Available online at www.sciencedirect.com
ScienceDirect
Journal of Business Research 60 (2007) 656–662
JOURNAL OF BUSINESS RESEARCH

Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression

John A. McCarty [a,*], Manoj Hastak [b,1]

[a] School of Business, The College of New Jersey, Ewing, New Jersey 08628, United States
[b] Kogod School of Business, American University, Washington, DC 20008, United States

Received 1 November 2005; received in revised form 1 April 2006; accepted 1 June 2006

Abstract

Direct marketing has become more efficient in recent years because of the use of data-mining techniques that allow marketers to better segment their customer databases. RFM (recency, frequency, and monetary value) has been available for many years as an analytical technique. In recent years, more sophisticated methods have been developed; however, RFM continues to be used because of its simplicity. This study investigates RFM, CHAID, and logistic regression as analytical methods for direct marketing segmentation, using two different datasets. It is found that CHAID tends to be superior to RFM when the response rate to a mailing is low and the mailing would be to a relatively small portion of the database, however, RFM is an acceptable procedure in other circumstances. The present article addresses the broader issue that RFM may focus too much attention on transaction information and ignore individual difference information (e.g., values, motivations, lifestyles) that may help a firm to better market to their customers.

Keywords: Database marketing; Data-mining; RFM; CHAID; Analytical procedures

1. Introduction

Segmentation in direct marketing has become more efficient in recent years because of the development of database marketing techniques. These data-mining approaches provide direct marketers with better ways to segment their current customers and develop marketing strategies tailored to particular segments and/or individuals. Over the recent years, database marketing techniques have evolved from simple RFM models (models involving recency of customer purchases, frequency of their purchases, and the amount of money they have spent with the firm) to statistical techniques such as chi-square automatic interaction detection (CHAID) and logistic regression. More recently, neural network models are employed in the database marketing arena (Yang, 2004).

In spite of recent statistical advances in data-mining, marketers continue to employ RFM models. A study by Verhoef et al. (2002) shows that RFM is the second most common method used by direct marketers, after cross tabulations, in spite of the availability of more statistically sophisticated methods. There are a couple of related reasons for the popularity of RFM. As Kahan (1998) notes, RFM is easy to use and can generally be implemented very quickly. Furthermore, it is a method that managers and decision makers can understand (Marcus, 1998). This is an important consideration in that a successful technique for a direct marketer is one that differentiates likely responders to a particular mailing from those who are unlikely to respond, yet does so in a way that is easy to explain to decision makers. However, it has been argued that the simplicity of RFM has been overemphasized, but its ability to differentiate, relative to statistical techniques, has not been considered to the extent that it should be (Yang, 2004).

Although the efficiency of RFM has been questioned, little research documents its ability relative to newer statistical techniques. This paucity of research is partly because RFM refers to a general approach to data-mining; there are a variety

* Corresponding author. Tel.: +1 609 771 3220.
E-mail addresses: mccarty@tcnj.edu (J.A. McCarty), mhastak@american.edu (M. Hastak).
[1] Tel.: +1 202 885 1973.

0148-2963/$ - see front matter
doi:10.1016/j.jbusres.2006.06.015

of ways of applying the use of recency, frequency, and monetary value. Research that has been conducted on the efficacy of RFM generally focuses on proprietary or judgmental models of RFM (e.g., Levin and Zavari, 2001; Magidson, 1988) and not on empirically based RFM models. More recently, research has moved away from RFM and has focused instead on newer, more sophisticated approaches to data-mining (c.f., Deichmann et al., 2002; Linder et al., 2004). The current study evaluates one popular, empirically based (as opposed to judgmental) approach to RFM. This RFM approach is compared to CHAID and logistic regression, in an effort to understand its capabilities as a database marketing analytical tool.

2. Analytical segmentation methods in data-mining

2.1. RFM analysis

Recency, frequency, and monetary (RFM) analysis has been used in direct marketing for a number of decades (Baier et al., 2002). This analytical technique grew out of an informal recognition by catalog marketers that three variables seem particularly related to the likelihood that customers in their house datafiles would respond to specific offers. Customers who recently purchased from a marketer (recency), those who purchase many times from a marketer (frequency), and those who spend more money with a marketer (monetary value) typically represent the best prospects for new offerings.

As noted, RFM analysis is utilized in many ways by practitioners, therefore, RFM analysis can mean different things to different people. One common approach to RFM analysis is what is known as hard coding (Drozdenko and Drake, 2002). Hard coding RFM is a matter of assigning a weight to each of the variables recency, frequency, and monetary value, then creating a weighted score for each person in the database. The assignment of weights is generally a function of the judgment of the database marketers with a particular database; for example, past experience may tell a marketer that recency should weigh twice as much as frequency and monetary value. Therefore, this application of RFM is often referred to as judgment based RFM. The weightings could also vary as a function of the particular mailing (Baier et al., 2002). The weights can, of course, be empirically derived based on offerings mailed to database members in the past, thus relying on previous data rather than judgments.

Regardless of the way that RFM is utilized, there are two common characteristics of RFM procedures. First, RFM is used to segment a house file (i.e., a company's current customers) using information related to recency, frequency, and monetary value. RFM is not applicable to the prospecting for new customers because a marketer would not have transaction information for prospects. Second, RFM analysis generally focuses on the three behavioral variables of recency, frequency, and monetary value. Although these variables are considered powerful predictors of future behavior, traditional RFM is limited to these three things.

A well known, empirically based RFM method is a procedure advocated by Arthur Hughes (2000). Hughes' approach is applicable in instances when a marketer intends to send a mailing to customers in its database and would like to find those in the database who are the most likely to respond to the specific mailing. Hughes recommends a test mailing to a sample of customers in the file; then the selection of the members of the rest of the file is made as a function of the results of the test. Thus, compared with hard coding RFM, Hughes' method is not arbitrary with respect to the weighting of recency, frequency, and monetary value. The importance of each of these is determined by the test mailing for the particular offer.

The first step in the method is for the marketer to sort the customer file according to how recently customers have purchased from the firm. The database is then divided into equal quintiles and these quintiles are assigned the numbers 5 to 1. Therefore, the 20% of the customers who most recently purchased from the company are assigned the number 5; the next 20% are assigned the number 4, and so on. The next step involves sorting the customers within each recency quintile by how frequently they purchase from the marketer. For each of these sorts, the customers are divided into equal quintiles and assigned a number of 5 to 1 for frequency. Each of these groups (25 groups) is sorted according to how much money the customers have spent with the company. These sorts are divided into quintiles and assigned numbers 5 to 1. Therefore, the database is divided into 125 roughly equal groups (cells) according to recency, frequency, and monetary value.

Hughes recommends conducting a test mailing to a randomly sampled subset of each cell (e.g., 10%). After the responses of the test mailing are received, the proportion of respondents in each cell can be calculated. The cells can then be ordered as a function of response percent. The marketer can then elect to mail to a certain portion of the remaining file (e.g., the top 20% of the cells). Alternatively, the marketer can elect to mail to the cells that are above a break even percent, given the cost of the mailing and the expected revenue for each return. For example, if a mailing costs \$1.50 and the revenue received is \$50.00 per order, the break even percentage would be 3%. Thus, for the 90% of the file that is left after the test mailing, the direct marketer would mail to the RFM cells that the test mailing predicted a 3% or better return.

It is important to note that Hughes' method does not assume a monotonic relationship between the dependent variable (responded/did not respond) with the variables of recency, frequency, and monetary value. Each cell is a discreet group that is considered individually in terms of its performance. Thus, if middle levels of one of the independent variables (e.g., frequency) are more related to response compared with higher or lower levels of this variable, then the procedure can accommodate the non-monotonic nature of the relationship.

2.2. CHAID

Chi Square Automatic Interaction Detector (CHAID) (see, for example, Sargeant and McKenzie, 1999) is a method of database segmentation that has been used for a number of years. Research has shown that CHAID is superior to judgment based RFM with respect to the identification of likely responders

(Levin and Zavari, 2001; Magidson, 1988). CHAID is similar to the RFM approach of Hughes because it creates groupings (nodes) of database members. The main difference is that these groupings are not created a priori as is the case with RFM. Rather, the file is split according to a statistical algorithm after a test mailing is conducted. After the returns of the test mailings are received, the procedure starts with a node that includes everyone in the test file. The procedure then searches for the independent variable (e.g., number of times purchased) that best discriminates among the file members with respect to a dichotomous variable (i.e., purchased/did not purchase on current mailing). It splits the original node on this independent variable into as many subgroups as are significantly different with respect to the dichotomous variable. The procedure then splits these new nodes according to the variables that discriminate each of them. The procedure continues until no other splits are significant. CHAID analysis is often called tree analysis because a trunk (original node) is split into branches, then more branches, etc. The terminal nodes are those that can not be split any further.

The analysis is similar to RFM because the terminal nodes can be evaluated according to which ones break even with respect to expected profit and mailing costs. The direct marketer can then use the rules that define the terminal nodes in the test mailing (i.e., levels of the independent variables that define each terminal node) to select the groups of people left in the file after the test that should receive the mailing. It is also similar to RFM in that CHAID can accommodate relationships between the dependent variable and the predictor variables that are non-monotonic. For example, if the number of times purchased relates to the dependent variable, CHAID may divide the file members into three nodes: those who purchase 1 to 3 times, those who purchase 4 to 8 times, and a third node of those who purchase 9 or more times. These three nodes represent discreet groupings.

An important difference between CHAID and RFM is that CHAID can accommodate a variety of independent variables. The independent variables could include recency, frequency, and monetary value, but could also include other transaction variables (e.g., used a credit card or not), as well as individual difference variables such as demographic and psychographic variables.

2.3. Logistic regression

Logistic regression is a modeling procedure where a set of independent variables are used to model a dichotomous criterion variable. Therefore, it is appropriate for direct marketers who would like to model the dichotomous variable of respond/don't respond to a mailing. Logistic regression is particularly useful in these circumstances in that the actual criterion variable is dichotomous; however, the predicted variable is the response probability, which varies from zero to one. Therefore, the model can provide a probability of response for everyone in the file, given the estimated parameters for a set of predictor variables.

After a test mailing similar to CHAID, logistic regression can be used to analyze the response variable as a function of several independent variables (e.g., number of times purchased) and provide an equation that can calculate the response probability for the entire house file. The marketer can then mail to everyone left in the file (excluding those in the test) who has a probability higher than the break even percent. Similar to CHAID, the independent variables are not restricted to recency, frequency, and monetary value.

Logistic regression differs from both RFM and CHAID in two important ways. First, logistic regression provides a response probability for individual members of the dataset rather than creating discreet groups of people. Therefore, in theory, each person in the dataset may have a different response probability. In practice, however, if few independent variables are used to construct the logistic function and each has a small number of different possible values, then there would be a relatively small number of different response probabilities across the people in the file. Second, for continuous predictor variables, logistic regression model relationships of the independent variables with the dichotomous dependent variable that are monotonic; both RFM and CHAID are distribution free. This has implications for the performance of logistic regression in instances where the relationship between a predictor variable and the response variable is neither continuously increasing nor decreasing. For example, when the relationship between recency of previous purchase and purchase on the test mailing is curvilinear, logistic regression may not be able to capture the relationship in ways similar to that for RFM or CHAID.

3. The studies

The viability of Hughes' approach to RFM as a method to segment a marketer's customers is evaluated with two datasets. The RFM method is compared with CHAID and logistic regression. For these comparisons, CHAID and logistic regression are limited to the same information used to create recency, frequency, and monetary value for the RFM method. Thus, the studies are designed to assess the discriminating characteristics of the three segmentation approaches given the same independent variables.

Both datasets were provided by the Direct Marketing Educational Foundation. One of these datasets is for a multi-division mail order company; the other is for a non-profit organization that solicits contributions from its members. Therefore, these two files provide somewhat different situations for which a database marketer would apply segmentation techniques to find likely responders to a mailing. Both datasets include information about how recently each person purchased (or contributed for the non-profit organization), how many times each person purchased (contributed), and the lifetime dollar amount of purchases (contributions). Each dataset also has the results of a recent mailing. These results include the percentage of the dataset that responded to the mailing. The mail order company's dataset includes 99,200 people and has a return rate of 27.4% for the recent mailing. Thus, the two datasets provide for tests of the sensitivity of the segmentation procedures at very different levels of response (i.e., less than 5% responding versus over one quarter of the file responding to the mailing).

The frequently used method of cross validation is employed to evaluate the three segmentation methods. Each of the datasets is randomly split in half. One half of each dataset is considered the test group; the second half of each set is a hold out sample. RFM (Hughes' method), CHAID, and logistic regression are applied on the test group of each dataset using the number of days since last activity, total number of purchases (contributions), and total dollar amount of purchases (contributions). The parameters developed from the test groups for each of the segmentation procedures are then applied to the hold out sample for each dataset. Thus, one could consider the test group to be analogous to the test mailing of a typical direct marketing situation and the hold out sample to represent the rest of a direct marketer's house file to which the results of the test mailing would typically be applied.

The viability of each procedure is evaluated in two ways. For these analyses, the percentages of all respondents who would be reached if a mailing is sent to only a subset of the individuals in the file (as opposed to the entire file) are calculated for the test and hold out samples; the subsets include those file members that each procedure assesses to be most likely to respond to the procedure. For example, if a procedure is used to select 10% (or 20%, or 30%, etc.) of the individuals in a file who are deemed most likely to respond, what percentage of actual respondents are included in the group that is selected? Note that if a procedure performs no better than chance, one would expect a selection of 10% of names in a file to yield 10% of all respondents. Database marketers typically refer to these proportions as gain percentages. The researchers evaluate the three segmentation procedures at four levels of depth in the file—20%, 30%, 40%, and 50%.

One analysis involves comparing the gain percentage for a particular depth of the file (e.g., 10%) in the test group with the gain percentage for the same file depth in the hold out sample. This approach provides information about the reliability of the model developed in the test group. If the model is reliable, it can be expected that gain percent in the hold out sample will not differ appreciably from gain percent in the test sample. Alternatively, a significant difference between the proportion in the test group and the proportion in the hold out sample would suggest that the particular segmentation method may be misleading at that level of file depth.

The second set of analyses involves comparing the gain percentage for a particular depth of the file across the three segmentation procedures. These analyses provide a head-to-head comparison between the approaches on their ability to discriminate between responders and non-responders.

3.1. Study 1

As noted, the first study involves data from a multi-division catalog marketer. The marketer had made a mailing to the entire dataset and a portion of the file responded to the offer in the mailing. The dataset includes 96,551 members; it is randomly split into a test sample of 48,275 people and a hold out sample of 48,276. The overall response rate to the mailing is 2.46%. The response rate for the test group is 2.44%; the response rate fort the hold out sample is 2.47%.

3.1.1. Results

3.1.1.1. Reliability of the segmentation methods. Table 1 shows the proportion of respondents captured for 10% increments of file depth from 20% to 50% of the file for each of the segmentation methods (RFM, CHAID, and logistic regression) for the test and hold out groups. The table also presents the difference in proportions for each method at each depth between the test and hold out samples. The difference measure provides an indication of the extent to which each segmentation method produces results in the test sample that can be reliably replicated in the hold out sample.

As Table 1 shows, the reliability of RFM is questionable at both the 20% and 30% levels of the file. For example, the 20% of the test sample that RFM indicates would be the most likely to respond captures 39.2% of actual responders. However, when the parameters of this test are applied to the hold out sample, the top 20% of the sample only captures 34.6% of all respondents. Therefore, the proportion of respondents captured in the hold out sample is significantly lower than the proportion captured in the test sample. A significant drop in proportion (4.7%) is also evident at the 30% of depth level. At the 40% and 50% levels of the file, the difference in proportions of test and hold out samples are not significantly different for the RFM segmentation method. As the table shows, the proportions of respondents in the test and hold out samples are not significantly different for the CHAID and logistic regression segmentation methods at all four levels of depth. Therefore, these analyses suggest that if a marketer elected to mail to a relatively small portion of their house file, a test mailing using RFM may result in over prediction of the number of respondents when the results of the test are applied to the rest of the file.

Table 1
Percent of total responses for various levels of depth of total file

	Data-mining technique		
	RFM (%)	CHAID (%)	Logistic (%)
20% depth of file			
Test sample	39.2_a	38.5_a	36.5_a
Hold sample	34.6_a	37.7_b	35.8_{ab}
Difference	4.6 **	0.8	0.7
30% depth of file			
Test sample	51.3_a	49.9_{ab}	47.9_b
Hold out sample	46.6_a	49.9_b	47.6_{ab}
Difference	4.7*	0.0	0.3
40% depth of file			
Test sample	61.1_a	60.0_a	58.1_a
Hold out sample	58.0_a	58.5_a	57.1_a
Difference	3.1	1.5	1.0
50% depth of file			
Test sample	71.2_a	68.8_a	67.6_a
Hold out sample	67.3_a	67.4_a	65.4_a
Difference	3.9	1.4	2.2

Note. In any row for the test or hold out sample, percentages that do not share a common subscript are significantly different at $p<.1$ (two-tailed).
** $p<.01$, * $p<.05$ (one-tailed).

3.1.1.2. Gain percent for the segmentation methods. Although reliability of a segmentation method is a crucial consideration, an equally important evaluation is the relative predictive performance of the three segmentation methods. As Table 1 shows, the gain percentages indicate that the three segmentation methods perform similarly within each of the four levels of file depth in the test samples. The one significant difference is at the 30% depth level where RFM captures a higher proportion of respondents than logistic regression did.

When the parameters estimated in the test sample are applied to the hold out sample, CHAID captures a significantly higher proportion of respondents than does RFM at both the 20% and 30% levels of depth. However, the three segmentation methods do not differ significantly with respect to performance in gain percent at the 40% and 50% depth levels for the hold out sample.

Overall, the results suggest that RFM underperforms, relative to CHAID, when a marketer elects to mail to a relatively small portion of the file. At first glance it may appear that the differences in gain percent between CHAID and RFM are relatively small and perhaps unimportant at a practical level even in instances where they are statistically significant. To put these differences in perspective in terms of potential profit, hypothetical profit and cost figures can be applied to the response information. Assume that a mailing for a direct marketer costs $1.50 and the expected revenue per response is $100. Further assume that the test sample represents a 10% test mailing and the hold out sample is the rest of the house file; therefore, the results of the hold out sample would be multiplied by a factor of nine. Applying these assumptions to the performance of RFM and CHAID segmentation methods with this file, the RFM method used in the test sample would predict a profit of $284,903 in the full file (the hold out sample multiplied by a factor of nine) whereas CHAID would predict a profit of $277,488. The actual mailing to the likely members of the full file as determined by the two methods would show a very different pattern. RFM would secure a profit of $241,469 while CHAID would provide a profit of $274,781. Thus, CHAID would outperform RFM by $33,313. Obviously, small differences in proportions can make a big difference when applied to large house files.

3.2. Study 2

The second study involves data from a non-profit organization that had made a recent solicitation for a donation from members in its house file. The dataset includes 99,200 members; it is randomly split into a test group of 49,600 people and a hold out sample of 49,600. The overall response rate to the solicitation is 27.4%. The response rate for the test group is 27.3%; the response rate for the hold out sample is 27.6%.

3.2.1. Results

3.2.1.1. Reliability of the segmentation methods. Table 2 shows the proportion of respondents captured for 10% increments of file depth from 20% to 50% of the file for each of the three segmentation methods for the test and hold out groups. The table also presents the difference in proportions between test and hold outs samples for each segmentation method at each depth.

As the table shows, the difference in proportions between test and hold out samples for the three methods are very small (less than a percentage point for all methods at all four levels of depth). Moreover, none of these differences are statistically significant. Therefore, this set of analyses suggests that for this particular file with a rather large response rate, all three methods are generally able to provide an accurate prediction of the response rate when the results of a test mailing are applied to the full file.

3.2.1.2. Gain percent for the segmentation methods. Table 2 also presents the gain percents for the three different segmentation methods for the four levels of depth of the test and hold out samples. As the table shows, there are no differences between the performance of RFM and CHAID at all four levels of depth for either the test or hold out samples. Therefore, for this dataset, RFM appears to be as accurate as CHAID in capturing likely responders when CHAID is confined to the same independent variables as RFM.

A rather unexpected finding with respect to gain concerns the performance of logistic regression. CHAID performed significantly better than logistic regressions at both the 20% and 30% levels of depth for both the test and hold out samples. The RFM method significantly outperforms logistic regression at the 30% depth. Therefore, at least for this dataset, logistic regression may have difficulties when a marketer would elect to mail to a small portion of the file.

The results of study 2 suggest that RFM may perform on a par with more sophisticated statistical techniques when the response level is fairly high. At all four levels of file depth

Table 2
Percent of total responses for various levels of depth of total file

	Data-mining technique		
	RFM (%)	CHAID (%)	Logistic (%)
20% depth of file			
Test sample	36.3_{ab}	36.6_{a}	35.7_{b}
Hold sample	35.6_{ab}	36.1_{a}	35.2_{b}
Difference	0.7	0.5	0.5
30% depth of file			
Test Sample	49.1_{a}	49.0_{a}	47.7_{b}
Hold Out Sample	48.9_{a}	48.8_{a}	47.7_{b}
Difference	0.2	0.2	0.0
40% depth of file			
Test Sample	60.4_{a}	60.4_{a}	60.0_{a}
Hold Out Sample	59.9_{a}	60.2_{a}	60.4_{a}
Difference	0.5	0.2	0.4
50% depth of file			
Test Sample	70.3_{a}	70.4_{a}	70.7_{a}
Hold Out Sample	70.5_{a}	70.2_{a}	70.8_{a}
Difference	0.2	0.2	0.1

Note. In any row for the test or hold out sample, percentages that do not share a common subscript are significantly different at $p < .1$ (two-tailed).

tested, Hughes' RFM method is able to achieve gain levels of likely responders as well as logistic regression at a similar level as CHAID.

4. Discussion

The two studies present some intriguing findings with respect to the performance of the three segmentation methods. The two datasets present very different circumstances that may be presented to database marketers and the results with respect to these datasets are somewhat different. Study 2 is a non-profit organization that solicits contributions in their house file; the recent mailing that is modeled provided for a response rate of roughly one quarter of the entire file. Given these features of the dataset and mailing, RFM is as successful as CHAID and logistic regression in capturing likely responders to the solicitation at all tested levels of depth of the file (20% to 50%). Furthermore, the parameters of the test for RFM appear to be as reliable as those of CHAID and logistic regression when applied to the hold out sample. Therefore, if one were to consider only the results of study 2, it would be concluded that RFM is generally a robust procedure that is similar to the other two segmentation procedures in its ability to segment likely respondents. It appears that RFM may be successful when the overall response rate is fairly high.

The characteristics of study 1, however, present a fairly common scenario for database marketers. This dataset is for a multi-division mail order company; the response rate for the offer is under 5%. Given these relatively common characteristics of a direct marketing situation, the results suggest that Hughes' RFM may not perform as well as CHAID when a marketer only mails to a small portion of the file (i.e., 30% or less). In these instances, CHAID outperforms RFM in terms of reliability and ability to capture likely responders. The actual performance of CHAID in the hold out sample is quite similar to its predicted performance in the test sample. By contrast, the actual performance of RFM in the hold out sample is significantly worse than its predicted performance in the test sample. Also, CHAID captures more respondents than RFM at both the 20% and 30% depths of the file. The superiority of CHAID in these instances suggests that the grouping of dataset members by a statistical algorithm, as in the case of CHAID, may be superior to the arbitrary and a priori groupings of RFM. In this study, CHAID creates fewer cells than the fixed and large number used in RFM. Given the low response rate to the offer and the large number of cells specified by RFM, there is a greater likelihood that chance fluctuation rather than systematic differences play a role in the outcome for RFM compared with CHAID. Thus, the predicted level of response in the test does not hold up when the parameters of the test are applied to the hold out sample.

The results across the two studies allow the researchers to consider the circumstances where RFM underperforms relative to CHAID. The findings suggest that RFM may have difficulties when the response rate is low (as in study 1) and the database marketer desires to send an offering to a relatively small portion of the entire file (30% or less). Under these circumstances, RFM may be less reliable than CHAID. Alternatively, when the response rate is relatively high (as in study 2) or the database marketer desires to mail to a relatively large portion of the file, RFM may provide results similar to CHAID and logistic regression. Overall, the study can conclude that Hughes' approach to RFM can perform at an acceptable level in many database marketing situations when a direct marketer is limited to using basic transaction variables. Given that statistical modeling can be more costly than RFM because of the need for highly trained personnel (Drozdenko and Drake, 2002), RFM can be considered an inexpensive and generally reliable procedure.

Two caveats or limitations should be considered with respect to the findings. First, the two datasets represent different sets of circumstances, both of which are relatively common in database marketing. The offers that are modeled in these data likely represent fairly common mailings in direct marketing. Therefore, the researchers assume that the characteristics of these files are not unusual circumstances in direct marketing. Having said this, the study must concede that these two examples may not generalize to all other database files. House files of different organizations may have their own peculiar characteristics and different offers may vary on a variety of dimensions. Therefore, the conclusions with respect to the performance of the three segmentation methods must be tempered with the understanding that they may not hold for all database marketing circumstances. For example, if there is a curvilinear relationship between a predictor (e.g., recency) and response, this would likely impact the performance of logistic regression, as this method models a monotonic relationship between predictors and response. Future research that tests these segmentation procedures under a variety of circumstances using simulated data would be useful. Simulating different possible relationships between predictors and response will allow the researchers to further understand the sensitivities of the three segmentation procedures to conditions that may arise in a variety of direct marketing situations.

Second, the analyses compared RFM to CHAID and logistic regression where each method is constrained to use the same independent variables of recency, frequency, and monetary value. This constraint is enforced to achieve a fair test of the analytical algorithms of the three methods. Therefore, the conclusions about the relative performance are made with the understanding that the researchers are considering them only in the context of these transaction variables. In practice, however, CHAID and logistic regression are not constrained with respect to the variables that can be used as predictors. Response to a mailing can be modeled with a variety of variables using these two methods. One would assume that more precise modeling could be achieved using other variables.

This second caveat raises a broader, and perhaps more important issue. The analysis of recency, frequency, and monetary value, whether by an RFM model or by a statistical technique such as CHAID, focuses entirely on the past behavior of individuals. Although social scientists recognize the power of past behavior as a useful predictor of future behavior, such a narrow focus likely limits the direct marketer in their ability to understand their customers. Zahay et al. (2004) raise this point in their discussion of transactional and relational data. They argue that an emphasis on transactional information is taking a very sales oriented approach to customers. Such an emphasis may aid sales in the short run, however, it does not add to the long term relationship with customers. A consideration of relational data such as information about the motivations, attitudes, values, and lifestyles is taking more of a marketing approach to customers. Although these variables may be less useful than transaction information in their ability to predict a response to an immediate marketing activity (i.e., a mailing), they may be enormously useful in understanding the underlying tendencies in customers. This consideration would favor analytical techniques such as CHAID and logistic regression that can accommodate a variety of personality and individual difference information.

References

Baier M, Ruf KM, Chakraborty G. Contemporary database marketing: concepts and applications. Evanston: Racom Communications; 2002.

Deichmann J, Eshghi A, Haughton D, Sayek S, Teebagy N. Application of Multiple Adaptive Regression Spines (MARS) in direct response modeling. J Interact Market 2002;16:15–27 [Autumn].

Drozdenko RG, Drake PD. Optimal database marketing: strategy, development, and data mining. Thousand Oaks: Sage Publications; 2002.

Hughes AM. Strategic database marketing. New York: McGraw–Hill; 2000.

Kahan R. Using database marketing techniques to enhance your one-to-one marketing initiatives. J Consum Mark 1998;15:491–4.

Levin N, Zahavi J. Predictive modeling using segmentation. J Interact Market 2001;15:2-22 [Spring].

Linder R, Geier J, Kolliker M. Artificial neural networks, classification trees and regression: which method for which customer base? Database Mark Cust Strategy Manag 2004;11:344–56 [July].

Magidson J. Improved statistical techniques for response modeling: progression beyond regression. J Direct Mark 1988;2:6-17 [Autumn].

Marcus C. A practical yet meaningful approach to customer segmentation. J Consum Mark 1998;15:494–501.

Sargeant A, McKenzie J. The lifetime value of donors: gaining insight through CHAID. Fund Rais Manag 1999;30:22–7 [March].

Verhoef PC, Spring PN, Hoekstra JC, Leeflang PS. December the commercial use of segmentation and predictive modeling techniques for database marketing in The Netherlands. Decis Support Syst 2002;34:471–81.

Yang AX. How to develop new approaches to RFM segmentation. J Target Meas Anal Mark 2004;13:50–60 [October].

Zahay D, Peltier J, Schultz DE, Griffin A. The role of transactional versus relational data in IMC programs: bringing customer data together. J Advert Res 2004;44:3−18 [March].