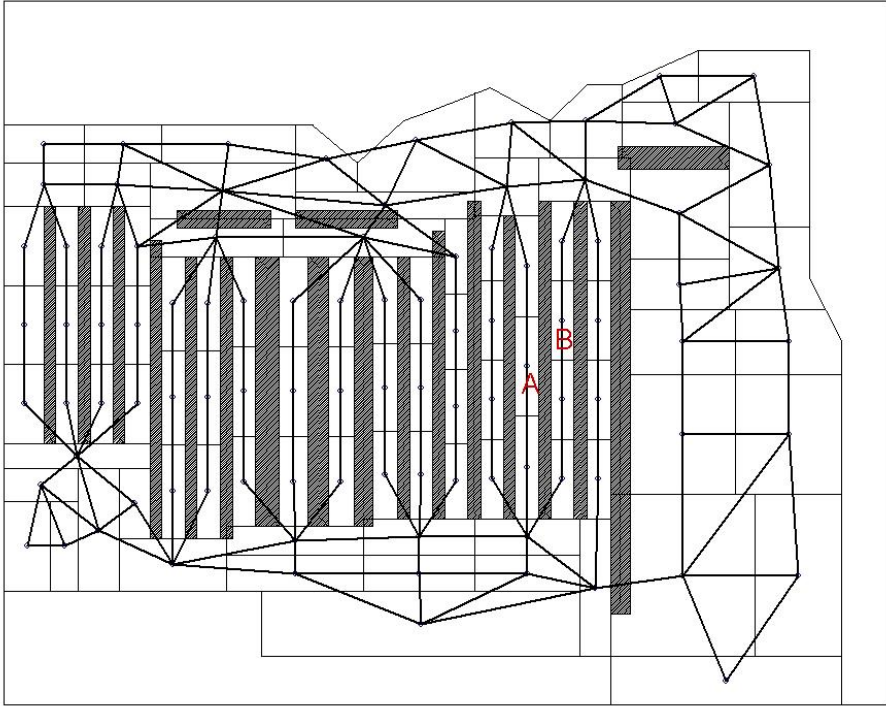
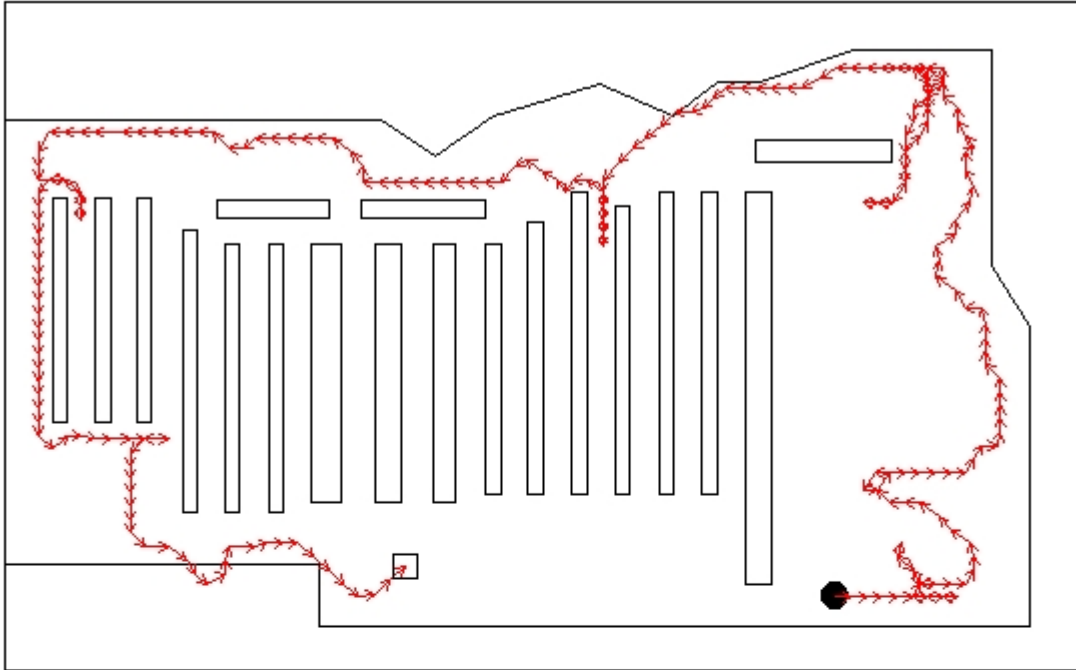


Figure 2. Grocery store represented by a graph of 96 nodes.

### Raw Path



### Path after discretization

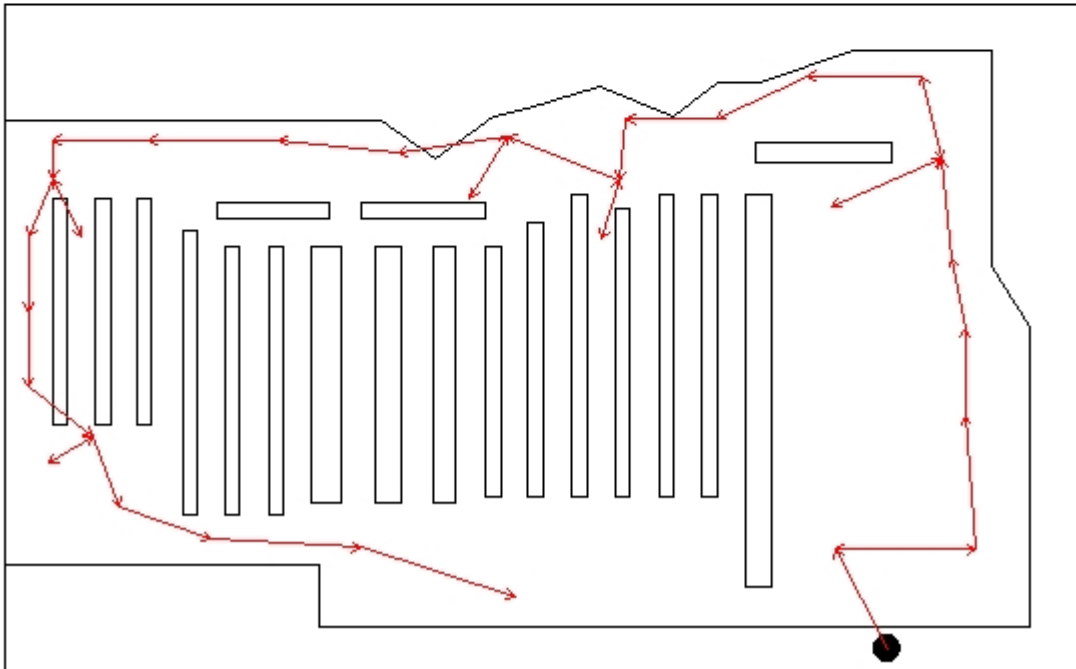


Figure 3. Example of path discretization.

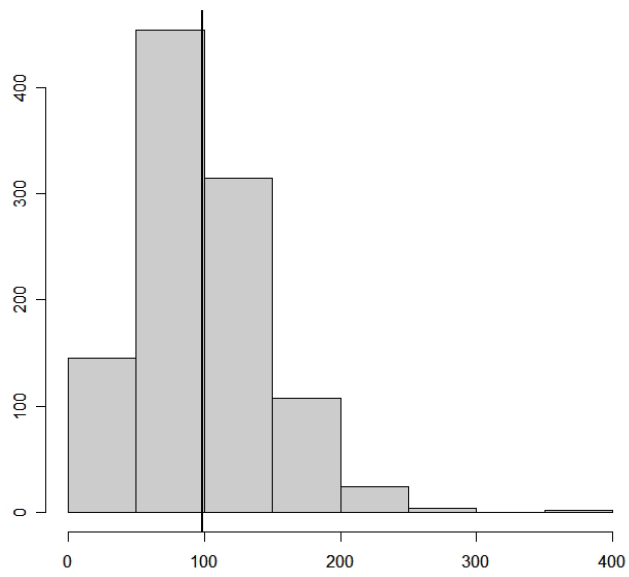


Figure 4. Histogram of number of steps (vertical line denotes the mean).

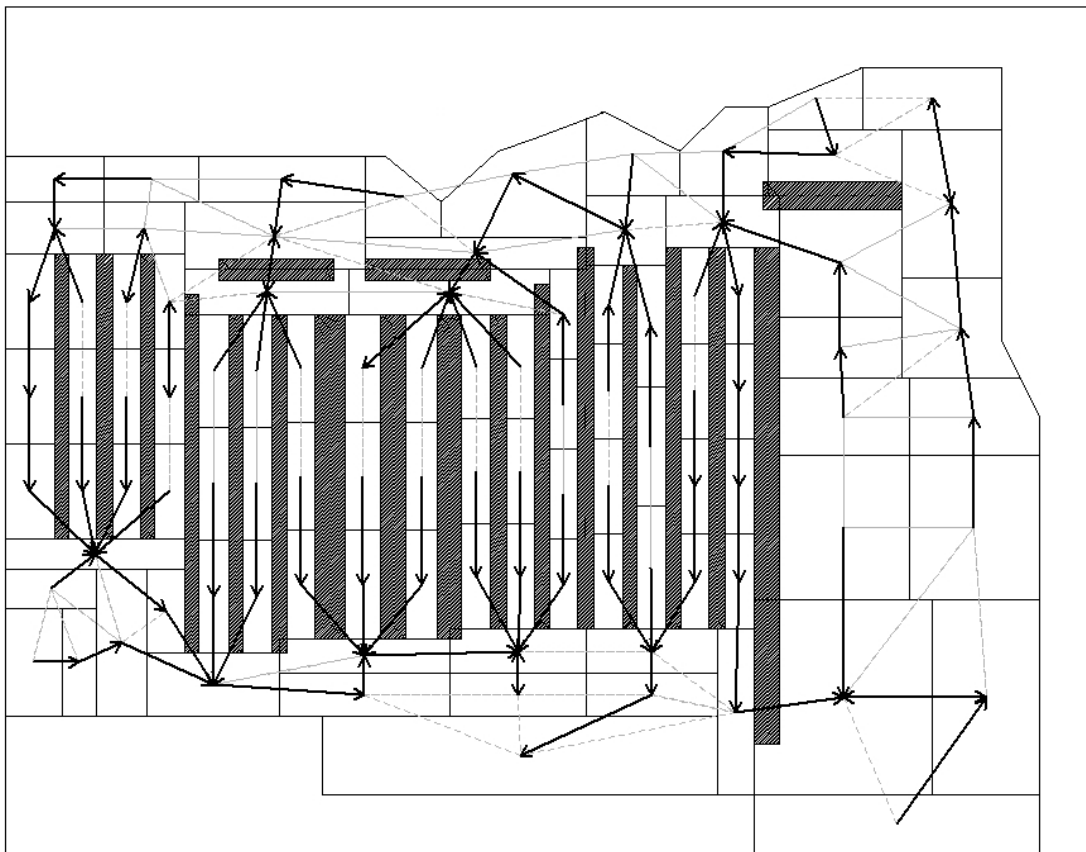


Figure 5. Most frequent transition out of each zone.



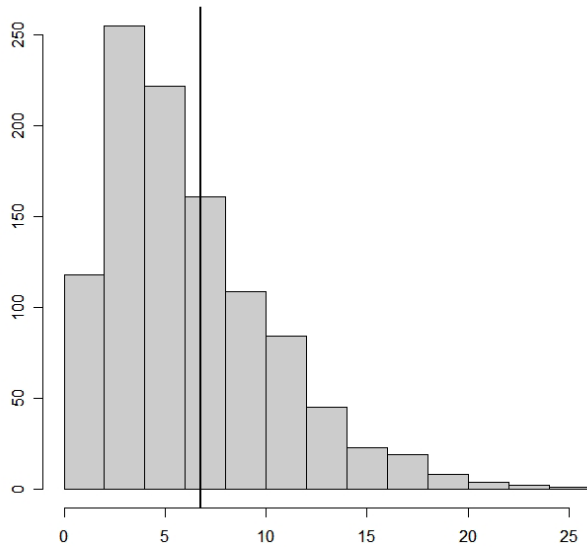


Figure 8. Histogram of the total number of product categories purchased (vertical line denotes the mean).

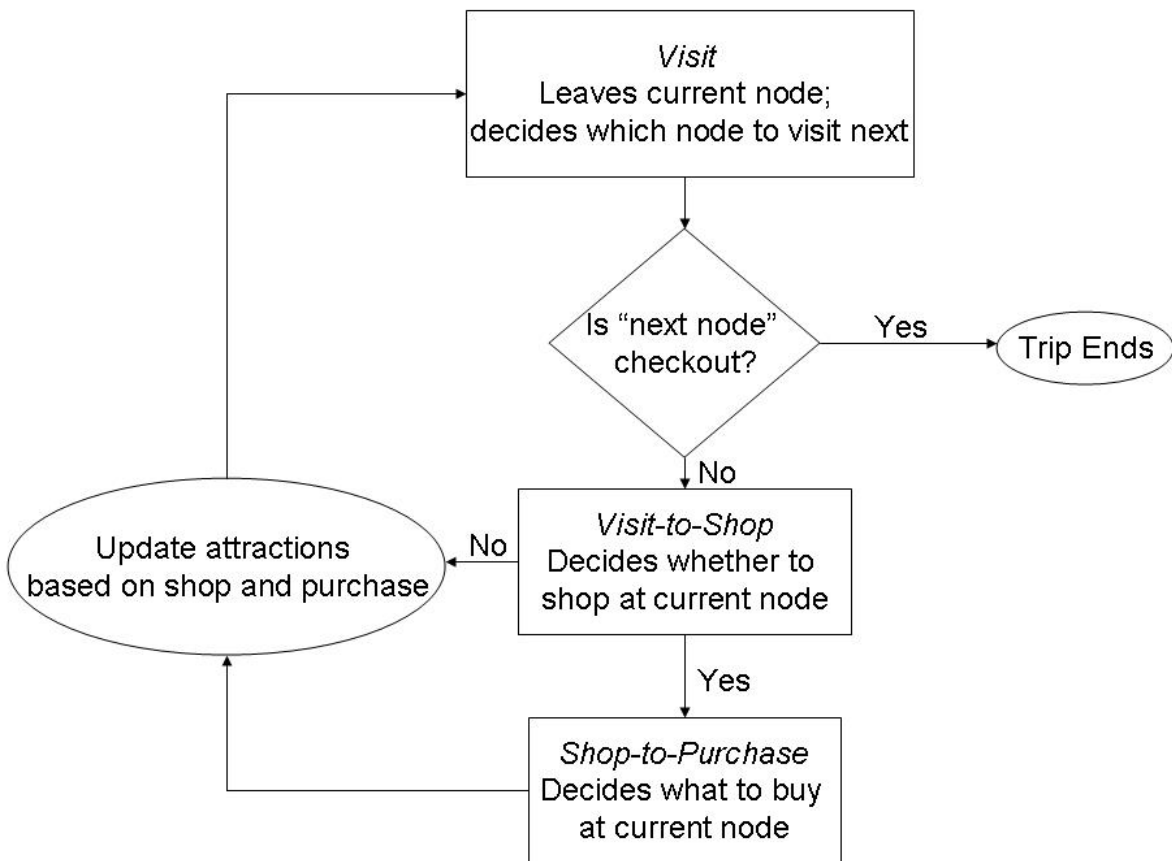


Figure 9. The shopper's in-store decision process.

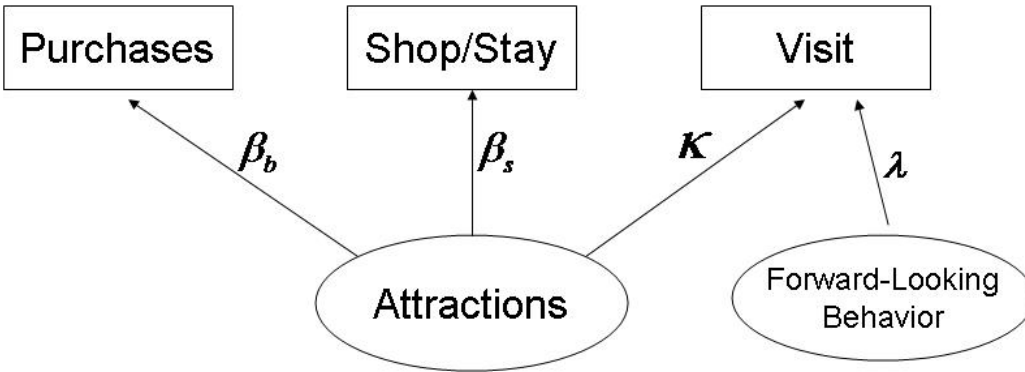


Figure 10. A schematic of the integrated model structure.

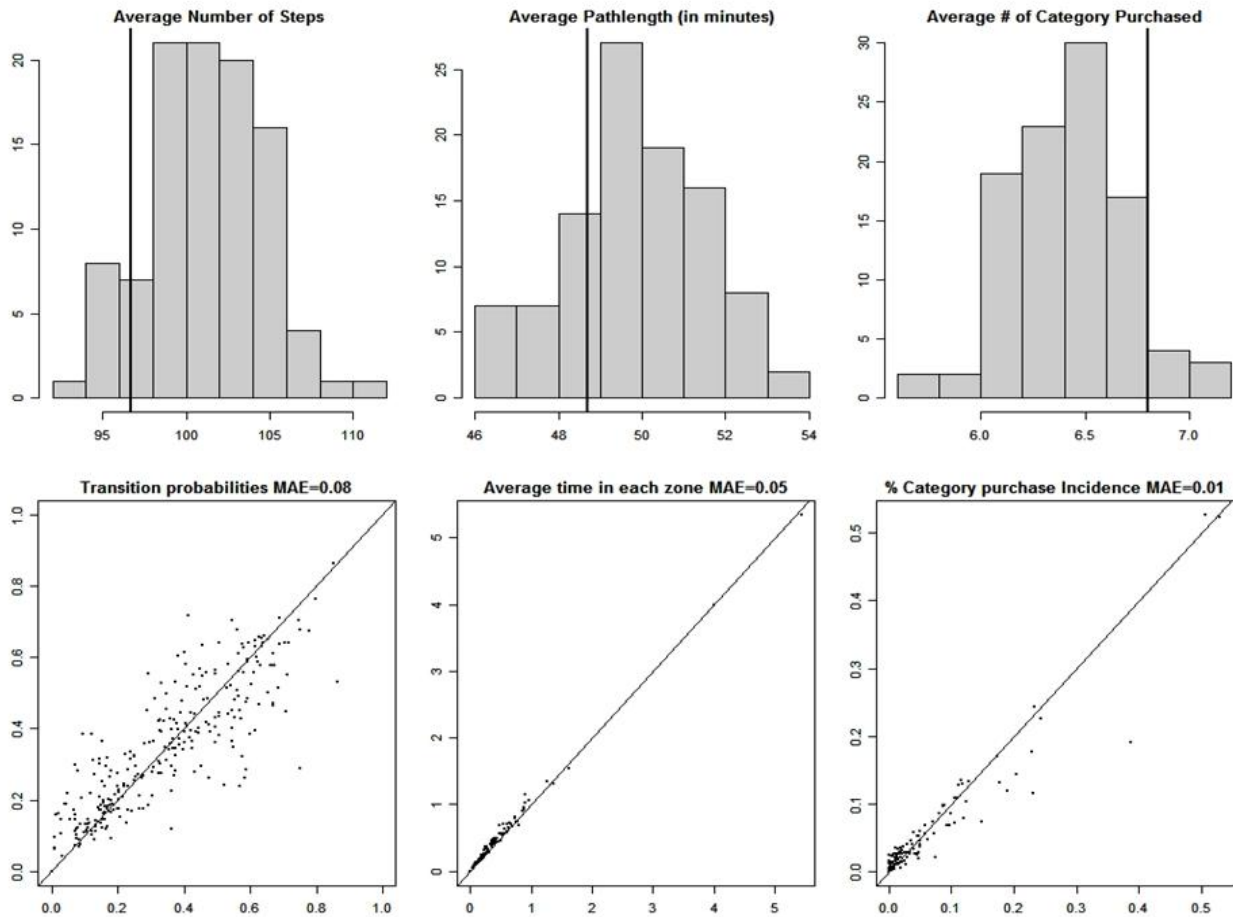


Figure 11. Posterior check for actual data. In the upper three panels, histograms of summary statistics are drawn with the solid vertical line for the actual (calibration) dataset. In the bottom three panels, the actual values of the summary statistics (calculated from the calibration dataset) are plotted on the x-axis; the mean from the posterior sample is plotted on the y-axis.

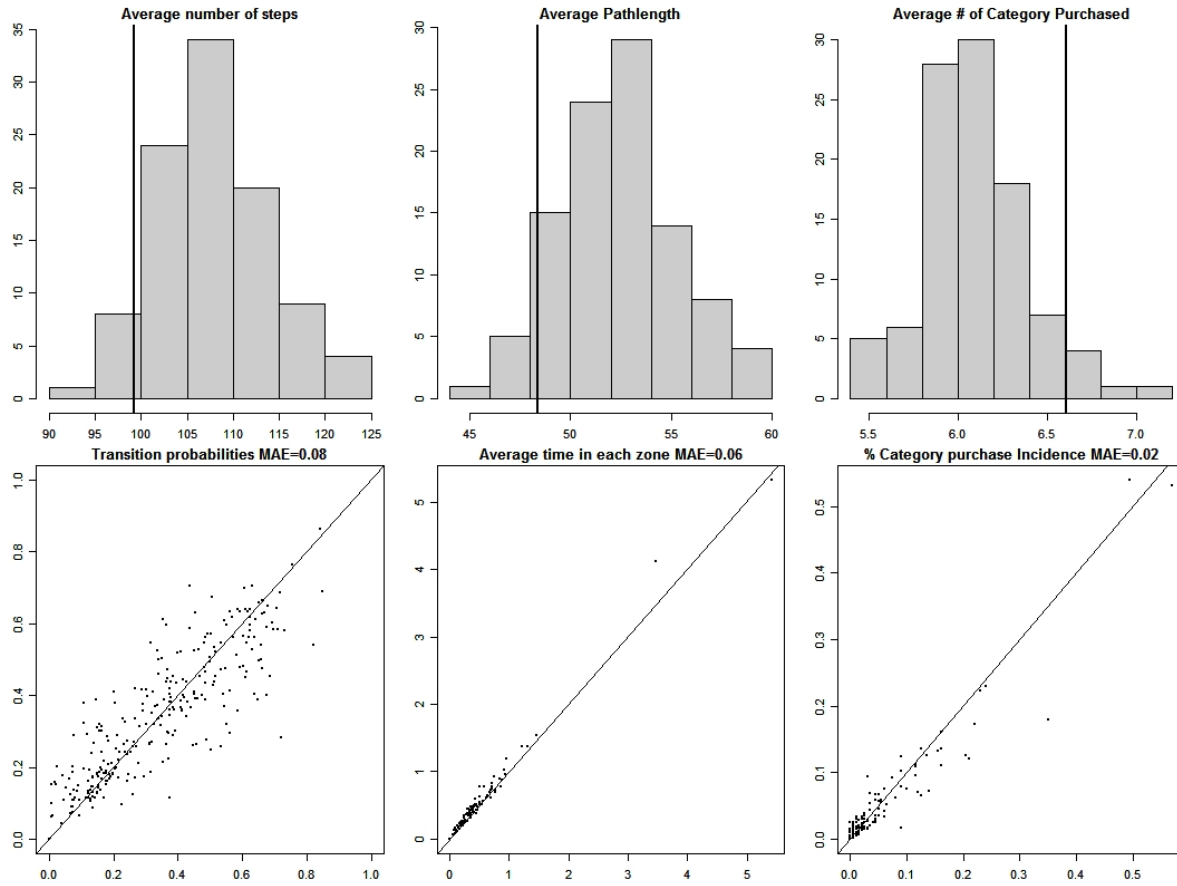


Figure 12. Holdout prediction posterior check. In the upper three panels, histograms of summary statistics are drawn with the solid vertical line for the holdout dataset. In the bottom three panels, the actual values of the summary statistics (calculated from the holdout dataset) are plotted on the x-axis; the mean from the posterior sample is plotted on the y-axis.

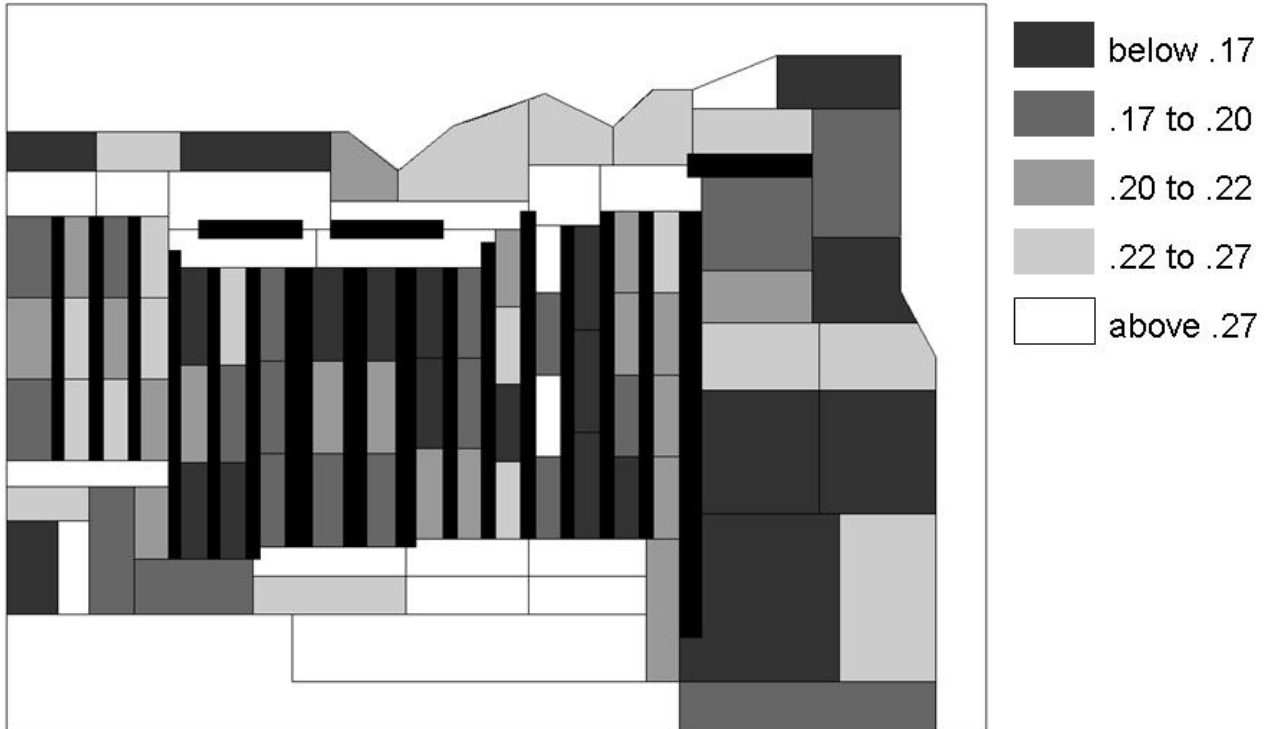


Figure 13.  $\tau^{shop}$  for each zone; zones with longer shopping time are shaded in darker gray.



Figure 14.  $Z_i$  for each zone; zones with higher  $Z_i$  are shaded in darker gray.