Customer-Base Analysis Using Repeated Cross-Sectional Summary (RCSS) Data

Kinshuk Jerath[a,1,*], Peter S. Fader[b,1,2], Bruce G.S. Hardie[c,1]

[a] *Columbia Business School, Columbia University, 521 Uris Hall, 3022 Broadway, New York, NY 10027, USA*

[b] *The Wharton School of the University of Pennsylvania, 749 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA*

[c] *London Business School, Regent's Park, London NW1 4SA, UK*

Abstract

We address a critical question that many firms are facing today: Can customer data be stored and analyzed in an easy-to-manage and scalable manner without significantly compromising the inferences that can be made about the customers' transaction activity? We address this question in the context of customer-base analysis. A number of researchers have developed customer-base analysis models that perform very well given detailed individual-level data. We explore the possibility of estimating these models using aggregated data summaries alone, namely *repeated cross-sectional summaries* (RCSS) of the transaction data (e.g., four quarterly histograms). Such summaries are easy to create, visualize, and distribute, irrespective of the size of the customer base. An added advantage of the RCSS data structure is that individual customers cannot be identified, which makes it desirable from a privacy viewpoint as well. We focus on the widely used Pareto/NBD model and carry out a comprehensive simulation study covering a vast spectrum of market scenarios. We find that the RCSS format of four quarterly histograms

[*] Corresponding author

Email addresses: jerath@columbia.edu (Kinshuk Jerath), faderp@wharton.upenn.edu (Peter S. Fader), bhardie@london.edu (Bruce G.S. Hardie)

URL: www.petefader.com (Peter S. Fader), http://www.brucehardie.com (Bruce G.S. Hardie)

[1] The authors thank David Bell for providing the Bonobos data used in this paper.

[2] The second author acknowledges the support of the Wharton Customer Analytics Initiative.

serves as an suitable substitute for individual-level data. We confirm the results of the simulations on a real dataset of purchasing from an online fashion retailer.

Keywords: OR in marketing, analytics, customer-base analysis, probability models, Pareto/NBD, scalability, data aggregation, data privacy, information loss

1. Introduction

With rapid increases in the technology for capturing and storing customer activity data, databases on customer behavior have grown tremendously in terms of richness and size. Many firms now have petabytes of data on their customers' offline and online transactions with them. Advancements in data-analysis tools, however, have not kept up with advancements in storage technology. So while it is possible for firms to enthusiastically collect huge amounts of data, many firms are unable to make any meaningful use of it. This is sometimes referred to as "data smog" or the "too-much-data problem" (Cox, 2013; Weil, 2011; Whitler, 2012). This problem is especially acute for medium-sized firms that find it cheap to invest in data collection and storage technology but are unable to invest sufficiently in data-analysis capabilities. Specifically, scalability of data analysis methodologies is a critical issue that many firms face. While examples of analytics-powered companies that use large datasets effectively certainty exist (e.g., Amazon and Google), such examples are the exception rather than the rule. Balasubramanian et al. (1998), Kettenring (2009) and Keller et al. (2012) lay out several statistical issues that arise in the analysis of large datasets, along with different approaches to resolve these issue. As one possible solution, they call for the development of parsimonious models and the aggregation of data. In line with this, IS practitioners also call for appropriate data aggregation methods to achieve scalability (e.g., discussants on the "webanalytics" forum on Yahoo! Groups suggest that "aggregation is 90 percent of scalability" (WA, 2008)).

In parallel, we have seen the development of a rich literature on statistical models for *customer-base analysis* (e.g., Abe, 2009; Batislam et al., 2007; Fader et al., 2005a,b, 2010; Jerath et al., 2011; Morrison et al., 1982; Schmittlein et al., 1987; Schmittlein and Peterson, 1994; Singh et al., 2009; Wu and Chen, 2000). These models use data on customers' past transactions with the firm to make forecasts about their future behavior, be it total purchasing by the whole customer base or individual-level predictions (conditional on the customer's past behavior). These forecast can also be used to derive formal metrics of the expected future value to the firm of individual customers (such as customer lifetime value, or CLV) and the cohort (such as customer equity).

As is the case with much of the work in Marketing, these models have been developed assuming the analyst has ready access to the raw customer-level transaction data and has the resources to process it. However, this is frequently not the case in practice. Be it for reasons of data security (e.g., Sarbanes-Oxley) or simply good data management practices, the IS group is typically unwilling to give unfettered access to the raw transaction data. (Furthermore, as many parts of the IS and analytics functions have been outsourced, the data-protection laws in many countries (particularly in Europe) complicate the process of transferring raw data across national borders (Carey, 2009; Singleton, 2006).) Even having received copies of the relevant data files, the task of data pre-processing is resource intensive. Furthermore, there is the computational burden associated with running these traditional statistical models on larger and larger datasets (i.e., scalability).

Any statistical model used for customer-base analysis is effectively a story about the data-generating process. Is there any reason why we have to assume we have access to the raw customer-level data when implementing our models? Can we continue to tell a granular "story," but specify the model likelihood function for the data presented in a more aggregated form? Our starting point is that the answers to these questions are no and yes, respectively, and therefore ask the following question: *Can customer data be stored and analyzed in an easy-to-manage and scalable format without significantly compromising the inferences that can be made about customer activity?* Working in a setting where we track the behavior of individual customers over time, we present a methodology that achieves scalability through a carefully designed data aggregation process.

The rest of the paper is organized as follows. In the next section, we describe the data structure we study, namely *repeated cross-sectional summaries* (RCSS). Then we briefly review the model that we focus on, namely, the Pareto/NBD model, and describe how its parameters can be estimated using RCSS data. We then discuss the theoretical foundations behind the method we use to determine the number of RCSS histograms required to give model performance comparable to that associated with the use of individual-level data. Following this, we carry out a comprehensive simulation study covering a vast spectrum of market scenarios characterized by various levels of customer-base "penetration" (i.e., individuals in the database making at least one transaction in a given time period) and mean transaction frequency. For each scenario, we simulate two years of individual-level data for a synthetic customer base. These data are used to examine the loss of information associated with the RCSS data structure and how best to create the RCSS data (i.e., the period of time associated with each cross-sectional summary and the number of such summaries required for estimation). Our (conservative) recommendation is that four 13-week histograms is an appropriate RCSS configuration. We then conduct the same analyses on a real dataset from the online fashion retailer Bonobos, which confirms the results of our simulations. Our results consistently establish that the model fit, parameter values, and forecasts associated with the use of RCSS data can closely match the corresponding estimates arising from the use of individual-level data. Next, we compare the performance of RCSS data with the performance of sampling-based methods in which individual-level data is sampled for a subset of the cohort. Interestingly, we find that RCSS analysis gives comparable (in fact, slightly better) performance while taking significantly lesser time than sampling-based methods. We conclude with some managerial perspectives on the use of RCSS data, including some other benefits that make it an attractive alternative to model implementation using individual-level data—even when there are no problems with data availability that may necessitate its use.

2. The RCSS Data Structure

Statistical models for customer-base analysis have been developed assuming the analyst has ready access to the raw customer-level transaction data. With reference to Figure 1a, it is assumed that we know the timing of each transaction for each customer (denoted by × on the individual time lines), along with its associated monetary value, etc. In some cases, the data may be stored in the firm's databases in such a way that we do not know the timing of each transaction, only how many transactions occurred in each pre-specified time interval (as illustrated in Figure 1b). Fader and Hardie (2010) derive the likelihood function for fitting the Pareto/NBD model to such interval-censored data. Going further, individual-level interval-censored data can be summarized across individuals for each time interval, resulting in what we call *repeated cross-sectional summary* (RCSS) data (as illustrated in Figure 1c). Each "summary" is a histogram and, for each histogram, the heights of the bars indicate the number of people in the database making $0, 1, 2, \ldots$ transactions in that period.

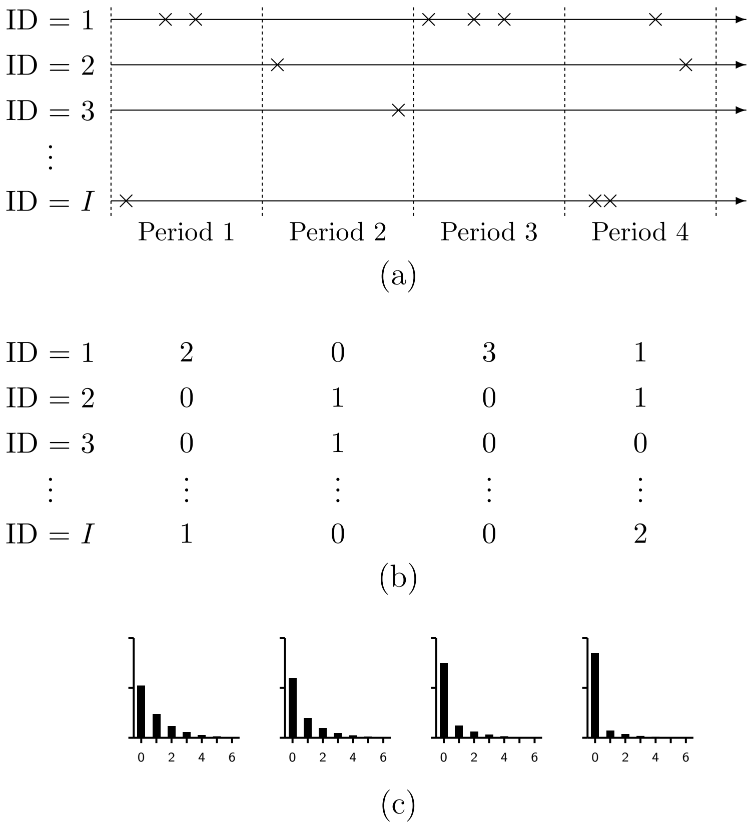

Figure 1: Panel (a) represents the individual-level transaction data for a customer database containing the records for I customers. Panel (b) reports the same data in an interval-censored form. Panel (c) shows the repeated cross-sectional summary (RCSS) data for the customer base.

Such RCSS data have a number of attractive properties. They are easy to create and distribute. Furthermore, they are inherently scalable. While the size of the dataset required by the analyst is a function of the number of customers in the transaction database (or the size of any sample) when the data are requested in raw transaction form (or as interval-censored data), the size of the RCSS dataset is effectively independent of the number of customers — as we go from summarizing the behavior of one thousand customers to one million customers, the only thing that changes is the scaling of the y-axis (i.e., the heights of the bars of the summary histograms in Figure 1c). An interesting by-product of creating these summaries is that, from the data structure that emerges, it is truly impossible to "reverse engineer" the behavior associated with any specific customer. This makes RCSS data an attractive option from a privacy-preservation viewpoint as well.

Of course, the downside to such a data summary is obvious: it is no longer possible to see the timing of the individual-level transactions. For instance, we would have no way of knowing that the customer with ID=1 above made $2, 0, 3$ and 1 transactions in periods $1, 2, 3$ and 4, respectively. This potentially reduces the information content of the data as we seek to estimate the parameters of the model intended to describe the customer buying behavior.

Some marketing managers and analysts already use RCSS data as they seek to characterize the behavior of their customers and compute basic performance metrics. For example, the *Tuscan Lifestyles* case (Mason, 2003) features a marketing director using summaries of customer behavior in the form of five annual histograms (of the number of orders per year) for the purpose of computing the value of a new customer. There is some evidence that such data can be used as a basis for more sophisticated analysis: Fader et al. (2007) illustrate how the Pareto/NBD can be fit to the RCSS data provided in the *Tuscan Lifestyles* case.

But while Fader et al. (2007) provide an initial "proof of concept" for the RCSS approach, they do not address three issues that may be critical for any analyst contemplating the use of such a data structure. First, they do not demonstrate that it is an adequate substitute for individual-level data. For example, are we able to draw the same model-based inferences we would make if we had access to the individual-level data? Second, they do not give any guidance for the creation of the repeated cross-sectional summaries of the transaction data. (They simply used the five annual histograms reported in Mason (2003).) For example, if we have a year of transaction data, can we fit the model using just one histogram that summarizes the year's transactions by the customer base, or is it better to use two six-monthly histograms, or four quarterly histograms? Such guidance is needed for any analyst considering the use of such a data structure. Third, the *Tuscan Lifestyles* case is purely a "backward-looking" analysis: it takes a historical dataset and "chops it up" into a RCSS format. Practitioners also need "forward-looking" advice; that is, how to create RCSS data "on the fly" as new transactions are being recorded. In this paper, we develop and explore this kind of analysis for simulated and real datasets.

With this background in mind, our objectives are as follows. Given RCSS data, we examine: (i) how much information is "lost" when repeated cross-sectional data summaries are used instead of the individual-level data, and (ii) how many cross-sections are required to minimize information loss. In order to address these issues, we need to pair up the RCSS data structure with a well-established model for customer-base analysis in a noncontractual setting; this gives us the benchmark required to understand how much information loss will occur under these circumstances. In this context, a natural benchmark is the Pareto/NBD model (Schmittlein et al., 1987). We know from the empirical validations of the model presented in Schmittlein and Peterson (1994) and Fader et al. (2005b), amongst others, that its predictive performance is impressive. Applications of this model include the work of Reinartz and Kumar (2000, 2003) on customer profitability, Hopmann and Thede (2005) on churn prediction, and Wübben and Wangenheim (2008) and Huang (2012) on managerial heuristics for customer-base analysis. Given its widespread and successful application, the Pareto/NBD is an appropriate model to use to evaluate the effectiveness of RCSS data in a customer-base analysis setting.

As an aside, we note that the term "transaction" used above is very general. To list a few examples, it can refer to purchasing from the firm through online and/or offline channels, visits to the firm's website, instances of content viewing on different media channels, advertisement exposures over time, and so on. Therefore, our study is relevant in a wide variety of practical situations (effectively, any situation in which customer activity is recorded over time). In the rest of the paper, we will use "transaction" and "purchase" interchangeably.

2.1. Related Literature

Constructing RCSS data is a data aggregation technique that groups continuous data into intervals. A large literature in statistics has addressed the question of how continuous data should be grouped into discrete intervals, and how much of the information content of the data is lost in this data-reduction exercise (e.g., Aghelvi and Mehran, 1981; Connor, 1972; Cox, 1957; Davies and Shorrocks, 1989; Gastwirth and Krieger, 1975; Krieger and Gastwirth, 1984; Parmigiani, 1998; Sawiris, 2000; Shaw et al., 1987; Tryfos, 1985). Each of the above papers addresses this question in a context specific to its motivating problem. For instance, several of the above papers study different variants of the problem of grouping income data with the objective that the values of an index of income inequality (e.g., the Gini coefficient) calculated from the continuous data and the grouped data are close enough (Aghelvi and Mehran, 1981; Davies and Shorrocks, 1989). Based on the specific problem of interest, an appropriate information-loss function and an acceptable degree of information loss are chosen; there is no general solution.

Our method is also related to other data-reduction alternatives such as *data squashing* (see DuMouchel (2002) for a review), in which a large dataset is summarized into a smaller dataset subject to some constraints, e.g., preserving the lower-order moments of the data to a specified level of accuracy. Other data reduction and pre-processing techniques are discussed in Han and Kamber (2006). Zheng et al. (2003) study some popular data-reduction methods and show that they are not without their pitfalls since different methods can lead to drastically different results, both in terms of characterizing the original data and out-of-sample predictions. In the analysis that follows, we show that our method of aggregation has no such problems (at least within the broad scope of our analysis).

A number of researchers have explored the issue of how to manage data for distribution so as to overcome privacy-related concerns (e.g., Menon and Sarkar, 2007). See Fung et al. (2010) for a review of current "privacy-preserving data publishing" practices. Common approaches include anonymizing individual-level data (by removing identifying information such as names, addresses and Social Security numbers) and perturbing micro-data such that the individual-level records look different but the distributions of original data values can be accurately estimated (Agrawal and Srikant, 2000; Gopal et al., 2002; Li and Sarkar, 2006a,b). However, there are serious concerns about the extent to which privacy is actually preserved in anonymized or perturbed micro-data (Malin and Sweeney, 2004; Mielikäinen, 2004; Narayanan and Shmatikov, 2008). Our approach is substantially different from all of the above because we summarize individual-level data using transaction frequency histograms. There are absolutely no privacy concerns in the data format we propose and we show that we can still accurately recover the key characteristics of the original data.

3. A Brief Review of the Pareto/NBD Model

The actual data-generating process that lies behind any observed customer behavior is without doubt very complex. Even if completely deterministic, it is not possible to measure all the variables that determine an individual customer's behavior. As such, any model of this behavior should be expressed in probabilistic terms so as to account for our ignorance regarding all the determinants (not to mention the lack of data to capture them). Rather than try to tease out the effects of various marketing, personal, and situational variables, a probability model acknowledges the fact that we can never completely describe the actual data-generating process. Thus we embrace the notion of stochasticity, viewing the behavior of interest as the outcome of some probabilistic process. Such a model typically has two components. First, the individual behavior of interest is characterized in terms of a probability distribution (or several distributions in the case of more complex models). Second, differences in the parameters of the individual-level distribution(s) are captured by additional probability distributions, resulting in a mixture distribution that characterizes the behavior of a randomly chosen individual. There is a long tradition of such models in the marketing literature (Fader et al., 2014).

The defining characteristic of what Reinartz and Kumar (2000) call a noncontractual setting is that the time at which an individual "dies" (i.e., should no longer be considered a customer) is unobserved by the firm. In their seminal model of buyer behavior in such settings, Schmittlein et al. (1987) proposed a latent-attrition/"buy till you die" model in which every customer is assumed to make transactions until he drops out of the cohort. More formally, they assumed that a customer's relationship with the firm has two phases: he is alive for an unobserved period of time, then becomes permanently inactive (i.e., dies). While alive, the customer's transaction activity is characterized by the NBD model (i.e., a customer's inter-arrival times are iid exponential and heterogeneity in the transaction rates is captured by a gamma distribution with shape parameter r and scale parameter α). The customer's unobserved "lifetime" (after which he is viewed as being dead) is treated as-if random, characterized by another exponential distribution; heterogeneity in the underlying death rate across customers is assumed to follow a gamma distribution with shape parameter s and scale parameter β. Noting that a gamma mixture of exponentials is also known as the Pareto (of the second kind) distribution, the resulting model of buyer behavior is called the Pareto/NBD.

Given the assumptions of the Pareto/NBD model, it turns out that we do not require information on when each of the customer's transactions occurred (as illustrated in Figure 1a). The only customer-level information required to estimate the four model parameters are "recency" and "frequency." The conventional notation used to represent this recency and frequency gwtoi4pi4information is $(x, t_x, \mathcal{T})$, where x is the number of transactions that occurred in the time interval $(0, \mathcal{T}]$ and t_x $(0 \leq t_x \leq \mathcal{T})$ is the time of the last transaction. (Note that while we do not need to know the exact time of each transaction, we do need to know the exact time of the last observed transaction.)

Given such "full information" data, the model likelihood function is

$$L(r, \alpha, s, \beta \,|\, x, t_x, \mathcal{T}) = \frac{\Gamma(r+x)\alpha^r \beta^s}{\Gamma(r)} \left\{ \frac{s}{r+s+x} \mathsf{A}_1 + \frac{r+x}{r+s+x} \mathsf{A}_2 \right\} \tag{1}$$

where

$$\mathsf{A}_1 = \begin{cases} \dfrac{{}_2F_1\big(r+s+x, s+1; r+s+x+1; \frac{\alpha-\beta}{\alpha+t_x}\big)}{(\alpha+t_x)^{r+s+x}} & \text{if } \alpha \geq \beta \\ \dfrac{{}_2F_1\big(r+s+x, r+x; r+s+x+1; \frac{\beta-\alpha}{\beta+t_x}\big)}{(\beta+t_x)^{r+s+x}} & \text{otherwise} \end{cases}$$

and

$$\mathsf{A}_2 = \begin{cases} \dfrac{{}_2F_1\big(r+s+x, s; r+s+x+1; \frac{\alpha-\beta}{\alpha+\mathcal{T}}\big)}{(\alpha+\mathcal{T})^{r+s+x}} & \text{if } \alpha \geq \beta \\ \dfrac{{}_2F_1\big(r+s+x, r+x+1; r+s+x+1; \frac{\beta-\alpha}{\beta+\mathcal{T}}\big)}{(\beta+\mathcal{T})^{r+s+x}} & \text{otherwise,} \end{cases}$$

where ${}_2F_1(\cdot)$ is the Gaussian hypergeometric function. (See Fader and Hardie (2005) for complete details of the derivation.)

For a sample of I customers, where customer i had x_i transactions in the period $(0, \mathcal{T}_i]$ with the last transaction occurring at t_{x_i}, the four Pareto/NBD model parameters (r, α, s, β) are estimated by maximizing the sample log-likelihood function

$$LL(r, \alpha, s, \beta \,|\, \text{data}) = \sum_{i=1}^{I} \ln \left[L(r, \alpha, s, \beta \,|\, x_i, t_{x_i}, \mathcal{T}_i) \right]. \tag{2}$$

This standard approach to parameter estimation assumes that we have the sufficient statistics (recency and frequency) for each and every customer; as such, it is of no use given RCSS data. However, as noted in the introduction, all we need to do is re-derive the model likelihood function for the data format at hand. Fader et al. (2007) demonstrate how this can be done in the case of RCSS data. We briefly review this approach here.

Schmittlein et al. (1987) derive an expressions for $P(X(0,\mathcal{T})=x)$, where the random variable $X(0,\mathcal{T})$ denotes the number of transactions observed in the time interval $(0,\mathcal{T}]$, as implied by the Pareto/NBD model assumptions. Following Fader et al. (2006), the probability of observing x transactions in the time interval $(\mathcal{T}_{j-1},\mathcal{T}_j]$, where $\mathcal{T}_{j-1}\geq 0$, is

$$\begin{aligned}
&P(X(\mathcal{T}_{j-1},\mathcal{T}_j)=x\,|\,r,\alpha,s,\beta)\\
&\quad=\delta_{x=0}\left[1-\Big(\frac{\beta}{\beta+\mathcal{T}_{j-1}}\Big)^s\right]\\
&\qquad+\frac{\Gamma(r+x)}{\Gamma(r)x!}\Big(\frac{\alpha}{\alpha+\mathcal{T}_j-\mathcal{T}_{j-1}}\Big)^r\Big(\frac{\mathcal{T}_j-\mathcal{T}_{j-1}}{\alpha+\mathcal{T}_j-\mathcal{T}_{j-1}}\Big)^x\Big(\frac{\beta}{\beta+\mathcal{T}_j}\Big)^s\\
&\qquad+\alpha^r\beta^s\frac{B(r+x,s+1)}{B(r,s)}\left\{\mathsf{B}_1-\sum_{k=0}^{x}\frac{\Gamma(r+s+k)}{\Gamma(r+s)}\frac{(\mathcal{T}_j-\mathcal{T}_{j-1})^k}{k!}\mathsf{B}_{2k}\right\}
\end{aligned}\tag{3}$$

where

$$\mathsf{B}_1=\begin{cases}\dfrac{{}_2F_1\big(r+s,s+1;r+s+x+1;\frac{\alpha-(\beta+\mathcal{T}_{j-1})}{\alpha}\big)}{\alpha^{r+s}} & \text{if } \alpha\geq\beta+\mathcal{T}_{j-1}\\[2ex] \dfrac{{}_2F_1\big(r+s,r+x;r+s+x+1;\frac{\beta+\mathcal{T}_{j-1}-\alpha}{\beta+\mathcal{T}_{j-1}}\big)}{(\beta+\mathcal{T}_{j-1})^{r+s}} & \text{otherwise}\end{cases}$$

and

$$\mathsf{B}_{2k}=\begin{cases}\dfrac{{}_2F_1\big(r+s+k,s+1;r+s+x+1;\frac{\alpha-(\beta+\mathcal{T}_{j-1})}{\alpha+\mathcal{T}_j-\mathcal{T}_{j-1}}\big)}{(\alpha+\mathcal{T}_j-\mathcal{T}_{j-1})^{r+s+k}} & \text{if } \alpha\geq\beta+\mathcal{T}_{j-1}\\[2ex] \dfrac{{}_2F_1\big(r+s+k,r+x;r+s+x+1;\frac{\beta+\mathcal{T}_{j-1}-\alpha}{\beta+\mathcal{T}_j}\big)}{(\beta+\mathcal{T}_j)^{r+s+k}} & \text{otherwise.}\end{cases}$$

This equation lies at the heart of any effort to estimate the parameters of the Pareto/NBD model using RCSS data. Suppose the calibration period $(0,\mathcal{T}]$ is split into J consecutive periods: $(0,\mathcal{T}_1],(\mathcal{T}_1,\mathcal{T}_2],\ldots,(\mathcal{T}_{J-1},\mathcal{T}_J]$.

(These periods do not have to be of equal length, but in practice we would expect them to be.) For each period, we determine how many people make 0, 1, 2, 3, ... transactions, giving us a histogram of transactions (as illustrated in Figure 1c).

Let us assume each histogram is right censored and has the bins $0, 1, 2, \ldots,$ $z-1, z+$ (i.e., the number of individuals who did not make a transaction, the number of individuals who transacted exactly once, exactly twice, ..., exactly $z-1$ times, and z or more times). Let $n(j, x)$ denote the number of people making x transactions in the jth time interval $(\mathcal{T}_{j-1}, \mathcal{T}_j]$ and $n(j, z+)$ denote the number of people making z or more transactions in this time interval.

The four Pareto/NBD model parameters (r, α, s, β) are estimated by maximizing the following log-likelihood function for RCSS data spanning J consecutive periods, with the first period starting at $\mathcal{T}_0 = 0$:

$$\begin{aligned} LL(r, \alpha, s, \beta \,|\, \text{data}) = \sum_{j=1}^{J} \Bigg\{ & \sum_{x=0}^{z-1} n(j, x) \ln(P(X(\mathcal{T}_{j-1}, \mathcal{T}_j) = x \,|\, r, \alpha, s, \beta)) \\ & + n(j, z+) \ln(P(X(\mathcal{T}_{j-1}, \mathcal{T}_j) \geq z | r, \alpha, s, \beta)) \Bigg\} . \quad (4) \end{aligned}$$

Once the parameters of the Pareto/NBD model are estimated, we can compute a number of different measures of interest to management. One such measure is customer lifetime value, which is the net present value of the future cashflows associated with a customer. The general explicit formula for the computation of customer lifetime value is

$$E(\text{CLV}) = \int_0^\infty E[v(t)]S(t)d(t)dt \,,$$

where $v(t)$ is the net cashflow associated with the customer at time t, $S(t)$ is the survivor function (i.e., the probability that the customer has remained alive to at least time t), and $d(t)$ is a discount factor that reflects the present value of money received at time t. Following Fader et al. (2005b), if we assume a constant net cashflow per transaction of $\bar{v}$, we have $v(t) = \bar{v}t(t)$, where $t(t)$ is the transaction rate at time t, and we have

$$E(\text{CLV}) = \bar{v} \int_0^\infty E[t(t)]S(t)d(t)dt \,.$$

The solution to the integral is called the Discounted Expected Transactions (DET) of the individual. For a just-acquired customer, DET measures the present value to the firm of all future transactions by the customer (accounting for the transactions while alive, and the death processes), with transactions at a future time point appropriately discounted to obtain their present values. Multiplying by $\bar{v}$ gives us an estimate of expected CLV.

When the flow of transactions is characterized by the Pareto/NBD model, DET of a randomly chosen customer is given by

$$\mathrm{DET}(r, \alpha, s, \beta, \delta) = \frac{r}{\alpha} \beta \Psi(1, 2 - s; \beta\delta)\,, \tag{5}$$

where δ is the continuous compound rate of interest and $\Psi(\cdot)$ is the confluent hypergeometric function of the second kind. (See the Appendix for details of the derivation.) Given the central importance of CLV (and thus DET) as both a managerially interesting metric as well as a long-run output of the Pareto/NBD model, we will examine it in our empirical investigations.

The question facing the analyst is: what level of granularity of RCSS data (i.e., the period of time associated with each cross-sectional summary and the number of such summaries required for estimation) is sufficient for it to be an adequate substitute for individual-level data? In the following section, we discuss the theoretical foundations of the method we use to answer this question.

4. Determining the Number of Histograms: Theoretical Foundations

The RCSS method we propose is basically a lossy data compression method in which individual-level data is aggregated into a set of histograms. To determine the RCSS configuration that will adequately substitute for individual-level data, we can invoke concepts from *rate-distortion theory*, a branch of information theory that provides theoretical foundations for lossy data compression. Rate-distortion theory states that, given the task of compressing data $\mathcal{X}$, the analyst must determine: (i) a distortion function $D(\mathcal{X}, \mathcal{Y})$, where $\mathcal{Y}$ is the compressed data, and $D(\mathcal{X}, \mathcal{Y})$ returns as output a scalar which indicates the degree of distortion, and (ii) a distortion threshold D^*. Note that the exact form of the distortion function and the value of D^* are chosen by the analyst, as appropriate for the application at hand. The compression of $\mathcal{X}$ to $\mathcal{Y}$ is acceptable if $D(\mathcal{X}, \mathcal{Y}) \leq D^*$.

Given the above constraint, rate-distortion theory provides a way to determine the compression method that ensures the most effificent message transfer (using the minimum number of bits of information) over a channel such that the source (input signal, $\mathcal{X}$) can be approximately reconstructed at the receiver (output signal, $\mathcal{Y}$). Specifically, this can be achieved by maximizing the mutual information between $\mathcal{X}$ and $\mathcal{Y}$. (See Cover and Thomas (2006) for more details.) It is well known, however, that this optimization problem typically cannot be solved analytically, except in the case of some simple textbook examples; thus a numerical approach is almost always employed.

In our case, we have already fixed the compression method — given the individual-level data as the input, we compress it into the RCSS data structure with a certain number of histograms as the output. The key decision that needs to be made is which configuration of the RCSS data structure is good enough (i.e., how many histograms should we use, and what should be the period of coverage of each histogram, such that the distortion after the lossy data compression is acceptable). In other words, we need to invoke only the distortion component of rate-distortion theory by defining a distortion metric when individual-level data is "compressed" into RCSS data.

As is done in applications of rate-distortion theory and in the literature on inference from grouped continuous data (as discussed above), we need to choose both the distortion metric and the distortion threshold. We define the distortion metric as follows. Suppose $\mathcal{X}$ is the individual-level data and $\mathcal{Y}_h$ is the RCSS data with h histograms. We first estimate the model parameters using the individual-level data and obtain the maximum value of the log-likelihood function in (2); we denote this by $LL_{\mathcal{X}}$. We then estimate the model parameters using a particular configuration of the RCSS data with h histograms by maximizing (4) and, using the parameters obtained, we calculate the value of (2); we denote this by $LL_{\mathcal{Y}_h}$. We define the distortion metric $D(\mathcal{X}, \mathcal{Y}_h)$ as the absolute percentage difference between $LL_{\mathcal{X}}$ and $LL_{\mathcal{Y}_h}$, which is defined as follows:

$$D(\mathcal{X}, \mathcal{Y}_h) = \left| \frac{LL_{\mathcal{Y}_h} - LL_{\mathcal{X}}}{LL_{\mathcal{X}}} \right| \times 100, \tag{6}$$

where $|x|$ denotes the absolute value of x. This is an appropriate distortion metric since the smaller it is, the closer $LL_{\mathcal{Y}_h}$ is to $LL_{\mathcal{X}}$, which is the maximum of the individual-level log-likelihood function.

With regard to the distortion threshold D^*, we use very strict distortion thresholds of the order of thousandths of one percent. However, given that we conduct a simulation study spanning a large number of different market scenarios, choosing one distortion threshold to use across all the different market scenarios does not seem appropriate. We therefore base our analysis on the trend in the values of the distortion function $D(\mathcal{X}, \mathcal{Y}_h)$ with different number of histograms (i.e., $h \in \{1, 2, 3, \dots\}$). We observe that, for every world, the distortion values stabilize to very small values beyond a certain number of histograms in the RCSS configuration; we use this to determine the appropriate RCSS configuration. More formally, we determine that $h \in \{2, 3, \dots\}$ is the appropriate number of histograms for the RCSS data structure if the distortion with h histograms is small enough, and there is a sharp reduction in distortion in going from $h-1$ to h histograms and a comparatively small reduction in distortion in going from h to $h+1$ and $h+1$ to $h+2$ histograms. In other words, we say that h is the appropriate number of histograms if, in the plot of distortion metric values against the number of histograms used, the "elbow" occurs at h, and the distortion at h is small enough.

5. Simulation Study

We approach the question of how best to create the RCSS data from two directions. First, we undertake a *"backward-looking"* analysis in which we consider how finely we need to partition a fixed period into cross-sectional summaries before model performance and parameter recovery stabilizes (if at all). Suppose we have 52 weeks of purchasing data: should we use one 52-week histogram, or two 26-week histograms, or three 17.3-week histograms, ..., or six 8.7-week histograms? Second, we undertake a *"forward-looking"* analysis in which we consider how many cross-sectional summaries (each of a fixed length) are required for a newly emerging dataset before model performance and parameter recovery stabilizes.

Before we present the simulation study, we offer a brief overview of what we expect to see as we increase the number of histograms used for the RCSS model.

Consider the case of a backward-looking analysis where the analyst already has 52 weeks of data and has to decide how many cross-sectional data summaries he should construct (one 52-week histogram, two 26-week histograms, etc.). In the case of one 52-week histogram, the recovery of the true underlying parameters (especially those associated with the death process) will be weak because we have clumped together all the data available for one year. Upon such aggregation, we lose the ability to track any time trends across the 52-week period that help identify the customer death process. Many different combinations of the parameters may lead to very similar 52-week histograms. For example, a large number of zero buyers could be due to a low underlying purchase rate while alive (i.e., r/α is small) or the fact that a large number of existing customers died early on in the year before they got around to making any purchases (i.e., s/β is large). Now consider the case where we have two 26-week histograms available. In this case, the ability to pin down the correct model parameters is better, since we can start to separate out patterns related to the death process through the growth in the number of zeroes from the first histogram to the second one. But it is still difficult to describe the nature of this growth pattern over time.

As the number of histograms increases, we can track the underlying dynamics in purchasing patterns more closely and we have more information to recover the true parameters. For instance, in the four histograms in Figure 1c, we can see that the number of customers making zero purchases increases over time and the number of customers making one, two or three purchases decreases over time, which is a strong indication of customer death. If we used one histogram for this length of time, we would not have been able to make any such inferences about customer death over time. If we used two histograms for this length of time, we would have been able to make such inferences but they would have been rather coarse.

But there may be limits to this logic if we take it too far. As we "chop" the data into a greater number of histograms (each covering a smaller number of weeks), we may lose the meaningful information content from each histogram. In the extreme case of, say, weekly histograms, we would have an enormous number of 0's and a very limited number of 1's in every histogram. Thus, we do not want to go too far in creating histograms to summarize the data; we want to find the "just right" balance of parsimony and information value.

Similar arguments can be extended to a forward-looking analysis, except that the problem here is even tougher. In this case, the analyst is constructing, say, quarterly histograms "on the go" and has to decide how many quarters of data are sufficient to yield stable (and valid) parameter estimation. As before, one histogram for one quarter is not expected to capture the purchasing behavior particularly well — the data are for a very short time period and we have only one histogram which makes it difficult to track any underlying time trends. If we use two histograms for the first two quarters, we can begin to gain some insight into the (latent) attrition patterns in the cohort, but the total period of time covered is still short and it will be hard to discern trends in the death process as well as heterogeneity in the purchasing while alive process. Histograms for three or four quarters start to capture behavior over a considerable length of time and are expected to further improve model performance. At some point, the incremental value of waiting for additional histograms will be quite limited for the purposes of determining the model parameters.

5.1. Simulation Design

We turn to an extensive simulation study that helps us answer the following questions: Can models built using RCSS data match the performance of traditional models based on individual-level data? If yes, how many histograms are required, both for backward-looking analysis and forward-looking analysis?

As noted in the previous section, the Pareto/NBD model has four parameters: r, α, s, β. The parameters r and α capture the heterogeneity in the latent purchase rates in the cohort, while the parameters s and β capture the heterogeneity in the latent death rates in the cohort. For the purchasing process, as α increases for a fixed value of r, the mean purchase rate for the cohort decreases, and as r decreases for a fixed value of α, the heterogeneity in purchase rates in the cohort increases. For the death process, as β increases for a fixed value of s, the median lifetime of customers in the cohort increases, and as s decreases for a fixed value of β, the heterogeneity in median lifetime increases. Hence, by manipulating these parameters we can construct a number of cohorts, each of which reflects a different pattern of customer purchasing and latent attrition behaviors.

We vary each of the four parameters at three levels, thus generating $3^4 = 81$ different "worlds." For the purchase process, we assign to r the values 0.5, 1 and 1.5 and we assign to α the values 5, 10 and 15; this results in average weekly purchase rates (while alive) ranging from 0.033 to 0.3. For the death process, we assign to s the values 0.5, 1 and 1.5, and to β the values 5, 10 and 15; this results in median lifetimes ranging from 2.9 to 45 weeks. (These parameter values are consistent with the range of values seen in many prior papers.) In the least active of the 81 worlds, the median customer life is a mere 2.9 weeks and the average customer is expected to make only 1.7 purchases if he lives for 52 weeks. This lower bound in our simulation is indeed a very low-activity world. In the most active of the 81 worlds, the median customer life is 45 weeks and the average customer is expected to make 15.6 purchases if he lives for 52 weeks. This upper bound represents a very high-activity world, with both the metrics (median customer life and average purchases while alive) being a full order of magnitude larger than in the “lower bound.” In terms of observable quantities, annual customer-base penetration ($100 \times P(X(0, 52) > 0)$) ranges from 13.1% to 85.7%, with an average of 47.9%, while average annual sales per member of the customer base ($E[(X(0, 52)]$) ranges from 0.2 to 9.9, with an average of 2.4.

For each of the 81 worlds, we simulate 104 weeks of individual-level data for 25 synthetic cohorts of 10,000 customers each. We use the first 52 weeks of data as the calibration sample and the last 52 weeks as the holdout sample. As explained above, using a 52-week data window gives a wide variation across cohorts in the purchasing and death processes. We now proceed to the backward-looking and forward-looking analyses.

5.2. Backward-Looking Analysis

In the backward-looking analysis, the analyst already has individual-level data for 52 weeks and wants to compress it into the RCSS format. The question is: How many data summaries should he construct? Is there a sufficient or optimal level of aggregation so that the resulting summaries recover the underlying story of customer purchasing dynamics (i.e., recover the model parameters)?

To answer these questions, for each of our 81 worlds, for each of the 25 simulation runs, we consider one 52-week histogram, two 26-week histograms, three 17.3-week histograms, four 13-week histograms, five 10.4-week histograms and six 8.7-week histograms. Each histogram has eleven “number of purchases” bins: $0, 1, \ldots, 9, 10+$.[3] For each of the configurations of the RCSS data, we calculate the distortion metric defined in (6) and determine the RCSS configuration that is an acceptable substitute for individual-level data. (Note that the distortion metric used to determine the appropriate RCSS configuration is based on a measure of in-sample fit.) Following this, we show that out-of-sample performance, parameter recovery, and recovery of DET are also very good for the appropriate RCSS configuration.

To illustrate the patterns in the results, we present in Table 1 the values of the distortion metric, $D(\mathcal{X}, \mathcal{Y}_h)$, for $h \in \{1, 2, 3, 4, 5, 6\}$ averaged across 25 simulation runs, for five “worlds.” These worlds are chosen to represent

[3] We also ran the complete simulation study with different numbers of bins, specifically, with the histograms ending at 13+, 16+, 19+, 22+ and 25+. In every case, we obtained virtually identical results in terms of all metrics considered.

worlds with a wide range of mean purchase rates and median lifetimes for the cohort. These are:

- World MM: specified by $r = 0.5, \alpha = 5, s = 0.5, \beta = 5$, with a medium mean purchase rate (0.1 purchases per week) and a medium median lifetime (15 weeks);

- World LL: specified by $r = 0.5, \alpha = 15, s = 1.5, \beta = 5$, with the lowest mean purchase rate (0.03 purchases per week) and the lowest median lifetime (2.94 weeks);

- World LH: specified by $r = 0.5, \alpha = 15, s = 0.5, \beta = 15$, with the lowest mean purchase rate (0.03 purchases per week) and the highest median lifetime (45 weeks);

- World HL: specified by $r = 1.5, \alpha = 5, s = 1.5, \beta = 5$, with the highest mean purchase rate (0.3 purchases per week) and the lowest median lifetime (2.94 weeks);

- World HH: specified by $r = 1.5, \alpha = 5, s = 0.5, \beta = 15$, with the highest mean purchase rate (0.3 purchases per week) and the highest median lifetime (45 weeks).

The numbers in Table 1 show that, for all the five worlds above, there is a substantial improvement in the value of the distortion metric as the number of histograms increases up to four, but beyond that point the drop is minuscule. The plot in Figure 2 shows this pattern visually for World MM. Furthermore, for all five worlds, the value of the distortion metric when four histograms are used is very small. Therefore, the RCSS configuration with $h = 4$ (i.e., four histograms of 13 weeks each) is an appropriate RCSS configuration. The pattern for these five worlds is highly representative of the patterns for all 81 worlds.

In Table 2, we present statistics to provide an idea of the distribution of the distortion metric across all 81 worlds (where the value used for each world is the average across the 25 runs for that world). Interestingly, we find that the pattern that holds for the five worlds discussed above holds for the average values of the distortion metrics across all worlds. To see this, consider the second column in Table 2, which shows the average values of the distortion metric, $D(\mathcal{X}, \mathcal{Y}_h)$, across all the worlds for different RCSS configurations. As before, the average of the distortion metrics stabilizes beyond four histograms

RCSS configuration	$D(\mathcal{X}, \mathcal{Y}_h)$ World MM	World LL	World LH	World HL	World HH
1 × 52 wks	0.160%	0.648%	0.298%	2.050%	0.392%
2 × 26 wks	0.006%	0.080%	0.005%	0.010%	0.001%
3 × 17.3 wks	0.004%	0.048%	0.003%	0.008%	0.001%
4 × 13 wks	0.002%	0.021%	0.003%	0.005%	0.001%
5 × 10.4 wks	0.002%	0.014%	0.004%	0.005%	0.001%
6 × 8.7 wks	0.002%	0.008%	0.003%	0.003%	0.000%

Table 1: Results of the backward-looking analysis for Worlds MM, LL, LH, HL and HH, averaged over 25 simulation runs.

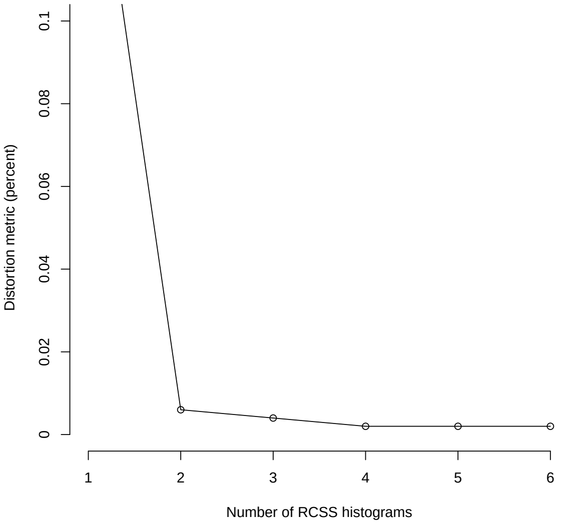

Figure 2: Plot of the distortion metric $D(\mathcal{X}, \mathcal{Y}_h)$ for $h \in \{1, 2, \ldots, 6\}$ for the backward-looking analysis for World MM, averaged over 25 simulation runs.

and is very small at four histograms (0.004%). This indicates that the RCSS configuration with four histograms is appropriate for all 81 worlds. The other columns show other statistics for the distribution of the average distortion metrics (the minimum, the three quartiles, and the maximum). Note that all the quantities follow the same pattern (dropping significantly on adding a histogram for up to four histograms, and stabilizing beyond four histograms) and support the conclusion that using four histograms of 13 weeks each is an

RCSS	$D(\mathcal{X}, \mathcal{Y}_h)$					
configuration	avg.	min.	1^{st} quart.	med.	3^{rd} quart.	max.
1 × 52 wks	0.391%	0.090%	0.166%	0.241%	0.392%	2.701%
2 × 26 wks	0.010%	0.001%	0.003%	0.006%	0.010%	0.080%
3 × 17.3 wks	0.006%	0.001%	0.002%	0.004%	0.007%	0.048%
4 × 13 wks	0.004%	0.000%	0.002%	0.003%	0.005%	0.021%
5 × 10.4 wks	0.003%	0.000%	0.002%	0.003%	0.004%	0.017%
6 × 8.7 wks	0.003%	0.000%	0.001%	0.002%	0.003%	0.010%

Table 2: Results of the backward-looking analysis for 81 worlds.

appropriate RCSS configuration.

The above analysis to determine four histograms as an appropriate substitute for individual-level data is based on a distortion metric that basically compares in-sample performance of RCSS and individual-level data by comparing the in-sample log-likelihood values. We now show that the four-histogram RCSS configuration gives performance comparable to individual-level data on other important metrics as well. We consider measures related to parameter recovery, out-of-sample predictions, and recovery of a forward-looking managerially relevant metric (DET). Note that since we have already determined the four-histogram RCSS configuration as the appropriate one, we present the above metrics only for this RCSS configuration.

A natural test of the performance of the RCSS approach is to compare the associated parameter estimates with those obtained using the individual-level data (i.e., parameter recovery). For each world, we know the original data-generating parameters and, for each of the 25 runs, we know the parameters recovered using individual-level data and the parameters recovered using the four-histogram RCSS configuration. For each parameter we compute the root mean square error (RMSE) across the 25 simulation runs for the world. A small value of RMSE indicates that the parameter recovery is good, with zero denoting perfect recovery. The RMSE values for parameter recovery for Worlds MM, LL, LH, HL and HH are provided in Table 3. These RMSE values for the individual-level analysis and the four-histogram RCSS analysis are comparable, and show that parameter recovery is good for both, though it is, understandably, less accurate for the RCSS case.

To obtain an idea of accuracy in parameter recovery across the 81 worlds, we present the RMSE values for the four parameters averaged across all

	World MM		World LL		World LH		World HL		World HH	
	Indiv	RCSS	Indiv	RCSS	Indiv	RCSS	Indiv	RCSS	Indiv	RCSS
r	0.000	0.032	0.063	0.089	0.032	0.032	0.055	0.118	0.032	0.045
α	0.134	0.173	1.362	1.509	0.623	0.581	0.179	0.338	0.122	0.155
s	0.032	0.045	0.095	0.212	0.077	0.122	0.055	0.084	0.032	0.032
β	0.721	1.288	0.842	1.925	4.514	7.686	0.338	0.619	1.105	1.452

Table 3: RMSE values for Worlds MM, LL, LH, HL and HH from an analysis with individual-level data and four 13-week RCSS histograms, averaged over 25 simulation runs.

81 worlds. For the individual-level analysis, these average RMSE values for r, α, s and β are 0.044, 0.451, 0.061 and 1.293, respectively, while for the four-histogram RCSS configuration, the average RMSE values are 0.069, 0.565, 0.100 and 2.332, respectively. As for the five worlds discussed above, average RMSE values across the 81 worlds are comparable for the individual-level analysis and the four-histogram RCSS analysis, and show that parameter recovery is good for both. However, as may be expected, it is less accurate for the RCSS case. Furthermore, recovery for the parameters of the death process (s and β) is less accurate than recovery for the parameters of the purchasing while alive process (r and α).

A second, more practical, way to understand how close the RCSS performance is to that associated with the individual-level data is to examine the quality of the forecasts created by the model (i.e., out-of-sample performance). Using the data from each simulation run, we construct the true histogram of purchases for the weeks 53–104 holdout period (i.e., one histogram for this 52-week period). We then generate the expected histogram of purchases for the same time period using the individual-level parameters and the RCSS parameters. The out-of-sample performance is based on how close the predicted histograms are to the true histogram. The metric we use to evaluate how closely a predicted histogram matches the true histogram is the standard χ^2 goodness-of-fit test statistic computed using these two histograms. A smaller value of this statistic corresponds to a better match between the true histogram and the predicted histogram, zero denoting a perfect match.[4]

[4] When assessing the in-sample "goodness of fit" of a model, a p-value is usually reported, which depends on the number of parameters used to estimate the model. In this

The χ^2 statistics for Worlds MM, LL, LH, HL and HH are provided in Table 4. Across the 81 worlds, the average χ^2 statistic from the individual-level-data parameters has the value 9.8, and the χ^2 statistic from the four-histogram RCSS configuration parameters has the value 10.7. The values of the χ^2 statistics from the individual-level estimation and the four-histogram RCSS estimation are close to each other; RCSS performance compares favorably to individual-level performance.[5]

	World MM	World LL	World LH	World HL	World HH
Ind χ^2	9.7	6.2	11.4	8.9	8.8
RCSS-4 χ^2	10.0	7.0	12.8	8.8	9.5

Table 4: χ^2 values for Worlds MM, LL, LH, HL and HH from an analysis with individual-level data and four 13-week RCSS histograms, averaged over 25 simulation runs.

Looking beyond one year, we now examine how well the RCSS method compares with individual-level data by comparing estimates of lifetime value, as captured by the Discounted Expected Transactions (DET) measure (defined in (5)) computed using the two sets of parameters. For our DET calculations, we use an annual discount rate of 15%, which corresponds to a continuously compounded rate of $\delta = 0.0027$. The DET values for Worlds MM, LL, LH, HL and HH are provided in Table 5. (The DET number for any given world is obtained by averaging the DET numbers from each of the 25 simulation runs for that world.) We see that the DET values from both individual-level analysis and RCSS analysis with four histograms are very close. Across the 81 worlds, the average percentage deviation in the DET obtained from the individual-level parameters and the four-histogram RCSS parameters is 0.6%. These values show that the four-histogram RCSS configuration provides DET values that are very close to those obtained from

case, however, we are comparing out-of-sample histograms: we report the χ^2 statistic only as a measure of the “match” between the original and predicted out-of-sample histograms, and not as a measure of the “goodness of fit” of the model. No parameters (or “degrees of freedom”) are associated with the holdout period, so it does not make sense to compute p-values here.

[5] Other metrics can also be used for evaluating out-of-sample performance. For instance, using the root mean square error (RMSE) between the predicted and the original out-of-sample histograms yields the same conclusions regarding out-of-sample performance in all cases.

an analysis of individual-level data.

	World MM	World LL	World LH	World HL	World HH
Ind DET	6.788	0.276	3.663	2.455	32.326
RCSS-4 DET	6.741	0.275	3.620	2.455	31.916

Table 5: DET values for Worlds MM, LL, LH, HL and HH from an analysis with individual-level data and four 13-week RCSS histograms, averaged over 25 simulation runs.

This DET comparison is not only very favorable for the RCSS approach but we also see it as the most diagnostic of the various analyses conducted in this section. Inaccuracies in parameter recovery can counterbalance each other, e.g., a high value of the r parameter can be nullified by a similarly high value of the α parameter, if the overall mean of the transaction rate (r/α) is unaffected. Likewise, a year-long forecast period is useful, but it is dominated by the infinite-horizon nature of the DET calculation. Thus, the very close matches seen in Table 5 are a very good indication about the validity of the RCSS estimation approach.

As we complete our detailed investigation of the backward-looking analysis, we offer a final comment on our assertion that four 13-week histograms are generally recommended for the RCSS method. This recommendation is a conservative one; looking across the 81 worlds, there are a number of scenarios in which the three-histogram configuration would be quite satisfactory as well. We used a variety of data-mining procedures to try to discover common characteristics of worlds that tend to support three versus four histograms, but did not come up with anything sufficiently systematic or robust. Thus for clarity and convenience, we stick with our global recommendation of four histograms, but we encourage future researchers to look more carefully for conditions under which alternative configurations may be preferable.

5.3. Forward-Looking Analysis

In the forward-looking analysis, imagine that the analyst receives data in the form of quarterly histograms. The key question is: How many quarterly histograms are needed before one can confidently uncover the story behind the purchasing process of the cohort, if at all? To answer this question, for each of our 81 worlds, we consider one 13-week histogram, two 13-week histograms, three 13-week histograms and four 13-week histograms.[6] As noted earlier, this is quite a different test as compared to the preceding backward-looking analysis, because we are changing the amount of data we use instead of varying the summary period for a fixed dataset. We should be able to better uncover the latent buying and dropout processes with more histograms, but it is possible that we can do this without all four quarters of data — the marginal improvement in model fit after, say, three quarterly histograms might be small.

In order to determine the number of RCSS histograms that adequately serve as a substitute for individual-level data, we analyze the values of the distortion metric $D(\mathcal{X}, \mathcal{Y}_h)$, for $h \in \{1, 2, 3, 4\}$. To illustrate the patterns in the results, we present in Table 6 these values for Worlds MM, LL, LH, HL and HH, as defined earlier, averaged across the 25 simulation runs of each world. (Note that in the forward-looking analysis, we cannot have more than four quarterly histograms for 52 weeks of customer data, while in the backward-looking analysis we could go beyond four histograms by reducing the time interval covered by each histogram.) These numbers show that there is a significant drop in the value of the distortion metric as we increase the number of histograms. By the time we get to four histograms, the value of the distortion metric is very small.

RCSS	$D(\mathcal{X}, \mathcal{Y}_h)$				
configuration	World MM	World LL	World LH	World HL	World HH
1 × 13 wks	1.739%	4.295%	2.727%	0.536%	0.133%
2 × 13 wks	0.015%	0.090%	0.017%	0.007%	0.005%
3 × 13 wks	0.004%	0.041%	0.004%	0.005%	0.001%
4 × 13 wks	0.002%	0.021%	0.002%	0.005%	0.001%

Table 6: Results of the forward-looking analysis for Worlds MM, LL, LH, HL and HH, averaged over 25 simulation runs.

We now consider all the 81 worlds. As in the backward-looking case, in Table 7, we present statistics that provide an idea of the distribution of the distortion metrics across all 81 worlds (where the value used for each world is the average across the 25 runs of that world).

[6] As in the backward-looking analysis, we have eleven "number of purchases" bins $(0, 1, \ldots, 9, 10+)$ for each histogram.

RCSS configuration	$D(\mathcal{X}, \mathcal{Y}_h)$ avg.	min.	1^{st} quart.	med.	3^{rd} quart.	max.
1 × 13 wks	2.771%	0.133%	1.412%	2.733%	3.770%	6.758%
2 × 13 wks	0.020%	0.004%	0.009%	0.015%	0.025%	0.090%
3 × 13 wks	0.007%	0.001%	0.003%	0.005%	0.008%	0.041%
4 × 13 wks	0.004%	0.000%	0.002%	0.003%	0.005%	0.021%

Table 7: Results of the forward-looking analysis for 81 worlds.

We note, once again, that our choice of four histograms is a conservative one, and the three-histogram RCSS configuration would be satisfactory in many cases. Using three quarterly histograms would imply that we can do without even using all four quarters of data. However, to be on the safe side when making any claims, we choose the four-histogram RCSS configuration.

For the other metrics for the four-histogram RCSS configuration (i.e., out-of-sample predictive accuracy, parameter recovery and recovery of managerially relevant metrics), we refer the reader to the previous section on backward-looking analysis. This is because the four-histogram RCSS configuration for the forward-looking analysis is exactly the same as the four-histogram RCSS configuration for the backward-looking analysis, which implies that the values of interest of these metrics, and thus the conclusions that we draw from them, are the same as those discussed in Section 5.2.

6. Validation on a Real Dataset

We now replicate our analysis on a dataset from Bonobos, a popular US online fashion retailer. This dataset has been used in Lee and Bell (2013). Beginning from October 2007, this dataset tracks the purchasing activity of 10,000 customers, starting with each customer's first-ever purchase at Bonobos. To simplify the analysis process, and to maintain consistency with our simulation study, we use exactly 52 weeks of purchase data for each customer (starting from the time of each customer's first-ever purchase with the company).

First, we fit the Pareto/NBD model on the individual-level data. We then proceed to the backward-looking analysis using RCSS data. Given 52 weeks of individual-level data, we construct one 52-week histogram, two 26-week histograms, . . . , and six 8.7-week histograms. For each of these configurations, we estimate the parameters of the Pareto/NBD model, and compute the in-sample log-likelihood for the individual-level data using these parameters. Table 8a shows the values of the distortion metric, $D(\mathcal{X}, \mathcal{Y}_h)$, for the different RCSS configurations. We see that the patterns from the simulations are confirmed for this dataset, and the RCSS configuration with four 13-week histograms is an appropriate choice. (There is a slight hint of degradation as we get to six histograms, suggesting that we are starting to "chop" the data into too many histograms and are therefore losing the meaningful information content from each histogram.)

For the forward-looking analysis, we construct one 13-week histogram, two 13-week histograms, three 13-week histograms, and four 13-week histograms and run the same analysis as above. Table 8b shows the values of the distortion metric, $D(\mathcal{X}, \mathcal{Y}_h)$, for the different RCSS configurations. We again see that the patterns from the simulations are confirmed, and the RCSS configuration with four 13-week histograms is an appropriate choice.

Finally, we show in Table 8c that the parameter estimates associated with the individual-level data are close to those associated with the RCSS data. For a randomly-chosen customer from this cohort, the DET values obtained from the individual-level data and the RCSS data with four 13-week histograms are 3.976 and 3.875, respectively, which implies a difference of only 2.55%.

This brief analysis shows that the patterns we find for in-sample fit and recovery of parameters from the simulation apply quite well to the "real world" Bonobos dataset. This is a strong indication that our simulation results are practical and robust.

7. Comparisons with Samples of Individual-Level Data

One advantage of constructing RCSS data from individual-level data is that it provides scalability when dealing with large datasets. But it is not the only way to make large datasets more manageable. Another (more common) approach is to utilize random samples from the complete dataset for model estimation. In the context of our study, the analyst could randomly sample the full purchase history data for a certain percentage of individuals in the dataset and estimate the model parameters on this smaller dataset. Estimation on the smaller dataset would take less time, while possibly providing performance close to that obtained with the full dataset. In this section, we compare the computational performance of estimation using a sample of the individual-level data versus the four-histogram RCSS configuration.

RCSS configuration	$D(\mathcal{X}, \mathcal{Y}_h)$
1 × 52 weeks	7.78%
2 × 26 weeks	0.27%
3 × 17.3 weeks	0.05%
4 × 13 weeks	0.03%
5 × 10.4 weeks	0.04%
6 × 8.7 weeks	0.10%

(a) Backward-looking analysis

RCSS configuration	$D(\mathcal{X}, \mathcal{Y}_h)$
1 × 13 weeks	63.94%
2 × 13 weeks	1.45%
3 × 13 weeks	0.19%
4 × 13 weeks	0.03%

(b) Forward-looking analysis

	Individual-level	RCSS
		(4 × 13 weeks)
r	0.887	0.753
α	28.784	24.884
s	0.241	0.242
β	2.241	2.222
DET	3.976	3.875

(c) Comparison of parameter estimates and DET

Table 8: Summary of the results of a backward- and forward-looking analysis of the Bonobos dataset.

We describe our analysis using World MM (with parameter values $r = 0.5$, $\alpha = 5$, $s = 0.5$ and $\beta = 5$) as an example. We simulate 52 weeks of data for 25 different cohorts of 100,000 customers. From each cohort of 100,000 customers, we sample individual-level data for subsets of 10,000, 20,000, 30,000, . . . , 100,000 customers (i.e., the last sample uses the full dataset). For each of these samples, we construct RCSS data with four histograms. Next, for each sample, we estimate the model parameters using the individual-level data and the RCSS data. This gives us two sets of parameter estimates for each sample, one from the individual-level data and one from the RCSS data. Using these two sets of parameter values, we calculate the value of log-likelihood expression in (2) using the individual-level data for *all* 100,000 customers; comparing the two log-likelihood values thus obtained for the full dataset but using different sets of parameter values provides a common basis for comparing the performance from the two estimation approaches. Note that the maximum of the log-likelihood expression in (2) obtained from the estimation on the individual-level data for the full cohort of 100,000 customers is the "benchmark" log-likelihood value against which the above log-likelihood values can be compared. We also record the time taken for estimation from the sampled individual-level data and from the RCSS data created from this sample.

In Table 9, we present the results of the analysis for the world under consideration. All values are averaged over 25 runs. The first column gives the sample sizes of the different cohorts, the second and third columns give the times taken for parameter estimation, and the fourth and fifth columns give the full-data log-likelihood values. Note that the log-likelihood value in the last row of the fourth column (i.e., −727,644.0) is the "benchmark" value for comparison, and the closer the other log-likelihood values are to this number, the "better" is the performance of that configuration. We find that the full-data log-likelihood numbers for both the individual-level data and the RCSS data are very close to the benchmark value (with percentage differences of the order of thousandths of one percent or smaller) and approach it as the sample size increases.

In Figure 3, we present the times taken by the estimations using individual-level data and RCSS data for cohorts of different sizes. The plot shows a striking pattern: we find that the time taken for the estimation with individual-level data increases approximately linearly with the size of the sample. On the other hand, the time taken for the estimation with RCSS data stays approximately constant (in fact, slightly decreases) with increasing sample size. Furthermore, the RCSS computation time drops below that of the individual-level data estimation time between a sample size of 10,000 and 20,000 customers.

We note in Table 9 that, for each sample size, the full-data likelihood evaluated using the individual-level data parameter estimates is slightly smaller than that associated with the RCSS parameter estimates. However, the full-data likelihood for the parameters estimated on the full-data RCSS sum-

	Estimation time (seconds)		Full-data LL	
Sample size	Individual	RCSS	Individual	RCSS
$10,000$	2.02	2.66	$-727,661.4$	$-727,672.1$
$20,000$	3.22	2.21	$-727,652.0$	$-727,662.7$
$30,000$	4.09	2.17	$-727,648.3$	$-727,653.0$
$40,000$	5.63	2.18	$-727,647.1$	$-727,653.0$
$50,000$	6.54	2.18	$-727,645.9$	$-727,650.8$
$60,000$	7.14	2.24	$-727,645.4$	$-727,648.7$
$70,000$	8.39	2.03	$-727,644.8$	$-727,647.4$
$80,000$	10.83	1.89	$-727,644.6$	$-727,647.6$
$90,000$	10.73	2.10	$-727,644.2$	$-727,647.2$
$100,000$	11.60	2.08	$-727,644.0$	$-727,646.7$

Table 9: Comparison of model performance with individual-level and RCSS data, with data samples of different sizes for the case where $r = 0.5, \alpha = 5, s = 0.5, \beta = 5$, averaged over 25 simulation runs.

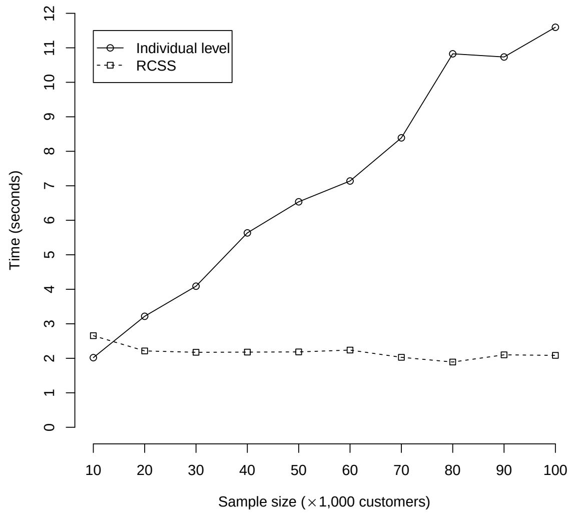

Figure 3: Plot of the time taken for estimation with individual-level data and RCSS data for data samples of different sizes, averaged over 25 simulation runs.

maries (i.e., all 100,000 individuals) is smaller than that associated with the individual-level data parameters estimated off a sample of 40,000 or fewer individuals, while taking less than half the time of the individual-level data sample of 40,000. Thus it makes sense to estimate the model parameters using full-data RCSS summaries than to estimate them using reasonably large samples of individual-level data.

We find that the patterns uncovered above extend to all 81 worlds. In the simulations for the other words, the time taken for individual-level estimation increased linearly with sample size, and the time taken for RCSS estimation was less than the time taken for individual-level estimation for 56 out of 81 worlds with a sample of 20,000 customers, and for 79 out of 81 worlds with a sample of 30,000 customers. In all cases, the full-data log-likelihood values, calculated as described above, were very close (with percentage differences of the order of thousandths of one percent or smaller). This analysis yields an interesting conclusion — individual-level data and RCSS data both perform well with sampling, but RCSS data provides a time advantage for larger samples.

8. Discussion

With advances in their IS capabilities, firms are able to finely track and record details on their customers' transactions with them. However, firms are getting inundated with the huge amounts of data generated, and are therefore struggling to extract useful insights about their customers from these data. A number of researchers have developed customer-base analysis models that perform very well given detailed individual-level customer data. In this paper, we explore the possibility of estimating these models using data summaries. We consider a simple data structure for recording the transaction activity of a cohort of customers, which we call repeated cross-sectional summary (RCSS) data. The RCSS data structure is an easy-to-manage, scalable, privacy-preserving format. Given RCSS data, we ask: Under what conditions, if at all, will a customer-base analysis model that has been shown to work well on individual-level data also work well with RCSS data? We focus on the Pareto/NBD model and answer this question based on a comprehensive simulation study covering a broad spectrum of market scenarios characterized by various levels of customer-base penetration and mean purchase frequency. Our results consistently establish that, for both backward-looking analysis and forward-looking analysis, the model fit (and parameter values) associated with the use of RCSS data can closely match the corresponding estimates associated with individual-level data. Beyond proving the viability of the RCSS data structure when using the Pareto/NBD model, we believe we have “raised the bar” for other analysts (and policy makers) who are searching for “best practices” in data storage that can leverage commercially available datasets while also being scalable and privacy-preserving.

The results we present in this paper are for a specific configuration (52 weeks for calibration and 52 weeks for holdout), which we have chosen to represent a typical situation faced by firms (i.e., given one year of data, forecast sales for the coming year). However, we expect our results to hold for other configurations with different time periods, as long as metrics such as penetration and purchase frequency (for that time period) fall in the range of those associated with the simulation. If customers purchase at a faster rate than considered in our simulations, then even shorter time periods should work. However, if customers purchase at a slower rate than considered in our simulations, a longer time period would be required. We provide a method, rooted in rate-distortion theory, for determining the optimal number of RCSS histograms in scenarios are not addressed by our analysis.

We use the Pareto/NBD model because it is has been used widely with great success for customer-base analysis in noncontractual settings using individual-level data. The aim of this paper, however, is not to study the capabilities of the Pareto/NBD model; rather it is to showcase the capabilities and advantages of the RCSS data structure (when the right kind of model is used). For this reason, we have characterized our simulated “worlds” using metrics such as penetration and purchase frequency, which are independent of the assumptions of the Pareto/NBD model. The success of the approach on the Bonobos dataset lends further support to this claim. We would expect our results to hold for similar customer-base analysis models, such as the BG/NBD model (Fader et al., 2005a) and the PDO model (Jerath et al., 2011).

As previously discussed, the use of the RCSS data structure has three attractive properties relative to the use of the raw customer-level data typically required by statistical models for customer-base analysis. First, the data summaries are easy to create and distribute. The analyst does not require access to the transaction database, and the process of creating the histograms is not too onerous a task for any IS group. Second, it is highly scalable — irrespective of the cohort’s size or level of activity, only a few histograms are required to summarize its buying behavior. Finally, the aggregated form of the data is such that there is no threat of customer privacy being compromised.

Beyond these reasons, the RCSS data format offers some additional advantages as well. First, the cross-sectional summaries need only be *repeated*, and not necessarily of equal length. For instance, for a 52-week period, the analyst could have one histogram for weeks 1 to 13 (13 weeks long), another for weeks 14 to 33 (20 weeks long) and another for weeks 34 to 52 (19 weeks long). Given these RCSS data, the underlying parameters could still be estimated in the same exact manner as described earlier in the paper. While we would expect it to work well in most cases, there might be some limits as to how short the period for a histogram can be. For instance, a 5-week period might be too short to have a sufficient number of customers who make one or more purchases, making the resulting histogram not sufficiently informative to help in parameter estimation.

The flexibility to use histograms of different lengths might offer an opportunity to fine-tune (and potentially shorten) the model estimation process. Recognizing the expected slowdown in purchasing (associated with the "buy till you die" perspective) suggests that early histograms will be relatively dense compared to later ones, so it may be possible to shorten the time window associated with them. There could be several potential improvements here, which we leave to future research.

Second, the cross-sectional summaries need not be immediately adjacent to one another. For instance, given a 52-week period, we might have one histogram for weeks 1 to 13, another for weeks 14 to 26, and another for weeks 40 to 52, while the histogram for weeks 27 to 39 might be missing. Our estimation procedure could still give accurate estimates of the underlying parameters. In contrast, such a scenario would make it very difficult to estimate the parameters of the Pareto/NBD model using the standard likelihood function (i.e., equation (2)) as it requires complete recency and frequency data. (Other customer-base analysis models would suffer from this same problem as well.) A comprehensive simulation study could inform us about how robust the estimation approach is to different types of missing histograms.

Finally, to construct the histograms for each period, the exact count of purchases is not required — the percentage of customers in each bin is sufficient. In other words, to calibrate the model, we only need the percentage of customers in the cohort making $0, 1, 2, \ldots$ repeat purchases in each period. Hence, if an analyst has a sufficiently accurate idea of these percentages and is able to construct the (approximately correct) histograms for (say) three quarters, the latent purchase and death characteristics of the customer base can be estimated, and future activity predicted.

The use of the RCSS data structure does not necessarily imply that firms have to give up on targeted marketing. Summary data for the entire cohort can be used for model estimation, but once the parameters are obtained, they can be applied to any purchase history (real or hypothetical) to obtain conditional expectations, CLV, and other forward-looking metrics commonly associated with customer-base analysis. For example, we could use the parameters estimated using the RCSS data to compute CLV over a recency×frequency grid, from which a set of rules could be derived for scoring the original transaction database.

Furthermore, we do not have to use a single set of histograms for the whole cohort; early-activity indicators can be used to construct segments of customers, the behavior of which is then summarized using segment-specific repeated cross-sectional summaries. The model can then be used to gain further insights into the behavior of the members of each segment. For instance, Mason (2003) provides RCSS data for two cohorts, where one cohort is composed of customers who had made a first purchase of less than $50 and the other is composed of customers who had made a first purchase of $50 or more. The analysis of these data presented in Fader et al. (2007) shows how the Pareto/NBD model can be used to identify the underlying factors that lie behind observed differences in expected CLV (e.g., differences in expected lifetime and purchasing while alive).

To conclude, we have established that, for customer-base analysis applications, RCSS data can work just as well as individual-level data in a wide variety of market scenarios. We have also laid out several promising research opportunities to be pursued in the future. However before incorporating these extensions, we encourage researchers and practitioners to contemplate the basic RCSS data structure and to begin to take advantage of its practical benefits for customer-base analysis.

Appendix: Derivation of DET$(r, \alpha, s, \beta, \delta)$

Our objective is to derive the expression for DET (discounted expected transactions) when buyer behavior is characterized by the Pareto/NBD model. This sees us specifying $t(t)$, $S(t)$, and $d(t)$ and solving the following integral:

$$\int_0^\infty E[t(t)]S(t)d(t)dt\,.$$

Conditional on λ and μ, $E[t(t)] = \lambda$ and $S(t) = e^{-\mu t}$. Since we are operating in continuous time, $d(t) = e^{-\delta t}$ where an annual discount rate of $(100 \times d)\%$ is equivalent to a continuously compounded rate of $\delta = \ln(1 + d)$. (If the data are recorded in time units such that there are k periods per year ($k = 52$ if the data are recorded in weekly units of time) then the relevant continuously compounded rate is $\delta = \ln(1 + d)/k$.) Therefore,

$$\begin{aligned}\mathrm{DET}(\lambda, \mu, \delta) &= \int_0^\infty \lambda e^{-\mu t} e^{-\delta t} dt \\ &= \frac{\lambda}{\mu + \delta}\,.\end{aligned}$$

To obtain the expression for DET for a randomly chosen customer, we remove the conditioning on λ and μ

$$\begin{aligned}\mathrm{DET}(r, \alpha, s, \beta, \delta) &= \int_0^\infty \int_0^\infty \mathrm{DET}(\lambda, \mu, \delta) g(\lambda|r, \alpha) g(\mu|s, \beta) d\lambda d\mu \\ &= \frac{r}{\alpha}\frac{\beta^s}{\Gamma(s)} \int_0^\infty \frac{\mu^{s-1}}{\mu + \lambda} e^{-\beta\mu} d\mu\end{aligned}$$

letting $z = \mu/\delta$ (which implies that $\mu = \delta z$ and $d\mu = \delta dz$)

$$= \frac{r}{\alpha}\frac{\beta^s \delta^{s-1}}{\Gamma(s)} \int_0^\infty \frac{z^{s-1}}{1 + z} e^{-\beta\delta z} dz$$

which, noting the integral representation of the confluent hypergeometric function of the second kind,

$$= \frac{r}{\alpha}\beta^s \delta^{s-1} \Psi(s, s; \beta\delta).$$

Given Kummer's transformation $\Psi(a, b; z) = z^{1-b}\Psi(a - b + 1, 2 - b; z)$, we can rewrite this as

$$\mathrm{DET}(r, \alpha, s, \beta, \delta) = \frac{r}{\alpha}\beta\Psi(1, 2 - s; \beta\delta).$$

References

Abe, Makoto (2009), “"Counting Your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model,” *Marketing Science*, **28** (May–June), 541–553.

Aghelvi, B.B. and F. Mehran (1981), “Optimal Grouping of Income Distribution Data,” *Journal of the American Statistical Association*, **76** (March), 22–26.

Agrawal, Rakesh and Ramakrishnan Srikant (2000), “Privacy-Preserving Data Mining,” *SIGMOD ’00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 439–450.

Balasubramanian, S., S. Gupta, W. Kamakura and M. Wedel (1998), “Modeling Large Data Sets in Marketing,” *Statistica Neerlandica*, **52** (3), 303–323.

Batislam, Meltem Denizel and Alpay Filiztekin (2007), “Empirical Validation and Comparison of Models for Customer Base Analysis,” *International Journal of Research in Marketing*, **24** (September), 201–209.

Carey, Peter (2009), *Data Protection: A Practical Guide to UK and EU Law*, 3rd edn, Oxford: Oxford University Press.

Connor, Robert J. (1972), “Grouping for Testing Trends in Categorical Data,” *Journal of the American Statistical Association*, **67** (September), 601–604.

Cover, Thomas M. and Joy A. Thomas (2006), *Elements of Information Theory*, 2nd edn, Hoboken, NJ: John Wiley & Sons, Inc.

Cox, D.R. (1957), “Note on Grouping,” *Journal of the American Statistical Association*, **52** (December), 543–547.

Cox, Kelsey (2013), “Infographic: The Cost of Too Much Data,” *Marketing Tech Blog*, May 31. http://www.marketingtechblog.com/infographic-the-cost-of-too-much-data/

Davies, J.B. and A.F. Shorrocks (1989), “Optimal Grouping of Income and Wealth Data,” *Journal of Econometrics*, **42** (1), 97–108.

DuMouchel, William (2002), “Data Squashing: Constructing Summary Data Sets”, in J. Abello et al. (eds.), *Handbook of Massive Data Sets*, Norwell, MA: Kluwer Academic Publishers, 579–591.

Fader, Peter S. and Bruce G. S. Hardie (2005), “A Note on Deriving the Pareto/NBD Model and Related Expressions.” http://brucehardie.com/notes/009/

Fader, Peter S. and Bruce G. S. Hardie (2010), “Implementing the Pareto/NBD Model Given Interval-Censored Data.” http://brucehardie.com/notes/011/

Fader, Peter S., Bruce G. S. Hardie, and Kinshuk Jerath (2006), “Deriving an Expression for $P(X(t, t + \tau) = x)$ Under the Pareto/NBD Model.” http://brucehardie.com/notes/013/

Fader, Peter S., Bruce G. S. Hardie, and Kinshuk Jerath (2007), “Estimating CLV Using Aggregated Data: The *Tuscan Lifestyles* Case Revisited,” *Journal of Interactive Marketing*, **21** (Summer), 55–71.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005a), “"Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model,” *Marketing Science*, **24** (Spring), 275–284.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005b), “RFM and CLV: Using Iso-value Curves for Customer Base Analysis,” *Journal of Marketing Research*, **42** (November), 415–430.

Fader, Peter S., Bruce G. S. Hardie, and Subrata Sen (2014), “Stochastic Models of Buyer Behavior,” in *The History of Marketing Science*, Russell S. Winer and Scott A. Neslin (eds.), Singapore: World Scientific Publishing, 165–205.

Fader, Peter S., Bruce G. S. Hardie, and Jen Shang (2010), “Customer-Base Analysis in a Discrete-Time Noncontractual Setting,” *Marketing Science*, **29** (November–December), 1086–1108.

Gastwirth, Joseph L. and Abba M. Krieger (1975), “On Bounding Moments from Grouped Data,” *Journal of the American Statistical Association*, **70** (June), 468–471.

Gopal, Ram, Robert Garfinkel and Paulo Goes (2002), "Confidentiality via Camouflage: The CVC Approach to Disclosure Limitation When Answering Queries to Databases," *Operations Research*, **50** (3), 501–516.

Fung, Benjamin C., Ke Wang, Rui Chen, and Philip S. Yu (2010), "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys*, **42** (4), Article No. 14.

Han, Jiawei and Micheline Kamber (2006), *Data Mining: Concepts and Techniques*, San Francisco, CA: Morgan Kaufmann.

Hopmann, Jörg and Anke Thede (2005), "Applicability of Customer Churn Forecasts in a Non-Contractual Setting," in Daniel Baier and Klaus-Dieter Wernecke (eds.), *Innovations in Classification, Data Science, and Information Systems* (Proceedings of the 27th Annual Conference of the Gesellschaft für Klassifikation e.V., Brandenburg University of Technology, Cottbus, March 12–14, 2003), Berlin: Springer-Verlag, 330–337.

Huang, Chun-Yao (2012), "To Model, or Not to Model: Forecasting for Customer Prioritization," *International Journal of Forecasting*, **28** (April–June), 497–506.

Jerath, Kinshuk, Peter S. Fader and Bruce G. S. Hardie (2011), "New Perspectives on Customer "Death" Using a Generalization of the Pareto/NBD Model," *Marketing Science*, **30** (5), 866–880.

Keller, Sallie A., Steven E. Koonin and Stephanie Shipp (2012), "News," *Significance*, **9** (4), 2–3.

Kettenring, Jon R. (2009), "Massive Datasets," *Wiley Interdisciplinary Reviews: Computational Statistics*, **1** (1), 25–32.

Krieger, Abba M. and Joseph L. Gastwirth (1984), "Interpolation from Grouped Data for Unimodal Densities," *Econometrica*, **52** (2), 419–426.

Lee, Jae Y. and David R. Bell (2013), "Neighborhood Social Capital and Social Learning for Experience Attributes of Products," *Marketing Science*, **32** (November–December), 960–976.

Li, Xiao-Bai and Sumit Sarkar (2006a), "Privacy Protection in Data Mining: A Perturbation Approach for Categorical Data," *Information Systems Research*, **17** (3), 254–270.

Li, Xiao-Bai and Sumit Sarkar (2006b), “Privacy Protection in Data Mining: A Tree-Based Data Perturbation Approach,” *IEEE Transactions on Data and Knowledge Engineering*, **18** (9), 1278–1283.

Mason, Charlotte H. (2003), “Tuscan Lifestyles: Assessing Customer Lifetime Value,” *Journal of Interactive Marketing*, **17** (Autumn), 54–60.

Malin, Bradley and Latanya Sweeney (2004), “How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Anonymity Protection Systems,” *Journal of Biomedical Informatics*, **37** (3), 179–192.

Menon, Syam and Sumit Sarkar (2007), “Minimizing Information Loss and Preserving Privacy,” *Management Science*, **53** (1), 102–116.

Mielikäinen, T. (2004), “Privacy Problems with Anonymized Transaction Databases,” in S. Arikawa and E. Suzuki (eds.), *Discovery Science: Proceedings of the 7th International Conference (DS 2004)*, Lecture Notes in Computer Science, 3245, Berlin: Springer, 219–229.

Morrison, Donald G. , Richard D. H. Chen, Sandra L. Karpis, and Kathryn E. A. Britney (1982), “Modelling Retail Customer Behavior at Merrill Lynch,” *Marketing Science*, **1**, (Spring), 123–141.

Narayanan, Arvind and Vitaly Shmatikov (2008), “Robust De-anonymization of Large Datasets,” *IEEE Symposium on Security and Privacy (SP2008)*, 111–125.

Parmigiani, Giovanni (1998), “Designing Observation Times for Interval Censored Data,” *Sankhyā: The Indian Journal of Statistics*, **60** (A-3), 446–458.

Reinartz, Werner 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** (October), 17–35.

Reinartz, Werner and V. Kumar (2003), “The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration,” *Journal of Marketing* **67** (January), 77–99.

Sawiris, Milad (2000), “Optimum Grouping and the Boundary Problem,” *Journal of Applied Statistics*, **27** (3), 363–371.

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.

Shaw, Dale G., Michael D. Huffman and Mark G. Haviland (1987), “Grouping Continuous Data in Discrete Intervals: Information Loss and Recovery,” *Journal of Educational Measurement*, *24* (2), 167–173.

Singh, Siddharth S., Sharad Borle, and Dipak C. Jain (2009), “A Generalized Framework for Estimating Customer Lifetime Value When Customer Lifetimes are Not Observed,” *Quantitative Marketing and Economics*, **7** (June), 181–205.

Singleton, Susan (2006), *Tolley’s Data Protection Handbook*, 4th edn, Croyden, Surrey: Lexis- Nexis UK.

Tryfos, Peter (1985), “On the Optimal Choice of Sizes,” *Operations Research*, **33** (3), 678–684.

WA (2008), “Home-grown analytics,” Discussion thread on *The Web Analytics Forum.* http://tech.groups.yahoo.com/group/webanalytics/message/16509.

Weil, Andrew (2011), “Why ‘Data Smog’ May Be Making You Depressed,” *Time: Ideas*, November 14. http://ideas.time.com/2011/11/14/why-data-smog-may-be-making-you-depressed/.

Whitler, Kimberly (2012), “The CEO/CMO Dilemma: So Much Data, So Little Impact,” *Forbes*, July 18. http://www.forbes.com/sites/kimberlywhitler/2012/07/18/the-ceocmo-dilemma-so-much-data-so-little-impact/

Wu, Couchen and Hsiu-Li Chen (2000), “Counting Your Customers: Compounding Customer’s In-store Decisions, Interpurchase Time, and Repurchasing Behavior,” *European Journal of Operational Research*, **127** (1) 109–119.

Wübben, Markus and Florian v. Wangenheim (2008). "Instant Customer Base Analysis: Managerial Heuristics Often "Get It Right"," *Journal of Marketing*, **72** (May), 82–93.

Zheng, Zhiqiang, Balaji Padmanabhan and Steven O. Kimbrough (2003), "On the Existence and Significance of Data Preprocessing Biases in Web-Usage Mining," *INFORMS Journal on Computing*, **15** (2), 148–170.