CAPTURING CUSTOMER EVOLVING BUYING-BEHAVIOUR IN CONSUMER PACKAGED GOODS DATA

Akomolafe Abayomi A*., Amahia G.N.**, and Chuckwu A.U.*
*** Department of Mathematics and Statistics, Joseph Ayo Babalola University, Ilesa, Osun State***
***** Department of Statistics, University of Ibadan, Nigeria***
e-mail: akomolafea@yahoo.com

Abstract

Many retailers monitor customer buying-behaviour as a measure of their stores' success. However, summary measures such as the total buying-behaviour per month provide little insight about individual-level shopping behaviour. Additionally, behaviour may evolve over time, especially in a changing environment like the Internet. Understanding the nature of this evolution provides valuable knowledge that can influence how a retail store is managed and marketed. This paper develops an individual-level model for store visiting behaviour based on juice drink and packed chicken buying-behaviour data. We capture cross-sectional variation in store-visit behaviour as well as changes over time as visitors gain experience with the store. That is, as someone makes more visits to an outlet, her latent rate of buying may increase, decrease, or remain unchanged as in the case of static, mature markets. So as the composition of the customer population changes (e.g., as customers mature or as large numbers of new and inexperienced Internet shoppers enter the market), the overall degree of buyer heterogeneity that each store faces may change. We also examine the relationship between visiting frequency and purchasing pro pensity. Previous studies suggest that customers who shop frequently may be more likely to make a purchase on any given shopping occasion. As a result, frequent shoppers often comprise the preferred target segment. We find evidence supporting the fact that people who visit a store more frequently are more likely to buy. However, we also show that changes (i.e., evolution) in an individual's visit frequency over time provides further information regarding which customer segments are more likely to buy. Rather than simply targeting all frequent shoppers, our results suggest that a more refined segmentation approach that incorporates how much an individual's behaviour is changing could more efficiently identify a profitable target segment.
Keywords**:** Consumer Packaged Goods (CPG) data, Buying Behaviour, Duration models, Heterogeneity, Nonstationarity

Introduction

Ever since shopping emerged as a viable medium for commercial purposes, analysts have closely tracked rate of buying as a principal yardstick to gauge the success of retail trade. Even after the unprecedented highs and lows that the world experienced recently, visit-related measures continue to be among the most widely used metrics by virtually all industry experts, including retail managers, journalists, consultants, and even investors (Demers and Lev 2000).

But the dramatic changes that took place in the year 2000 led to changes in the way that rate of buying measures are viewed and utilized. Measures of visitor retention and loyalty have proven to be more closely linked to the health of their businesses.

In Figure 1 we show the distribution of the number of visits per visitor for two leading Consumer Packaged Goods stores, juice drink and packed chicken, over a period of eight months in 2003. This aggregate "snapshot" shows some clear, systematic differences across the customer base, e.g., the high proportion of infrequent visitors, a pattern very similar to what has been observed for a variety of offline behavioural patterns. But unlike many well-established, relatively mature markets, rate of buying patterns are still in a state of transition. The static summary of differences in buying behaviour across customers shown in Figure 1 may be masking some significant changes over time at the individual level. When we acknowledge the existence of both of these sources of heterogeneity (i.e., differences across individuals and over time), it can become difficult – yet critically important – to separate them out from each other (as in Figure 1 in the appendix).

As a specific example, Table 1 provides an aggregate summary of the dynamics in visit patterns underlying the histograms shown in Figure 1. For the two CPG products, not only do we see growth in the total number of visits over time, but there also appears to be an increase in the number of visits per visitor. On the surface, these aggregate measures seem like great news for the store managers. However, these numbers may be misleading as the customer base is changing with the influx of new visitors (perhaps with relatively high visit rates) and the exit of more experienced users, potentially masking the true visit dynamics that exist.

Table 1. Summary of Visit Data Over Time

	Juice Drink		**Packed chicken**	
Total Number of Visits	5402	5899	1729	1890
Number of Unique Visitors	2693	2717	988	920
Visits/Visitors	2.01	2.17	1.75	2.05

Figure 2 provides a deeper view of visit trends with a histogram of the average intervisit time and how it varies across visit cycles. Again, deceptively, this view of the data indicates that as a customer repeatedly visits a store, intervisit times decrease slightly, i.e., visits become more frequent over time, consistent with the data presented in Table 1. However, this is also an inaccurate view of the actual trends in buying behaviour. Shoppers who have had the opportunity to make more buying in a fixed period of time will, by definition, have shorter average intervisit times than those shoppers for whom we could only observe a small number of visits in the same time interval. These frequent shoppers are the ones that dominate the right side of the histogram shown in Figure 1, and as such, the apparent downward trend in rate of buying times across visit cycles is unavoidable and highly nonrepresentative of the full set of visitors to each store.

In short, any attempt to summarize buying-behaviour times directly from observed data will be unable to provide accurate estimates of the true, underlying rates of repeat buying-behaviour. Likewise, any attempt to uncover dynamics in buying rates by splitting the sample into groups of "early" versus "late" visits will run into similar problems. The only way to overcome these selection biases while still ensuring the representativeness of the entire set of visitors is to use a well-specified individual level model (with suitable assumptions about heterogeneity) to obtain inferences about differences in buying-behaviour patterns across people and over time.

Therefore, one objective of this paper is to develop a probabilistic model that carefully sorts out all of these issues. An important (and unique) aspect of this model is the manner in which we allow for *evolving behaviour* in purchasing. Traditional stochastic models of purchasing behaviour assume that purchase rates are unchanging over time (e.g., Morrison and Schmittlein 1988). When these models are tested in stable and mature markets, such an assumption may indeed hold. But many new markets go through a state of flux for quite some time (Bronnenberg, Mahajan, and Vanhonacker 2000). In other words, an individual's shopping behaviour often changes as she continually adapts to a new environment. The model presented in this paper will relax the usual assumption of stationarity. More specifically, the evolutionary component of the model allows shoppers to return to the store either more or less frequently as they gain experience, while also accommodating customer attrition for those individuals who never return after one or more initial visits to the store.

From our repeat buying-behaviour model, we can estimate how likely (and when) a given buyer will return to the store as she gains more experience. Do buying times tend to speed up or slow down over a person's history, and how do these changes vary across people? Answers to these questions will give us the ability to forecast future store visits in order to better anticipate and manage the rate at which customer buys the product. We will show that our repeat model of buying behaviour forecasts rate of buying significantly better than an equivalent static model. Additionally, the evolutionary component of the model will offer useful diagnostics that will help shed light on other aspects of shopping behaviour.

Though the data is rich with behavioural information such as duration of visits, items bought, rate of buying over time, etc., we examine only the timing and frequency of store visits, as understanding buyer behaviour at this level is an important managerial issue in itself. However, despite the limited data that we use, we find that simple buying rates (and trends in these rates) are strong indicators of an individual's buying propensity. As we better describe customers in terms of their repeat buying behaviour, we relate visiting frequency to purchasing propensity. Previous studies suggest that people who shop frequently may be more likely to make a purchase on any given shopping occasion (Bellenger, Robertson, and Hirschman 1978, Jarboe and McDaniel 1987, Roy 1994). As a result, frequent shoppers are often the preferred target segment. Our data analysis strongly confirms this hypothesis, and then extends it by showing that *changes* (i.e., evolution) in an individual's buying frequency over time provide even better information. Rather than simply targeting all frequent shoppers, our results suggest that a more refined segmentation approach that incorporates how much an individual's buying behaviour is changing can more efficiently identify profitable customers for targeting purposes.

In the next two sections, we develop the model and address some of the key estimation issues that arise from the model. We then describe the CPG data that we will be using. In §5, we will present the results of the model when applied to two leading retailers and briefly discuss some of the managerial implications of the results. We validate the model by demonstrating its forecasting ability over a six-month period. Finally, we will illustrate how purchasing behaviour varies across shoppers as a function of their latent buying rates as well as changes in these rates over time.

Model Development

To understand the overall pattern of buying behaviour, let us imagine that each shopper tends to return to a store at a latent rate inherent to that individual. *When* that individual will visit the store next is driven largely by this rate of visit. Additionally, since customers are heterogeneous, this rate of visit varies from person to person. Some people may visit the store fairly frequently while others may not. But in addition to varying rates of visit across individuals, behaviour may also change over time for a given individual. As shoppers mature, perhaps as a result of increased knowledge and experience, their behaviour may evolve thereby changing their rates of buying over time.

To capture the processes described above, our model has three main components:

(1) A timing process governing an individual's rate of buying,
(2) A heterogeneity distribution that accommodates differences across people, and
(3) An evolutionary process that allows a given individual's underlying buying rate to change from one visit to the next.

Timing process with heterogeneity

As an appropriately robust starting point, repeat buying behaviour can be modeled as an exponential-gamma (EG) timing process. That is, each individual's rate of buying behaviour time is assumed to be exponentially distributed governed by a rate, λ_i. Furthermore, these individual rates of visit vary across the population. This heterogeneity can be captured by a gamma distribution with shape parameter, *r*, and scale parameter, ∙∙∙These distributions are given by the following distributions:

$$f\left(t_{ij};\lambda_i\right)=\lambda_i e^{-\lambda_i\left(t_{ij}-t_{h(j-1)}\right)} \quad \text{and} \quad g(\lambda_i;r,\alpha)=\frac{\lambda_i^{r-1}\alpha^r e^{-\alpha\lambda_i}}{\Gamma(r)} \qquad (1)$$

where λ_i is individual *i*'s latent rate of buying, *tij* is the day when the *j*th repeat buying occurred, and *ti0* is the day of their first trial. For a single visit occasion, this leads to the following familiar exponential-gamma mixture model:

$$f\left(t_{ij};r,\alpha\right)=\int_0^\infty f\left(t_{ij};\lambda_i\right).g(\lambda_i;r,\alpha)d\lambda=\frac{r}{\alpha}\left(\frac{\alpha}{\alpha+\left(t_{ij}-t_{i(j-1)}\right)}\right)^{r+1} \qquad (2)$$

While the exponential-gamma may be an excellent benchmark model, it fails to capture nonstationarity over time. To account for nonstationarity, extensions of this model are described next.

Evolving Behaviour

In the relatively new and fast-paced market environment, it is particularly important to address the issue of evolving behaviour as people are continually updating their behaviour, retailers must adapt to keep up with their customers. For example, studies have shown that as a shopper's knowledge and familiarity increase over time, the extent of shoppimg she undertakes may change, either increasing or decreasing depending on the situation (see Alba and Hutchinson 1987, Johnson and Russo 1984, Park, Iyer, and Smith 1989). Typically, increased store knowledge and familiarity lead to more efficient search behaviour. This increased shopping efficiency may have one of two effects on future store buying behaviour. One, the amount of explicit search required to make a purchase decision decreases as shoppers have more internal knowledge from which to draw (Bettman 1979, Johnson and Russo 1984, Park, Iyer, and Smith 1989). This may lead to *less* frequent store visits as the customer adapts to the shopping situation.

Though these studies suggest that store visiting behaviour may evolve as a function of experience, they are inconclusive in identifying the direction of this trend. Therefore, our model is a descriptive one that captures and characterizes any trend that may exist. We develop a flexible model that will accommodate varying magnitudes and directions of the behavioural change and offer a method to characterize the nature of this trend.

Researchers in marketing have used several different mechanisms to introduce these time-varying effects into the traditional stochastic modeling framework. For instance, Sabavala and Morrison (1981) incorporated nonstationarity by introducing a renewal process into a probability mixture model in accordance with the "dynamic inference" framework first set out by Howard (1965). Sabavala and Morrison applied this model to explain patterns of advertising media exposure over time; further applications of a similar type of renewal-process approach can be seen in Fader and Lattin (1993) as well as Fader and Hardie (1999).

While these renewal models provide one mechanism to introduce nonstationarity into a timing process, they do not offer any appealing way to capture the type of *evolutionary* process that we have described above. The aforementioned renewal models operate under the assumption that each shopper probabilistically discards their old rate parameter and draws an entirely new one from the original heterogeneity distribution, independent of previous values. This process allows for drastic changes in an individual's behaviour while maintaining the same heterogeneity distribution for the population as a whole. While this may be a powerful and effective way to capture large changes over time, it is not consistent with the type of gradual, evolutionary behavioural changes that are likely to occur from visit to visit. Furthermore, in an evolving market environment, it may be incorrect to assume that the overall heterogeneity distribution is not changing over time. The EV model that we propose allows for the population heterogeneity distribution to change as the customers that comprise the population gradually update their buying rates.

Specifically, our behavioural assumption is that customers' underlying rates of buying are continually and incrementally changing from one visit to the next. As individuals adapt to and gain experience with the new retail environment, they may return to the store at a more frequent rate, a less frequent rate, or perhaps at the same rate for the next buying. By assuming that each individual will update her latent rate, λ_i, after each buying, a very simple way to specify this updating process is as follows:

$$\lambda_{i(j+1)} = \lambda_{ij}.c \qquad (3)$$

where λ_{ij} is the rate associated with individual *i*'s *j*th repeat visit and ***c*** is a multiplier that will update this rate from one visit to the next. If the updating multiplier, ***c***, equals one, visiting rates are considered unchanging, and the stationary exponential-gamma would remain in effect. But if ***c*** is greater than one, shoppers are visiting more frequently as they gain experience, and if ***c*** is less than one, shoppers are visiting less frequently as they gain experience.

However, using a constant multiplier to update the individual λ's would be a very restrictive (and highly unrealistic) way of modeling evolutionary behaviour in a heterogeneous environment. A more general approach is to replace the scalar multiplier, ***c***, with a random variable *cij* in order to acknowledge that these updates can vary over time and across people. Each individual visit will lead to an update that may increase, decrease, or retain the previous rate of visit, depending on the stochastic nature of the updating multiplier.

To generalize (3) in this manner, we assume that these probabilistic multipliers, *cij*, arise from a gamma distribution, common across individuals and visits, with shape parameter *s* and scale parameter β. We choose the gamma distribution to describe the updating multiplier for the same reasons why we used it to describe the heterogeneity in λ. It is a very flexible distribution that accommodates a variety of shapes. This gamma distribution essentially describes the nature of the behavioural evolution faced by a given store. The updated $\lambda_{i(j+1)}$ then becomes a product of two independent gamma-distributed random variables: the previous rate, λ *ij*, and the multiplier, *cij*. The overall model, therefore, uses four parameters to simultaneously capture cross-sectional heterogeneity and evolving buying behaviour: two parameters (*r* and α) govern the gamma distribution that describes the initial heterogeneity in visiting rates, and another two parameters (*s* and β) govern the gamma distribution that describes the updating process. This is the entire model specification.

Regardless of whether the multiplier is increasing (*cij* > 1) or decreasing (*cij* < 1) a particular buying rate at a particular point in time, we expect that an individual's value will evolve relatively slowly over time. This suggests that the updating gamma distribution, *u*(*cij*; *s*, β), should have a mean fairly close to 1.0 but should also allow for more extreme increases or decreases in λ at any given update opportunity. The spread of this updating distribution is directly tied to the magnitude of the *s* and β parameters. As both of these parameters become large, the distribution degenerates towards a spike located at s/β. Taken to the extreme (i.e., *s* and β get extremely large), this model would then collapse into the deterministic updating model (3) with c= s/β.

Finally, another interesting characteristic of the updating distribution is that it allows for customer attrition, since the gamma distribution can yield a draw of *cij* extremely close to 0. When this situation arises, the customer effectively drops out and is unlikely to return to the store. Such attrition may be very common for websites and has been the centerpiece of other types of models in this general methodological area (Reinartz and Kumar 2000; Schmittlein, Morrison, and Colombo 1989). The fact that we can accommodate attrition in such a simple, natural manner is an appealing aspect of the proposed modeling approach.

Likelihood Specification

When estimating the ordinary (stationary) exponential-gamma model, there are two ways of obtaining the likelihood function for a given individual. The usual approach is to specify the individual-level likelihood function, conditional on that person's (unobserved) value of λ *i*. This likelihood is the product of *Ji* exponential timing terms, where *Ji* is the number of repeat buying made by panelist *i*, plus an additional term to account for the right-censoring that occurs between that customer's last arrival and the end of the observed calibration period (at time *T*):

$$L_i/\lambda_i = \lambda_i e^{-\lambda_i(t_{i1}-t_{10})}.\lambda_i e^{-\lambda_i(t_{12}-t_{i1})}\cdot\ldots\cdot\lambda_i e^{-\lambda_i\left(t_i J_i - t_{i(J_i-1)}\right)}.e^{-\lambda_i(T-t_i J_i)} \qquad (4)$$

To get the unconditional likelihood we then integrate across all possible values of λ, using the gamma distribution as a weighting function:

$$L_i|r,\alpha = \int_0^\infty L_i|\lambda_i.gamma(\lambda_i;r,\alpha)d\lambda_i \qquad (5)$$

Where gamma(λ_i; *r*, α) denotes the gamma distribution as shown in (1). This yields the usual exponential-gamma likelihood, which can be multiplied across the *N* panelists to get the overall likelihood for parameter estimation purposes:

$$L = \prod_{i=1}^{N} \frac{\Gamma(r+J_i)}{\Gamma(r)} \left(\frac{\alpha}{\alpha+T-t_{i0}}\right)^{r} \left(\frac{1}{\alpha+T-t_{i0}}\right)^{J_i} \qquad (6)$$

An alternative path that leads to the same result is to perform the gamma integration separately for each of the *Ji*+1 exponential terms, and then multiply them together at the end. This involves the use of Bayes Theorem to refine our "guess" about each individual's value of λ *i* after each arrival occurs. Specifically, it is easy to show that if someone's first repeat visit occurs at time *tij*, then:

$$g(\lambda_{i2}|\text{arrival at } t_{i1}) = gamma(r+1,\alpha+t_{i1}-t_{i0}) \qquad (7)$$

The gamma distribution governing the rate of buying for subsequent arrivals follows:

$$g\left(\lambda_{i(j+1)}|\text{arrival at } t_{ij}\right) = gamma(r+j,\alpha+t_{ij}-t_{i0}) \qquad (8)$$

Using this logic, we can re-express the likelihood as the product of separate EG terms

$$L = \prod_{i=1}^{N}\prod_{j=1}^{J}\left(\frac{r+j+1}{\alpha+t_{i(j-1)}-t_{i0}}\right)\left(\frac{\alpha+t_{i(j-1)}-t_{i0}}{\alpha+t_{ij}-t_{i0}}\right)^{r+j}.S\left(T-t_{iJ_i}\right) \qquad (9)$$

which collapses into the same expression as (6).

When we introduce the nonstationary updating distribution, the multipliers (*cij*) change the value of λi from visit to visit, thereby requiring us to use the sequential approach given in (9) to derive the complete likelihood function. We need to capture two forms of updating after each visit: one due to the usual Bayesian updating process (which is associated with stationary behaviour given by (8)) and the other due to the effects of the stochastic evolution process. Therefore, the distribution of buying rates at each repeat visit level is the product of two gamma distributed random variables – one associated with the updating multiplier and one capturing the previous visiting rate. For the case of panelist *i* making her *j*th repeat visit at time *tij*:

$$G\left(\lambda_{i(j+1)} \middle| \text{arrival at } t_{ij}\right) = gamma\left(r + j, \alpha + t_{ij} - t_{i0}\right).\, gamma(s, \beta) \quad (10)$$

One issue with this approach is that the product of two gamma random variables does not lend itself to a tractable analytic solution. However, there is an established result (see, e.g., Kendall and Stuart 1977, p. 248) suggesting that the product of two gamma distributed random variables can be approximated by yet another gamma distribution, obtained by multiplying the first two moments about the origins of the original distributions:

$$\begin{gathered} m_1^{(\lambda_{i(j+1)})} = m_1^{(\lambda_{ij})} \text{ x } m_1^{(c_{ij})} \\ \text{and} \\ m_2^{(\lambda_{i(j+1)})} = m_2^{(\lambda_{ij})} \text{ x } m_2^{(c_{ij})} \end{gathered} \quad (11)$$

As shown in Appendix A, this moment-matching approximation, used in conjunction with Bayesian updating, allows us to recover the updated gamma parameters that determine the rate of buying, λij, for panelist *i*'s *j*th repeat visit as follows:

$$r(i, j+1) = \frac{[r(i,j)+1].s}{[r(i,j)+2].(s+1) - [r(i,j)+1].s} \quad (12)$$

$$\alpha(i, j+1) = \frac{\lfloor \alpha(i,j) + t_{ij} - t_{i(j-1)} \rfloor.\beta}{[r(i,j)+2].(s+1) - [r(i,j)+1].s} \quad (13)$$

where *r*(*i*, 1) and ·(*i*, 1) are equal to the initial values of *r* and ··as estimated by maximizing the likelihood function specified in (8).

We performed 20 separate simulations to verify the accuracy of using such a moment-matching approximation. In each simulation, we first generated 1000 random draws from a gamma distribution with randomly determined shape and scale parameters to represent initial λ·values. Then, a matrix of updating multipliers was also simulated for a series of five updates (i.e., five future repeat visits). Each 1000x5 matrix was generated by taking draws from a gamma distribution, again with randomly determined shape and scale parameters, where columns one through five represented the updates after one to five visits. The updated λ·series after five repeat visits was calculated using two methods (1) direct (numerical) multiplication of the 1000 initial λ's and the five updating series or (2) randomly drawing 1000 values from the distribution resulting from the moment-matching approximation across all five updates. A Kolmogorov-Smirnov test of fit indicated that, for each of the 20 simulations, the distribution of values resulting from the moment-matching approximation is not significantly different from that resulting from the direct multiplication of these random variables. Therefore, we are confident that the moment-matching approximation accurately captures the gamma distributed updating process we wish to model.

After incorporating the evolution process into our model, the likelihood function to be maximized follows:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J} \left(\frac{r(i,j)}{\alpha(i,j)} \right) \left(\frac{\alpha(i,j)}{\alpha(i,j) + t_{ij} - t_{i(j-1)}} \right)^{r(i,j)+1} . S\left(T - t_{iJ_i}\right) \qquad (14)$$

where *r*(*i, j*) and α (*i, j*) are defined in equations (12) and (13) while the survival function, *S*(*T-tij*), is defined as:

$$S\left(T - t_{ij}\right) = \left(\frac{\alpha(i, J_i + 1)}{\alpha(i, J_i + 1) + T - t_{iJ_i}} \right)^{r(i, J_i + 1)} \qquad (15)$$

For the special case in which behaviour is not evolving and the nonstationary updating distribution degenerates to a spike at 1.0 (i.e., $s = \beta =$ M, where M approaches infinity), then this equation collapses down exactly to the ordinary (stationary) exponential-gamma model.

Data

We apply the models described in the previous section to fruit juice data collected from a Departmental store retail outlet in Ibadan. For our purposes, we are interested in the dates of the visits each panelist makes to a given store. To consolidate the data just a bit, we aggregated buying behaviour to weekly level. For example, a shopper may leave a store briefly and return later that week. However, this second visit is unlikely to be considered a repeat visit but rather an extension of the first visit. Therefore, if a given panelist were to visit a particular store multiple times in a single week, we would encode that behaviour as just one visit for the day when the session began. Since we are interested in the timing and frequency of repeat buying, our dataset describes each customer as a sequence of weeks when buyers were made. All customers that have visited the store of interest at least once during the observation period were included in this dataset. We use data from two stores – selling juice drink and packed chiken. Fruit juice attracted 2,140 unique buyers during this six-month period totaling 9,260 visits, while packed chicken had 975 visitors making 3,210 visits (refer back to Figure 1 and Table 1 for more detailed summaries of the data).

Model Results

Before estimating the evolving visit model developed in §3, we first examine the static exponential-gamma timing model as a benchmark. When the static, two-parameter model is applied to the eight months of fruit juice buying behaviour data, we find that the mean rate of visit (E[λ]=r/α) is 0.0112. In other words, the expected inter visit time ($1/\lambda$) is

89.3 days, which is high, but reasonably consistent with the summary statistics mentioned earlier. But beyond their ability to capture the mean of the heterogeneous visiting process, the model parameters also provide useful information about the nature of the distribution of visit rates across the population. With a shape parameter of 0.271 and a scale parameter of 42.955, the distribution of visiting rates can be described by the gamma distribution in Figure 3. This distribution has a large proportion of the population with very low rates of buying. The median rate, according to this model, is 0.005, corresponding to an buying rate- time of 200 days. This distribution of rates is very consistent with the observed histogram of visit frequency (Figure 1), suggesting that the stationary EG model provides a very good benchmark model for visit behaviour.

A principal reason for these high expected buying rate is the fact that the stationary model does not allow shoppers to drop out and never return. As a result, a customer who has actually dropped out would be seen by the model as still being "alive," but having a very slow visiting rate, since she would not have yet returned to the store by the end of the observation period.

In Table 2, we contrast the parameter estimates and fit statistics for the static EG model with those from our four-parameter model of evolving visiting behaviour. Not only does the latter model fit the data better, but it also has more intuitively appealing results. While the basic shape of the gamma distribution for initial buying rates (shown in Figure 4a) may appear to be similar to that of the static EG model, it is less dominated by low-frequency shoppers, leading to a substantially lower mean buying rate (52 days, E[λ] = 0.019). Likewise, the median buying rate shrinks to 167 days (median λ $\cdot$= 0.006). These differences reflect the fact that dropout – or other types of evolution – can take place as the customer becomes more familiar with the product.

Table 2. *Model Results for Amazon*

	Stationary EG Model	Evolving Visit Model
R	0.483	0.324
α	42.955	16.857
S		2.299
β		2.304
LL	-34,347.2	-33,648.0
No. of parameters	2	4
CAIC	68,711.17	67,296.0

According to the evolutionary model, the mean update for any given visit (s/β) is very close to one (0.998) suggesting, perhaps, that it is a fairly stationary process. However, a closer look at the distribution (see Figure 4b) shows that there is significant variance about this mean. Though the mean update is close to one, the distribution is quite skewed. With a median value of cij=0.858, buyers tend to return to the store at slower rates from visit to visit. The implications of these results are in stark contrast to the measures summarized in Table 1 that implied increased visiting frequency over time.

Though other models have acknowledged the issue of nonstationarity, many of them have focused primarily on dropout (Eskin 1973, Kalwani and Silk 1980, Schmittlein, Morrison, and Colombo 1989). These models allow for individuals to make several purchases, become disenchanted, and never purchase again. To test if the evolving visit model is capturing evolving behaviour over time in addition to a dropout phenomenon, we also estimated an exponential-gamma model with a dropout component similar to that specified by Eskin (1973) and Fader and Hardie (1999)

In the EG model with dropout, the probability of buying given that you are an active buyer is modeled as an exponential-gamma process. However, the probability of being an active buyer after the *j*th visit, π_j, is determined by the following:

$$\pi_j = \Phi\left(1 - e^{-\theta j}\right) \quad (16)$$

where ϕ is the long run probability of a customer remaining active, and θ is the rate at which the π approaches this long run probability. Though the EG model with dropout provides a significant improvement in fit over the stationary EG model (LL = -33,804.7), it does not approach the performance of the evolving visit model which has the same number of parameters.
This suggests that the evolving visit model is capturing a phenomenon in addition to just dropout.

Validation

While we have discussed the fact that the evolving visit model fares well on a relative basis compared with various benchmark models, we have yet to show that it performs sufficiently well on an absolute basis. In this section, we validate the evolving visit model by examining the accuracy of longitudinal forecasts. Because the evolving visit model relies on an approximation (11) to specify and estimate the model, we need to perform simulations to generate data for tracking/forecasting purposes. This is a straightforward and computationally efficient task. For each iteration of the simulation, we create a simulated panel that matches the actual panel in terms of its size and the distribution of its initial visit times. We then generate a sequence of repeat visits using the parameter estimates from the model. This requires us to maintain a time-varying vector of λ's for each panelist, which starts with random draws from the initial (r, α) gamma distribution, and then gets updated using the (s, β) gamma distribution after each simulated exponential arrival occurs. We continue this process until every simulated panelist gets past the tracking/forecasting horizon of interest to us. It is then a simple matter to count up the number of visits on a week-by-week basis for each iteration of the simulation. We then average across 1000 iterations to generate the tracking and forecasting plots. Using the MATLAB programming language, each of these iterations takes only a few seconds on a standard PC, and we see very consistent convergence properties after a few dozen iterations.

Before creating the forecasts, we re-estimate both models (stationary and evolving EG) using only the first half (i.e., four months) of the dataset. (It is worth noting that the evolving model parameters are quite robust to this changing calibration period, while the stationary model has a noticeably higher visit rate over the shorter period – clear evidence of the slowdown discussed earlier). To generate the forecasts for the evolving visit model, we use the simulation procedure described above. For the stationary EG model, the expected number of repeat visit per week can be calculated directly as follows:

$$E[repeat\ \text{visits}_w] = N_w t\left(\frac{r}{\alpha}\right) \quad (17)$$

where *Nw* is the number of eligible repeat buying in week *w* and *t* is the time period of interest, i.e., seven days in this case. Figure 5 shows cumulative forecasts as well as actual visits for the juice drink buying behaviour.

Both models seem to track the data quite well over the initial four-month calibration period. However, as we enter the forecasting period, the stationary EG model begins to diverge, ultimately over predicting by 37% for juice drink buying behaviour at the end of the eight month period. It overestimates the number of visits per week as it does not recognize that shoppers are returning less frequently over time. The evolving visit model, however, forecasts quite accurately, well within 5% of the actual sales line throughout the forecast period. This is an impressive achievement and serves as a strong testimonial to the validity of the assumptions, structure, and parameter estimates associated with the proposed model.

Results for PACKED CHICKEN

The same set of models and analyses were also applied to packed chicken buying behaviour data (results in Table 3). We see a remarkably similar set of patterns as in the case of juice drink. In moving from the static EG model to the evolving specification, we see significantly shorter intervisit times, since the latter model can accommodate customer dropout. We also see, once again, that the mean update is close to 1.0 (0.991), but with a median of 0.837, customer shopping frequency is more likely to decrease than increase after each visit. We emphasize once more that these results contradict the summary statistics from Table 1, which seemed to imply that shopping frequency is increasing from one visit cycle to the next.

Table 3. Model Results for packed chicken buying-behaviour

	Stationary EG Model	Evolving Visit Model
r	0.255	0.165
α	28.305	8.889
s		2.084
β		2.104
LL	-9,459.6	-9,120.7
No. of parameters	2	4
CAIC	18,934.1	18,271.0

Other benchmark models (involving dropout and/or constant updates) proved once again to be vastly inferior to the evolving visit model. Finally, our forecast validation led to encouraging results with projected visits only 2% above the actual number at the end of the eight month period, compared to a 40% over-forecast for the stationary model. While we are very encouraged by these strong initial results, we are also surprised at the degree of similarity seen for these two Consumer Packaged Goods. We certainly do not want to suggest that the specific patterns captured here will generalize to all online (or offline) retailers, but this should be ample motivation for future studies to find and describe a broader range of buying-behaviour.

Discussion and Conclusions

Many skeptics claim that the offline or online shopping is nothing more than a new distribution channel, and thus it should not change the way we examine customer behaviour. While this may be true in certain respects, this paper highlights some of the uniquely different research perspectives that we gain from examining Consumer Packaged Goods buying-behaviour data. Thanks to rich new sources of data , we can now examine behavioural phenomena that would be impossible to study using more traditional sources, such as grocery store scanner data.

The detailed, disaggregate data available to us make it possible to study the evolution of visit behaviour at a retail store. The model developed here is tailored specifically to both online and offline buying.

We posit a behaviourally plausible – and highly parsimonious – model that allows visiting behaviour to evolve gradually over time, although it also allows for more abrupt changes, such as permanent dropout from the store. And indeed, our empirical analysis reveals the fact that the average update in household visiting rates is a multiplier close to 1.0, but there is significant spread around this value. Additionally, the manner in which we implement this updating scheme – a gamma distribution to capture the different values of these multipliers – is a new methodological contribution, which merits consideration for other types of non-stationary modeling contexts.

Use of the model reveals that individual-level behaviour patterns appear to contradict the perspective that one would obtain from examining the aggregate data alone. Specifically, the aggregate data seem to indicate an acceleration of visiting behaviour at each of two leading consumer packaged goods products, yet our model parameters suggest that the typical shopper is experiencing a gradual slowdown in her buying rate over time. The difference here is that an increasing number of new visitors are coming to each store over time, masking the slowdown that may be occurring for many experienced visitors. This effect could have dramatic implications for managers who neglect to examine their data at a sufficiently fine level of disaggregation.

Beyond the intuitive appeal of the model specification and its estimated parameters, we also show that it has excellent validity from an out-of-sample forecasting perspective. For both retail stores, the model tracks future visiting patterns extremely well, remaining within 5% of the actual data over the entire duration of a four-month holdout period. While this model was not constructed with forecasting in mind as a principal objective, this result certainly speaks well about its overall versatility.

Perhaps the most dramatic demonstration of the model's validity and usefulness is its ability to delineate highly significant differences in purchasing behaviour across shoppers. There is a significant amount of past literature suggesting that customers who visit a particular store frequently also tend to buy something during a relatively high proportion of those shopping trips. We provide strong confirming evidence of this hypothesis. But the evolutionary nature of our model allows us to test an equally compelling complementary hypothesis: people who experience increases in their buying rates over time are more likely to purchase something at any given visit than those who are slowing down.

Limitations and Future Research

Since this paper is among the first attempts to carefully examine buying-behaviour using CPG data, we have deliberately kept the model as clear and simple as possible in order to highlight the chief phenomena that we have observed in these datasets. However, one limitation is the fact that the data does not reveal when each customer first started visiting each store. As a result, the model is only able to provide a description of customer buying rates and how they are changing during the data period being examined. In time, as all potential customers have adopted and become accustomed to the online store environment, perhaps no evolution will be detected. However, the evolving model presented in this paper will allow retailers to monitor trends until that time comes and also know *when* that time has arrived.

But as the types of data and methods employed here become more commonplace, we can see several extensions to the model that may be worth pursuing. Because we have been emphasizing the importance of evolution in a new marketplace (such as online sales of books) we have paid little attention to the fact that these markets might eventually shift towards a more steady-state nature, i.e., with updates occurring less frequently and with smaller magnitudes. It is unlikely that the same distribution of updating multipliers (chj) will stay in place over a long period of time. Perhaps this distribution starts to collapse towards a spike at 1.0 as the market matures. The excellent performance of our holdout forecasts does not seem to indicate any such pattern in our datasets, but as our observation window extends to several years' worth of data in the future, we might see more benefits from such a specification.

Finally, it is important to acknowledge the need for further process-oriented research to better explain and extend the psychological mechanisms underlying our findings concerning the relationships between conversion rates and buying behaviour. While there is ample theoretical reasoning behind the well-established frequency hypothesis, it would be useful to establish an equally solid base of explanations and controlled experimental evidence for the effect of positive vs. negative evolution, as well as the substantial interaction effect we have observed.

REFERENCES

Bellinger, Danny N., Dan H. Robertson, and Elizabeth C. Hirschman (1978), "Impulse Buying Varies by Product," Journal of Advertising Research, 18 (December), 15-18.

Bronnenberg, Bart J., Vijay Mahajan, and Wilfried R. Vanhonacker (2000), "The Emergence of Market Structure in New Repeat-Purchase Categories: The Interplay of Market Share and Retailer Distribution," Journal of Marketing Research, 37 (February), 16-31.

Eskin, Gerald J. (1973), "Dynamic Forecasts of New Product Demand Using a Depth of Repeat Model," Journal of Marketing Research, 10 (May), 115-129.

Fader, Peter S. and Bruce G.S. Hardie (1999), "Investigating the Properties of the Eskin/Kalwani & Silk Model of Repeat Buying for New Products," in Lutz Hildebrandt, Dirk Annacker, and Daniel Klapper (eds.), Marketing and Competition in the Information Age, Proceedings of the 28th EMAC Conference, May 11-14, Berlin: Humboldt University.

Fader, Peter S. and James M Lattin (1993), "Accounting for Heterogeneity and Nonstationarity in a Cross-Sectional Model of Consumer Purchase Behaviour, Marketing Science, 12 (3), 304-317.

Howard, R.A.(1965), "Dynamic Inference," Operations Research, Vol. 13 (2), 712-733.

Jarboe, Glen R. and Carl D. McDaniel. (1987). "A Profile of Browsers in Regional Shopping Malls," Journal of the Academy of Marketing Science, 15 (Spring): 46-53.

Kalwani, Manohar and Alvin J. Silk (1980), "Structure of Repeat Buying for New Packaged Goods," Journal of Marketing Research, 17 (August), 316-322.

Morrison, Donald G. and David C. Schmittlein (1988), "Generalizing the NBD Model for Customer Purchases: What Are the Implications and Is It Worth the Effort?" Journal of Business & Economic Statistics, 6 (2), 1

Reinartz, Werner J. and V. Kumar (2000), "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," Journal of Marketing, 64 (4), 17

Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987), "Counting Your Customers: Who Are They and What Will They Do Next?" Management Science, 33, 1-24.

APPENDIX A

Moment-Matching Approximation of the Product of Two Gamma Distributions

If x and y are two gamma distributed random variables,

x ~ Gamma (r, a)

y ~ Gamma (s, b)

then the product, z = xy, can be assumed to be a gamma distributed random variable

z ~ Gamma (R, A)

with shape and scale parameters, R and A, such that the first two raw moments of the z-distribution is the product of the moments of the x- and y-distributions.

$$m_1^x = \frac{r}{\alpha} \qquad\qquad m_2^y = \frac{r(r+1)}{\alpha^2}$$

$$m_1^y = \frac{s}{\beta} \qquad m_2^y = \frac{s(s+1)}{\beta^2}$$

$$m_2^z = m_1^x . m_1^y = \frac{rs}{\alpha\beta} \qquad m_2^z = m_2^x . m_2^y = \frac{r(r+1)s(s+1)}{\alpha^2 \beta^2}$$

Since the first moment of the z-distribution, m_1^z*, is R/A and the second moment,* m_2^z *z, is R(R+1)/A*^{2}*, we can solve for R and A with the following two equations:*

$$\frac{R}{A} = \frac{rs}{\alpha\beta} \qquad \frac{R(R+1)}{A^2} = \frac{r(r+1)s(s+1)}{\alpha^2 \beta^2}$$

Therefore, the gamma distribution describing the product of two independently distributed gamma random variables has shape and scale parameters that can be calculated from the parameters of the multiplying distributions.

$$R = \frac{rs}{(r+1)(s+1) - rs} \qquad A = \frac{\alpha\beta}{(r+1)(s+1) - rs}$$

with Bayesian updating after observing one arrival at time t…

$$R = \frac{(r+1)s}{(r+2)(s+1) - (r+1)s} \qquad A = \frac{(\alpha + t)\beta}{(r+2)(s+1) - (r+1)s}$$