

Iowa State University

From the Selected Works of Meltem Denizel

September, 2007

Empirical validation and comparison of models for customer base analysis

Emine Persentili Batislam, *Sabanci University*

Meltem Denizel, *Sabanci University*

Alpay Filiztekin, *Sabanci University*



Available at: <https://works.bepress.com/meltem-denizel/2/>

**Empirical Validation and Comparison of
Models for Customer Base Analysis**

Emine (Persentili) Batislam^{*}, Meltem Denizel, Alpay Filiztekin

Sabanci University, Orhanli-Tuzla, 81474 Istanbul, Turkey

^{*}Corresponding author. Tel: 90.216.483.9663; fax: 90.216.483.9670; e-mail:
batislam@sabanciuniv.edu

Abstract**Empirical Validation and Comparison of Models for Customer Base Analysis**

The benefits of retaining customers lead companies to search for means to profile their customers individually and track their retention and defection behaviors. To this end, the main issues addressed in customer base analysis are identification of customer active/inactive status and prediction of future purchase levels. We compare the predictive performance of Pareto/NBD and BG/NBD models from the customer base analysis literature — in terms of repeat purchase levels and active status — using grocery retail transaction data. We also modify the BG/NBD model to incorporate zero repeat purchasers. All models capture the main characteristics of the purchase and dropout process of individual customers and produce similar forecasts. There are some deviations in the cumulative purchase estimates of the models, which may be due to the characteristics of grocery purchasing.

Keywords: Customer Base Analysis, Probability Models, Pareto/NBD, BG/NBD, Customer Lifetime

1. Introduction

Companies with a strategic focus on establishing long-term customer relationships build databases to identify their customers, track customer transactions, and predict changes in customer purchase patterns at an individual level. They can also leverage the purchase information available in these databases to target the right customers to retain through customer base analysis.

Customer base analysis is concerned with distinguishing active customers from already defected ones, and predicting their lifetime and future level of transactions considering their observed past purchase behavior (Schmittlein, Morrison, & Colombo, 1987). Developing a valid measurement framework that adequately describes the process of birth, purchase activity, and defection is not a trivial task due to various behavioral aspects of the purchasing process (Jain & Singh, 2002; Reinartz & Kumar, 2000). Due to the randomness of individual purchase behavior and customer heterogeneity, conducting individual level analysis and prediction is subject to a great deal of error (Mulhern, 1999). Customer base analysis becomes even more difficult in noncontractual relations where customers do not notify companies when they drop out; so identifying active and inactive customers in the database at a given time requires a systematic investigation (Schmittlein & Peterson, 1994).

A highly regarded model in the literature addressing these issues is the Pareto/Negative Binomial Distribution (NBD) model proposed by Schmittlein et al. (1987). The Pareto/NBD model has been widely cited and researchers praised its empirical performance and managerial implications (see, for instance, Balasubramanian, Gupta, Kamakura, & Wedel, 1998; Colombo & Jiang, 1999; Fader & Hardie, 2001; Jain & Singh, 2002; Mulhern, 1999; Fader, Hardie, & Lee, 2005). However, only a very few studies have actually documented its empirical validation.

Schmittlein and Peterson (1994) applied the model using the customer database of an office supplies firm, and presented its predictive performance. Researchers employed the model in other studies; but they have not provided model validation or statistical results about its predictive performance (Wu & Chen, 2000; Reinartz & Kumar, 2000 and Reinartz & Kumar, 2003).

The low number of articles implementing the Pareto/NBD model might arise from its sophisticated nature and computational complexity (Jain & Singh, 2002; Fader & Hardie, 2001; Fader et al., 2005). A second empirical validation of the Pareto/NBD model is presented by Fader et al. (2005) together with a new statistical model, the beta-geometric/NBD (BG/NBD), developed to overcome the computational complexities of Pareto/NBD. The computational burden is significantly reduced in the BG/NBD model so that it becomes possible to estimate parameters even in a spreadsheet environment. The study also provides a comparative analysis of the two models in terms of fit and prediction of customer purchasing patterns, using online CD purchasing data. The authors propose BG/NBD as an easier alternative to Pareto/NBD, since they yield similar results.

Given the facts that (i) managerial interest in how to manage the customer-firm relationship is even more pronounced in current practice and (ii) technological advances in computer technology enable firms to maintain large longitudinal databases, researchers have focused increasingly on modeling and empirically measuring the customer-firm relationship (Balasubramanian et al., 1998; Reinartz & Kumar, 2003). However, there are not enough empirical applications reported in the literature to prove the applicability and validity of these models in different purchasing settings and to encourage managers to use them to take better advantage of the information already contained in the customer databases. As stated in Leeflang

and Wittink (2000), successful applications of models in new contexts, using different data sets, may also accelerate the demand for the models. Moreover, reporting model deficiencies when they are applied in real settings may lead to further research and extension of ideas and inspire researchers to improve the existing models or to propose new ones to overcome the observed difficulties.

In our study, we empirically validate both the Pareto/NBD and BG/NBD models using grocery purchase data of individual customers. Grocery purchase data characteristics provide more insights and understanding about the performance of the aforementioned models in different settings. We carry out a comparative analysis to evaluate the performance of the models in predicting customer purchases and active status. Moreover, we modify the BG/NBD model by incorporating one more dropout condition to handle zero repurchasers more realistically. We denote the modified model as MBG/NBD.

The rest of the paper is organized as follows: In the next section we describe the grocery retailer customer base and sampling issues. In section 3, we provide short summaries of the Pareto/NBD, BG/NBD, and MBG/NBD models, and in section 4 we discuss parameter estimation. We present an empirical validation study of both models and compare their performance in predicting purchases in Section 5 and in predicting active status in Section 6. Section 7 discusses the main findings and further research questions.

2. Grocery Retail Customer Base

The customer base used in our analysis comes from a specific store of a large grocery retail chain in Turkey. Due to confidentiality, we can only disclose that the store is

located in a busy metropolitan area. The store offers a broad assortment of grocery products. It has been issuing store-cards since the beginning of its operations in mid-1999. Initially, store-cards provided only product discounts for the cardholders. With the launch of a loyalty program in the last quarter of the year 2000, incentives to use the store-card increased. Cardholders collect cash-points from all their purchases from the chain stores and can redeem the collected points as cash whenever they like. The loyalty program has significantly increased the number of cardholders and changed the sales composition of the store. In a very short time after the launch of the program, cardholders accounted for 80% of store revenue. In our analysis, we use only cardholders' information, as they are identifiable on an individual basis and constitute the majority of customers.

Scanner data includes the transaction details, including date of purchase, items and quantity purchased, amount paid, and promotions redeemed by cardholders. Daily transactions are contaminated with significant noise, since some customers visit the store more than once in a given week; but the values of purchases after the first visit in a given week are typically very small. Most probably these subsequent visits are to complete the weekly list of purchases forgotten early on. Considering that people usually do their grocery shopping on a weekly basis, we aggregated daily purchases by individuals to a weekly frequency.

The store supplied us with transaction data for a period of 146 weeks, starting from July 2000 to April 2003. Both the Pareto/NBD and BG/NBD models require tracking customer transactions starting with their initial purchases. Since the store had not kept records of the initial purchases of cardholders, we left-filtered customer records with transactions before August 2001 (within the first 13 months) to guarantee that the

customers we include in our analysis are newcomers with known initial purchase times.

Figure 1 provides the two-year cycle of the total weekly purchases decomposed as initial and repeat purchases starting from May 2001. Both initial and repeat purchases start to increase in September in both years, purchases fluctuate until the summer, and they are low during the summer. The store management explains that the fluctuations are due to promotional activities that continue throughout the year at different scales. Major promotional activities are held at the end of the summer, before and during religious holidays, and before the new year. Promotional activities are very common in grocery stores to increase sales and to recruit new customers since the majority of grocery customers are very sensitive to such activities. Promotional activities and low shop switching costs are some of the reasons for customer heterogeneity in grocery shopping.

After left-filtering and weekly aggregation, our final observation window covers 91 weeks, starting from August 2001 until the end of April 2003; and within the observation window, the customer base includes 33,544 cardholder customers with 124,097 weekly purchase records. In the observation window, we identified customers who made their initial purchases within the second quarter (August-October 2001) and third quarter (November 2001-January 2002) as Cohort1 and Cohort2, respectively. We considered 52-week and 78-week observation periods for Cohort1 and only a 52-week observation period for Cohort2. Some statistics related to Cohort1 and Cohort2 for different observation periods are given in Table 1. The statistics on the number of repeat purchases, mean inter-purchase time, and duration between final and initial purchase (tenure) are similar in Cohort1 and Cohort2. Extending the length of the observation period to 78 weeks results in an increase in

the number of repurchases and tenure, though the mean inter-purchase time does not change. In the cohorts, the median of the inter-purchase times, even after excluding zero repurchasers, is around 5 weeks. In the previous applications of the Pareto/NBD model, median inter-purchase time was 7 months for office supplies (Schmittlein et al., 1994), 17 weeks for catalog sales (Reinartz et al., 2000), and 25 weeks for computer-related products (Reinartz et al., 2003). Compared to these customer bases, grocery shopping has very short purchase cycles, since grocery items are non-durable and require frequent replenishment.

Besides short purchase cycles, the number of customers and their heterogeneity are high in the grocery retail customer base. For instance, in the online CD customer base from Fader et al. (2005), the majority of customers (approximately 85% of the total) make 0, 1 or 2 repurchases. Among them, 60% are zero repurchasers. On the other hand, approximately 40% of grocery retail customers are zero repurchasers and customers with 0, 1, or 2 repurchases make around 65% of total grocery retail customers. High heterogeneity in grocery purchases decreases the precision of the models.

3. Pareto/NBD, BG/NBD and MBG/NBD Models

Unlike the previous efforts on modeling repeat buying, the Pareto/NBD model proposed by Schmittlein et al. (1987) and the BG/NBD model proposed by Fader et al. (2005) take into account not only the purchasing pattern of customers, but also the dropout probability of customers. The key purchasing conditions are the same in both models. Customers can make purchases from the store and can drop out randomly whenever they like. Both models allow customer heterogeneity, that is, they assume

that customers may differ in their purchase and dropout behaviors as well. The same transaction history data is used in our analysis of both models, including the customer's first transaction time, his number of transactions during the observation period (x), and his last transaction time (t_x) within the observation period (T).

The number of purchases while a customer is active follows the NBD (Poisson-gamma mixture) counting process in both models. Accordingly, the major underlying assumptions in modeling the purchase process are as follows:

- While active, the number of transactions made by a customer follows a Poisson process with transaction rate λ .
- Heterogeneity in transaction rates across customers follows a gamma distribution with shape parameter r and scale parameter α .

Modeling of the dropout process is the major difference between the Pareto/NBD and BG/NBD models. Since the dropout time of a customer is not directly observed, the only evidence that a customer may have become inactive is a suspiciously long period of time without any transaction after the last observed purchase. Hence, in the Pareto/NBD model, the time to drop out is modeled using the Pareto (exponential-gamma mixture) timing model with the following assumptions:

- Each customer has an unobserved lifetime starting from his initial purchase (birth) until the time he becomes inactive (death). The lifetime of any customer is distributed exponentially with dropout rate μ .
- Heterogeneity in dropout rates across customers follows a gamma distribution with shape parameter s and scale parameter β .
- The purchase rate λ and the dropout rate μ vary independently across customers.

Differing from the Pareto/NBD model, the BG/NBD model assumes that a dropout can occur only immediately after a purchase. Hence the authors model the dropout process using the beta-geometric model with the following assumptions.

- After any transaction, a customer becomes inactive with probability p . Therefore the point at which the customer drops out is distributed across transactions according to a (shifted) geometric distribution.
- Heterogeneity in p follows a beta distribution, with parameters a and b .
- The transaction rate λ and the dropout probability p vary independently across customers.

In this model, the assumption that a customer dropout can occur only after a transaction leads to treating the customers with zero repeat purchase during the observation period as active at time T and thereafter until they make a transaction. To deal with this issue, we modified the BG/NBD model by including an additional chance of dropout at time zero, i.e., immediately after the first purchase of a customer. All other assumptions are similar to those of BG/NBD. Including one more dropout condition may improve the model flexibility, especially for zero repurchasers.

In the rest of this section, we present the modified model for three main quantities of interest at the individual level: the likelihood of the observed transaction data, the probability of a certain number of purchases, and the expected number of purchases in a given time period. We refer the reader to Fader et al. (2005) for details on the BG/NBD model, and we provide in the current paper the modified formulas for MBG/NBD.

In MBG/NBD, a customer may drop out at time zero with probability p . This leads to the following individual-level likelihood function.

$$L(\lambda, p|X = x, t_x, T) = (1-p)^{x+1} \lambda^x e^{-\lambda T} + p(1-p)^x \lambda^x e^{-\lambda t_x}. \quad (1)$$

The first term specifies the case where the customer is still alive at time T , with x transactions made until that time and the last transaction occurring at time t_x . The second term is for the case where the customer drops out at time t_x after her last transaction. Note that $t_x = 0$ when $x = 0$.

Let $X(t)$ denote the number of transactions in a time period of length t . Based on the assumption that the transactions follow a Poisson process with parameter λ , the probability that an individual customer makes x transactions in a time period t can be written as:

$$P(X(t) = x|\lambda, p) = (1-p)^{x+1} \frac{(\lambda t)^x e^{-\lambda t}}{x!} + p(1-p)^x \left(1 - e^{-\lambda t} \sum_{j=0}^{x-1} \frac{(\lambda t)^j}{j!} \right). \quad (2)$$

In (2), the first term considers the case where a customer does not drop out at time zero, makes exactly x transactions during the t time units, and is still active after the last transaction. The second term specifies the case where the customer drops out after the last transaction, makes x transactions during t time units, and the last transaction occurs before or at t . The latter is the Erlang- x cumulative density function.

Following the derivation in Fader et al. (2005), the expected value of the number of transactions in a period of t time units becomes:

$$E(X(t)|\lambda, p) = \frac{(1-p)}{p} (1 - e^{-\lambda p t}). \quad (3)$$

Note that the difference between the above and the corresponding formula in Fader et al. (2005) is the $(1-p)$ term in the product, which is the probability that the customer is active at time zero.

Equations (1)–(3) are generated for an individual customer with a specific transaction rate and dropout probability. Considering customer heterogeneity requires incorporation of the probabilistic nature of these parameters across customers. We present the associated derivations in the Appendix.

4. Parameter Estimation

Parameter estimation of the Pareto/NBD model is regarded as being somewhat complex and demanding from a computational standpoint (Reinartz & Kumar, 2003; Fader et al., 2005). In particular, the maximum likelihood estimation (MLE) approach to estimate key parameters of the model requires a numerical search algorithm that must evaluate the Gauss hypergeometric function (Schmittlein et al., 1994) and the precision of some numerical procedures varies substantially over the parameter space; this causes problems for numerical optimization routines (Fader et al., 2005). As an alternative method for parameter estimation, Schmittlein et al. (1994) present a two-step method-of-moments (MOM) estimation procedure that is claimed to be more tractable than MLE. Reinartz and Kumar (2003) compare the parameter estimation results of both methods and report that the results are similar. In contrast, Fader et al. (2005) find that the BG/NBD model can be easily estimated using the MLE method. Indeed, they provide a simple Excel file to show the simplicity of estimation in their setting.

In this research, we estimate the BG/NBD and the MBG/NBD parameters using MLE as outlined by Fader et al. (2005). In Pareto/NBD estimation, we used both the MOM procedure of Schmittlein et al. (1994) and MATLAB code for MLE provided by

Fader, Hardie, and Lee¹. Our intention is not to evaluate different estimation methods and their properties, but rather to compare the predictive power of the two models in a grocery retailer customer base. As a result, we chose the MLE parameter estimates for Pareto/NBD since the resulting log-likelihood function value is lower than that of MOM, and we then have the same parameter estimation platform for the comparison of the two models.

To explore the robustness of the two models, we estimated model parameters, using first Cohort1 then Cohort2 data. Parameter estimation was conducted using data collected within the first 52 weeks for each cohort (hereafter, the observation period) and 78 weeks for Cohort1 to see effect of the length of customer histories on model estimation. Forecasts of future purchases were made for the remaining 39 weeks for Cohort1, 26 weeks for Cohort2, and 13 weeks for Cohort1 with a 78-week observation period. We refer to these remaining weeks as the forecast period. The parameter estimation results of the models are presented in Table 2. While there are differences in the point estimates of parameters across different cohorts, the estimated transaction rates and dropout rates do not vary significantly within each model.

5. Purchase Prediction Validity

Empirical analysis was carried out for both Cohort1 and Cohort2 customers, for both the 52 and 78-week observation periods for Cohort1. Comparison of repeat purchase estimates for Cohort1 in different observation periods shows that the length of customer history affects the results in our case. The forecast results of Cohort1 for 78 weeks better match the actual figures compared to the results for 52 weeks; we

¹ A Note in Implementing the Pareto/NBD Model in MATLAB, March 2005. Available from <http://brucehardie.com/notes/008>.

believe that this is due to the fact that the 78-week observation period includes the peak season in quarters six and seven in Figure 1. As a result of the altering nature of grocery purchases, observation periods should be long enough to cover the periods representing regular sales fluctuations, and forecast periods should be short. Here, we present only the results of Cohort1 for the 78-week observation period.

Figure 2 presents the estimated and actual frequencies of repeat purchases. The expected numbers of customers making 0,1,...,10+ repeat purchases using the three models in the 78-week observation period are compared to the actual frequency distribution. Around 40% of customers had no repeat purchases in the observed period and the Pareto/NBD and MBG/NBD models produce results very close to the actual frequency for zero repurchasers. However, deviations of both models from the actual frequency for customers with 1, 2, and 10+ repeat purchases lead to significant differences between model estimates and the actual frequencies of repeat purchases in chi-square goodness of fit tests (Pareto/NBD: observed, $\chi^2_7 = 59.94$ and MBG/NBD: observed $\chi^2_7 = 52.92$, both with p-values < 0.0001). A chi-square test to compare the Pareto/NBD and MBG/NBD models leads to $\chi^2_7 = 9.79$ (p-value = 0.20), implying that the two models are “close”.

The BG/NBD model estimates the frequency of zero repurchasers lower than does the MBG/NBD. This can be attributed to BG/NBD’s dropout requirement of a purchase before dropping out. BG/NBD seems to drop some of the zero repurchasers after one, two, or three repeat purchases as indicated by their higher frequency estimates in Figure 2. Nonetheless, both the BG/NBD and MBD/NBD models yield almost identical estimates for the expected number of weekly and cumulative repeat

purchases (see Figures 3 and 4). Hereafter, we will continue the comparative analysis only with MBG/NBD results.

Both the Pareto/NBD and MBG/NBD models perform well in predicting the weekly repeat purchases, as shown in Figure 3. They track actual purchases well, and they converge to actual repeat purchases in the forecast period. Weekly sales rise during the first 13 weeks due to the entry of new customers in Cohort1 and their repeat purchases. After week 13, we observe only repeat purchases of Cohort1 customers. Both models miss the peak in weekly actual purchases in the initial weeks. This might be a result of an increasing trend in repeat purchases due to promotions in the same period.

The deviation of weekly estimates from actual purchases during the initial weeks leads to underestimation in cumulative repeat purchases in the initial weeks as well (see Figure 4). Later in the observation period, both models match the actual purchases, and in the forecast period they slightly under-forecast to the tune of 2% in the Pareto/NBD model and 7% in the MBG/NBD model. While the transaction and dropout rates are independent in the Pareto/NBD model, they have an explicit association in BG/NBD (Fader et al., 2005). Due to the manner in which the dropout process is modeled, the higher transaction rates in grocery shopping (compared to other shopping environments considered in the literature) may result in higher dropout rates in the BG/NBD model, in turn leading to lower than actual expected sales in the BG/NBD model.

The relative forecasting performance of models at the individual-level is examined in Figure 5 in terms of conditional expectations for the forecast period, broken down by the number of repeat purchases (x) during the observation period. For each value of x in the observation period, we take the average of the actual number of purchases in

the forecast period and compare them with the forecast figures. Both models forecast the number of transactions of individuals almost identically to each other, and slightly more than the actual figures for customers with a “high” number of purchases. The individual purchase forecasts of both models are very close to each other and to the actual figures. This is confirmed by a three-group ANOVA ($F_{2, 16434} = 0.328$, with p -value = 0.72), which is not significant.

Compared to predicting purchase frequency and weekly repeat purchases, forecasts of individual purchases include more customer information and are expected to provide higher accuracy in individual level forecasts. As stated previously, it is difficult to model individual purchase behavior and it is even more complex to do so for the highly heterogeneous purchase behavior encountered in grocery sales. Since the major distinctive advantage of both models is to allow individual level analysis, prediction validity at the individual level is crucial in supporting the strengths of these models.

6. Active Status Prediction Validity

Besides forecasting the individual customer purchases, calculating their active status helps companies to choose the right customers to invest in for retention activities. The customer-specific probability of being active (Schmittlein et al., 1987) indicates the probability that a customer, who is still active at time T , will make at least one repeat purchase after T . We calculated the probability of being active at the end of week 78 for each customer given the purchase history of that individual along with the estimated model parameters.

Figure 6 compares the two models in terms of the average probability of being active for customers with the same number of repeat purchases (x) in the observation period. Both models forecast the average probability of being active very close to each other, except for the customers with zero repurchases in the observation period. For such customers, MBG/NBD assigns higher probability values than Pareto/NBD, but these values are still small enough to declare such customers inactive. The negligible number of actual repeat purchases observed for these customers in the forecast period serves as verification for declaring zero repurchasers as inactive during the forecast period.

Other differences between two models in the values of probability of being active for individual customers might arise due to the following reasons:

- According to Pareto/NBD, the probability decreases as the time period without any purchases increases; therefore, Pareto/NBD assigns low probability when the average elapsed time without purchase exceeds the mean inter-purchase time of customers.
- MBG/NBD assigns a dropout probability after each purchase, and this probability does not change until the next purchase. According to MBG/NBD, the customers who have a high probability of being active in their last purchase remain with this active status if no other purchase is observed.

Misclassification of the active status of customers might arise either from classifying some customers as active and continuing to invest in them although they are inactive or classifying some customers as inactive and stopping investment in them while they actually continue shopping at the store. A major misclassification of active status might take place for customers with a very low number of repeat purchases in the

observation period (say, 1–5 repeat purchases), which might not be sufficient to specify their purchase pattern. Still, this misclassification is not very important, since for such customers, the average transaction size is observed as very low, less than 0.5, in the forecast period. Both models assign high probabilities to customers with high repeat purchases, and most of these customers continue to purchase from the store during the forecast period. The majority of these customers are probably the most valuable customers for the store, so that the store should continue to invest in them.

7. Conclusion

The empirical application of the Pareto/NBD, BG/NBD, and MBG/NBD models in a grocery purchase environment demonstrates that these models capture the main characteristics of the purchase and defection processes of grocery customers and successfully forecast their future purchase patterns. They have high degrees of validity especially in individual purchase forecasting.

The models show a similar performance in tracking the weekly purchases. The BG/NBD model underestimates sales slightly more than does Pareto/NBD. However, when the computational requirements of the two models are compared, the BG/NBD model provides an advantage of ease and less ambiguity in parameter estimation and calculations. Some deviations observed in estimating weekly purchases arise mainly in periods with high promotional activities. In fact, the models seem to have difficulty in handling high transaction rates during promotion periods. As promotion activities are quite common in several industries, a new model or modifications of the existing models that explicitly controls for promotion periods may be required.

Acknowledgement

The authors thank an anonymous referee for her/his valuable comments, David Goldsman for his important suggestions, and Ayca Azgin Hintoglu for her helps in data processing.

References

- Balasubramanian, S., Gupta, S., Kamakura, W., & Wedel, M. (1998). Modeling Large Datasets in Marketing. *Statistica Neerlandica*, 52(3), 303-323.
- Berger, P.D., & Nasr, N.L. (1998). Customer Lifetime Value: Marketing Models and Applications. *Journal of Interactive Marketing*, 12(1), 17-30.
- Colombo, R., & Jiang, W. (1999). A Stochastic RFM Model. *Journal of Interactive Marketing*, 13(3), 2-12.
- Fader, P.S., & Hardie, B.G.S. (2001). Forecasting Repeat Sales at CDNOW: A Case Study. *Interfaces*, 31(3), 94-107.
- Fader, P.S., Hardie, B.G.S., & Lee, K.L. (2005). Counting your Customers the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, 24(2), 275-284.
- Jain, D., & Singh, S.S. (2002). Customer Lifetime Value Research in Marketing: A Review and Future Direction. *Journal of Interactive Marketing*, 1(2), 34-46.
- Leeflang, P.S.H., & Wittink, D.R. (2000). Building Models for Marketing Decisions: Past, Present and Future. *International Journal of Research in Marketing*, 17, 105-126.
- Mulhern, F.J. (1999). Customer Profitability Analysis: Measurement, concentration, and Research Directions. *Journal of Interactive Marketing*, 13(1), 25-49.
- Reinartz, W.J., & Kumar, V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. *Journal of Marketing*, 64(October), 17-35.

- Reinartz, W.J., & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, 67(January), 77-99.
- Schmittlein, D.C., Morrison, D.G., & Colombo, R. (1987). Counting Your Customers: Who Are They And What Will They Do Next? *Management Science*, 33(1), 1-24.
- Schmittlein, D.C., & Peterson, R.A. (1994). Customer Base Analysis: An Industrial Purchase Process Application. *Marketing Science*, 13(1), 41-66.
- Wu, C. & Chen, H. (2000). Counting Your Customers: Compounding Customer's In-store Decisions, Interpurchase Time, and Repurchasing Behavior. *European Journal of Operational Research*, 127(1), 109-119.

Appendix

In this appendix, we provide the model for a randomly chosen individual, based on the individual-level model presented by equations (1)–(3), which are conditioned on the transaction rate λ and the dropout probability p . For a randomly chosen customer, we need to take into account the probability distributions for λ and p , which are Gamma (with parameters r and α) and Beta (with parameters a and b), respectively.

Integrating the individual likelihood function (1) over the distributions of λ and p we obtain:

$$L(r, \alpha, a, b | X = x, t_x, T) = \frac{B(a, b + x + 1) \Gamma(r + x)}{B(a, b) \Gamma(r)} \frac{\alpha^r}{(\alpha + T)^{r+x}} + \frac{B(a + 1, b + x) \Gamma(r + x)}{B(a, b) \Gamma(r)} \frac{\alpha^r}{(\alpha + t_x)^{r+x}}$$

Similar to Fader et al. (2005) we maximize the log-likelihood to estimate r , α , a , b .

To find probability of x transactions in t time units for a randomly chosen customer, we integrate (2) over λ and p .

$$P(X(t) = x | r, \alpha, a, b) = \frac{B(a, b + x + 1) \Gamma(x + r)}{B(a, b) \Gamma(r)x!} \left(\frac{\alpha}{\alpha + t} \right)^r \left(\frac{t}{\alpha + t} \right)^x + \frac{B(a + 1, b + x)}{B(a, b)} \left(1 - \left(\frac{\alpha}{\alpha + t} \right)^r \left\{ \sum_{j=0}^{x-1} \frac{\Gamma(r + j)}{\Gamma(r)j!} \left(\frac{t}{\alpha + t} \right)^j \right\} \right)$$

Similarly to finding the expected value of transactions in a period of t time units for a random customer, we integrate (3) over λ and p to obtain

$$E(X(t) | r, \alpha, a, b) = \frac{b}{a-1} \left(1 - \left(\frac{\alpha}{\alpha + t} \right)^r {}_2F_1 \left(r, b+1; a+b; \frac{t}{t+\alpha} \right) \right), \quad (4)$$

where ${}_2F_1(\)$ is Euler's integral for the Gaussian hypergeometric function.

Finally we present the formula for the probability that the customer is active at time T

as:

$$P(\text{Active}|x, t_x, T, r, \alpha, a, b) = \frac{1}{1 + \frac{\Gamma(a+1)\Gamma(b+x)}{\Gamma(a)\Gamma(b+x+1)} \left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}} \quad (5)$$

Based on (4) and (5), the expected number of transactions in the period $(T, T+t]$ for an individual with an observed history of x transactions with the last transaction occurring at time t_x can be written as:

$$E(Y(t)|x, t_x, T, r, \alpha, a, b) = \frac{\frac{b+x}{a-1} \left(1 - \left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x} {}_2F_1\left(r+x; b+x+1; a+b+x; \frac{t}{t+T+\alpha}\right) \right)}{1 + \frac{a}{b+x} \left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}}$$

where the Gamma function, $\Gamma(a)$ from Equation (5), is approximated by $(a-1)!$.

Table 1

Summary of customer groups used in parameter estimation

	Total number of customers	Number of 0 repurchasers (% in total)	Number of repurchases	Interpurchase time (excluding 0 repurchasers)	Last-Initial purchase weeks (tenure)
Cohort1 (52 wks)	5479	2452 (45%)	Mean: 3 StDev: 6	Mean: 4 wks StDev: 6	Mean: 13 wks StDev: 16
Cohort2 (52 wks)	6276	3379 (54%)	Mean: 2 StDev: 5	Mean: 5 wks StDev: 7	Mean: 11 wks StDev: 16
Cohort1 (78 wks)	5479	2221 (40%)	Mean: 5 StDev: 9	Mean: 5 wks StDev: 9	Mean: 23 wks StDev: 27

Table 2

Parameter estimates

		r	α	r/α	s	β	s/β	a	b	$a/$ $(a+b)$	LL
Cohort1 (52 wks)	<i>Pareto/NBD</i>	0.49	4.06	0.12	1.42	42.2	0.03				48,870
	<i>MBG/NBD</i>	0.89	4.28	0.21				0.45	0.75	0.40	48,919
	<i>BG/NBD</i>	0.27	2.15	0.13				0.44	2.07	0.17	49,027
Cohort2 (52 wks)	<i>Pareto/NBD</i>	0.40	4.45	0.09	0.76	15.6	0.05				43,329
	<i>MBG/NBD</i>	0.74	4.19	0.18				0.49	0.59	0.45	43,342
	<i>BG/NBD</i>	0.21	2.17	0.10				0.45	1.65	0.21	43,421
Cohort1 (78 wks)	<i>Pareto/NBD</i>	0.48	4.38	0.11	0.57	17.6	0.03				67,926
	<i>MBG/NBD</i>	0.67	3.92	0.17				0.40	0.90	0.31	67,931
	<i>BG/NBD</i>	0.27	2.34	0.12				0.40	2.09	0.16	68,007

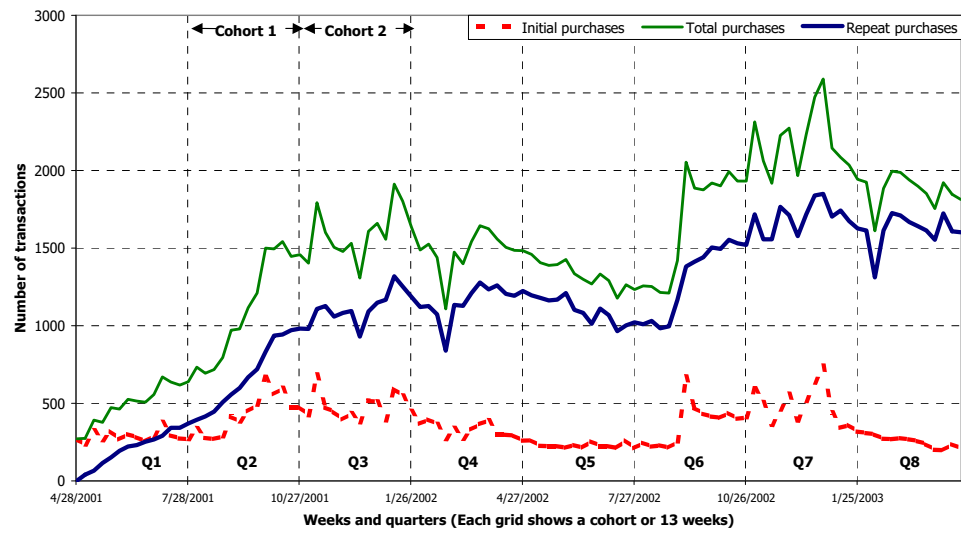


Figure 1. Purchase composition of the customer base

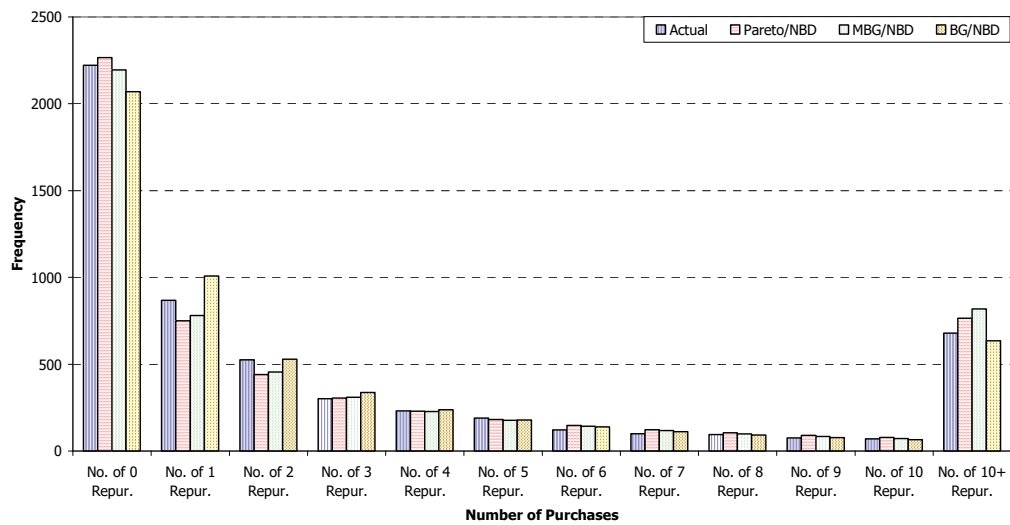


Figure 2. Frequency of repeat purchases

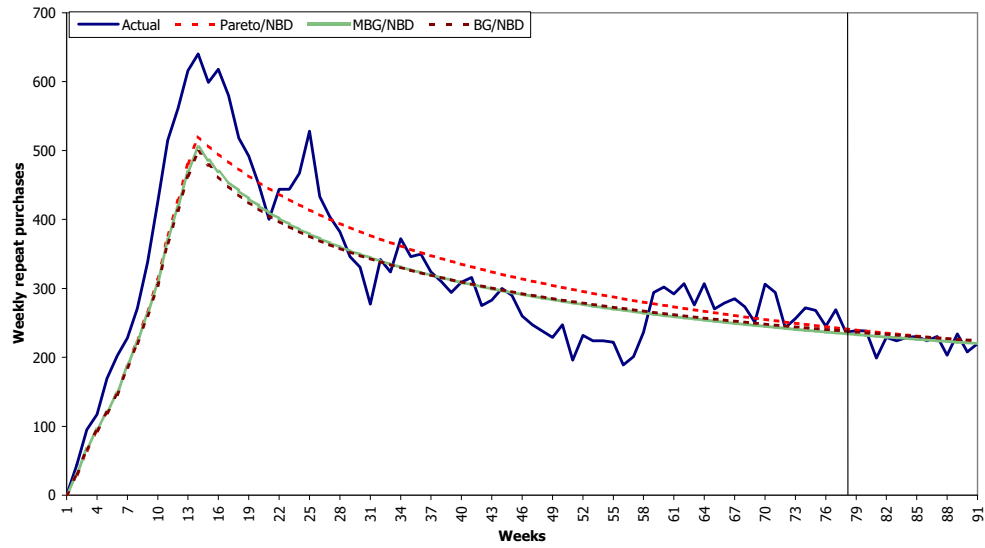


Figure 3. Weekly repeat purchases

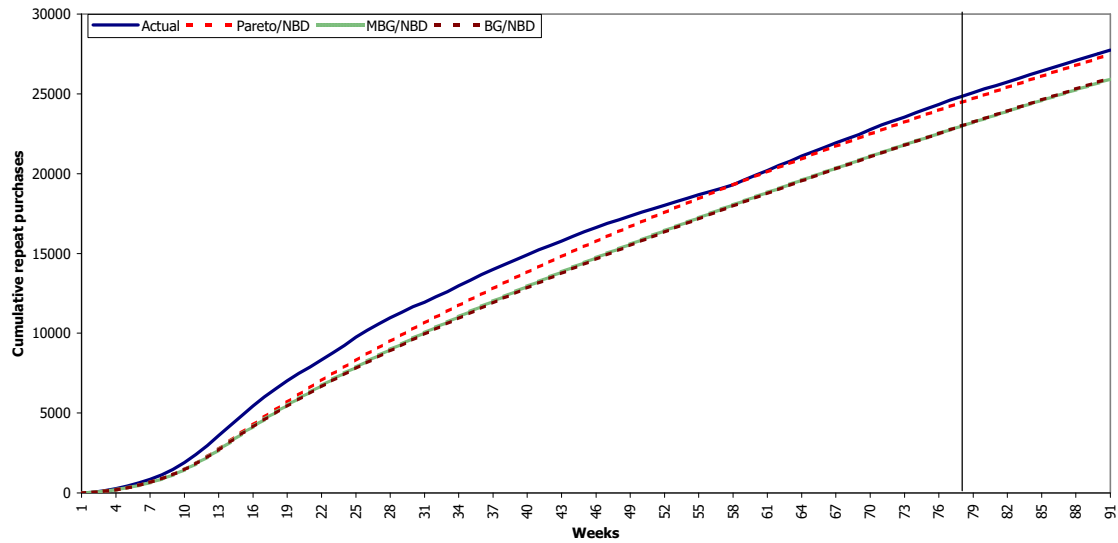


Figure 4. Cumulative repeat purchases

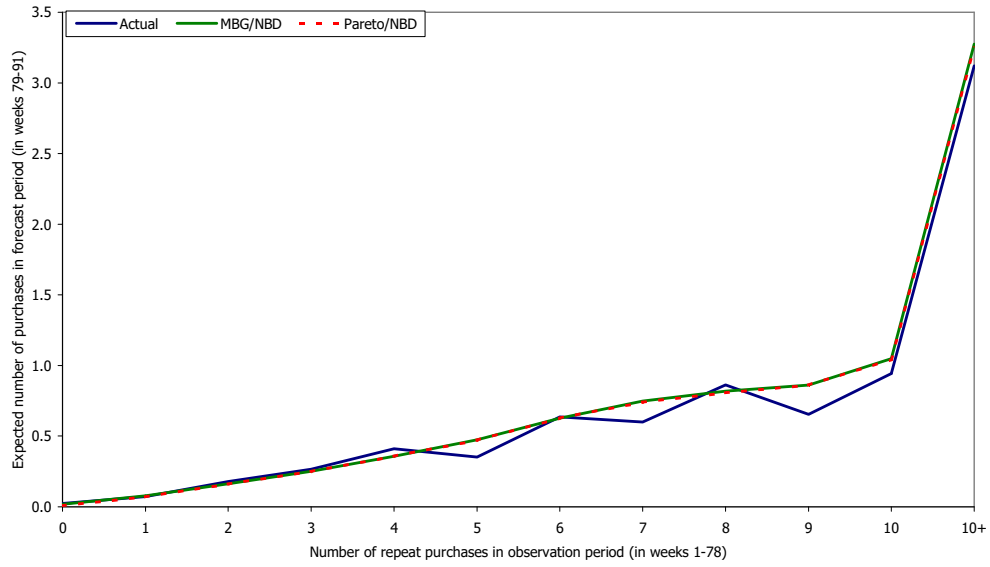


Figure 5. Conditional expectations

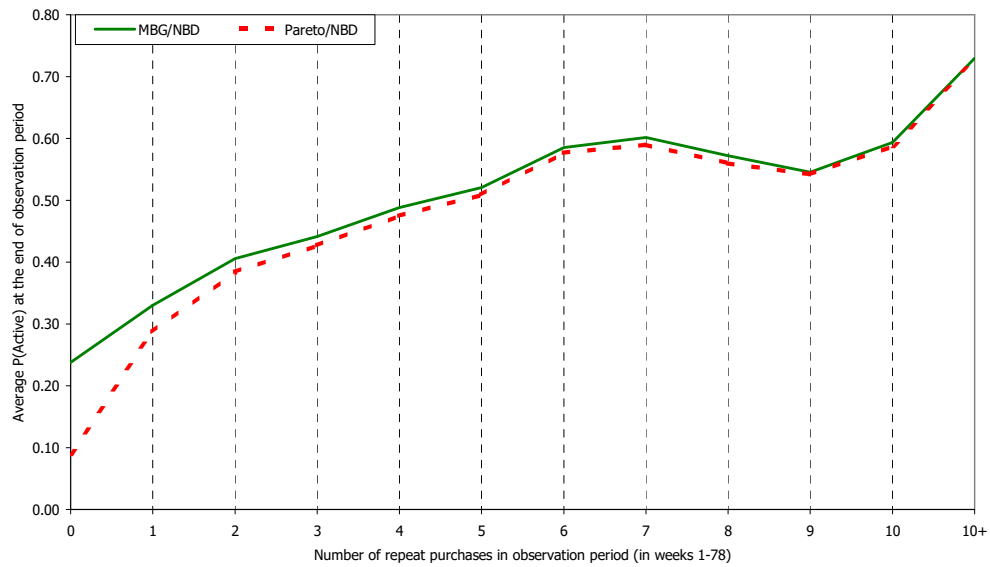


Figure 6. Probability of active at the end of week 78