Journal of Business Research 67 (2014) 2751–2758

Contents lists available at ScienceDirect

Journal of Business Research

Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees

Kristof Coussement [a,*], Filip A.M. Van den Bossche [b,c], Koen W. De Bock [a]

[a] IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 8179), Expertise Center for Database Marketing (ECDM), Department of Marketing, 3 Rue de la Digue, F-59000, Lille, France
[b] Hogeschool-Universiteit Brussel, Faculty of Economics and Management, Warmoesberg 26, B-1000 Brussels, Belgium
[c] Katholieke Universiteit Leuven, Faculty of Business and Economics, Naamsestraat 69, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 1 February 2012
Received in revised form 1 July 2012
Accepted 1 September 2012
Available online 5 October 2012

Keywords:
Customer segmentation
Direct marketing
Data quality
Data accuracy
RFM
Decision trees

ABSTRACT

Companies greatly benefit from knowing how problems with data quality influence the performance of segmentation techniques and which techniques are more robust to these problems than others. This study investigates the influence of problems with data accuracy – an important dimension of data quality – on three prominent segmentation techniques for direct marketing: RFM (recency, frequency, and monetary value) analysis, logistic regression, and decision trees. For two real-life direct marketing data sets analyzed, the results demonstrate that (1) under optimal data accuracy, decision trees are preferred over RFM analysis and logistic regression; (2) the introduction of data accuracy problems deteriorates the performance of all three segmentation techniques; and (3) as data becomes less accurate, decision trees retain superior to logistic regression and RFM analysis. Overall, this study recommends the use of decision trees in the context of customer segmentation for direct marketing, even under the suspicion of data accuracy problems.

1. Introduction

Nowadays, increased digitization of transactions results in a boost of customer information stored in large transactional databases (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). This evolution has led to the emergence of the database marketing domain as a popular discipline in academic research and business practice (Ko, Kim, Kim, & Woo, 2008). A prominent database marketing application is customer segmentation for direct marketing, where the analyst tries to find homogeneous groups of customers with respect to their response behavior by means of so called data-mining tools (Akaah, Korgaonkar, & Lund, 1995; Cortinas, Chocarro, & Villanueva, 2010; McCarty & Hastak, 2007; Merrilees & Miller, 2010; Morganosky & Fernie, 1999). The usage of data-mining tools in direct marketing is subject to the knowledge discovery in databases (KDD) process, of which the growing importance is reflected by the large number of publications and applications in both academia and business (e.g. Bose & Mahapatra, 2001).

KDD prescribes a multi-level process to derive valuable top-level strategic insights from low-level raw data (Fayyad et al., 1996). A typical KDD process consists of the following five consecutive steps: (1) problem identification, in which the application domain is defined and objectives are formulated; (2) data preparation, or selecting, preprocessing, reducing, and transforming the data; (3) data mining, or choosing and applying an appropriate analysis technique; (4) the analysis, evaluation, and interpretation of results; and (5) presentation, assimilation, and use of knowledge (Han & Kamber, 2006; Martínez-López & Casillas, 2009).

Although the success of implementing a KDD process depends on the value of each of its five constituent steps (Crone, Lessmann, & Stahlbock, 2006; Fayyad et al., 1996), a significant proportion of recent research in direct marketing has unilaterally focused on the data-mining phase and its tools for segmenting customers (e.g. RFM (recency, frequency, and monetary value) analysis McCarty & Hastak, 2007, logistic regression McCarty & Hastak, 2007, decision trees Haughton & Oulabi, 1993 and more advanced techniques, such as artificial neural networks Zahavi & Levin, 1995, 1997, support vector machines Viaene et al., 2001, and genetic fuzzy systems Martínez-López & Casillas, 2009).

In addition to the choice of the best segmentation tool, data quality (DQ) is an equally important concept in customer analytics (Feelders, Daniels, & Holsheimer, 2000; Ko et al., 2008). Prior research shows that bad data yield bad analytical results, often referring to this process as the "garbage in, garbage out" principle (Baesens, Mues, Martens, & Vanthienen, 2009). Consider a marketing department of a direct marketing company that wants to profile its segmented customers according to their monetary value, i.e. their past total money spent at the company. If parts of the monetary value figures are not

* Corresponding author. Tel.: +33 320545892.
E-mail addresses: K.Coussement@ieseg.fr (K. Coussement), Filip.VandenBossche@hubrussel.be (F.A.M. Van den Bossche), K.DeBock@ieseg.fr (K.W. De Bock).

0148-2963/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbusres.2012.09.024

correct, the uncertainty of calculating the correct average monetary value per segment increases, and consequently the information quality and the segmentation performance decrease. DQ is often considered as a multi-dimensional construct having four subcategories (Wang and Strong, 1996); (1) *intrinsic DQ* denoting that data have quality in their own right, (2) *contextual DQ* referring to the fact that DQ should be considered within the context of the task at hand, (3) *representational DQ* and (4) *accessibility DQ* both linked to the importance of the information system(s).

Although many DQ attributes in each of these four subcategories have been introduced in the literature, this study focuses on how segmentation performance is impacted by the intrinsic DQ attribute *accuracy* which is defined as conformity with the real world (Wand & Wang, 1996). Three arguments are given to motivate the need for investigating the impact of data accuracy on segmentation performance. First, data accuracy is one of the well-documented attributes in the DQ literature. Second, data inaccuracy can be simulated and its impact is measureable in an objective manner, something which is impossible for other more subjective dimensions of data quality. Third, no research is available that investigates the impact of data accuracy problems upon segmentation performance in a direct marketing setting.

In the KDD process, DQ results from choices made in the data-preparation phase (Fayyad et al., 1996). The data-preparation phase consists of the following sub-steps: (i) data selection, aimed at the selection of relevant information while minimizing noise; (ii) data preprocessing; and (iii) data reduction and transformation. Previous research mainly focuses on strategies to improve DQ within the preprocessing and transformation phases and discusses topics such as feature selection (Kim, Street, Russell, & Menczer, 2005), re-sampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), outlier detection (Van Gestel et al., 2005), the discretization of continuous attributes (Berka & Bruha, 1998), and the mapping and scaling of categorical variables (e.g., Zhang, Zhang, & Yang, 2003). However, the data selection phase has a significant impact on DQ, and thus on its impact on segmentation performance. Problems that may arise here are missing values, outdated data values and inaccurate data (Even, Shankaranarayanan, & Berger, 2010). Several authors investigate the merits of missing value imputation in the KDD process (e.g., Batista & Monard, 2003; Brown & Kros, 2007), but research on the direct impact of inaccurate and outdated data on the performance of segmentation models for direct marketing is not available. In summary, the objectives of this study are to assess the impact of data accuracy problems on the quality of customer segmentation approaches and to uncover whether some segmentation techniques are more resistant to these problems than others.

The paper has the following organization. Section 2 describes the segmentation approaches and the evaluation metric. Section 3 describes the experimental setup of this study, and Section 4 describes the impact of data accuracy problems on the segmentation performance for real-life direct marketing data. Section 5 revises the managerial implications of the impact of poor data accuracy, and finally Section 6 summarizes the results and offers suggestions for further research.

2. Methodology

2.1. Segmentation approaches

This paragraph details the segmentation techniques employed throughout this study. The impact of data accuracy issues on the performance of RFM analysis, decision trees, and logistic regression is evaluated. These techniques are chosen given their popularity in business and their extensive use in the direct marketing literature (e.g. McCarty & Hastak, 2007).

2.1.1. RFM analysis

RFM analysis originates from the practice of direct marketing in catalog sales companies in the 1960s (Blattberg, Kim, & Neslin, 2008). Such analysis prescribes a segmentation of customers in the company's database based on past behavior (Bitran & Mondschein, 1996; Hughes, 2000). Three variables represent this past behavior: (1) the period elapsed since the customer's last purchase (i.e., *recency*; R), (2) the number of purchases in an arbitrary period in the past (i.e., *frequency*; F), and (3) the total monetary value of past purchases (i.e., *monetary value*; M). Customers who purchased recently, frequently, and spent large amounts of money are more likely to respond to mailings and therefore represent more attractive prospects for future marketing campaigns. The objective of RFM analysis is to identify a segment of customers who have a high probability of responding to a marketing campaign. By focusing on these customers and avoiding spending resources on customers who would not have responded anyway, a company makes its marketing actions more targeted.

Although several RFM model variations exist in literature, this research relies on the RFM procedure proposed by Hughes (1996, 2000). In this approach, the RFM variables transform into discrete codes that take values in the set {1, 2, 3}. Thus, every customer receives one code for every RFM variable. In detail, the attributed codes come from the following procedure. The first step summarizes customers' purchase histories in RFM variables. This historical purchase information is easily derivable from a company's transactional database in which all past customer purchases before a direct mail campaign are recorded. The second step sorts customers on recency, divides them into three equal customer groups, and assigns them one of the three discrete codes. For example, the 33.33% of customers who purchased most recently receive code 3. The third step, within each recency group, sorts customers on their purchase frequency, attributing the codes in a similar way. Finally, the fourth step sorts each frequency group on monetary value and again attributes the codes to the subsets. From this procedure, every customer receives three codes that indicate membership to one of 27 ($3 \times 3 \times 3$) groups of equal size. The analyst then concatenates different codes for recency, frequency, and monetary value of customers and uses them to rank the customers. This ranking allows marketing decision makers to formulate a set of rules that helps to identify customers who should be targeted in a direct mail campaign. In real-life, several criteria can guide the selection of the correct number of customers to target. For example, the analyst could take an arbitrary proportion of the customer file or determine an optimal number of target customers to maximize the company's profits (Blattberg et al., 2008).

On the one hand, RFM analysis is a popular approach in database marketing because of its simplicity and reasonable performance. The relationship between the response and the RFM variables is not assumed to be monotonic or known in advance (McCarty & Hastak, 2007). On the other hand, several important disadvantages exist. First, the discretization procedure introduces a loss of explanatory information. Second, the customer coding procedure is arbitrary. Depending on the case, more or fewer categories might be more appropriate. For instance, depending on the budget, finer or cruder RFM coding schemes could be employed (Blattberg et al., 2008; Hughes, 1996). Finally, the technique is not suited to add other features that might relate to a customer's future response behavior (Blattberg et al., 2008).

2.1.2. Decision trees

Because of their combination of simplicity, transparency, and strong performance, decision trees are a popular modeling technique in business (Duda, Hart, & Stork, 2001). In the context of customer segmentation, the analyst constructs a decision tree by subsequently splitting the entire group of heterogeneous customers into smaller and more homogeneous subsets of customers. The top of the decision tree, or the node in which all customers enter the model, is the “root node”. Splits are made into two or more child nodes according to the values of one or more independent variables. In particular, the algorithm identifies a

splitting criterion that represents the best possible separation between responding and non-responding customers. This process repeats until some criterion is met that indicates that the tree has been fully grown. Nodes within the tree are "internal nodes", and final child nodes are known as "terminal nodes".

This study considers the chi-square automatic interaction detection (CHAID) decision tree (Kass, 1980), a decision tree algorithm that has been used in several studies for the purpose of customer database segmentation (Galguera, Luna, & Mendez, 2006; McCarty & Hastak, 2007). A CHAID decision tree subsequently splits customers into subgroups based on the chi-square statistic, identifying which variable splits the data best and whether further splitting induces a statistically significant improvement.

CHAID has some similarities to the RFM approach. First, as with RFM analysis, CHAID also represents a strategy to group customers into smaller groups. CHAID enables terminal nodes and their respective rules to be evaluated and ranked in terms of the response rate to come up with desirable segments. Second, the nature of the relationship between response and the RFM variables is not assumed to be known in advance or subject to a monotonicity constraint. The most important difference with RFM is that CHAID is more flexible in terms of input variables. In addition to RFM, other relevant variables exist, such as demographic and psychographic variables, textual information (Coussement & Van den Poel, 2008), or variables related to the social network of the customer (Macskassy & Provost, 2007).

Because of the widespread adoption and favorable performance of CHAID in previous studies on customer segmentation in the context of direct mail (Levin & Zahavi, 2001; McCarty & Hastak, 2007), this study assesses its robustness to data accuracy problems.

2.1.3. Logistic regression

Logistic regression is a well-known technique for predicting a binary dependent variable (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). Thus, logistic regression is a suitable technique for discriminating between responders and non-responders. In running a logistic regression analysis, the likelihood function is produced and maximized to achieve an appropriate fit to the data (Hosmer & Lemeshow, 2000). A logistic regression model outputs a probability of future response to each and every customer in the database based on the historic customer characteristics. Logistic regression is a popular segmentation technique and occurs frequently in predictive marketing settings mainly for three reasons. Logit modeling (1) is a conceptually simple technique (Bucklin & Gupta, 1992), (2) offers a closed-form solution for the posterior probabilities rather than creating discrete groups of individuals as in RFM and decision trees, and (3) provides quick and robust results compared to the other techniques (Neslin et al., 2006).

2.2. Evaluation metric

After applying a segmentation approach, the direct marketer wants to objectively evaluate the performance of a proposed methodology. To accommodate an objective measurement of segmentation performance, the marketer uses different parts of the data to create the model and uses others to objectively measure its performance. Practically, a segmentation model is first built on a training set. Afterwards, the model is validated using previously unseen data, the test set (Witten & Frank, 2000). Evaluating the performance of a segmentation approach on a separate test set is essential to verify whether the segmentation model performs well for customers who were not included in the data set used to create the model.

The evaluation metric used in the current research uses is gain percentage (McCarty & Hastak, 2007). The gain percentage corresponds to the percentage of actual responders that would be reached by contacting a subset of the whole population. For example, if a segmentation approach targets 10% of the people who are most likely to respond to a certain mailing, the percentage of actual responders included in this specific subset reflects the gain percentage. Note that if the direct marketer selects 10% of the cases at random, he or she expects to capture 10% of the actual responders from the total data set (McCarty & Hastak, 2007). Because segmentation approaches should outperform random guessing, gain percentages surpassing 10% of the actual respondents are desired. The percentage of customers considered to calculate the gain percentage reflects the file depth. Because different file depths might be relevant in practice, in function of particular business settings and desired project outcomes, the experiments in this study consider several alternative levels of depth. In particular, gain percentages are calculated for 10 levels of file depth, ranging from 5% to 50% with incremental steps of 5%. File depths of 5% and 10% are the most commonly used in practice, because customer target sizes are restricted as a result of limited company budgets.

3. Experimental setting

This study assesses the sensitivity of the three considered analytical methods for customer segmentation to the different levels of data accuracy using two empirical direct marketing data sets provided by the Direct Marketing Educational Foundation (DMEF). The data are available through the educators section on the foundation's website (see http://www.directworks.org/). The first data set originates from a multi-division mail-order catalog business and contains information on 96,551 customers of which 2.46% responded to the mailing. The second data set contains 99,200 members of a non-profit organization representing a response rate of 27.43%. As such, these two data sets are an ideal framework to test the sensitivity of the three segmentation techniques to data accuracy problems at varying levels of response (low: 2.46% versus high: 27.43%).

These data sets are typical direct marketing data sets. The purpose of the marketer is to find a good segmentation model that discriminates well between responders and non-responders to a catalog mailing. The data sets contain actual response behavior (i.e., whether or not a customer responded to the recent mailing) and information on the elapsed time since the last purchase, the total number of purchase occasions, and the total amount the customer spent. For a fair comparison among the discrimination power of the RFM method, logistic regression, and CHAID, this study limits the information to the previously described independent variables, similar to other segmentation studies (e.g., McCarty & Hastak, 2007).

To investigate the impact of data accuracy on the performance of alternative segmentation techniques, the original DMEF data sets are manipulated to mimic the presence of different levels of data inaccuracies in the data selection phase. As it is not possible to obtain an indication of the real-life data accuracy level, this research paper defines a variable (i.e. recency, frequency, or monetary value) as accurate if the values of this variable equal the data values of the original data set. Various levels of data inaccuracy are created by randomly selecting a given percentage of the values of a variable and assigning them a randomly generated value that lies within the range of the original variable. This process repeats for every variable in the data set. Several levels of data accuracy problems are investigated. This study defines 11 different levels of data inaccuracy, with percentages of randomly replaced values equal to 0% (i.e., the original data set), 1%, 2%, 3%, 4%, and 5%, and from 10% to 50% in steps of 5%.

To methodologically correctly compare the different segmentation approaches, this study applies the cross-validation procedure as suggested by Hastie, Tibshirani, and Friedman (2001) and as implemented in a variety of response modeling settings (e.g., Cui, Wong, & Lui, 2006). In particular, this study implements a fivefold cross-validation. That is, the study splits the complete data set into five subsets of equal size. Iteratively, one part is left out for testing, while the remaining four parts are used for training. In the end, each case in the data set is predicted once. In summary, the

Table 1
Parameter estimates. Note. [a] and [b] indicate coefficient significance levels. [a]: p<0.05; [b]: p<0.10.

File depth	Parameter estimates											
	β_0		β_1		β_2		β_3		β_4		β_5	
	Data set 1	Data set 2	Data set 1	Data set 2	Data set 1	Data set 2	Data set 1	Data set 2	Data set 1	Data set 2	Data set 1	Data set 2
5	[a] −2.0589	[a] −2.1389	[a] −0.0786	[a] −0.1138	0.0254	[a] −0.0522	[a] −0.0026	[a] −0.0041	[a] −0.0039	[a] −0.0022	[a] −0.0070	−0.0001
10	[a] −1.5021	[a] −1.5569	[a] −0.0894	[a] −0.0783	0.0031	[a] −0.0178	[a] −0.0020	[a] −0.0026	[a] −0.0014	[a] −0.0012	[a] −0.0043	[a] −0.0019
15	[a] −1.1934	[a] −1.2329	[a] −0.0906	[a] −0.0453	[a] −0.0499	[a] −0.0253	[a] −0.0020	[a] −0.0015	0.0004	[a] −0.0019	[a] −0.0010	[a] −0.0006
20	[a] −1.0039	[a] −1.0157	[a] −0.0728	[a] −0.0601	[a] −0.0331	[a] −0.0551	[a] −0.0012	[a] −0.0011	0.0001	[a] −0.0008	0.0000	[a] 0.0010
25	[a] −0.8533	[a] −0.8545	[a] −0.0666	[a] −0.0352	[a] −0.0380	[a] −0.0509	[a] −0.0007	[a] −0.0009	−0.0003	[b] −0.0003	0.0001	[a] 0.0007
30	[a] −0.7257	[a] −0.7192	[a] −0.0539	[a] −0.0309	[a] −0.0658	[a] −0.0492	[a] −0.0008	[a] −0.0006	0.0001	0.0000	[a] 0.0006	[a] 0.0006
35	[a] −0.6189	[a] −0.6035	[a] −0.0525	[a] −0.0204	[a] −0.0490	[a] −0.0236	[a] −0.0007	[a] −0.0006	0.0004	0.0000	0.0001	0.0001
40	[a] −0.5367	[a] −0.5045	[a] −0.0326	[a] −0.0120	[a] −0.0236	[a] −0.0099	[a] −0.0004	[a] −0.0005	0.0002	−0.0001	[a] −0.0005	[a] −0.0005
45	[a] −0.4626	[a] −0.4171	[a] −0.0277	[a] −0.0083	[b] −0.0055	[a] −0.0198	[a] −0.0003	[a] −0.0004	0.0001	[a] −0.0003	[a] −0.0010	[a] −0.0002
50	[a] −0.3948	[a] −0.3444	[a] −0.0281	[a] −0.0136	−0.0033	[a] −0.0265	[a] −0.0004	[a] −0.0003	0.0001	[a] −0.0004	[a] −0.0005	[a] 0.0001

cross-validated performance better reflects the true capabilities of a segmentation approach in validating unseen data because it reduces the variability of the validation results by building five models on slightly different training sets compared with a random one-shot split of the original data set into a training and a test set (Malthouse, 2001). This study reports the cross-validated gain percentage, which is the average of the gain percentages within the five folds in the cross-validation.

4. Results

This study attempts to investigate the influence of different levels of data inaccuracy on the performance of three techniques for customer segmentation in terms of gain percentages. The study evaluates gain percentage performance at several levels of desired file depth. Thus, the experiments investigate the influence of two factors on gain performance, for each of the 10 levels of file depth. The combination of segmentation algorithms and levels of data accuracy allow for ten two-factor analysis of variance models without cell replication. In general, these models help investigate relationships between a dependent variable and one or more explanatory factors. These models also prove useful in the evaluation of various aspects of KDD modeling, such as classification or pruning methods (Curram & Mingers, 1994; Han, Chandler, & Liang, 1996), multiple data sets (Curram & Mingers, 1994; Mingers, 1989), the measurement scale and correlation structure of variables (Han et al., 1996), and the level of missingness and imputation methods (Brown & Kros, 2007). In these studies, the dependent variable is often a measure of classification accuracy (percentage of miss-classifications, percentage of correctly predicted holdout cases) or the root mean squared error.

The main problem with the current setting is that no degrees of freedom are available for estimating an experimental error, because only one measurement of gain percentage is available for each combination of segmentation type and data accuracy. Solutions are available either by assuming that the two factors in the analysis do not interact (Kutner, Nachtsheim, Neter, & Li, 2005) or by treating one of the factors as a numeric (continuous) variable (Ramsey & Schafer, 2002). Because experiments on the data show that segmentation type and data accuracy level may indeed interact, the current study opts for the second solution and treats the data accuracy as a numerical variable. This manipulation is justified because the meaning of each data accuracy category is unambiguous and the distance between the different levels of data accuracy is measurable in accordance with the values assigned to the levels. Although the distances are smaller between the higher levels of data accuracy than between the lower levels, the power of information obtained from an analysis in which the data accuracy is treated numerically strongly outweighs the possible technical inaccuracies (Rea & Parker, 2005). Using a combination of a categorical and a continuous explanatory variable, the segmentation performance is analyzed with general linear models (Dobson & Barnett, 2008), and results are interpreted as in ordinary regression. The models help disentangle the performance for the three segmentation techniques under consideration (RFM, logistic regression, and CHAID decision trees) for different levels of data accuracy.

Practically, this study defines dummy regressors *logit* and *rfm* for the corresponding segmentation types, which are equal to 1 if the segmentation type is, respectively, logistic regression or RFM and equal to 0 if otherwise. Thus, for CHAID, both dummy regressors are equal to 0. The continuous variable *rand* indicates the second factor under consideration (i.e., the data accuracy), represented by the percentage of randomized values for each variable in the data set. The study also includes interaction terms in the model, defined as the product of a segmentation type dummy regressor and the continuous data accuracy variable. Interaction allows the data accuracy effects to differ over the segmentation types. Finally, the dependent variable in

Table 2
Segmentation technique comparison: statistical test summary (*Algorithm 1* vs. *Algorithm 2*). Note. '<' designates that *Algorithm 1* is significantly outperformed by *Algorithm 2* (p<0.05); '.' indicates no significant difference.

File depth	Logistic regression vs. CHAID H_0: $\beta_1=0$						RFM vs. CHAID H_0: $\beta_2=0$						Logistic regression vs. RFM H_0: $\beta_1=\beta_2$					
	Data set 1			Data set 2			Data set 1			Data set 2			Data set 1			Data set 2		
	t	p	Sign.	t	p	Sign.	t	p	Sign.	t	p	Sign.	*F*	p	Sign.	*F*	p	Sign.
5	−2.417	0.020	<	−5.129	0.000	<	1.486	0.145	.	−7.603	0.000	<	11.147	0.002	<	7.255	0.010	<
10	−4.604	0.000	<	−7.763	0.000	<	0.290	0.774	.	−6.002	0.000	<	24.519	0.000	<	36.457	0.000	<
15	−6.326	0.000	<	−9.769	0.000	<	−6.592	0.000	<	−7.946	0.000	<	8.574	0.006	<	13.983	0.001	<
20	−4.630	0.000	<	−6.865	0.000	<	−5.123	0.000	<	−34.287	0.000	<	7.124	0.011	<	0.334	0.566	.
25	−4.163	0.000	<	−8.824	0.000	<	−5.231	0.000	<	−29.654	0.000	<	3.753	0.060	.	17.754	0.000	>
30	−5.014	0.000	<	−12.087	0.000	<	−11.548	0.000	<	−45.292	0.000	<	1.684	0.202	.	61.988	0.000	>
35	−6.594	0.000	<	−7.849	0.000	<	−11.313	0.000	<	−14.848	0.000	<	0.264	0.610	.	1.377	0.248	.
40	−7.350	0.000	<	−6.328	0.000	<	−8.173	0.000	<	−10.108	0.000	<	5.566	0.023	<	1.352	0.252	.
45	−6.301	0.000	<	−3.943	0.000	<	−1.916	0.063	.	−16.910	0.000	<	28.102	0.000	<	26.800	0.000	>
50	−5.136	0.000	<	−4.833	0.000	<	−1.053	0.299	.	−40.213	0.000	<	23.999	0.000	<	20.468	0.000	>

Table 3
Influence of data accuracy problems on segmentation technique performance: statistical test summary. Note. '—' designates that parameters have a significantly negative relationship to gain percentage ($p<0.05$); '.' shows no significant relationship to gain percentage.

File depth	CHAID H_0: $\beta_3=0$						Logistic regression H_0: $\beta_3+\beta_4=0$						*RFM* H_0: $\beta_3+\beta_5=0$					
	Data set 1			Data set 2			Data set 1			Data set 2			Data set 1			Data set 2		
	t	p	Sign.	t	p	Sign.	*F*	p	Sign.	*F*	p	Sign.	*F*	p	Sign.	*F*	p	Sign.
5	−7.270	0.000	—	−21.024	0.000	—	59.520	0.000	—	55.795	0.000	—	399.759	0.000	—	392.438	0.000	—
10	−7.323	0.000	—	−36.988	0.000	—	29.562	0.000	—	84.577	0.000	—	348.286	0.000	—	3232.020	0.000	—
15	−6.205	0.000	—	−39.654	0.000	—	15.342	0.000	—	467.286	0.000	—	148.704	0.000	—	658.195	0.000	—
20	−2.906	0.006	—	−21.861	0.000	—	6.074	0.018	—	35.541	0.000	—	111.898	0.000	—	1.503	0.228	.
25	−2.148	0.038	—	−17.519	0.000	—	4.762	0.035	—	71.134	0.000	—	51.364	0.000	—	4.887	0.033	—
30	−4.192	0.000	—	−15.588	0.000	—	6.351	0.016	—	54.809	0.000	—	40.313	0.000	—	102.518	0.000	—
35	−4.800	0.000	—	−14.722	0.000	—	2.507	0.121	.	44.750	0.000	—	370.324	0.000	—	118.454	0.000	—
40	−4.754	0.000	—	−13.848	0.000	—	2.683	0.109	.	87.332	0.000	—	236.849	0.000	—	1661.980	0.000	—
45	−3.455	0.001	—	−17.718	0.000	—	4.112	0.049	—	96.757	0.000	—	131.881	0.000	—	405.499	0.000	—
50	−3.272	0.002	—	−10.186	0.000	—	4.675	0.037	—	41.913	0.000	—	273.117	0.000	—	20.526	0.000	—

the analysis, *gain*, indicates the gain percentage or the percentage of customers who respond to the catalog mailing for a given file depth. Generalized for each level of file depth, this regression model appears as follows:

$$\log(gain_i) = \beta_0 + \beta_1 logit_i + \beta_2 rfm_i + \beta_3 rand_i + \beta_4 logit_i \times rand_i + \beta_5 rfm_i \times rand_i + \varepsilon_i, \quad (1)$$

where the indicator i denotes the file depth and ranges from 5% to 50% by increments of 5%. The parameters β_0, β_1, and β_2 reflect the performance of the segmentation techniques (in the case of optimal data accuracy and, thus, if $rand_i=0$), and β_3, β_4 and β_5, measures how the gain percentage varies for each additional decrease in data accuracy (i.e., for every additional increase in the percentage of randomized data values in the data set) and for each segmentation type.

Initial data investigation indicates outlying behavior for the logistic regression when $rand_i=0$. Therefore, this study deploys the regression analysis using White's robust standard errors (Hill, Griffiths, & Lim, 2012; Verbeek, 2008). Table 1 provides the parameter estimates of the models for both data sets.

4.1. The performance of segmentation techniques under optimal data accuracy

A first consideration is to check whether the three segmentation techniques perform differently without data accuracy problems. When the data set is not affected by a loss of data accuracy ($rand_i=0$), the parameters β_0, β_1, and β_2 show how the segmentation techniques differ in terms of gain percentage and, thus, segmentation performance. Table 2 shows the results of the relevant significance tests.

Overall, three conclusions emerge. First, the use of CHAID over logistic regression and RFM is recommended when no data inaccuracy is known or suspected. Specifically, the estimated (log-transformed) gain percentage is always significantly higher for a segmentation based on CHAID than for one based on the logistic regression model, at all levels of file depth (H_0: $\beta_1=0$ is rejected and β_1 is consistently negative) independently of the level of the response rate in the data set. Moreover, a comparison between RFM and CHAID (H_0: $\beta_2=0$) on both data sets suggests a superior performance for CHAID confirming the findings of McCarty and Hastak (2007). Only for the data set with the low response rate (data set 1), no differences exist between RFM and CHAID segmentation at the lowest (5% and 10%), and highest (45% and 50%) levels of file depth. Second, the preference of using RFM over the logistic regression approach (H_0: $\beta_1=\beta_2$) is confirmed for the low response rate data set, which suggests that the RFM approach captures at least as much respondents as the logistic regression segmentation. Third, the comparison between logistic regression and RFM for the high response rate data set (data set 2) is summarized along a two-tier structure. For small file depths (5%–15%), the overarching performance of RFM over logistic regression is consistent with the conclusion for the low response rate data set. However, logistic regression becomes the preferred segmentation technique for medium to high levels of file depth as it identifies responding customers at least as well as the RFM segmentation.

4.2. Impact of data inaccuracy on segmentation technique performance

This section investigates the impact of inaccurate data on segmentation performance. The objective is to assess the impact of imperfect data on the segmentation performance of the three techniques under consideration. A first question involves whether a decrease in the level of data accuracy (i.e., an increase of the percentage of randomized values in the data set) has a harmful impact on the gain percentage for the three segmentation techniques. A second question is whether some segmentation techniques are more robust against changes in data accuracy than others.

First, consider the influence of data accuracy problems on segmentation performance for both data sets. The robustness of the segmentation types toward changes in data accuracy is quantified by means of the slope component in the relationship between gain percentage and the level of data accuracy. Table 3 summarizes the significance *t*-test and *F*-test results for linear restrictions.

Overall, these results demonstrate that all three segmentation techniques are sensitive to problems with data accuracy. The decreases in performance due to decreased data accuracy are significant across all file depths for CHAID (H_0: $\beta_3=0$), logistic regression (H_0: $\beta_3+\beta_4=0$), and RFM (H_0: $\beta_3+\beta_5=0$) segmentation.

The second question involves comparing segmentation techniques with respect to their sensitivity to data accuracy problems, and their overall performance under the assumption of imperfect data quality. Note that in terms of Eq. (2), for every segmentation technique, the relationship between data accuracy ($rand_i$) and gain percentage ($gain_i$) is characterized by an intercept and a slope component. Hence, differences in performance between segmentation techniques can occur on three levels: (i) the intercept level, if their gain percentage performance differs under optimal data accuracy (i.e., $rand_i=0$), (ii) the slope level, if their sensitivity to deteriorating data accuracy differs, or (iii) the overall performance level, if their gain percentage levels differ due to an interplay of levels (i) and (ii). Even though Table 1 analyzed the differences in terms of the intercept, Table 4 presents the significance test results from comparing the segmentation techniques' performance in terms of sensitivity to (columns 1–3) and overall performance under data accuracy problems (column 4). Note that results in column 4 are obtained using the *average distance measure* proposed by Rogosa (1980).

Table 4
Comparison of segmentation techniques' sensitivity to data accuracy problems: statistical test summary. Note 1 (column 1–3). '<' designates that *Algorithm 1* is significantly outperformed by *Algorithm 2* ($p<0.05$); '.' indicates no significant difference. Note 2 (column 4). coding scheme: A1 = CHAID, A2 = Logistic and A3 = RFM. For instance, "A1 ≥ A3" ("A1 ≤ A3") indicates that algorithm A1 has a higher, yet not statistically different (lower) average log(*gain*) value than A3. "A1 > A3" ("A1 < A3") indicates that Algorithm A1 has a statistically higher (lower) average log(*gain*) value ($p<0.05$).

File depth	Logistic regression vs. CHAID $H_0: \beta_4 = 0$						RFM vs. CHAID $H_0: \beta_5 = 0$						Logistic regression vs. RFM $H_0: \beta_4 = \beta_5$						Overall segmentation performance	
	Data set 1			Data set 2			Data set 1			Data set 2			Data set 1			Data set 2			Data set 1	Data set 2
	t	p	Sign.	t	p	Sign.	t	p	Sign.	t	p	Sign.	*F*	p	Sign.	*F*	p	Sign.		
5	−4.306	0.000	<	−2.574	0.014	<	−11.810	0.000	<	−0.298	0.767	.	10.282	0.003	>	6.074	0.018	<	A1 ≥ A3 ≥ A2	A1 > A3 > A2
10	−2.065	0.046	<	−2.822	0.008	<	−9.837	0.000	<	−18.062	0.000	<	15.908	0.000	>	3.135	0.084	.	A1 ≥ A3 ≥ A2	A1 ≥ A3 ≥ A2
15	0.695	0.491	.	−11.658	0.000	<	−2.481	0.018	<	−6.706	0.000	<	8.031	0.007	>	51.470	0.000	<	A1 > A3 ≥ A2	A1 ≥ A3 > A2
20	0.157	0.867	.	−2.385	0.022	<	0.036	0.971	.	8.529	0.000	>	0.030	0.863	.	27.833	0.000	<	A1 > A3 > A2	A1 > A3 > A2
25	−0.494	0.624	.	−1.837	0.074	.	0.238	0.813	.	9.814	0.000	>	0.604	0.442	.	47.524	0.000	<	A1 > A3 > A2	A1 > A3 ≥ A2
30	0.227	0.822	.	0.335	0.739	.	3.254	0.002	>	14.661	0.000	>	3.644	0.064	.	48.461	0.000	<	A1 > A2 ≥ A3	A1 > A2 ≥ A3
35	1.562	0.126	.	0.180	0.858	.	0.695	0.491	.	0.961	0.343	.	1.957	0.170	.	0.195	0.661	.	A1 > A2 ≥ A3	A1 > A2 ≥ A3
40	1.622	0.113	.	−1.497	0.143	.	−5.227	0.000	<	−12.237	0.000	<	37.920	0.000	>	37.298	0.000	>	A1 > A2 ≥ A3	A1 > A2 ≥ A3
45	0.443	0.660	.	−3.562	0.001	<	−6.868	0.000	<	−5.173	0.000	<	41.141	0.000	>	0.611	0.439	.	A1 > A3 ≥ A2	A1 > A2 > A3
50	0.443	0.660	.	−3.445	0.001	<	−3.839	0.000	<	2.970	0.005	>	14.757	0.000	>	21.533	0.000	<	A1 > A3 > A2	A1 > A2 ≥ A3

The following conclusions emerge from Table 4. First, CHAID tends to be significantly less or equally vulnerable to the introduction of data accuracy problems than logistic regression (H_0: $\beta_4 = 0$; column 1). Second, the sensitivity differences between RFM and CHAID (H_0: $\beta_5 = 0$; column 2) are less consistent. For both response data sets, CHAID is less sensitive to data accuracy problems for smaller mailing depths (5%–15%), while no clear-cut conclusions emerge for medium and high file depth levels. However, the overall average segmentation performance of CHAID for all file depth levels, including the medium and high levels 20% till 50%, stays significantly superior to the average segmentation performance of logistic regression and RFM (column 4). Third, the sensitivity of logistic regression and RFM to the data inaccuracies depends on the level of the response rate. Although RFM is the preferred segmentation technique for the low response data set for all file depth levels under optimal data accuracy (see Table 2), the sensitivity results show that logistic regression is never more vulnerable to data inaccuracies than RFM. This leads to a shuffle in the preferences where the choice between logistic regression and RFM becomes interchangeable for medium and high file depths levels under data accuracy problems. For the high response rate data set, RFM has a tendency to be less or equally sensitive to the deterioration of data accuracy than logistic regression, but this leads to no shifts in the segmentation choice as seen under the optimal data accuracy setting (see Table 2).

5. Managerial implications

In the presence or suspicion of inaccurate data, a marketer should take into account two implications from this study. The first pertains to the importance of implementing an information quality system that automatically assesses and benchmarks the customer information accuracy in the data warehouse. When the number of observations with inaccurate attributes is relatively small and/or random, the probability that the poor data quality will heavily affect the segmentation performance is small. However, when the pattern in the inaccurate observations is relatively large and/or systematic in nature, the segmentation results may be heavily distorted (Feelders et al., 2000). Consequently, given the findings of this research study that reveal the deteriorating impact of data accuracy problems on segmentation performance, companies are encouraged to introduce an alert system that warns them when serious negative deviations of the acceptable data accuracy level occur. Prior academic research proposes several information quality alert systems (Delone & McLean, 1992; Lee, Strong, Kahn, & Wang, 2002; Wang & Strong, 1996; Zmud, 1978), as do practitioner's solutions created by companies such as AT&T (Redman, 1992) and Diamond Technology Partners (Matsumura & Shouraboura, 1996).

A second implication pertains to the importance of carefully choosing the appropriate segmentation technique. Fig. 1 shows for both data sets and for the different file depth levels to what extent the gain performance for a segmentation technique deteriorates for each additional decrease in data quality.

Given the multiplicative nature of the model, the figures on the Y-axis can be interpreted as "correction factors" that indicate how gain percentage on average suffers from a unit decrease in data accuracy. Consider, for example, the CHAID segmentation, a file depth of 5%, and a gain level equal to g_0. In the model for the low response data set, the estimate of the slope parameter (β_3) is −0.0026 (see Table 1), and $\exp(\beta_3) = 0.9974$. For a unit increase in the level of data inaccuracy, the new percentage gain g_1 is estimated to be $0.9974 \times g_0$. This gain implies that the model identifies fewer true respondents. Note that the distortions in the percentage gain are larger for limited file depths, and the effect decreases when a larger portion of the file is mailed. The correction factors are generally close to one (99% and more), reflecting fairly small reductions in percentage gain. However, small changes in captured response may result in significant cost differences with large customer files (McCarty & Hastak, 2007).

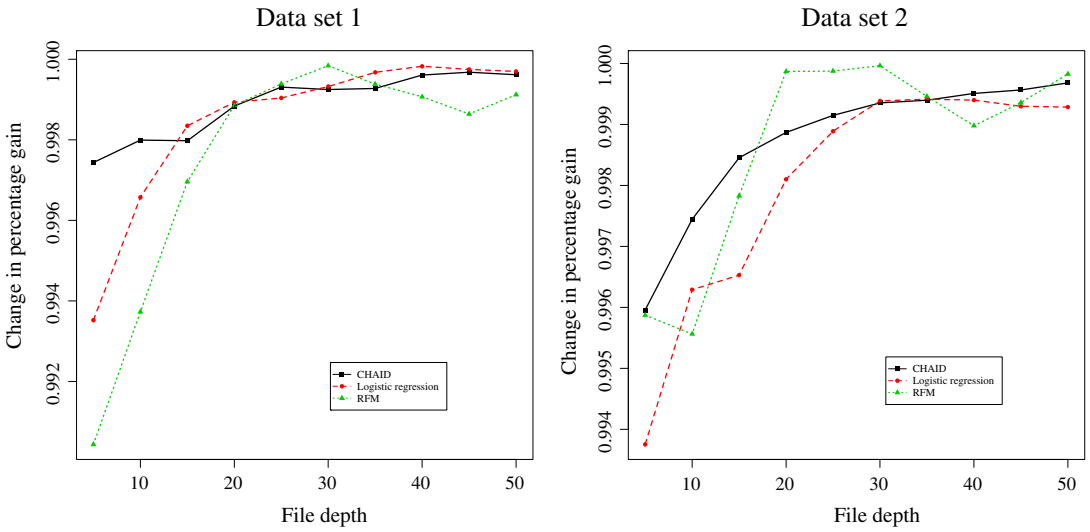

Fig. 1. Sensitivity of gain percentage to changes in data accuracy for various file depth levels.

The analyses show that observed, suspected or expected presence of data inaccuracy should influence the marketer's choice between segmentation algorithms as different techniques are affected differently. For instance, for low levels of file depth (5% and 10%) in the low response data set. RFM and CHAID have similar performance, and they both outperform the logistic regression segmentation. However, in terms of the *reduction* in percentage captured when the data accuracy decreases, the performance of CHAID decreases slower with deteriorating data accuracy than the performance of RFM and logistic regression. In sum, the CHAID decision tree segmentation is most robust against changes in data accuracy for this specific example.

6. Conclusion and directions for further research

The goal of this research is to investigate the impact of the level of data accuracy on the performance of three segmentation algorithms. Using two real-life direct marketing data sets, this study treats a specific number of observations as inaccurate by subsequently replacing a fraction of the values of a variable with randomly generated values from the variable range. The general linear model demonstrates that in the absence of (additional) data accuracy issues, CHAID is the superior choice compared to RFM and logistic regression, while RFM is preferred over logistic regression for the low response rate data set, and for small file depth levels having a high response incidence. The introduction of data accuracy problems deteriorates the performance of all three techniques. Overall, CHAID is considered as the best choice for both low and high response data sets under data accuracy problems, because (1) CHAID is less sensitive to increasing problems in data accuracy than logistic regression or RFM which enforces the superiority of CHAID, or (2) under the condition that CHAID is more sensitive to data inaccuracies than RFM, CHAID's performance stays much stronger than RFM which makes CHAID a safe choice for direct marketing analysts under all conditions.

In addition, the sensitivity towards data accuracy issues between RFM and logistic regression depends on the height of the response rate, i.e. logistic regression seems less or equally sensitive to data inaccuracies in a low response data set, while the inverse is true for settings with high response rates. This leads to a minor change in the preference as observed under the optimal data accuracy situation; logistic regression becomes a competitor of RFM for medium to high file depths levels for low response data sets.

Several avenues for further research are identified. As this study compares the impact of the three most common segmentation approaches in a direct marketing setting, further research should investigate the sensitivity of other segmentation models. In addition, this study uses the same information for the different algorithms (i.e., RFM). This limitation enables a fair comparison among the different segmentation approaches and helps inform the conclusions. However, in practice, direct marketers are not constrained to these three variables.

Thus, research should incorporate other transactional and sociodemographic information in CHAID and logistic regression to increase the quality of discrimination between responders and non-responders. Finally, we truly belief that considerable research opportunities exist to establish research methods that mimic the problem of data accuracy. The merits and drawbacks of these new methods could be compared with this study's implementation of the data inaccuracy simulation.

References

Akaah, I. P., Korgaonkar, P. K., & Lund, D. (1995). Direct marketing attitudes. *Journal of Business Research*, *34*(3), 211–219.

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: upcoming trends and challenges. *Journal of the Operational Research Society*, *60*, 16–23.

Batista, G., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533.

Berka, P., & Bruha, I. (1998). Empirical comparison of various discretization procedures. *International Journal of Pattern Recognition and Artificial Intelligence*, *12*(7), 1017–1032.

Bitran, G. R., & Mondschein, S. V. (1996). Mailing decisions in the catalog sales industry. *Management Science*, *42*(9), 1364–1381.

Blattberg, R. C., Kim, B. -D., & Neslin, S. A. (2008). *Database marketing: Analyzing and managing customers*. New York: Springer.

Bose, I., & Mahapatra, R. K. (2001). Business data mining — A machine learning perspective. *Information Management*, *39*(3), 211–225.

Brown, M. L., & Kros, J. F. (2007). A comparison of imputation methods in the presence of imprecise data when employing a neural network sigmoid function. *International Journal of Business Intelligence and Data Mining*, *2*(3), 292–310.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation — An integrated modeling approach. *Journal of Marketing Research*, *29*(2), 201–215.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research*, *16*, 321–357.

Cortinas, M., Chocarro, R., & Villanueva, M. L. (2010). Understanding multi-channel banking customers. *Journal of Business Research*, *63*(11), 1215–1221.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information Management*, *45*(3), 164–174.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781–800.

Cui, G., Wong, M. L., & Lui, H. -K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612.
Curram, S. P., & Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis — An empirical comparison. *The Journal of the Operational Research Society*, *45*(4), 440–450.
Delone, W. H., & McLean, E. R. (1992). Information system success: The quest for the dependent variable. *Information Systems Research*, *3*(1), 60–95.
Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd edition). Boca Raton: Chapman & Hall/CRC Press.
Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification.* New York: Wiley.
Even, A., Shankaranarayanan, G., & Berger, P. D. (2010). Evaluating a model for cost-effective data quality management in a real-world CRM setting. *Decision Support Systems*, *50*(1), 152–163.
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34.
Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information Management*, *37*(5), 271–281.
Galguera, L., Luna, D., & Mendez, M. P. (2006). Predictive segmentation in action — Using CHAID to segment loyalty card holders. *The International Journal of Marketing Research*, *48*(4), 459–479.
Han, I., Chandler, J. S., & Liang, T. P. (1996). The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods. *Expert Systems with Applications*, *10*(2), 209–221.
Han, J., & Kamber, M. (2006). *Data mining — Concepts and techniques.* San Francisco: Morgan Kauffman.
Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction.* New York: Springer-Verlag.
Haughton, D., & Oulabi, S. (1993). Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing*, *7*(3), 16–26.
Hill, R. C., Griffiths, W. E., & Lim, G. C. (2012). *Principles of econometrics.* Hoboken: Wiley.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd edition). Chichester: Wiley.
Hughes, A. M. (1996). *The complete database marketer: Second generation strategies and techniques for tapping the power of your customer database.* Chicago: Irwin.
Hughes, A. M. (2000). *Strategic database marketing.* New York: McGraw-Hill.
Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*(2), 119–127.
Kim, Y., Street, W. N., Russell, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, *51*(2), 264–276.
Ko, E., Kim, S. H., Kim, M., & Woo, J. Y. (2008). Organizational characteristics and the CRM adoption process. *Journal of Business Research*, *61*(1), 65–74.
Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th edition). New York: McGraw-Hill.
Lee, W. L., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information Management*, *40*, 133–146.
Levin, N., & Zahavi, J. (2001). Predictive modeling using segmentation. *Journal of Interactive Marketing*, *15*(2), 2–22.
Macskassy, S. A., & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, *8*, 935–983.
Malthouse, E. C. (2001). Assessing the performance of direct marketing scoring models. *Journal of Interactive Marketing*, *15*(1), 49–62.
Martínez-López, F. J., & Casillas, J. (2009). Marketing Intelligent Systems for consumer behaviour modelling by a descriptive induction approach based on genetic fuzzy systems. *Industrial Marketing Management*, *38*(7), 714–731.
Matsumura, A., & Shouraboura, N. (1996). Competing with quality information. *Proceedings of the Conference on Information Quality*. (pp. 72–86)MA: Cambridge University Press.
McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656–662.
Merrilees, B., & Miller, D. (2010). Brand morphing across Wal-Mart customer segments. *Journal of Business Research*, *63*(11), 1129–1134.
Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, *4*(2), 227–243.
Morganosky, M. A., & Fernie, J. (1999). Mail order direct marketing in the United States and the United Kingdom: Responses to changing market conditions. *Journal of Business Research*, *45*(3), 275–279.
Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. X., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211.
Ramsey, F., & Schafer, D. (2002). *The statistical sleuth: A course in methods of data analysis.* Belmont: Duxbury Press.
Rea, L. M., & Parker, R. A. (2005). *Designing and conducting survey research: A comprehensive guide.* San Francisco: Jossey-Bass.
Redman, T. C. (1992). *Data quality: Management and technology.* New York: Bantam Books.
Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*(2), 307–322.
Van Gestel, T., Baesens, B., Van Dijcke, P., Suykens, J., Garcia, J., & Alderweireld, T. (2005). Linear and nonlinear credit scoring by combining logistic regression and support vector machines. *Journal of Credit Risk*, *1*(4), 31–60.
Verbeek, M. (2008). *A guide to modern econometrics.* West Sussex: Wiley.
Viaene, S., Baesens, B., Van Gestel, T., Suykens, J. A. K., Van den Poel, D., Vanthienen, J., et al. (2001). Knowledge discovery in a direct marketing case using least squares support vector machines. *International Journal of Intelligent Systems*, *16*(9), 1023–1036.
Wand, Y., & Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95.
Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–34.
Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations.* San Francisco: Morgan Kaufmann.
Zahavi, J., & Levin, N. (1995). Issues and problems in applying neural computing to target marketing. *Journal of Direct Marketing*, *9*(3), 33–45.
Zahavi, J., & Levin, N. (1997). Applying neural computing to target marketing. *Journal of Direct Marketing*, *11*(1), 5–22.
Zhang, S. C., Zhang, C. Q., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381.
Zmud, R. (1978). Concepts, theories and technologies: An empirical investigation of the dimensionality of the concept of information. *Decision Sciences*, *9*(2), 187–195.