Segmenting Customers from Population to Individuals: Does 1-to-1 Keep Your Customers Forever?

Tianyi Jiang and Alexander Tuzhilin

Abstract—There have been various claims made in the marketing community about the benefits of 1-to-1 marketing versus traditional customer segmentation approaches and how much they can improve understanding of customer behavior. However, few rigorous studies exist that systematically compare these approaches. In this paper, we conducted such a study and compared the predictive performance of aggregate, segmentation, and 1-to-1 marketing approaches across a broad range of experimental settings, such as multiple segmentation levels, multiple real-world marketing data sets, multiple dependent variables, different types of classifiers, different segmentation techniques, and different predictive measures. Our experiments show that both 1-to-1 and segmentation approaches significantly outperform aggregate modeling. Reaffirming anecdotal evidence of the benefits of 1-to-1 marketing, our experiments show that the 1-to-1 approach also dominates the segmentation approach for the frequently transacting customers. However, our experiments also show that segmentation models taken at the best granularity levels dominate 1-to-1 models when modeling customers with little transactional data using effective clustering methods. In addition, the peak performance of segmentation models are reached at the finest granularity levels, skewed towards the 1-to-1 case. This finding adds support for the microsegmentation approach and suggests that 1-to-1 marketing may not always be the best solution.

Index Terms—Personalization, clustering, 1-to-1 marketing, segmentation, microsegmentation.

1 Introduction

Customer segmentation, such as customer grouping by the level of family income, education, or any other demographic variable, has been considered as one of the standard techniques used by marketers for a long time [20]. Its popularity comes from the fact that segmented models usually outperform aggregated models of customer behavior [3], [6]. More recently, there has been much interest in the marketing and data mining communities in learning *individual* models of customer behavior within the context of 1-to-1 marketing [28] and personalization [2], [7], when models of customer behavior are learned from the data pertaining only to a particular customer. Although there have been many claims made about the benefits of 1-to-1 marketing [28], including that 1-to-1 helps to retain your customers forever [29], there have been few scientific studies systematically comparing individual and segmented models of customer behavior in the marketing and the personalization literature. We review this related work in Section 2.

It is a nontrivial problem to do these types of comparisons because of the trade-off between the sparsity of data for individual customer models and customer heterogeneity in aggregate models: Individual models may suffer from sparse data resulting in high variance of performance measures of predictive models [14], while aggregate models suffer from high levels of customer heterogeneity, resulting in high prediction biases [14]. Depending on which effect dominates the other, it is possible that models of individual customers dominate the segmented or aggregated models, and vice versa.

In this paper, we address this issue and provide a systematic empirical study across a broad spectrum of experimental settings in which we compare the performance of individual, aggregate, and segmented models of customer behavior, where customer models are learned from the transactional data pertaining to individual customers, the whole customer base, and customer segments, respectively. Our experiments show that individual-level models statistically outperform aggregate models of customer behavior, even for sparse data having only a few transactions per customer. Our experiments also show that for the highly transacting customers or poor segmentation techniques, individual-level customer models outperform segmentation models of customer behavior. These two results reaffirm the anecdotal evidence about the advantages of personalization and the 1-to-1 marketing stipulated in the popular press [4], [28]. However, our experiments also show that segmentation models, taken at the best granularity level(s) and generated under effective clustering methods, dominate individual-level customer models when modeling customers with little transactional data. Moreover, this best granularity level is significantly skewed towards the 1-to-1 case and is usually achieved at the finest segmentation levels. This finding provides additional support for the case of *microsegmentation* [20], [24]—when customer segmentation is done at a very "fine-grained"

• *The authors are with the Department of Information, Operations, and Management Sciences, Stern School of Business, New York University, 44 West 4th Street, Room 8-185 New York, NY 10012. E-mail: {tjiang, atuzhili}@stern.nyu.edu.*

Manuscript received 15 Sept. 2005; revised 8 Apr. 2006; accepted 24 May 2006; published online 18 Aug. 2006.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0377-0905.

level, such as grouping all the undergraduate computer science majors living in the university dorms. The virtues of microsegmentation have been known to marketers and CRM practitioners [24], such as Capital One Bank, that practices precision marketing campaigns targeted at ever finer customer segments [31]. In this paper, we support this anecdotal evidence with a systematic study across multiple data sets, methods, and experimental conditions. Finally, we show that some of the popular clustering techniques could lead to poor performance results for the predictive models built on the generated segments and that these performance results are comparable to that of the random clustering. This finding stresses the importance of selecting good segmentation methods for producing better predictive models of customer behavior. Consequently, we examine why these popular clustering methods produce poor segmentation results, and present modifications to these algorithms that improve their performance in the context of our segmentation problem.

2 Problem Formulation

We model customer behavior by building predictive data mining models either at the individual, segment, or aggregate level; each model tries to predict different aspects of customer behavior, such as the total price of purchase or the time of the day of the purchase. These predictive models are learned from the data sets tracking transactional histories of individual customers, such as online or in-store groceries purchasing transactions, over time. This type of cross-sectional time series data is termed panelist data within the marketing and economic research community [20].

More formally, let C be the customer base consisting of N customers, each customer C_i is defined by the set of m demographic attributes $A = \{A_1, A_2, \ldots, A_m\}$, k_i transactions $Trans(C_i) = \{TR_{i1}, TR_{i2}, \ldots, TR_{ik_i}\}$ performed by customer C_i, where transaction TR_{ij} is defined by its schema $T = \{T_1, T_2, \ldots T_p\}$; and, the set of transaction values $\{t_{ij_1}, t_{ij_2}, \ldots t_{ij_p}\}$, each value t_{ij_q} corresponding to attribute T_q of schema T. Moreover, we have h summary statistics $S_i = \{S_{i1}, S_{i2}, \ldots, S_{ih_i}\}$ for customer C_i, each S_{ij}, defined as a statistics [25] on some of the attributes in T across the transactions $Trans(C_i)$. For example, a customer C_i can be defined by attributes

$$A = \{\text{Name, Age, Income, and other demographic attributes}\},$$

by the set of purchasing transactions $Trans(C_i)$ she made at a Web site, each transaction defined by such transactional attributes T as an item being purchased, when it was purchased, and the price of an item. Finally, summary statistics S_i can be computed for all of customer C_i's purchasing sessions and can include such statistics as the average amount of purchase, the average number of items bought, and the average time spent per purchase session. Given this data, we learn predictive models of customer behavior of the form

$$Y = \hat{f}(X_1, X_2, \ldots, X_p), \quad (1)$$

where $X_1, X_2, \ldots, X_p$ are some of the demographic attributes from A and some of the transactional attributes from T. Function $\hat{f}$ is a predictive model learned via different types of machine learning classifiers from the transactional and demographic data described above, as will be explained below.

Various models of customer behavior can be built at different levels of analysis as customers can be grouped into different *segments* based on some of their demographic and behavioral characteristics. Moreover, we can have different levels of analysis depending on how finely we want to partition the customer base into various segments. In this paper, we consider the following three levels of analysis:

- *Aggregate level*—when the unit of analysis is the *whole* customer base, and only *one* predictive model of customer behavior (1) is built for the whole customer base. This model is learned from *all* the transactional and demographics data of all the customers contained in the data set.
- *Segmentation level*—when "similar" customers are grouped into *segments*, and the model(s) of customer behavior are built for *each* segment based on the transactions and the demographic data of *that* particular grouping of customers. In this case, we still use the model of (1) but learn it from the data pertaining *only* to the selected segment of customers and do this for each customer segment. We group customers into segments using different clustering methods to be described below.
- *Individual (or 1-to-1) level*—when the unit of analysis is an individual customer, the model of customer behavior is built based *only* on the purchase transactions of that particular customer. In other words, we build a model of type (1) for *each* customer C_i in the customer base C using the transactional history $Trans(C_i)$ of that customer. Such customer-specific models capture idiosyncrasies of the purchase behavior of individual customers.

As we refine models from the aggregate to the segmented and then to the individual models of customer behavior, as described above, we would create increasingly more "homogeneous" customer groups for which predictive models have a reduced bias and therefore should be more accurate. However, as we consider progressively more refined segments containing fewer customers, we decrease the sample space of individual transactions resulting in an increased variance of the estimation of function $\hat{f}$ in (1). This trade-off between the specificity of the model resulting in reduced bias versus the sparsity of the data resulting in higher variance leads to the following general research problem: *determine which level of analysis would provide better prediction of customer behavior*, as defined by some measure of predictive performance of models of type (1). To solve this problem, we can compare predictive models across the three levels of customer partitioning and address the questions of which of the following modeling approaches is better:

1. aggregate versus segmentation modeling,
2. segmentation versus individual modeling, and
3. aggregate versus individual modeling.

Question 1 has been extensively studied by marketers who show that segmentation models that account for customer heterogeneity usually outperform aggregated models [3], [6]. In particular, the traditional segmentation modeling approach used by marketers is a fixed-effects approach that estimates individual customer level parameters without regard to particular probability distributions of heterogeneity [9], [37]. Marketers have also studied various individual and segmentation mixture models such as the random-effects models which stochastically pool data across customers so that individual customer level parameters can be estimated from a mixed distribution of all customer data [15]. These segmentation approaches allow marketers to differentiate product offerings for different market segments and identify niche segments and target appropriate products for them. While the individual level model can be viewed as a specialized case of the segmented level modeling, where the group of customers in each segment is just an individual customer, past marketing research has not focused on the individual level modeling, where each model is learned from the data pertaining *only* to this particular customer. This is the case because of the sparseness of individual-level data, the computational expense of evaluating exact posterior distribution of individual effects [3], and because existing segmentation methods, such as fixed-effects approaches, break for the 1-to-1 case since they only yield aggregate summaries of heterogeneity [3], hence calling for new analysis methods.

Although comparison questions 2 and 3 have not been extensively studied before, certain prior work is related to these questions, including the work of some of the marketing researchers. In particular, most of the marketing literature compares segmentation and individual modeling approaches in terms of the discrete and continuous mixture distribution methods of modeling customer heterogeneity, where continuous distributions estimate individual parameters of customer models [3], [34], [35]. Usually, these individual parameters are estimated using Hierarchical Bayes (HB) and Markov Chain Monte Carlo (MCMC) simulation methods [34]. Moreover, it was shown that these continuous distributions computed using HB and MCMC methods tend to outperform the mixture model approaches [3], [22], thus providing some evidence of the advantages of individual customer modeling. However, despite all this progress, [35] points out that "we need more simulation—and more empirical studies—to fully understand the strengths and weaknesses of the several methods for estimating individual-level parameters." Moreover, such empirical studies can be applied not only to the mixture models and the HB method estimating individual-level parameters but also to a broad range of other statistical and data mining methods of modeling customer behavior, such as decision trees, Naive Bayesian methods and Support Vector Machines [14], with the purpose of comparing segmented and individual models of customers. Such empirical studies are even more crucial for these data mining models since it is even more difficult to compare segmentation and individual approaches analytically because of the nonlinearity of these models and the complexity of the performance measures. Finally, although the HB approach estimates individual parameters of customer models, it pools transactional data across different customers to estimate these parameters, which is different from the 1-to-1 method that utilizes only the customer data when building an individual model.

The problem of building individual and segmented models of customer behavior is also related to the work on user modeling and customer profiling in data mining. In particular, customer profiles can be built in terms of simple factual information represented as a vector or as a set of attributes. For example, in [27], a user profile is defined as a vector of weights for a set of certain keywords. Customer profiles can be defined not only as sets of attributes but also as sets of rules defining behavior of the customer [1], sets of sequences such as sequences of Web browsing activities [14], [26], [33] and signatures, used to capture the evolving behavior learned from data streams of transactions [11]. There has also been some work done on modeling personalized customer behavior by building appropriate probabilistic models of customers. For example, [8] builds customer profiles using finite mixture models and [23] uses maximum entropy and Markov mixture models for generating probabilistic models of customer behavior. However, all these approaches focus on the task of building good profiles and models of customers and do not compare the performance of individual versus segmented and versus aggregate models of customer behavior.

Questions 2 and 3 are also related to the work on clustering that partitions the customer base and their transactional histories into homogeneous clusters of customers [38]. In this paper, we use the clustering method for the very same purpose, but we also go beyond this partitioning and compare the performance of aggregated versus segmented and versus individual models of customer behavior. Also, in his keynote address at the KDD 2000 conference, Papadimitriou reviewed mass customization and remarked that, although the field has certain advantages, it may be worse off than segmentation if uncertainty is above a certain threshold. In this paper, we systematically study this general remark made by Papadimitriou. Last, the problem of building local versus global models in data mining and statistics [5], [12], [14] is also related to questions 2 and 3. Rather than building one global aggregated model of customer behavior, it is often better to build several local models that would produce better performance results. Furthermore, this method can be carried to the extreme when a local model is built for *each* customer, resulting in 1-to-1 customer modeling. In this paper, we follow this local approach and compare the performance of aggregate, segmented, and individual models of customer behavior.

In summary, although some of the prior work on personalization and 1-to-1 marketing reviewed in this section is related to questions 2 and 3, little prior work directly addressed these two questions and provided definitive answers to them. In this paper, we address questions 2 and 3 by conducting empirical studies that are described below.

3 Experimental Setup

To answer questions 2 and 3 stated in Section 2, we build individual, segment-based, and aggregate predictive models of type (1) across the experimental settings (dimensions) of:

1. types of data sets,
2. types of customers (high versus low-volume),
3. types of predictive models (classifiers),
4. dependent variables,
5. performance measures, and
6. segmentation techniques (clustering algorithms).

We explain each of these dimensions below:

3.1 Types of Data Sets

In our study, we worked with the following data sets:

- Two popular real-world marketing data sets containing panel data[1] of online browsing and purchasing activities of Web site visitors and panel data on beverage purchasing activities of "brick-and-mortar" stores. The specific characteristics of each data set are provided below:
  - The first data set contains ComScore data from Media Metrix on Internet browsing and buying behaviors of 100,000 users across the United States for a period of six months (available via Wharton Research Data Services at http://wrds.wharton.upenn.edu/).
  - The second data set contains Nielsen panelist data on the beverage shopping behavior of 1,566 families for the years 1992 and 1993.

  The ComScore and Nielsen marketing data sets are very different in terms of the type of purchase transactions (Internet versus physical purchases), variety of product purchases, number of individual families covered, and the variety of demographics. Compared to Nielsen's beverage purchases in local supermarkets, the ComScore data set covers a much wider range of products and demographics and is more representative of today's large marketing data sets.
- Two synthetic data sets representing two types of customers defined in item 2 below: high-volume customers (*Syn-High*) having many transactions and low-volume customers (*Syn-Low*) having few transactions. Within each data set, customer differences are defined by generating a different customer summary statistic vector S_i for each customer i. All subsequent customer purchase data can then be generated from the set of summary statistic vectors S_i.

  The Syn-Low and Syn-High data sets were generated as follows: First, we generated 2,048 unique summary statistics vectors of length 12,

$$\{\mu_1, \sigma_1, skew_1, \mu_2, \sigma_2, skew_2, \ldots, \mu_4, \sigma_4, skew_4\},$$

  representing four sets of three summary statistics measures (mean, standard deviation, and skewness) on four transactional variables,

$$\{TR_1, TR_2, TR_3, TR_4\}.$$

  The mean and skewness parameters were generated from a Gaussian distribution with values ranging from -1 to 1, and the variance parameters were generated from a Gaussian distribution with values ranging from 0 to 1. Each synthetic customer, C_i, has one summary statistics vector, S_i, which is used to generate customer C_i's purchase data, $Trans(C_i)$. The transaction generation process works as follows: For each transaction in $Trans(C_i)$, the values of the four element tuple, $\{tr_1, tr_2, tr_3, tr_4\}$, are generated from a Gaussian distribution function with its corresponding mean, standard deviation, and skewness parameters specified in S_i. Another set of 2,048 unique demographic vectors of length 11, $\{A_1, A_2, \ldots, A_{11}\}$, was generated and appended to each customer's generated transactional tuples. Each customer in the Syn-High data set has 100 purchase transactions and each customer in the Syn-Low data set has 10 purchase transactions.
- Two pseudosynthetic data sets representing high-volume customers (*PSyn-High*) and low-volume customers (*PSyn-Low*), respectively, where, unlike the previous two sets of synthetic data sets, 2,048 unique customer summary statistics were generated by sampling from ComScore customer summary statistics distributions, which is then used to generate the purchase transactions with four transactional variables. The number of transactions per customer is also determined from ComScore customer transaction distributions. This data set is used to better simulate real-world transactional data sets.

3.2 Types of Customers

In order to study the effect of data sparsity, we partitioned our ComScore and Nielsen data sets into high and low-volume customers, where "volume" is defined by the frequency of transactions performed by a customer. Ideally, we would like to experiment across the entire customer population for both ComScore and Nielsen data sets, but the sheer size of the ComScore data set and our computational requirements across all dimensions of analysis make this task impossible to accomplish. Thus, we sorted all the customers from the ComScore and Nielsen data sets by their transactional volumes and selected the top and bottom 5 percent of the customers from the ComScore data set and the top and bottom 10 percent of the customers from the Nielsen data set for our experiments.

In order to perform 10-fold crossvalidations of our various predictive models, we restricted our low-volume customers to have at least 10 transactions. We note that this minimum amount of transactions per customer does reduce the variance of individual level models and may impact our findings when comparing individual level models against aggregate and best segment models for low-volume

1. Panel data [20], also called longitudinal or crosssectional times series data, when used in the context of marketing means that the data about a preselected group of consumers on whom a comprehensive set of demographic information is collected is also augmented with the complete set of their purchases. Panel data provides a comprehensive view of purchasing activites of a preselected set of consumers.

TABLE 1
Customer Types, Independent Variables, and Transaction Counts

DataSet	Customer Type	Independent Variables	% of Total Population	Families	Total Transactions	Average Transactions Per Household
ComScore	High	EducationLevel,CensusRegion, HouseHoldSize,OldestAge, HouseHoldIncome,HasChild, Ethenicity,BroadBand,Country	5%	2,230	137,157	62
ComScore	Low		5%	2,230	24,344	11
Nielsen	High	ShopperAge,StoreType, CouponValue,AdType, AdDisplayLocation,HouseHoldSize, Ethenicity,HasChild,Income, HasKitchen, NumberOfTV AgeOfHeadOfHouseHold	10%	156	28,985	186
Nielsen	Low		10%	156	5,007	32
Nielsen	All		100%	1,566	132,210	84
Syn-High	High	11 Generated Demographic Variables	100%	2,048	204,800	100
Syn-Low	Low		100%	2,048	20,480	10
PSyn-High	High		100%	2,048	133,120	65
PSyn-Low	Low		100%	2,048	20,480	10

customers. However, as we argue in Sections 4.2 and 5.2, this restriction does not affect most of our findings. Moreover, few customers in ComScore and Nielsen data sets have fewer than 10 transactions. Therefore, this restriction does not significantly affect our samples of customers. This observation can also be generalized to several other high-frequency applications, such as some of the banking, credit card, and other financial services applications.

In summary, we created nine data sets in total: high and low-volume customers for ComScore; high, low, and all-volume customers for Nielsen; and high and low-volume synthetic and pseudosynthetic data sets. Table 1 shows the key characteristics of the nine data sets used in our studies, including all the independent variables used in each data set. Note that the average transaction frequencies are different among the low-volume customers of different data sets; this will play a role in our results below.

3.3 Types of Predictive Models

We use four different types of classifiers for building predictive models: C4.5 decision tree [30], Naïve Bayes [18], rule-based RIPPER [10], and nearest neighbor-based NNge [32]. The four classifiers are chosen because they represent different and popular approaches to predictive model building, and they are fast to generate. We generated a total of 2,170,344 unique predictive models across all dimensions of analysis in this study, and the amount of computational effort to build a classifier is a practical concern. Computational time constraint is also a critical factor behind our decision to not use other high-performance classifiers, such as support vector machines and neural networks. Since our goal in this research is to compare the relative performance of individual, segmentation-based, and aggregate models of customer behavior, it is crucial to do this comparison under the same set of experimental settings. Hence, it is less important to select the best possible classifiers for the study since their performance comparisons are done not with each other, but across individual, segmentation-based, and aggregate models.

3.4 Dependent Variables

We selected different dependent variables from the transaction variables when building predictive models in order to avoid variable-specific effects when comparing the discussed approaches across different experimental settings. In particular, as dependent variables in our models we used: 1) eight ComScore transactional attributes: Internet purchase session duration, number of Web page viewed, time of the day, day of the week, category of the Web site, product category, product price, and basket total price; 2) five Nielsen transactional attributes: category of drinks bought, primary shopper's gender, day of the week, quantity of drinks bought, and total price; and 3) four dependent variables for the synthetic and pseudosynthetic data sets that were generated from synthetic customer summary statistic vectors.

The independent variables used in these models include customer demographic (as listed in Table 1) and all other transactional variables besides the one selected as the dependent variable.

3.5 Performance Measures

We used Weka 3.4, from the University of Waikato [36], for all predictive modeling tasks. Each classifier generates a model via 10-fold cross validation. The predictive power of the model is then evaluated via three performance measures: percentage of correctly classified instances, root mean squared error, and relative absolute error which are defined as [36]:

- Correctly classified instances (CCI) $= \frac{\sum_{1}^{n} TP_i}{n}$, where

$$\left\{ \begin{array}{ll} TP_i = 1 & if\ p_i = a_i \\ TP_i = 0 & if\ \ p_i \neq a_i \end{array} \right\},$$

p_i is the predicted value, a_i is the actual value, and n is the total number of observations in a customer segment.

- $$\text{Root Mean-Squared Error (RME)} = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}.$$

- Relative Absolute Error (RAE) $= \frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \bar{a}| + \ldots + |a_p - \bar{a}|}$, where $\bar{a}$ is the average value of predicted class.

3.6 Segmentation Techniques

We segment the customer base using hierarchical clustering techniques.[2] In particular, we generate progressively smaller groupings of customers via five levels of segment/subsegment hierarchy. With the exception of random clustering, all clustering algorithms try to find similarity among customers from customer summary statistics variables S and demographics variables A. To split jth customers group g_{lj} in subsegment level l, we input into a clustering algorithm the set of summary statistics and demographic information XC_{lj} for all customers C_i in group g_{lj}:

$$\mathrm{XC}_{lj} = \sigma_{C_i \in g_{lj}}(S) \bowtie \sigma_{C_i \in g_{lj}}(A),$$

where σ denotes the select operator and $\bowtie$ the join operator. For example, XC_{32} defines a set of customer summary statistics vectors along with their respective demographic vectors for customer group 2 in level 3 of the segment hierarchy. For each customer i in group 2, we would have C_i's summary statistic vector, S_i, such as mean, standard deviation, and distribution skewness of purchase price and C_i's demographic vector, A_i, such as education level, household income, number of children, etc. To generate the fourth level of subsegments for customers within group 2, we would then use a clustering algorithm to further split data points contained within XC_{32}. Random clustering, where the customer base gets segmented into random groups regardless of customer "similarity," is used as the control group. Predictive models of customer behavior based on random clustering of customers are used as a benchmark to measure the effectiveness of a particular clustering technique. A segmentation technique is considered "poorly behaved" if the resultant performance measures are statistically equivalent to or worse than those of random clustering of customers. On the other hand, a segmentation technique is considered "well-behaved" if the resultant performance measures are statistically better than those of random clustering of customers.

For each new level $l+1$ in the segment/subsegment relationship in a hierarchical clustering, we created k new customer groups at level $l+1$ from a single customer group g_{lj} at level l of the hierarchy. The branching factor k is used to control the granularity of clusters. If k is set too high, we would approach near individual level clustering, as we create increasingly smaller customer groupings at subsequent subsegment levels. Due to different data set sizes, we used branching factors of $k = 3$ for ComScore and synthetic data sets, and $k = 2$ for Nielsen data sets.

We used the following clustering methods to create different segmentations of the customer base (methods 2, 4, and 5 below are supported by Weka and are described in [36]):

1. *Random Clustering (Random)*—To create k new groups on subsegment level $l+1$ from a set of customers in group j on subsegment level l, the probability of customer C_i belonging to a new group g_j out of possible k new groups in subsegment level $l+1$ is $1/\mathrm{k}$ and is the same for all g_js.
2. *SimpleKMeans (SMean)*—k local minimum cluster centers in the XC instance space are chosen via a random start iterative approximation strategy. Completely different clusters centers can be returned due to the initial random cluster center selections [36].
3. *SMean_Mod*—Modified version of *SMean* in that we rerun the *SMean* algorithm multiple times with random starting seeds to generate more balanced clusters in terms of customer group sizes. This method is used as a smoothing mechanism in order to avoid highly skewed cluster sizes as described in Section 5.2.
4. *FarthestFirst (FFirst)* [16]—A greedy k-center algorithm that is guaranteed to produce clustering results within a constant factor of two of the optimum.
5. *Expectation Maximization (EM)*—An iterative approach to approximate the cluster probabilities and distribution parameters that converges to a local maximum.
6. *EM_Mod*—Similar to *SMean_Mod*, this is a modified version of *EM* in that we rerun the *EM* algorithm multiple times with random starting seeds to generate more balanced clusters in terms of customer group sizes.

In summary, we have described six dimensions of our experimental settings in this section: different types of data sets, customers, predictive models, dependent variables, performance measures, and segmentation techniques. In the rest of this paper, we describe how predictive models of customer behavior vary across these six dimensions and report our findings.

4 Comparing Individual versus Aggregate Levels of Customer Modeling

In this section, we compare individual versus aggregate levels of customer modeling. More specifically, we compare the predictive accuracy of function (1) estimated from the transactional data $TRANS(C_i)$ for all the *individual* customer models and compare its performance with the performance of function (1) estimated from the transactional data for the *whole* customer base. In particular, we explore the aforementioned trade-off between the hetero-

2. Note that we could have used other clustering methods, such as semisupervised clustering methods based on partitioning data on a selected attribute (variable). However, because our objective is to compare predictive performances of aggregate, segmented, and individual models of customers across a wide range of experimental conditions (optimal, as well as suboptimal), rather than determine the best settings that maximize the predictive performance of a particular type of a model, we have chosen the unsupervised hierarchical clustering approach for consistent analysis across different data sets.

geneity of the customer base affecting model bias and the sparsity of data affecting model variance.

4.1 Experimental Setup

As a first step, we discretized Nielsen, ComScore, and the two different types of synthetic data in order to build predictive models with the four types of classifiers described in Section 3.3. We discretized the continuous-valued dependent attributes, such as price and Internet browsing durations, based on entropy measures via our implementation of Fayyad's recursive minimal entropy partitioning algorithm [13], resulting in roughly equal representation in sample data to avoid overly optimistic classification due to highly skewed class priors.

To determine whether individual modeling performs statistically better than aggregate level modeling, we use a variant of the nonparametric Mann-Whitney rank test [25] to test whether the accuracy score of the one aggregate model is statistically different from a random variable with a distribution generated from individual accuracy results of the individual level models. The null Hypothesis I for each of the performance measures of CCI, RME, and RAE is then:

- H_0: The aggregate level performance measure *is not* different from the set of individual level performance measures.
- H_1+: The aggregate level performance measure *is* different from the set of individual level performance measures in the *positive* direction.
- H_1-: The aggregate level performance measure *is* different from the set of individual level performance measures in the *negative* direction.

To illustrate what we have done, consider the following example.

Example. Consider the aggregate NaiveBayes model α for predicting the day of the week of a purchasing transaction for the set of 156 Nielsen low-volume families (customers), learned from all their demographic and purchasing data via 10-fold crossvalidation. When doing 10-fold crossvalidation generating model α, we use three performance measures for α: CCI_α, RME_α, and RAE_α.

To compare the performance of the aggregated model α against individual level models, we generate 156 separate NaiveBayes models β_i for each of the 156 low-volume families predicting the day of the week family C_i would do shopping. Let $\beta = \{\beta_1, \beta_2, \ldots, \beta_{156}\}$ be the set of these models. As explained before, for each model β_i in β, we compute three performance measures CCI_i, RME_i, and RAE_i.

Let CCI_β, RME_β, and RAE_β be three random variables having distributions corresponding to the sets $\{CCI_i\}_{i=1,\ldots,156}$, $\{RME_i\}_{i=1,\ldots,156}$, and $\{RAE_i\}_{i=1,\ldots,156}$, respectively. Then, to test for H_0 (I) for the performance measure CCI, we would compare CCI_α against CCI_β and determine whether CCI_α is statistically different from CCI_β using a variant of the Mann-Whitney rank test mentioned earlier.

TABLE 2
Aggregate versus Individual Level Customer Models for Hypothesis Test (I)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	96	0	42
ComScore	Low	96	0	7
Nielsen	High	60	0	7
Nielsen	Low	60	0	3
Nielsen	All	60	0	7
Syn-High	High	48	0	26
Syn-Low	Low	48	0	5
PSyn-High	High	48	0	23
PSyn-Low	Low	48	0	4

(Numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0.)

The above scenario is repeated for all customer type data sets listed in Table 1, across eight ComScore transactional variables, five Nielsen transactional variables, four generated synthetic and pseudosynthetic transactional variables, respectively, listed in Section 2, four different classifiers, and three different performance measures.

4.2 Results

Table 2 lists the number of statistical tests that rejects the null hypothesis (I) at 95 percent significance level for all customer type data sets.[3] From Table 2, we can conclude that:

- None of the statistical tests accepts H_1+, which means that the performance measures at the aggregate level is *never* greater than that of the individual level for all the tests.
- The number of significant results drops as we go from the high-volume customer data set to the low-volume data set: from 42 to 7 for ComScore, 7 to 3 for Nielsen, 26 to 5 for the synthetic data, and 23 to 4 for the pseudosynthetic data.
- ComScore data, which has 10 times more families in each customer type data set as compared to Nielsen data, has the highest number of significant results in the high-volume data set and the greatest discrepancies between the high and low-volume data sets; there are also significant discrepancies between the high and low-volume synthetic and pseudosynthetic data.

We can conclude from the analysis summarized in Table 2 that *for high-volume customers, modeling customer behavior at the individual level yields significantly better results than for the aggregate case*. Moreover, we can also conclude that modeling low-volume customers at the individual level will not be worse off than for the aggregate case. However, we note that as the minimum number of customer specific transactions decreases from $k = 10$ transactions per customer to a significantly smaller number, it is entirely possible for the aggregate model to outperform individual level

3. Note that by counting the number of statistically significant distribution tests on results generated from 10-fold crossvalidation, we do not perform undesirable multiple comparison procedures [17] because we are comparing distributions of performance scores rather than choosing the highest score for statistical significance testing.

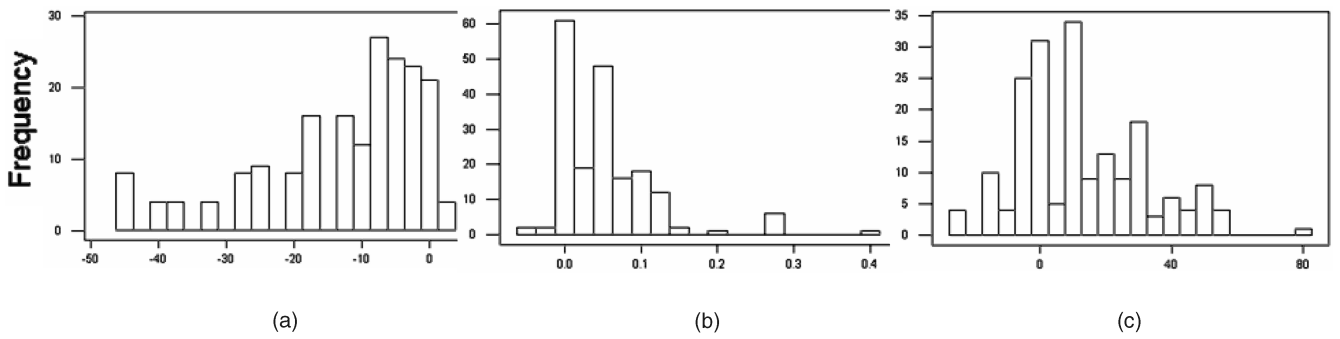


Fig. 1. Median difference distributions of aggregate vs. individual models. (a) Aggregate versus individual CCI. (b) Aggregate versus individual RME. (c) Aggregate versus individual RAE.

models since the variance of the models increases with the smaller values of k.

To get a sense of the magnitude of the dominance that individual level models have over aggregate models, we computed the difference of each performance measure between aggregate models and the medians of each performance measure distribution across the individual level models for a particular combination of data set, dependent variable, classifier, and performance measure. We repeated this process for all the 564 comparisons across the nine data sets, 47 total dependent variables, four classifiers, and three performance measures, and plotted out the histograms of the median differences for the CCI, RME, and RAE measures in Fig. 1, respectively.

The negative values for the CCI measure and positive values for the RME and RAE measures in Fig. 1 show that individual models significantly outperform the aggregate model across most of the experimental conditions, thus providing additional visual evidence and the quantitative extent of the dominance of individual models over aggregate models that was already statistically demonstrated with the Mann-Whitney tests.

5 Comparing Individual versus Segmentation versus Aggregate Levels of Customer Modeling

In this section, we compare individual versus segmentation and aggregate versus segmentation levels of customer modeling. More specifically, we compare the predictive accuracy of (1) estimated from the transactional data $TRANS(C_i)$ for the segmentation level models with the performance results for individual models obtained in Section 4.

As explained in Section 3, we generate progressively finer customer subsegment levels using different hierarchical clustering techniques. In our studies, we generated five levels of subsegments for the six hierarchical clustering methods across different experimental settings, where the number of customer groups in each subsegment level is determined by a branching factor k.

As was also explained in Section 3, the factors that influence the prediction accuracies of different subsegment levels include the quality of segmentation, the levels of refinements, data sparsity (i.e., model variance), and customer heterogeneity (i.e., model bias). We examine these factors now.

5.1 Segmenting Customer Base Using Clustering Methods

Once we determined the new groupings of our customers within each of the five subsegment levels, we generate predictive models for each of the groups as described in Section 3.

To compare the clustering algorithms against aggregate and individual level models along each of the three performance measures, we first determine the *best performing segmentation level* for a clustering algorithm among the five segmentation sublevels described in Section 5. In particular,

$$\text{Best Segment Level}_{\text{CCI}} = \arg \max_{l=1,\ldots,5} (CCI_l), \tag{2}$$

$$\text{Best Segment Level}_{\text{RME}} = \arg \min_{l=1,\ldots,5} (RME_l), \tag{3}$$

$$\text{Best Segment Level}_{\text{RAE}} = \arg \min_{l=1,\ldots,5} (RA_l), \tag{4}$$

where these best segment levels are determined from pairwise Mann-Whitney rank test comparisons of corresponding performance measure distributions generated from predictive models of various customer groupings across all five levels for a specific combination of dependent variable, classifier, data set, customer type, and clustering algorithm.

Then, for each of the three performance measures, we compare aggregate model to this best segment level determined by (2), (3), and (4) in the same manner as we compared aggregate versus individual models in Section 4 (by comparing the individual performance measures from the aggregate model against the distribution of performance measures of individual segments). The null Hypothesis II for comparing the best clustering level for each clustering algorithm against the aggregate model becomes:

- H_0: The aggregate level performance measure *is not* different from the set of best segment level performance measures.
- H_1+: The aggregate level performance measure *is* different from the set of best segment level performance measures in the *positive* direction.

TABLE 3
Aggregate versus Best Segment Level Models
Hypothesis Test (II)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	576	0	260
ComScore	Low	576	3	214
Nielsen	High	360	12	87
Nielsen	Low	360	27	60
Nielsen	All	360	7	90
Syn-High	High	288	0	178
Syn-Low	Low	288	0	134
PSyn-High	High	288	0	151
PSyn-Low	Low	288	5	123

(Numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0.)

- H_1- : The aggregate level performance measure *is* different from the set of best segment level performance measures in the *negative* direction.

To compare best segment level against individual models across the three performance measures, we again use the Mann-Whitney rank test as our statistical comparator [25] because of the nonnormal distribution of performance measures and of the different sample sizes across segment levels. The null Hypothesis III for comparing the best segment level for each clustering algorithm against individual level models then becomes:

- H_0: The distribution of the individual model performance measure *is not* different from that of the best segment level model.
- H1+: The distribution of the individual model performance measure *is* different from that of the best segment level model in the *positive* direction.
- H1−: The distribution of the individual model performance measure *is* different from that of the best segment level model in the *negative* direction.

Hypotheses II and III do performance comparisons for each clustering algorithm. However, it is also useful to do these comparisons only for the "good" clustering algorithms. One measure of such goodness, adopted in this paper, is whether a clustering algorithm outperforms Random clustering. Therefore, we need to compare performance of our clustering algorithms SMean, FFirst, EM, SMean_Mod, and EM_Mod against Random clustering across various experimental settings. The null Hypothesis IV for comparing best segment level for each clustering algorithm against random clustering is:

- H_0: The distribution of the performance measure from the best segment level generated under Random clustering *is not* different from the best segment level generated under the clustering algorithm.
- H1+: The distribution of the performance measure from the best segment level generated under Random clustering *is* different from the best segment level generated under the clustering algorithm in the *positive* direction.
- H1−: The distribution of the performance measure from the best segment level generated under Random clustering *is* different from the best segment level generated under the clustering algorithm in the *negative* direction.

5.2 Results

Table 3 lists the number of statistical tests that reject the null hypothesis (II) at 95 percent significance level for all customer type data sets. Similar to our analysis for aggregate versus individual level models, there are 96 statistical comparisons for each ComScore data clustering scheme, which gives us a total of 576 comparisons aggregated across all six clustering algorithms. Likewise, the 60 statistical comparisons for each Nielsen data clustering scheme and 48 statistical comparisons for each synthetic and pseudosynthetic data clustering scheme give us 360 and 288 comparisons aggregated across all six clustering algorithms, respectively.

From Table 3, we can draw the following conclusions, consistent with prior marketing results [3], [6]:

- Best Segment Level significantly dominates aggregate level models across all customer types.
- There is a small number of instances where the aggregate level models performed better than best segment level models across the Nielsen data sets. We will see below that this occurred because some of the clustering algorithms resulted in relatively poor performance.

Similar to Fig. 1, we plotted in Fig. 2 the performance differences of aggregate models against the median of

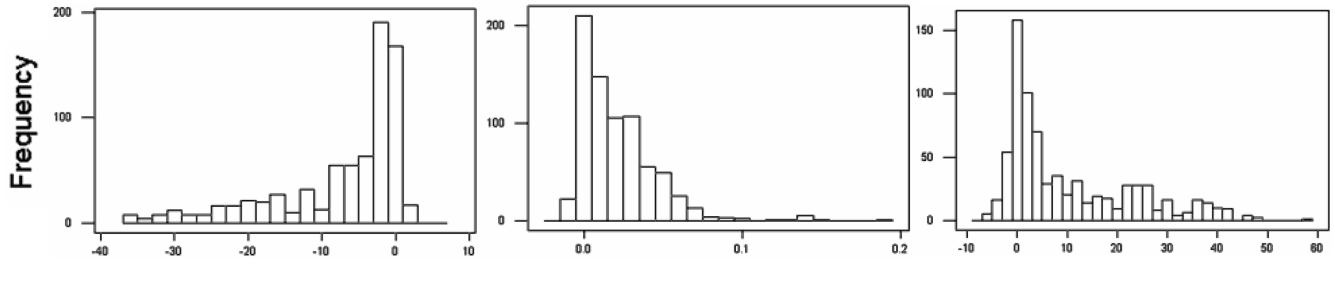


Fig. 2. Median difference distributions of aggregate versus best segment level models. (a) Aggregate versus Best Level CCI. (b) Aggregate versus Best Level RME. (c) Aggregate versus Best Level RAE.

TABLE 4
Individual versus Best Segment Level Models for Hypothesis Test (III)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	576	394	57
ComScore	Low	576	235	134
Nielsen	High	360	228	125
Nielsen	Low	360	201	128
Nielsen	All	360	237	99
Syn-High	High	288	203	19
Syn-Low	Low	288	151	87
PSyn-High	High	288	198	30
PSyn-Low	Low	288	111	65

(Numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0.)

performance measure distributions of the Best Segment Level models across all three different performance measures. As in the case of Fig. 1, the median difference for the CCI measure is skewed to the left and for the RME and RAE measures to the right from 0. As in the case of aggregate versus individual models, this skewness in Fig. 2 clearly demonstrates that the Best Segment Level models dominate the aggregate models.

When comparing individual versus best segment approaches, Table 4 lists the number of statistical tests that reject the null hypothesis (III) described in Section 5.1. From Table 4, we see that individual level models significantly dominate best segment level models across all data sets. This is somewhat surprising since we expected that the effects of data sparsity (i.e., model variance) to dominate the effect of customer heterogeneity (i.e., model bias) for the individual models of *low-volume* customers. This expectation is based on that there may not be enough transactions per customer to build adequate customer models. However, our results for ComScore, synthetic, and pseudosynthetic data show that the discrepancy in significance counts between the individual level models and best segment models decreases from high to low-volume customers (e.g., 394 and 57 for row 1 versus 235 and 134 for row 2 in Table 4). We will see later that this is indeed due to the data sparsity effect (i.e., model variance) among the low-volume customer data sets.

The above result reaffirms one of the main tenets of the 1-to-1 marketing approach—that it pays to treat each customer individually—popularized in various popular press publications, including [28]. However, this is not *always* true: In some cases, such as when the number of individual transactions is low, as argued in Section 4.2, it pays to build segmented models of customers, as we will show below.

The above analysis was done across different clustering algorithms. As explained earlier, we then study "good" clustering methods that outperform Random clustering by testing Hypothesis (IV) described in Section 5.1. In particular, we compare performance measure distributions for the best segment levels taken from each clustering algorithm, data set, customer type (high versus low), classifier, and predictive variable. Table 5 lists the number of statistical tests that reject the null hypothesis (IV) at 95 percent significance level for various data sets, customer types and clustering algorithms. These tests are made across different classifiers, predictive variables, and performance measures. For example, the first row in Table 5 lists 96 tests, where the number 96 represents the number of statistic comparisons made between EM and Random clusterings' $Best\_Segment\_Level_{CCI}$, $Best\_Segment\_Level_{RME}$, and $Best\_Segment\_Level_{RAE}$ distributions across four classifiers and eight dependence variables for ComScore high-volume data set. As discussed in Section 5.1, each of the Best_Segment_Level is independently determined for each performance measure, classifier, and dependent variable combination.

We can draw the following conclusions from these tests:

- Random clustering dominates clustering algorithms SMean and EM across most of the experimental settings.
- FFirst, SMean_Mod, and EM_Mod performed better than Random clustering for the high-volume customer data sets.
- FFirst, SMean_Mod, and EM_Mod clustering algorithm performed the best among all unsupervised distance-based clustering algorithms used in our study.

The surprising result that such well-known clustering algorithms as SMean and EM do not outperform Random clustering in our experiments can be explained as follows: As will be discussed in Sections 5.3 and 5.4, the peak performance of all the clustering algorithms is usually achieved at the customer segmentation levels of high granularity (representing *microsegments*—see Fig. 5). Due to the splitting process, Random clustering achieves even grouping of customers at these high levels of granularity, with customer group sizes varying from 2 to 17 customers for Nielsen, Table 5. Random versus other clustering algorithms (CA) for Hypothesis Test (IV) (numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0) ComScore and the two types of synthetic data sets at the end of the splitting process. This means that at the finest segment level, Random clustering becomes a close approximation of the individual (1-to-1) modeling approach, which explains the good performance of Random clustering. In contrast to Random clustering, SMean and EM fail to generate balanced customer groups at the finest granularity levels because of the random start and unsupervised nature of SMean and EM algorithms. In particular, SMean and EM often find large clusters in the early splitting stages of the hierarchical segmentation process and subsequently fail to split these large clusters when going from one segmentation level to the next. This results in customer groups with wildly different sizes at the finest granularity level of subsegmentation under SMean and EM clustering algorithms. As demonstrated through the cluster size distribution diagrams for ComScore low-volume customers in Fig. 3, at the finest segment level, Random clustering achieves the smallest cluster groups while SMean and EM clustering result in highly skewed cluster distributions. Note that these distributions are typical for the four clustering methods.

TABLE 5
Randon versus Other Clustering Algorithms (CA) for Hypothesis Test (IV)

DataSet	Customer Type	CA	Tests	H_1+	H_1-
ComScore	High	EM	96	40	7
ComScore	Low	EM	96	15	3
Nielsen	High	EM	60	11	4
Nielsen	Low	EM	60	9	0
Nielsen	All	EM	60	20	4
Syn-High	High	EM	48	11	13
Syn-Low	Low	EM	48	7	8
PSyn-High	High	EM	48	20	12
PSyn-Low	Low	EM	48	22	8
ComScore	High	EM_Mod	96	15	36
ComScore	Low	EM_Mod	96	21	5
Nielsen	High	EM_Mod	60	9	11
Nielsen	Low	EM_Mod	60	19	5
Nielsen	All	EM_Mod	60	23	21
Syn-High	High	EM_Mod	48	9	20
Syn-Low	Low	EM_Mod	48	4	8
PSyn-High	High	EM_Mod	48	9	15
PSyn-Low	Low	EM_Mod	48	8	4
ComScore	High	FFirst	96	7	40
ComScore	Low	FFirst	96	20	3
Nielsen	High	FFirst	60	0	8
Nielsen	Low	FFirst	60	20	4
Nielsen	All	FFirst	60	28	13
Syn-High	High	FFirst	48	8	17
Syn-Low	Low	FFirst	48	3	9
PSyn-High	High	FFirst	48	6	20
PSyn-Low	Low	FFirst	48	11	5
ComScore	High	SMean	96	47	8
ComScore	Low	SMean	96	17	4
Nielsen	High	SMean	60	0	5
Nielsen	Low	SMean	60	0	3
Nielsen	All	SMean	60	19	16
Syn-High	High	SMean	48	15	8
Syn-Low	Low	SMean	48	7	3
PSyn-High	High	SMean	48	13	9
PSyn-Low	Low	SMean	48	9	4
ComScore	High	SMean_Mod	96	13	33
ComScore	Low	SMean_Mod	96	16	11
Nielsen	High	SMean_Mod	60	5	7
Nielsen	Low	SMean_Mod	60	17	13
Nielsen	All	SMean_Mod	60	29	16
Syn-High	High	SMean_Mod	48	7	19
Syn-Low	Low	SMean_Mod	48	1	11
PSyn-High	High	SMean_Mod	48	5	16
PSyn-Low	Low	SMean_Mod	48	4	11

(Numbers in column H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0.)

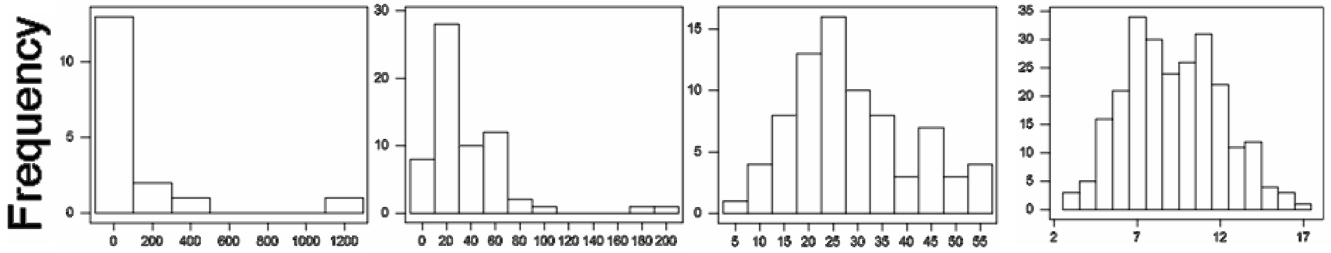


Fig. 3. Segment level 5 cluster size distributions across (a) SMean, (b) EM, (c) FFirst, and (d) Random hierarchical clustering.

TABLE 6
Individual versus Best FFirst, SMean_Mod, and EM_Mod Segment Level Models for Hypothesis Test (III)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	288	135	35
ComScore	Low	288	5	151
Nielsen	High	180	17	3
Nielsen	Low	180	27	4
Nielsen	All	180	57	21
Syn-High	High	144	61	9
Syn-Low	Low	144	33	64
PSyn-High	High	144	88	14
PSyn-Low	Low	144	3	93

(Numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0.)

Because we have found earlier that individual models generally dominate segmentation models, it becomes no surprise that the best segment level under Random clustering generally dominates the best segment levels for the SMean and EM clustering algorithms as the Random clustering method approximates individual models at the finest granularity levels. In contrast to SMean and EM, the FFirst algorithm makes good splits in the early stages of the hierarchical segmentation process and, hence, produces customer groups of more uniformly distributed sizes, as shown in Fig. 3c. We note that SMean_Mod and EM_Mod also performed well against that of Random clustering as both methods use multiple random seeds to generate more balanced customer groupings during every split decision. This results in better performance of FFirst, SMean_Mod, and EM_Mod over that of SMean and EM methods and in generally better results than Random clustering, especially for high-volume customer data sets, as shown in Table 5.

Because FFirst, SMean_Mod, and EM_Mod algorithms usually outperform the other three clustering algorithms, including Random clustering, we selected these three as the best hierarchical clustering algorithms for comparing individual versus best segmentation methods for null hypothesis (III). Table 6 presents the results of this comparison across all data sets under the FFirst, SMean_Mod, and EM_Mod clustering schemes *only*. Here, we note a clear reversal of performances for low-volume ComScore and synthetic customers in comparison to the results presented in Table 4 taken across *all four* clustering methods. In summary, the results from Table 6 state that, while individual level models dominate the best segment level models for the *high-volume* customers, the best segment level models clearly outperform individual level for the *low-volume* customers, which is in line with the earlier discussion on the trade-off between customer heterogeneity and data sparsity when building customer segmentation models.

We would also like to point out that this reversal of performance dominance among low-volume customers, as shown in Table 6, was not evident with the Nielsen data sets. This is the case because customer types within the Nielsen data sets are not as distinctive as those in ComScore and synthetic data sets: As shown in Table 1, low-volume customers in the Nielsen data set conducted an average of 32 transactions per customer versus 11 and 10 transactions per low-volume ComScore and synthetic customer, respectively. Thus, the fact that we do not see a similar reversal of relative performance for Nielsen low-volume customers is attributed to the fact that the Nielsen low-volume customer data set is not really that sparse, and, therefore, individual level models would still outperform segmentation level models, as is the case of high-volume customers (see Tables 4 and 6). We note that this finding of segmentation level models outperform individual models for low-volume customers is generalizable to data sets where individual customer transactions are much lower than our experimental limit of 10 minimal transactions per customer. Therefore, for truly sparse data sets, where we have a very high level of variance among the individual level models, segmentation models will clearly dominate.

In summary, our results show that, while individual level modeling is appropriate for most customer transaction data sets, there are still situations where customer behavior could be better modeled at the best segment levels—the so-called "microsegments" [20], using a good segmentation approach.

5.3 Performance Curves

In Sections 4 and 5, we compared performance of aggregate, segmented, and individual models of customers across different experimental conditions and drew several conclu-

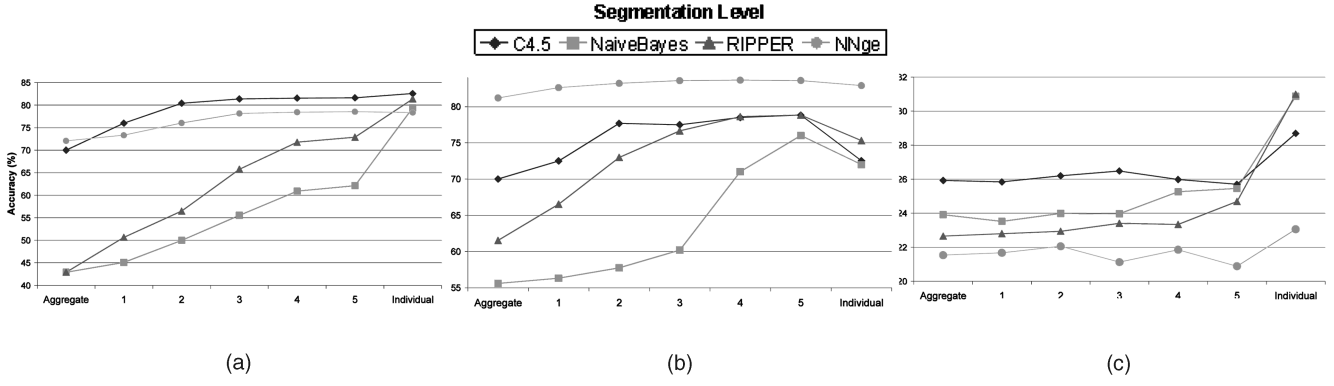


Fig. 4. (a) ComScore high-volume customer, SMean clustering on predicated product category. (b) ComScore low-volume customer, EM clustering on expected shopping duration. (c) Nielsen low-volume customer, FFirst clustering on day of the week.

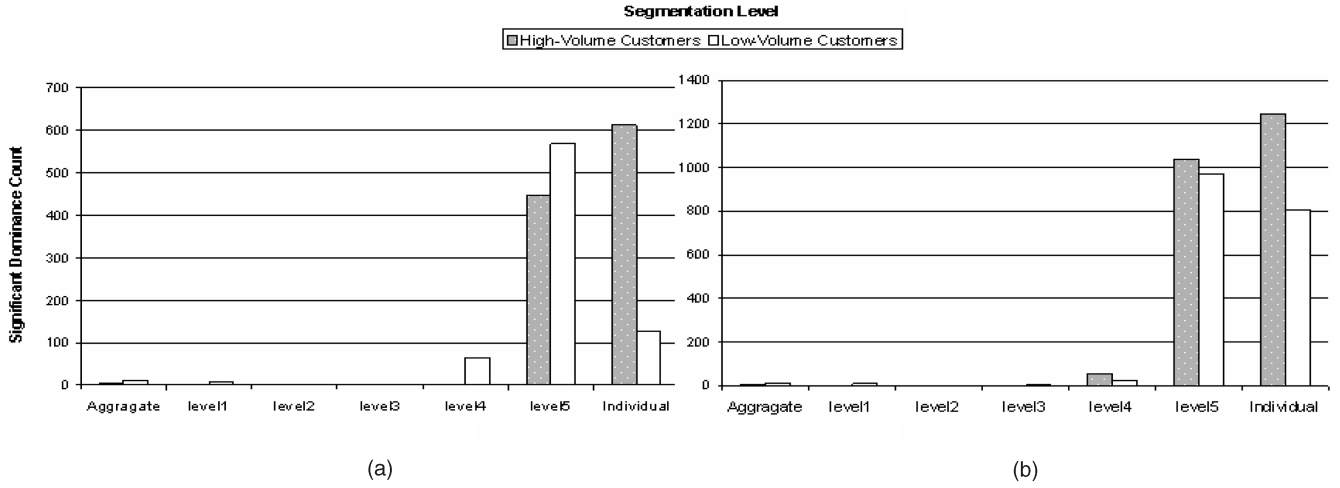


Fig. 5. (a) Histogram of statistically significant comparisons counted across all data sets and all null hypothesis tests for the FFirst, SMean_Mod, and EM_Mod clustering algorithms. (b) Histogram of statistically significant comparisons counted across all data sets and all null hypothesis tests.

sions from these studies. In this section, we will gain additional insights into these comparisons by studying the *performance curves* plotting performance measures across different segmentation levels, as shown in Fig. 4. An important question is how would the *shapes* of the curves change under different experimental conditions, e.g., would the performance grow monotonically when the customer segments are refined, or would the performance curves have a set of alternative shapes in different experimental settings?

We generated many such curves for different types of parameters. For example, Fig. 4 shows the performance curve predicting primary product category, based on the CCI measure, for high-volume ComScore data under SMean clustering. The X-axis denotes the level of segmentation, which runs from the aggregate level near the origin, through the five segmentation levels, to the individual family level on the far right. The Y-axis specifies the $\overline{CCI}$ (average CCI) measure. The four different curves represent the performance of the four classifiers across all the levels of predictive models.

We plotted such performance curves for all nine types of customer data sets, across six clustering schemes and a total of 47 transactional dependent attributes. From the 1,128 performance curves of $\overline{CCI}$, we observed three dominating patterns. For *high-volume* customers and "well-behaved" clustering algorithms (as defined in Section 3), we see a monotonically increasing curve as represented by Fig. 4a. This is the dominant shape observed across all data sets, as the individual level approach dominates both the aggregated and segmentation approaches in our experimental settings. This occurs primarily for high-volume customer data sets because with sufficient data, we can build models of idiosyncratic customer behavior all the way to the individual level without running into the problem of data sparsity.

Fig. 4b shows the second general pattern, that of convex performance curves. This observation is especially true for *low-volume* customer data sets and "well-behaved" clustering algorithms. This pattern shows that for low-volume customers, we will eventually run into the problem of data sparsity while trying to build progressively finer segmentation models of customer behavior. Our discussion of FFirst, SMean_Mod, and EM_Mod clustering algorithms for low-volume ComScore and synthetic customers presented in Section 5.2 fits well into this category.

Fig. 4c shows the third general pattern, that of concave performance curves. This pattern is observed mainly with "poorly behaved" clustering algorithms. This "concave" pattern occurs because heterogeneous customers are grouped into the same segments by noneffective clustering algorithms.

5.4 Significant Comparison Trends

As shown in Section 5.3, in cases of convex curves where segmentation level models dominate, we observe that the maximum performance is usually reached at Level 5 in Fig. 4b for all the four classifiers. We further studied the question at which segmentation level the predictive performance reaches its peak. In particular, we plotted the statistically significant comparison counts listed in Table 6, where we only examine comparisons done with the "well-behaved" FFirst, SMean_Mod, and EM_Mod clustering algorithms, across the aggregate, best performing segmentation, and individual levels. The results, presented in Fig. 5a, show that the majority of the peak performance is reached *at the finest segmentation level for low-volume* customers, whereas the peak performance is reached at the *individual level for high-volume* customers, which is consistent with the results from Section 5.2. The performance peak at the finest granularity level provides additional support for the case of *microsegmentation* [20].

We extend this investigation to comparisons made across all clustering methods, including Random clustering, for all data sets and plotted the statistically significant comparison counts in Fig. 5b. We note that the reverse performance trends of low-volume customers peaking at level 5 and high-volume customers peaking at the individual level is still evident, although the effect is less pronounced than in Fig. 5a. This confirms our hypothesis that, while individual level models dominate for high-volume customers, segmented models outperform individual level models for low-volume customers.

6 Conclusions

In this paper, we compared aggregate, segmentation, and individual-level modeling to determine which of these three levels of analysis would provide better predictions of customer behavior. This question translates into the following types of comparisons between the modeling approaches: 1) aggregate versus segmentation, 2) segmentation versus individual, and 3) aggregate versus individual modeling.

Since marketers answered the first question by showing that segmentation models usually outperform the aggregate models, we focused on the remaining two questions. We conducted a comparative study of the three types of customer modeling across multiple dimensions of analysis, such as different types of data sets, customers, predictive models, dependent variables, performance measures, and segmentation techniques and identified the following factors significantly influencing outcomes of customer behavior models: customer heterogeneity, data sparsity, quality of segmentation techniques, and levels of segmentation. We studied how these factors affect the performance of customer models.

Our experiments show that individual-level models statistically outperform aggregate models of customer behavior, even for sparse data having only a few transactions per customer. Our experiments also show that for the high-transaction customers or poor clustering techniques, individual-level customer models statistically outperform segmented models at all the segmentation levels, as demonstrated by the performance curve presented in Fig. 4a. This shows that for the high-volume customers, we can improve performance of our applications by refining customer segments all the way to the individual level without running into the data sparsity problem. These two results reaffirm many informal claims made by the popular press about the benefits of the 1-to-1 marketing approach. However, our experiments also show that for low-volume transaction customers *and* good clustering techniques, there is an optimal segmentation level that outperforms the individual-level customer models, as demonstrated by the performance curve presented in Fig. 4b. This means that the effect of data sparsity tends to dominate customer heterogeneity, as customer segments become too fine-grained and reach the limit of 1. Moreover, this optimal segmentation level is significantly skewed towards the 1-to-1 case and is usually achieved at the finest segmentation levels, as Figs. 4b, 5a, and 5b demonstrate this. This result, together with previous observations about the advantages of individual modeling makes a strong case for *microsegmentation* of customer bases. Finally, we showed that some of the popular clustering techniques could lead to poor performance results comparable to the random segmentation method, as demonstrated by the performance curve in Fig. 4c. This result stresses the importance of selecting good segmentation methods for producing better predictive models of customer behavior.

The results stated above can be explained in terms of the bias-variance trade-off. When the customer base is initially split into few coarse segments using effective clustering methods, this initial partitioning reduces customer heterogeneity and, therefore, bias of the resulting predictive models, while each segment has plenty of data for good predictive purposes, thus minimally affecting variance. Once we continue splitting customers into progressively smaller segments, we get diminishing returns in terms of bias reduction, while the variance is increasing because of the lack of data to fit against the model. Once the effect of variance increase dominates the effect of bias decrease, the performance of customer models begins to diminish. For the data considered in our experiments, this effect happens only at the finest levels of customer segmentation or even in the 1-to-1 limit depending on whether these are high or low-volume transaction customers, thus providing a strong support for microsegmentation.

Our results have significant implications for the fields of CRM and personalization for the following reasons: First, they provide insights into how to grow and refine customer segments and when and where to stop in this process. Second, they reaffirm prior anecdotal evidence of the superiority of the 1-to-1 marketing approach advocated in the popular press [4], but only under certain conditions described in the paper. Third, our results stress the importance of good segmentation methods that effectively partition customer bases into high-performing segments.

As an extension to our research, we plan to explore the formulation of automated segmentation algorithms that operate as a mixture of 1-to-1 and segmentation models. The goal is to reduce the variance by a form of smoothing, from 1-to-1 models towards segmented models. Another direction of future research could also leverage research on Naive-Bayes Trees [19] and recent advances in semisupervised clustering algorithms [21] in developing methodology that automatically selects an appropriate mixture of segmentation and 1-to-1 modeling for any sample customer population.

Finally, our results are based on the empirical studies conducted on several data sets. To be able to generalize our conclusions, it is important to perform a theoretical analysis of the observed phenomena using the bias-variance trade-off models applicable to customer segmentation problems. We believe that this analysis is quite complicated in the realistic industrial settings where few simplifying assumptions can be made about the nature of these models. Nevertheless, it is important to do this analysis, and we hope that some researchers will be able to pursue this work in the future.

References

[1] G. Adomavicius and A. Tuzhilin, "Expert-Driven Validation of Rule-Based User Models in Personalization Applications," *Data Mining and Knowledge Discovery*, vol. 5, nos. 1/2, pp. 33-58, 2001.

[2] G. Adomavicius and A. Tuzhilin, "Personalization Technologies: A Process-Oriented Perspective," *Comm. CAM*, Oct. 2005.

[3] G.M. Allenby and P.E. Rossi, "Marketing Models of Consumer Heterogeneity," *J. Econometrics*, 1999.

[4] D. Armbruster, "Experiential Marketing Comes Alive," *1to1 Magazine*, 2005.

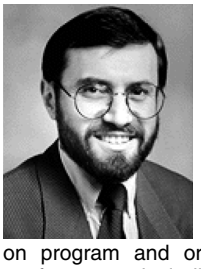
[5] C.G. Atkeson, A.W. Moore, and S. Schaal, "Locally Weighted Learning," *Artificial Intelligence Rev.*, 1997.

[6] D. Besanko, J.-P. Dube, and S. Gupta, "Competitive Price Discrimination Strategies in a Vertical Channel Using Aggregate Retail Data," *Management Science*, vol. 49, no. 9, pp. 1121-1138, 2000.

[7] *Comm. ACM*, special issue on personalization, 2000.

[8] I.V. Cadez, P. Smyth, and H. Mannila, "Predictive Profiles for Transaction Data Using Finite Mixture Models," technical report, UC Irvine, www.datalab.uci.edu/papers/profiles.pdf, 2001.
[9] "Heterogeneity, Omitted Variable Bias, Duration Dependence," *Longitudinal Analysis of Labor Market Data,* J.J. Heckman, B. Singer, G. Chamberlain, ed., Cambridge Univ. Press, 1985.
[10] W.W. Cohen, "Fast Effective Rule Induction," *Proc. Int'l Conf. Machine Learning,* 1995.
[11] C. Cortes, K. Fisher, D. Pregibon, A. Rogers, and F. Smith, "Hancock: A Language for Extracting Signatures from Data Streams," *Proc. ACM SIGKDD,* 2000.
[12] J. Fan and R. Li, "Local Modeling: Density Estimation and Nonparametric Regression," *Advanced Medical Statistics,* J. Fang and Y. Lu, eds., pp. 885-930, World Scientific, 2003.
[13] U.M. Fayyad and K.B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proc. Int'l Joint Conf. Artificial Intelligence,* 1993.
[14] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining,* Section 6.3.2-6.3.3. MIT Press, 2001.
[15] J. Heckman, "Statistical Models for Analysis of Discrete Panel Data," *Structural Analysis of Discrete Data,* C.F. Manski and D. McFadden, pp. 114-178. Cambridge: MIT Press, 1982.
[16] S.D. Hochbaum and B.D. Shmoys, "A Best Possible Heuristic for the K-Center Problem," *Math. Operational Research,* 1985.
[17] D.D. Jensen and P.R. Cohen, "Multiple Comparisons in Induction Algorithms," *Machine Learning,* vol. 38, 2000.
[18] G.H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," *Proc. 11th Conf. Uncertainty in Artificial Intelligence,* pp. 338-345, 1995.
[19] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," *Proc. Int'l Conf. Knowledge Discovery in Databases and Data Mining,* 1996.
[20] P. Kotler, *Marketing Management,* 11th ed. Prentice Hall, 2003.
[21] B. Kulis, S. Basu, I. Dhillon, and R. Mooney, "Semi-Supervised Graph Clustering: A Kernel Approach," *Proc. Int'l Conf. Machine Learning,* 2005.
[22] P.J. Lenk, W.S. DeSarbo, P.E. Green, and M.R. Young, "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science,* vol. 15, no. 2, 1996.
[23] E. Manavoglu, D. Pavlov, and C.L. Giles, "Probabilistic User Behavior Models," *Proc. IEEE Int'l Conf. Data Mining (ICDM),* 2003.
[24] S. McDonnell, "Microsegmentation," *ComputerWorld,* 2001.
[25] W. Mendenhall and R.J. Beaver, *Introduction to Probability and Statistics.* Thomson Pub, 1994.
[26] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Using Sequential and Non-Sequential Patterns for Predictive Web Usage Mining Tasks," *Proc. IEEE Int'l Conf. Data Mining (ICDM),* 2002.
[27] M. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," *Machine Learning,* 1997.
[28] D. Peppers and M. Rogers, *The One-to-One Future.* New York: Doubleday, 1993.
[29] B.J. Pine, D. Peppers, and M. Rogers, "Do You Want to Keep Your Customers Forever?" *Harvard Business Rev.* 1995.
[30] R. Quinlan, *C4.5: Programs for Machine Learning.* 1993.
[31] W.J. Reinartz, "Customer Relationship Management at Capital One," case report, UK: INSEAD, 2003.
[32] S. Roy, "Nearest Neighbor with Generalization," Univ. of Canterbury, Christchurch, New Zealand, 2002.
[33] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," *INFORMS J. Computing,* vol. 15, no. 2, 2003.
[34] M. Wedel and W. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations,* 2nd ed. Kluwer Publishers, 2000.
[35] M. Wedel, W. Kamakura, N. Arora, A. Bemmaor, J. Chiang, T. Elrod, R. Johnson, P. Lenk, S. Neslin, and C.S. Poulsen, "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling," *Marketing Letters,* vol. 10, no. 3, pp. 219-232, 1999.
[36] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann, 2000.
[37] "Alternative Approaches to Unobserved Heterogeneity in the Analysis of Repeatable Events," *Sociological Methodology,* K. Yamaguchi and B. Tuma, ed. pp. 213-219, Washington, DC: Am. Sociological Assoc., 1986.
[38] Y. Yang and B. Padmanabhan, "Segmenting Customer Trans. Using a Pattern-Based Clustering Approach," *Proc. Int'l Conf. Data Mining,* 2003.

Tianyi Jiang received BS and the MEng degrees from Cornell University, School of Electrical and Computer Engineering. He is a fifth year PhD student in the Department of Information, Operations, and Management Sciences, Stern School of Business, New York University (NYU). His current research interests include personalization, customer segmentation, consumer profiling, and credit risk modeling. Prior to starting his PhD studies at NYU, he worked as a senior software consultant for large investment banks in the area of globally distributed financial trading systems.



Alexander Tuzhilin received the PhD in computer science from the Courant Institute of Mathematical Sciences, New York University. He is a professor of information systems at the Stern School of Business, New York University (NYU). His current research interests include knowledge discovery in databases, personalization, and CRM technologies. He has been published widely in leading CS and IS journals and conference proceedings. Dr. Tuzhilin served on program and organizing committees of numerous CS and IS conferences, including as a program cochair of the Third IEEE International Conference on Data Mining. He also serves on the editorial boards of the *IEEE Transactions on Knowledge and Data Engineering*, the *Data Mining and Knowledge Discovery* journal, the *INFORMS Journal on Computing*, and the *Electronic Commerce Research* journal.