Forecasting Repeat Sales at CDNOW: A Case Study

Peter S. Fader
Bruce G. S. Hardie[1]

December 1999
Revised July 2000

[1]Peter S. Fader is Associate Professor of Marketing at the Wharton School, University of Pennsylvania (email: fader@wharton.upenn.edu; web: www.petefader.com). Bruce G. S. Hardie is Assistant Professor of Marketing, London Business School (email: bhardie@london.edu; web: www.brucehardie.com).
The authors thank the editors and reviewers for an unusually encouraging set of comments on an earlier draft of this paper.

Abstract

We describe a modeling exercise, conducted in conjunction with the online music retailer CDNOW, where the goal was to develop a simple stochastic model of buyer behavior capable of generating a medium-term forecast of aggregate CD purchasing by a cohort of new customers.

Weekly sales are modeled using a finite mixture of beta-geometric distributions with a separate time-varying component to capture nonstationarity in repeat buying. The resulting model can easily be implemented within a standard spreadsheet environment (e.g., Microsoft Excel).

We demonstrate that the model does a good job at describing the underlying sales patterns, as well as generating an excellent medium-term forecast.

1 Introduction

With the growth of e-commerce, many companies are now capturing transaction data of such detail that could have only been dreamed of ten years ago. The challenge facing them is what to do with the data that are rapidly accumulating within their databases. While the literature on database marketing and one-to-one marketing talks of the use of models to gain managerial insights (Mulhern 1999, Forrester Report 1999), actual examples (especially on an enterprise-wide scale) are relatively limited.

In this paper, we describe an exploratory study undertaken in conjunction with CDNOW, a leading online music retailer. The objective was to develop an easily implementable model of buyer behavior capable of generating a medium-term forecast of aggregate CD purchasing by a cohort of CDNOW customers. Aggregate-level forecasts are a critical input to any attempt to value a customer base, and serve as a diagnostic to help gauge the effectiveness of various short-term marketing programs (e.g., the provision of a baseline sales estimate against which the performance of a promotion can be evaluated).

At the time of this study, many commentators felt that the Internet was still at its nascent stage and therefore any forecasting exercise would have been futile. For example, Buchanan and Lukaszewski (1997, p. 143) made the comment that:

> At this stage of the Internet's evolution, accurate sales forecasts are as much of an oxymoron as "military intelligence".

However, it was — and still is — our view that the underlying patterns of buying behavior are consistent across purchasing channels — including the Internet — and therefore the development of a forecasting model is a fruitful exercise.

The paper is organized as follows. First, we discuss the background to the modeling exercise and present the dataset on which this work is based. We then develop the stochastic model of buyer behavior which is used to generate forecasts of future CD purchasing. This is followed by the empirical analysis where we examine the fit of the proposed model and its forecasting ability. We finish with a discussion of this work and outline several avenues for future research.

2 Background

CDNOW is one of the oldest and largest online retailers, having sold different forms of music (and related products) on the World Wide Web since 1994. They carry approximately 500,000 different albums — about ten times the size of the typical "bricks and mortar" megastore — and report store traffic of over 200,000 visitors per day. During their first five years of operations, CDNOW attracted over 700,000 unique customers who made purchases at the website.

For the purposes of this work, we focus on a single cohort of new customers who made their first purchase at the CDNOW website in the first quarter of 1997. We have data covering their initial ("trial") and subsequent ("repeat") purchases for the three month period (1/97–3/97) during which over 23,000 individuals purchased nearly 70,000 units (CDs). We are interested in forecasting the future (repeat) purchasing of these customers using a model calibrated with these quarter 1 data. Furthermore, our goal is to develop a model that can be implemented with minimal difficulty in the target organization.

Faced with these data, the analyst may choose to apply an existing model specifically designed for customer-base analysis (e.g., Allenby, Leone, and Jen 1999; Colombo and Jiang 1999; Schmittlein, Morrison, and Colombo 1987; Schmittlein and Peterson 1994). Some of these models treat underlying buyer behavior as if it were stationary. For example, Schmittlein, Morrison, and Colombo (1987) model individual-level purchasing via a Poisson counting process and overlay an exponential "death process" to capture customer attrition; customer heterogeneity in these two elements is captured via gamma distributions. Aggregate-level predictions, as well as individual inferences, can be derived from such a model.

The problem with these approaches is that they require an analyst who is used to dealing with large individual customer-level datasets — all of the above models are estimated using the detailed customer-level data — and who has relatively sophisticated model-building skills and access to appropriate computational software. However, such people are rare in most organizations. Furthermore, the sophistication of most of these models is such that it can be difficult to explain their logic to managers, which serves as a further barrier to their implementation.

We therefore seek to develop a simple forecasting model that can easily be implemented using readily available software with which most business people will be familiar — ideally a common

spreadsheet package such as Microsoft Excel. Central to this goal is structuring the raw data in an aggregate form that is easy for the analyst to manage, while at the same time still allowing for the development of a well-specified model of repeat purchasing.

For the cohort of customers who first purchased at the CDNOW website in the first quarter of 1997, we choose to work with the summary of total purchasing as presented in Table 1. This gives the distribution of the number of units purchased for each of the twelve weeks, along with details of total purchasing and the number of new customers (triers) in each week. In estimating our model, we will use no further information beyond the aggregate numbers shown in this table. Later on in the paper, we will use some more disaggregate measures to help gauge the quality of the model's tracking and forecasting capabilities. We wish to emphasize that the simplicity of this data structure is an important contribution of our work.

Units	Week											
Purchased	1	2	3	4	5	6	7	8	9	10	11	12
0		1478	3033	4763	6608	8616	10829	12716	14698	16774	18881	20902
1	750	852	984	1066	1237	1262	1204	1278	1397	1444	1387	1148
2	383	387	456	484	566	649	592	606	644	659	677	663
3	191	214	270	267	293	320	302	343	365	374	355	367
4	95	120	114	161	163	196	156	195	179	187	199	182
5	55	72	68	89	96	96	80	100	95	118	94	120
6	36	40	42	40	51	54	65	45	75	71	72	54
7	18	12	27	30	36	40	39	31	41	37	30	43
8	12	15	9	21	19	21	20	24	23	29	24	32
9	9	9	8	9	21	14	21	8	14	9	12	16
10+	25	17	27	32	36	55	39	35	48	42	50	43
Total units	3627	3857	4512	5054	5843	6456	5906	6077	6757	6848	6770	6781
Incr. triers	1574	1642	1822	1924	2164	2197	2024	2034	2198	2165	2037	1789
Cum. triers	1574	3216	5038	6962	9126	11323	13347	15381	17579	19744	21781	23570

Table 1: Raw Data

While this is a very convenient summary of the customers' purchasing, it suffers from two critical shortcomings: (1) we have no explicit information on the breakdown of trial vs. repeat sales in each week, and (2) we cannot see the longitudinal series of purchase events at the household level, thereby making it impossible to construct a standard model of repeat purchasing (i.e., depth of repeat or counting). We must therefore develop a model of week-by-week repeat purchasing whose parameters can be estimated using the above data. We now turn our attention to the development of such a model.

3 Model Development

Our objective to is develop a simple stochastic model of buyer behavior capable of producing a medium-term forecast of unit purchases by the cohort of new customers whose total purchasing during the first quarter of 1997 is summarized in Table 1. Let us consider these data more carefully, focusing first on the column corresponding to week 2. This reports the distribution of purchase quantity for that week by the 3216 customers who could have made a purchase. This set is comprised of 1642 customers who made their first purchases at the CDNOW website in week 2, as well as 1574 customers who first purchased at it in week 1 and who therefore may be back in the market for additional (repeat) purchases in week 2. By definition, the 1642 week 2 triers must have purchased at least one unit. This implies that the 1478 people who made no purchase at the website in week 2 must be customers who made a trial purchase in week 1. Thus we have $1574 - 1478 = 96$ week 1 triers who made repeat purchases in week 2. In other words, this observed distribution of week 2 purchasing represents a *mixture* of purchases by those whose first purchase occasion occured in week 2 and repeat purchases by those who tried in week 1. Therefore, the probability of observing someone purchasing x units in week 2 is simply a weighted average of the probability that a week 2 trier bought x units during her initial week, and the probability that a week 1 trier bought x units on at least one *repeat* purchase occasion in week 2. The weights are determined by the number of triers in weeks 1 and 2, i.e.,

$$P(X_2 = x) = \frac{1642}{1574 + 1642} \times P(T_2 = x) + \frac{1574}{1574 + 1642} \times P(R_{2|1} = x)$$

where $P(T_2 = x)$ is the probability that a randomly chosen customer making her first purchase(s) at CDNOW in week 2 buys x units, and $P(R_{2|1} = x)$ is the probability that a randomly chosen customer who first purchased in week 1 purchases x units in week 2.

Similarly, the column corresponding to week 3 reports the distribution of purchase quantity for the 5038 customers who could have made a purchase that week: 3216 of these customers made their first purchase in weeks 1 or 2, and 3033 of these people made no (repeat) purchase in week 3. We therefore have 183 week 1 and week 2 trialists making repeat purchases in this week, but we do not observe the specific number of week 1 versus week 2 triers, nor each of

these groups' respective distribution of units purchased. Extending the same logic from above, however, we can express the probability of observing x purchases in week 3 as a weighted average of the probability that a week 3 trier made x purchases, the probability that a week 2 trier made x repeat purchases in week 3, and the probability that a week 1 trier made x repeat purchases in week 3, i.e.,

$$P(X_3 = x) = \frac{1822}{1574 + 1642 + 1822} \times P(T_3 = x) + \frac{1642}{1574 + 1642 + 1822} \times P(R_{3|2} = x) + \frac{1574}{1574 + 1642 + 1822} \times P(R_{3|1} = x)$$

where the weights are determined by the number of triers in weeks 1–3.

More generally, the distribution of purchases in week w can be modeled using a finite mixture model with known mixing weights:

$$P(X_w = x) = \frac{1}{\sum_{i=1}^{w} n_i} \left[n_w P(T_w = x) + \sum_{i=1}^{w-1} n_i P(R_{w|i} = x) \right] \tag{1}$$

where n_i is the number of triers in week i (i.e., customers making their first purchase(s) at the CDNOW website), $P(T_w = x)$ is the probability that a randomly chosen customer making her first purchase(s) at CDNOW in week w buys x units, and $P(R_{w|i} = x)$ is the probability that a randomly chosen customer who first purchased in week i buys x units in week w. We therefore need to develop submodels for $P(T_w = x)$ and $P(R_{w|i} = x)$.

Modeling Trial Purchases

Let the random variable T_w denote the number of units purchased in week w by a customer whose trial purchase occurs in week w. (Note that, by definition, T_w is a zero-truncated discrete random variable.) Our submodel for the distribution of T_w is based on the following two assumptions:

i. At the level of the individual customer, T_w is distributed according to a shifted geometric distribution with parameter q_T and probability mass function

$$P(T_w = x \,|\, q_T) = \begin{cases} q_T(1 - q_T)^{x-1} & x = 1, 2, \ldots ;\ 0 < q_T < 1 \\ 0 & x = 0 \end{cases}$$

ii. q_T is distributed across the population according to a beta distribution with parameters α_T and β_T, and pdf

$$g(q_T) = \frac{1}{B(\alpha_T, \beta_T)} q_T^{\alpha_T - 1}(1 - q_T)^{\beta_T - 1}, \quad 0 < q_T < 1;\ \alpha_T, \beta_T > 0.$$

The intuition associated with these two assumptions is as follows. The geometric distribution corresponds to purchasing following a "coin-flipping" process in which the individual customer keeps buying until she tosses a "head". The beta distribution is simply a means of allowing $P(\text{"heads"})$ to vary across the customer base.

It follows that the aggregate distribution of the number of units purchased by a week w trialist is given by

$$\begin{aligned} P(T_w = x) &= \int_0^1 P(T_w = x \mid q_T) g(q_T)\, dq_T \\ &= \begin{cases} \dfrac{B(\alpha_T + 1, \beta_T + x - 1)}{B(\alpha_T, \beta_T)} & x = 1, 2, \ldots \\ 0 & x = 0 \end{cases} \end{aligned} \tag{2}$$

which we call the shifted beta-geometric distribution. Elsewhere in the marketing literature, this distribution was used by Morrison and Perry (1970) as a quantity submodel in their NBD-based model of purchase frequency and purchase quantity. (The validity of this distribution as a model of trial week purchasing is explored in Appendix A.) The mean of this distribution is given by

$$E(T_w) = \frac{\alpha_T + \beta_T - 1}{\alpha_T - 1}\ . \tag{3}$$

Modeling Repeat Purchases

Let the random variable $R_{w|i}$ denote the number of (repeat) purchases made in week w by a customer who made her trial purchase in week i ($w > i$). Specifying an appropriate model for the distribution of $R_{w|i}$ is the single most important step in this modeling effort. To do so, we will start with the assertion that the purchasing by a new customer at an established store (or website) is analogous to a consumer's purchasing of a new product. We know that repeat buying rates for new products tend to be nonstationary — at least early in a new product's

life (Fader and Hardie 1999a) — with the purchase rate declining (towards an equilibrium level) over time. One way to capture this pattern is to assume that, for a given cohort, the number of people making zero purchases in a given week grows (at a decreasing rate), which means that the observed average number of units purchased decreases over time.

Our submodel for the distribution of $R_{w|i}$ is based on the following three assumptions:

i. In week w, existing customers are either out of the market, i.e., definitely not going to make a repeat purchase that week, or a possible repeat buyer. (The notion that someone is a "possible repeat buyer" does not ensure that she will actually purchase any units that week; it merely conveys the fact that she will consider purchasing with some non-zero probability.) The probability of a week i trialist being out of the market in week w, which we denote by $\pi_{w|i}$, is assumed to be governed by the following time-dependent distribution:

$$\pi_{w|i} = 1 - \gamma(w-i)^{\delta}\,, \quad w > i\,.$$

When $\delta < 0$, $\pi_{w|i}$ grows at a decreasing rate as $w - i$ increases; consequently, the number of week i triers making zero purchases in week w increases over time. Likewise, δ can also be positive, allowing for the possibility that the number of repeat buyers actually increases over time. (Note that while this fraction ($\pi_{w|i}$) of buyers is definitely not going to make a repeat purchase in week w, we are *not* assuming that they are permanently out of the market, i.e., they may consider buying again in future weeks.)

ii. For an individual who has been classified as a "possible repeat buyer" in week w, the number of units purchased, R_w, is distributed according to a geometric distribution with parameter q_R and probability mass function

$$P(R_w = x \,|\, q_R) = q_R(1-q_R)^x\,, \quad x = 0, 1, \ldots\,;\ 0 < q_R < 1\,.$$

iii. q_R is distributed across the population according to a beta distribution with parameters α_R and β_R, and pdf

$$g(q_R) = \frac{1}{B(\alpha_R, \beta_R)} q_R^{\alpha_R - 1}(1 - q_R)^{\beta_R - 1}, \quad 0 < q_R < 1\,;\ \alpha_R, \beta_R > 0.$$

Qualitatively, the same type of "coin-flipping" story as discussed earlier for the trial submodel applies here as well. Note, however, that there are two differences. First, there is no longer a truncation at zero, i.e., the first coin-flip determines whether a "possible repeat buyer" actually chooses to purchase one unit (or more). Second, the stopping probability ($P(\text{"heads"})$) is governed by a different beta distribution than that used for the trial purchasing process.

It follows that the aggregate distribution of the number of units purchased in week w by a week i trialist ($w > i$) is given by:

$$\begin{aligned} P(R_{w|i} = x) &= \delta_{x=0}\pi_{w|i} + (1 - \pi_{w|i}) \int_0^1 P(R_w = x \mid q_R) g(q_R)\, dq_R \\ &= \delta_{x=0}\pi_{w|i} + (1 - \pi_{w|i}) \frac{B(\alpha_R + 1, \beta_R + x)}{B(\alpha_R, \beta_R)} \end{aligned} \tag{4}$$

where $\delta_{x=0}$, the Kronecker delta, equals 1 if $x = 0$, and 0 otherwise. We call this the "time-dependent, zero-inflated beta-geometric" distribution. The mean of this distribution is

$$E(R_{w|i}) = \gamma(w - i)^{\delta} \frac{\beta_R}{\alpha_R - 1}\,. \tag{5}$$

Parameter Estimation

Given the data presented in Table 1, maximum likelihood estimates of the six model parameters $(\alpha_T, \beta_T, \alpha_R, \beta_R, \gamma, \delta)$ are found by maximizing the following log-likelihood function:

$$\begin{aligned} LL = \sum_{x=1}^{9} n_{1x} \ln\big[P(T_1 = x)\big] + \left(n_1 - \sum_{x=1}^{9} n_{1x}\right) \ln\left[1 - \sum_{x=1}^{9} P(T_1 = x)\right] + \\ \sum_{w=2}^{12} \left\{ \sum_{x=0}^{9} n_{wx} \ln\big[P(X_w = x)\big] + \left(n_w - \sum_{x=0}^{9} n_{wx}\right) \ln\left[1 - \sum_{x=0}^{9} P(X_w = x)\right] \right\} \end{aligned} \tag{6}$$

where n_{wx} is the number of people making x purchases in week w.

In order to evaluate the log-likelihood function, we must be able to compute $P(T_w = x)$ and

$P(R_{w|i} = x)$, as given in (2) and (4), respectively. While it is feasible to employ these equations directly, significant advantages in coding and estimating the model can be achieved by utilizing very simple recursive relationships that exist for both components of the model. For instance, $P(T_w = x)$ can be re-expressed as follows:

$$P(T_w = x) = \frac{P(T_w = x)}{P(T_w = x-1)} P(T_w = x-1)\,.$$

Fortunately many terms in the ratio on the right-hand side can be canceled out, including all of the beta functions (which are quite inconvenient to evaluate in any common spreadsheet environment). This leaves us with the much simpler expression:

$$P(T_w = x) = \begin{cases} 0 & x = 0 \\ \dfrac{\alpha_T}{\alpha_T + \beta_T} & x = 1 \\ \dfrac{\beta_T + x - 2}{\alpha_T + \beta_T + x - 1} P(T_w = x-1) & x \geq 2 \end{cases} \tag{7}$$

Similarly, probabilities associated with the "time dependent, zero-inflated beta-geometric" distribution (4) can be computed using the following forward recursive relationship:

$$P(R_{w|i} = x) = \begin{cases} 1 - \gamma(w-i)^\delta \left(\dfrac{\beta_R}{\alpha_R + \beta_R}\right) & x = 0 \\ \gamma(w-i)^\delta \dfrac{\alpha_R \beta_R}{(\alpha_R + \beta_R)(\alpha_R + \beta_R + 1)} & x = 1 \\ \dfrac{\beta_R + x - 1}{\alpha_R + \beta_R + x} P(R_{w|i} = x-1) & x \geq 2 \end{cases} \tag{8}$$

Combining these simplified expressions back into (1) and then into (6) completes our description of the model as actually implemented.

While the log-likelihood function (6) appears to be rather complicated — it involves the evaluation of 131 terms — each of these calculations is very simple and actually constructing this function in a spreadsheet can be done quite easily using basic "cut and paste" techniques. (The interested reader can obtain a copy of this spreadsheet, as well as a detailed appendix describing its construction, from `http://brucehardie.com/pmnotes.html`.)

Predicting Unit Sales

Let the random variable N_w be the total number of units (CDs) purchased by the (eligible) cohort members in week w. Our best estimate of weekly unit sales is given by

$$E(N_w) = \begin{cases} n_w E(T_w) + \sum_{i=1}^{w-1} n_i E(R_{w|i}) & w \leq 12 \\ \sum_{i=1}^{12} n_i E(R_{w|i}) & w > 12 \end{cases} \tag{9}$$

where $E(T_w)$ and $E(R_{w|i})$ are calculated using (3) and (5), respectively.

Summary

In summary, the objective of our modeling effort is to develop a simple model for generating a medium-term forecast of unit purchases by a cohort of customers who made their first purchases at CDNOW during the first 12 weeks of 1997. The expression used to create such forecasts is given in (9). Central to this are expressions for the number of units purchased on a trial occasion and the number of units purchased in week w, given trial in week i; expressions for these are given in (2) and (4), respectively. However, the nature of reported data is such that we do not observe these separate components of sales — we only observe the overall distribution of total purchases by all customers in a given week. Using the finite mixture model presented in (1), we are able to estimate the parameters of the submodels for trial and repeat purchasing using the aggregated data, which are then used to create the sales forecast.

Readers familiar with the marketing literature on stochastic models of buyer behavior may wonder why the beta-geometric distribution is being used as the underlying counting distribution, as opposed to the more common NBD model, which is widely used within marketing (Morrison and Schmittlein 1988). The primary reason is one of communication. Successful implementation of a modeling exercise requires management's acceptance of the underlying model. The chance of management's acceptance increases with their ability to understand the workings of the model, at the very least at an intuitive level. The logic of the beta-geometric distribution is easy to communicate to a managerial audience, using the "coin-flipping" story discussed above. This has proven to be much easier to explain than the rationale for the NBD model — Poisson purchasing at the individual-level with gamma heterogeneity. Two secondary reasons

for using the beta-geometric distribution as opposed to the NBD are: (1) the ease of handling the zero-truncated distribution for trial purchasing, and (2) a marginally better fit associated with the beta-geometric distribution.

4 Empirical Analysis

Given the dataset presented in Table 1, the above model was implemented entirely within the Microsoft Excel spreadsheet package; the parameter estimates were obtained by using its Solver add-in to maximize the log-likelihood function given in (6). The estimation procedure is extremely fast (requiring only a few seconds on a standard PC) and highly robust (always converging very close to the global optimum). The maximum value of the log-likelihood function is $LL = -112{,}923.9$, which occurs when the model parameters take on the following values:

α_T	β_T	α_R	β_R	γ	δ
6.901	7.185	5.024	5.595	0.122	−0.291

Using (3), we find that the mean number of CDs purchased during a new customer's trial week is 2.22 units. For any customer who is a "possible repeat buyer" in a given week, the expected number of CDs purchased, as computed using (5), is 1.39 units. Being classified as a "possible repeat buyer" by no means implies that a purchase actually occurs — there is a 47% chance that such a person make no purchase at all. The fact that $\hat{\delta} < 0$ implies that $\pi_{w|i}$ increases over time (at a decreasing rate). Consequently, the number of buyers who will definitely not make a repeat purchase in a given week grows as we move further from their trial week, and therefore the observed number of repeat purchases will decline with time.

4.1 Model Fit

The predicted distribution of weekly purchases, obtained using (1), produces a good fit to the observed data (Table 1), as judged by the chi-square goodness-of-fit test ($\chi^2_{113} = 129.21$, $p = 0.141$). This is further demonstrated in Figure 1.

We first see that the expected total sales, as computed using (9), tracks observed total sales (Table 1) very well. As discussed above, our model allows us to decompose these total sales figures into separate trial and repeat components. When we compare these *estimates* to

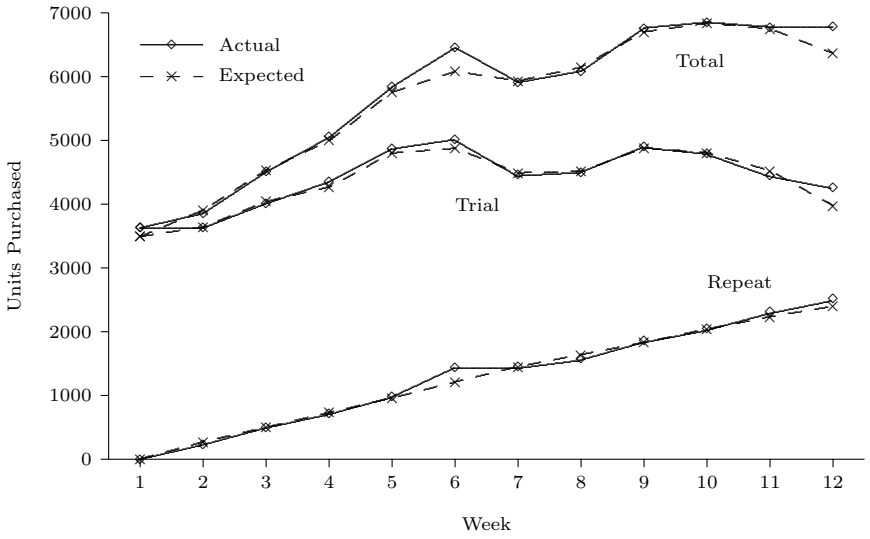

Figure 1: Model Fit: Trial/Repeat Decomposition

the *actual* trial and repeat numbers determined using the original (disaggregate) dataset, it is clearly evident that the model recovers these underlying components (which were not separately identified in Table 1) very well. The ability to do this provides further support for the validity of our model, and a high degree of confidence in our ability to extrapolate beyond the 12 week calibration period.

Looking closely at the repeat sales numbers in Figure 1, we see that the curve appears to be linear. It grows over time as the base of eligible repeaters expands, but there are no obvious indications that the number of repeats per repeater is changing over time, at least within the calibration period. Since the number of new triers drops to zero after week 12, we might expect this curve to stop growing, perhaps remaining close to its final level of approximately 2500 units per week for the entire cohort. We call this the "linear projection" forecast. In the next section, we contrast this simple benchmark with our model's forecast, and compare both to the actual sales numbers.

4.2 Model Forecasting Performance

For this cohort of 23,570 people who first purchased in quarter 1 of 1997, we were able to extract records of their total purchasing for the 40 weeks beyond the calibration period (i.e., 4/97–12/97). The forecasting performance of the proposed model is evaluated against these actual purchasing

numbers. Given the parameter estimates, the aggregate sales forecast is computed using (9).

Figure 2 reports our model's forecast, and that of the linear projection, along with the actual repeat purchase numbers for the cohort through 12/97. It is clear that our model provides a much more realistic picture of future repeat purchasing than does the linear projection. In sharp contrast to the linear projection, our model predicts a decreasing level of repeat purchasing by this cohort as it ages, and the number of "possible repeat buyers" in a given week shrinks over time.

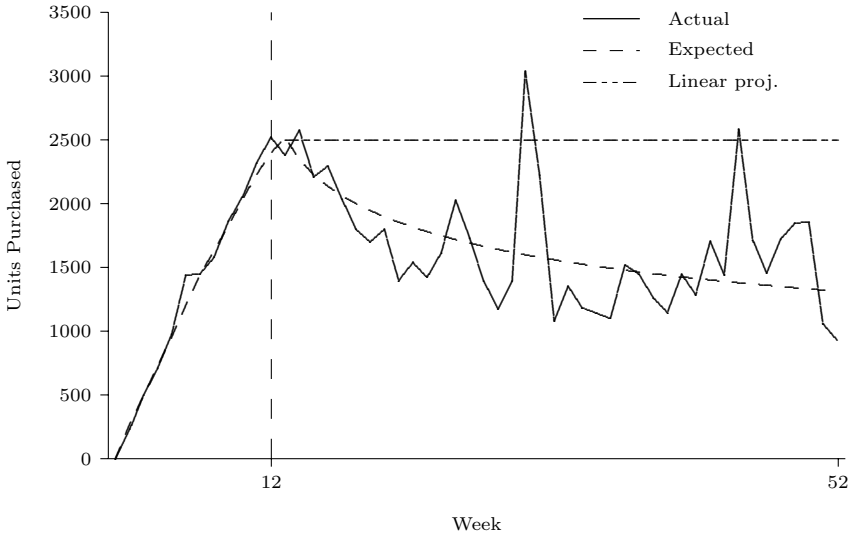

Figure 2: Forecast Weekly Repeat Sales

The first observed major deviation from our forecast corresponds to a mid-year promotion run by CDNOW, while the second spike corresponds to the Christmas season. The model's projections seem to serve as an accurate and potentially valuable benchmark to understand what expected sales levels would have been in the absence of these special events.

The performance of our model appears to be even more apparent when we examine the same data in cumulative form, as shown in Figure 3. At the end of the year (week 52), the forecast index (relative to actual) for our model is 98.7%, while the index associated with the linear projection is 140.7%. The under-prediction associated with our model should come as no surprise as the actual sales numbers contain promotional and seasonal events not captured within our model.

Re-examining Figure 2, we may be tempted to assert that post-trial sales did at first decline

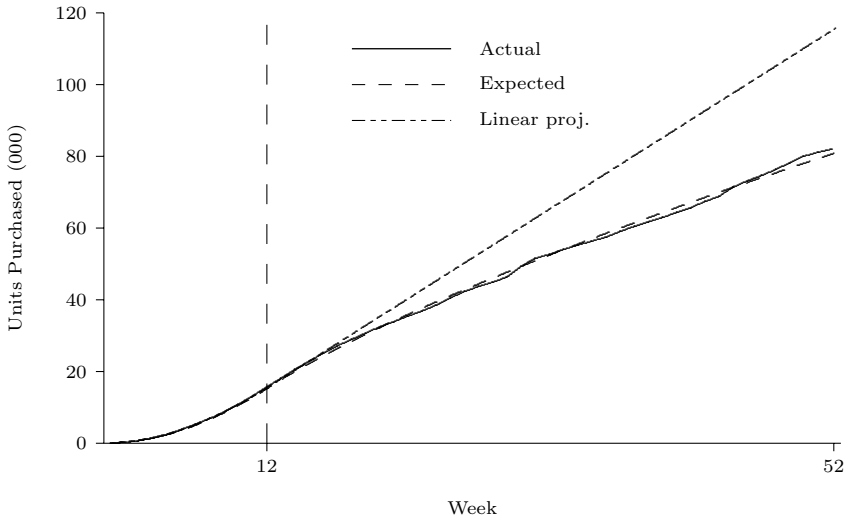

Figure 3: Cumulative Repeat Sales Forecast

but then leveled-off around the time of the mid-year promotion; perhaps the sales for the second-half of the year could be represented by a horizontal line. If this were the case, the model would severely under-predict future sales, as the expected sales curve will continue to decline. To examine this hypothesis, we obtained the same cohort's purchasing data for the first six months of 1998, and report in Figure 4 our model's forecast along with the actual repeat purchase numbers through 6/98. It is very clear that the declining trend predicted by our model holds, albeit with obvious deviations due to promotional activities. The ability of a simple six parameter model, calibrated on twelve weeks of data, to forecast the underlying trend of future purchasing 66 weeks into the future is quite remarkable.

5 Discussion and Conclusions

We have developed a new stochastic model of repeat buying behavior at an e-commerce site, which is capable of generating medium-term forecasts of repeat purchasing by a fixed cohort of customers. First, we carefully structured the individual customer-level data into an appropriate aggregate form that is easy for the analyst to manage. Second, we specified a simple model that can accurately decompose the observed total sales into the underlying trial and repeat components (while at the same time capturing underlying dynamics in repeat buying). As a

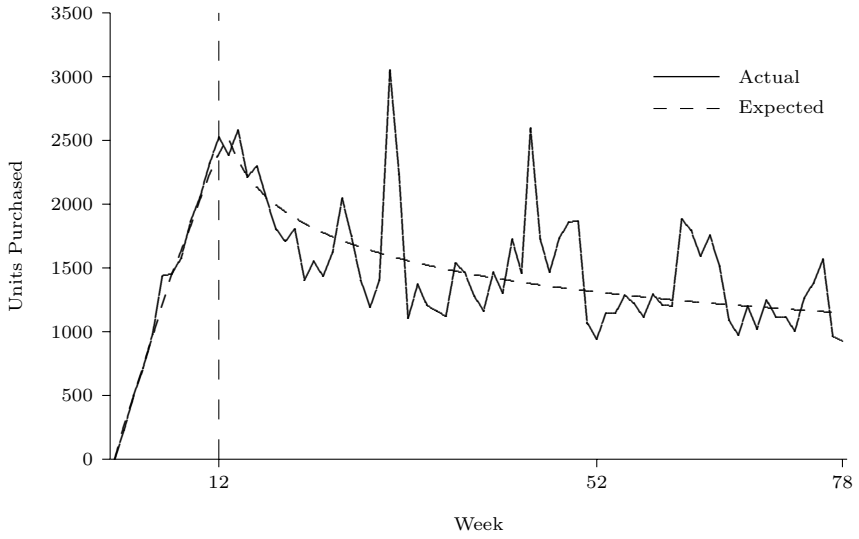

Figure 4: Forecast Weekly Repeat Sales

result, we have a model that can easily be completely implemented in a common spreadsheet environment — both for parameter estimation and generation of forecasts.

The corresponding medium-term sales forecast is quite precise, providing the manager with a tool for determining the overall value of a customer base and/or a baseline sales estimate which can then be used to generate an estimate of the incremental sales associated with a firm's marketing activities — which is a fundamental input to any promotion-event evaluation.

5.1 Future Research

Staying within the modeling framework developed in this paper, an obvious next step would be to apply this model to other cohorts (e.g., customers making their first purchases at CDNOW in the second quarter of 1997) and to examine the stability of the associated model parameters. Furthermore, it would be useful to develop a model for the arrival of new customers to the website. Coupling these two models would enable us to forecast a site's *overall* sales (as opposed to those for a given cohort, as in the analysis presented in this paper).

As noted earlier, the model developed here does not take full advantage of the richness available in the individual-level transaction data (i.e., customer purchase histories). Therefore a second stream of future research would be to model more formally both the separate components of purchasing and the dynamics of buyer behavior, using the disaggregate panel data. There

are several aspects to such a modeling exercise. First, the model presented in this paper focuses on total units purchased in a given week. This is the result of two processes — the number of transactions made by an individual in a given week and the number of units purchased on each transaction. A more sophisticated model would explicitly recognize this decomposition of total purchasing. Second, the model treats the data as a series of cross-sections; we do not model the longitudinal series of purchase events at the level of the individual customer. Consequently, we are unable to make customer-level predictions of future behavior and/or profile individual customers — activities that are vital to many database marketing efforts. In future work, transactions could be modeled using an individual-level counting process (e.g., # of transactions across unit time intervals) or an inter-transaction timing model.

Finally, we should consider the $\pi_{w|i}$ term, which we label as the probability of a week i trialist being "out of the market" in week w. While this captures the nonstationarity in buying, as evident through the decline in repeat purchasing, it cannot tell us whether this is due to a slowdown among active repeat customers and/or customers dropping out of the market. A model that captures nonstationarity at the individual customer-level, such as Fader and Hardie's (1999a) NSEG model, would provide such insights.

While these extensions all lead to a more "correct" model of buying behavior, they come at a cost. The resulting models become quite complex and must be calibrated using individual customer-level data. Furthermore, the analyst is required to have relatively advanced modeling skills and access to the appropriate computational software. These two factors, combined with the challenges of explaining the logic of the models to managers, poses a serious threat to their implementation, especially when compared to a model such as that developed in this paper. It would be very useful to compare the aggregate forecasting performance of such models to that of our far simpler model, which captures the sales patterns but cannot properly diagnose the causes of the observed behavior. Based on a similar comparison undertaken for new product sales forecasting models (Fader and Hardie 1999b), it is our expectation that the aggregate forecasting performance of our model would be on par with that of any more complex model.

Furthermore, from the perspective of introducing marketing models to an organization, it is also good to start with simple models and then evolve towards more complete (and complex) models as the key personnel become more comfortable with making use of marketing models

and are more willing to commit the resources that the more complex models require (Urban and Karash 1971). We view our model as a suitable first step down this path and hope that some readers are genuinely able to proceed to more advanced levels in an appropriate and productive manner.

Appendix A: Validating the Trial Purchasing Model

The overall fit of the model, along with its ability to decompose trial and repeat sales from the aggregate data, suggests that the assumptions underlying our model are reasonable. However, this conclusion is based on an analysis that utilizes a mixture of beta-geometric distributions. In this appendix we explore the validity of the beta-geometric distribution itself.

Looking closely at Table 1, it is clear that the column corresponding to week 1 presents trial week only purchases by a group of 1574 customers. These data enable us to examine in greater detail the assumptions underlying the beta-geometric model or, more correctly, the *shifted* beta-geometric model (to acknowledge the truncation at zero) associated with trial purchases, as given in (2).

In modeling trial, we may first be tempted to fit the homogeneous shifted-geometric distribution to the trial purchases data; i.e., assume all customers have the same value of the latent trial purchasing rate parameter, q_T. Fitting the shifted geometric distribution to the week 1 purchase data, the maximum likelihood estimate of q_T is 0.444. On the basis of the chi-square goodness-of-fit test, the fit of this model to the observed purchase data for week 1 is poor ($\chi^2_8 = 51.8,\ p < 0.001$).

An obvious potential cause for the poor fit of this model is the fact that it ignores differences in people's propensity to make multiple purchases (q_T). Under the assumption of beta heterogeneity, we have the shifted beta-geometric distribution of trial week purchase quantities. Fitting (2) to the week 1 purchase data, the maximum likelihood estimates of α_T and β_T are 5.912 and 6.283, respectively. On the basis of the chi-square goodness-of-fit test, the fit of this model to the observed purchase data for week 1 is now excellent ($\chi^2_7 = 3.3,\ p = 0.86$). This is visually confirmed in Figure A. Using (3), we see that $E(T_1) = 2.28$. This implies that $E(N_1)$ is 3587, which is within 1.1% of the actual number of units purchased (Table 1), thus providing further evidence of the validity of the beta-geometric distribution.

It is interesting to compare the underlying distribution of q_T estimated using the week 1 data ($\hat{\alpha}_T = 5.912$ and $\hat{\beta}_T = 6.283$) with that derived using all 12 weeks worth of data ($\hat{\alpha}_T = 6.901$ and $\hat{\beta}_T = 7.185$). Note that the 12 week estimate is derived using the mixture of trial and repeat purchasing distributions, i.e., (1); it is not based on "clean", trial-only data. The mean of the distribution of q_T derived using all of the data is slightly higher than that associated with

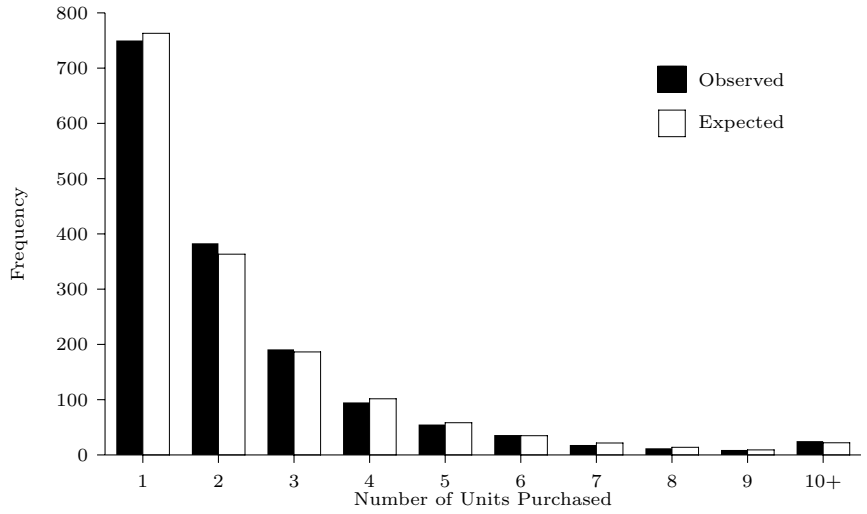

Figure A: Fit of Week 1 Trial Model

the week 1 only data (0.490 vs. 0.485) and the corresponding variance is slightly lower (0.017 vs. 0.019). However, the differences are negligible; this is clearly evident in Figure B, which plots the underlying distributions of q_T based on the week 1 data and all 12 weeks worth of data. This similarity only provides more support for the assumptions underlying our basic model.

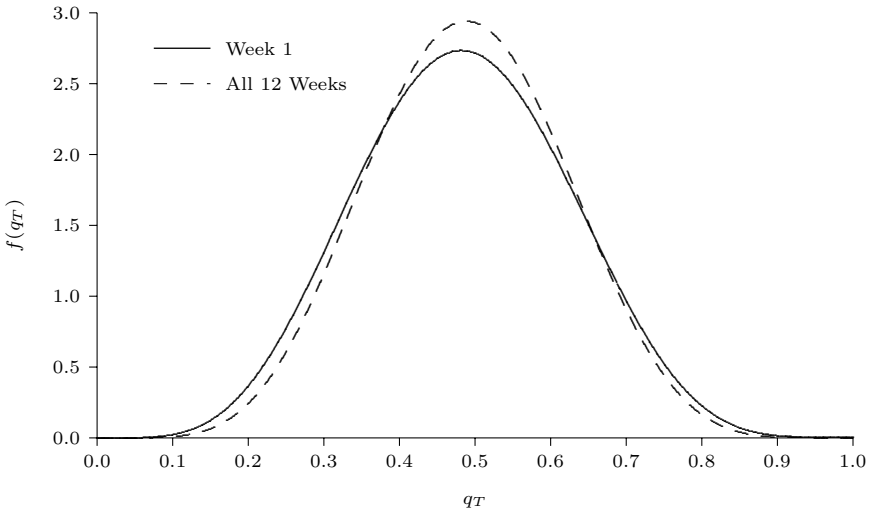

Figure B: Distribution of q_T: Week 1 Only vs. All 12 Weeks

References

Allenby, Greg M., Robert P. Leone, and Lichung Jen (1999), "A Dynamic Model of Purchase Timing with Application to Direct Marketing," *Journal of the American Statistical Association*, **94** (June), 365–74.

Buchanan, Robert W., Jr. and Charles Lukaszewski (1997), *Measuring the Impact of Your Web Site*, New York: John Wiley & Sons, Inc.

Colombo, Richard and Weina Jiang (1999), "A Stochastic RFM Model," *Journal of Interactive Marketing*, **13** (Summer), 2–12.

Fader, Peter S. and Bruce G. S. Hardie, (1999a), "Modeling the Evolution of Repeat Buying," Wharton Marketing Department Working Paper.

Fader, Peter S. and Bruce G. S. Hardie (1999b), "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.

Forrester Report (1999), *Measuring Web Success*, November, Cambridge, MA: Forrester Research, Inc.

Morrison, Donald G. and Arnon Perry (1970), "Some Data Based Models for Analyzing Sales Fluctuations," *Decision Sciences*, **1** (July–October), 258–74.

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 and Economic Statistics*, **6** (April), 145–59.

Mulhern, Francis J. (1999), "Customer Profitability Analysis: Measurement, Concentration, and Research Directions," *Journal of Interactive Marketing*, **13** (Winter), 25–40.

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** (January), 1–24.

Schmittlein, David C. and Robert A. Peterson (1994), "Customer Base Analysis: An Industrial Purchase Process Application," *Marketing Science*, **13** (Winter), 41–67.

Urban, Glen L. and Richard Karash (1971), "Evolutionary Model Building," *Journal of Marketing Research*, **8** (February), 62–6.