The Proof of the Pudding is in the Eating
Data Fusion: An Application in Marketing

Pascal van Hattum*and Herbert Hoijtink

Utrecht University, The Netherlands

November 18, 2008

* *Address for correspondence:* Pascal van Hattum, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. Box 80140, NL-3508 TC Utrecht, The Netherlands, Tel.: +31 30 2537983, Fax.: +31 30 2535797.
E-mail: p.vanhattum@uu.nl

The Proof of the Pudding is in the Eating
Data Fusion: An Application in Marketing

Abstract

Data fusion, or combining multiple data sets in one data set, is not a new concept. However, due to the increasing desire of differentiated direct marketing strategies, it is getting more popular in marketing. This paper shows how marketing information can be fused to a company's customer database. Using a real marketing application, two traditional data fusion methods, that are, polytomeous logistic regression and nearest neighbor algorithms, are compared with two model based clustering approaches. Finally, the results are evaluated using internal and external criteria.

Key words: Data Fusion, Differentiated Marketing, Nearest Neighbor, Logistic Regression, Model Based Clustering, Internal Validation, External Validation.

1 Introduction

In this paper, the following problem is addressed: a marketeer has knowledge and information about a small group of customers. Because the marketeer would like to have one-to-one communication with his customers, he would like to get the same knowledge and information for all the customers in his database. For reasons, like time, money, non-response, et cetera[1–3], obtaining the required knowledge and information using a single source questionnaire[4], is not an option. However, an attractive and practical solution is data fusion.

In this paper, data fusion is used in a marketing application. In the application, a Dutch energy supplier wants to send differentiated questionnaires to all the customers in the database. However, for only a fraction of the customers it is known what kind of differentiated questionnaire is preferred. Using data fusion techniques, the information about the preferred differentiated questionnaire becomes known for all the customers in the database.

The general problem of data fusion can best be illustrated using the schematic representation in Figure 1. In this representation data set $\mathbf{A}$ is the customer database and contains knowledge and information (represented by J items) from all customers. Data set $\mathbf{B}$ contains knowledge and information (represented by $J+1$ items) from a small group of customers. The first amount of knowledge and information (represented by the first J items) for a single customer is the same in each data set. However, from the small group of customers in data set $\mathbf{B}$ there is some additional knowledge and information, that is, item $J+1$. The goal of this paper is to fuse the extra knowledge and information in data set $\mathbf{B}$, that is, item $J+1$, to data set $\mathbf{A}$. As a result of this data fusion, the knowledge and information about item $J+1$ becomes 'known' for all customers in the database, data set $\mathbf{A}$.

Throughout the world, different terminologies are used for above described fusion or integration of two (or more) data sets, for example: multi-source imputation, data attribution, data fusion, statistical record linkage, statistical matching, micro data set merging, et cetera[3,5,6]. Since the 1980s a discussion has been going on about a clear and unambiguous terminology[6]. As in European marketing literature and practice, data fusion is the most commonly used term today[6–8], the terminology of data fusion will be used in this paper.

Not only is there some discussion about the terminology of integrating multiple data sets, there also is '*terminological confusion*'[3] about the different (statistical) procedures of data set integration. The focus of this paper is on integrating (or fusing) one single categorical item to another data set, whereas other papers[2,3,5,6] focus on integrating (or fusing) multiple (categorical) items.

The structure of this paper is as follows. Section 2 describes the concept of data fusion and how data fusion can be used in the context of marketing. This section also describes two more traditional algorithms, that are, nearest neighbor methods[8−11] and polytomeous logistic regression methods[12], versus two newly made data fusion algorithms, that are, methods based on model based clustering[13], which are used in this paper. Finally, this section shows how the four data fusion methods can be evaluated using internal and external validation criteria. This section also explains why we choose to work with real data sets, rather than simulated data sets, in order to validate the four data fusion algorithms. Section 2 illustrates data fusion with the use of the marketing application about the Dutch energy supplier. For the application the marketing goals are described, how data fusion is used to obtain the marketing goals and what the results are of applying the proposed data fusion algorithms. This paper concludes with a discussion in Section 3.

2 Data fusion

2.1 Introduction

The concept of data fusion is not new. Although there has always been resistance to do data fusion[6,14,15], there has been a great diversity in data fusion applications since the 1960s (see Rässler[6] and D'Orazio[3] for an overview of applications in Europe and the United States). It is since the 1980s that data fusion is also used for marketing purposes[2,3,8]. The most commonly used data fusion method in these applications, is based on nearest neighbor methods.

Section 2.2 shows how data fusion can be used for differentiated marketing purposes. Section 2.3 describes the two existing and the two new data fusion methods, that will be evaluated in this paper. Section 2.4 explains why we prefer to use real data sets and cross-validation over simulated data sets, in order to evaluate the performance of the fusion methods under consideration in this paper. Finally, Section 2.5 shows how data fusion procedures can be evaluated using two validation criteria.

2.2 Data fusion in marketing

'*Differentiated marketing builds greater loyalty and repeat purchasing by considering customer needs and wants. Differentiated marketing creates more total sales with a concentrated marketing effort in selected areas. Concentrated or target marketing gains market position with specialized market segments. Target marketing of products or services reduces the cost of production, distribution, and promotion.*'[16]. It is because of these benefits that differentiated marketing is getting more popular[5,17–19]. Instead of targeting customers with the same marketing strategy, companies want to target customers as individually as possible. Or, in other words, the company may be trying to sell exactly the same product or service, but it will change, for example, its promotional methods for (a group of) individuals.

In order to target (groups of) customers individually, it is important to know how they react on different marketing mix strategies. How do customers want their products or services to be packed? Where do they shop? Do they read advertisements? How do they react on discounts? All interesting facts companies need to know about their customers in order to set up good direct marketing strategies.

Information about customers can be found everywhere. An example is a company's customer database. Also market research is a powerful tool to get information about customers. In the past years, these sources of customer data have grown exponentially[5]. It seems that collecting all the desired customer information in one single source market research questionnaire[4] is the best solution. But as time and money[2,3] is limited in most marketing companies, this is often not realized. An attractive and practical solution is data fusion, or, in other words, integrating different data sets.

Data fusion is used in the following marketing application. A Dutch energy supplier wants to know their customer's interests in energy products and services; what is their interest in: information about energy savings, solar panels, custom-made advice, government grants for energy, et cetera. The energy supplier wants to send a questionnaire to all their 1,133,405 customers in their customer database (data set **A** in Figure 1) in order to obtain the desired knowledge about their customers.

To get the highest response, the energy supplier decides to send differentiated written questionnaires. From past experiences the supplier knows what the responses are with regular (or undifferentiated) written questionnaires. Using the differentiated questionnaires the supplier hopes to trigger the interest of the customers and, consequently, to improve the response. Furthermore, from past experiences, the energy supplier knows, on average, in how many energy products and services, customers are interested. Using the differentiated approach the supplier hopes that the interests in energy products and services will increase[20,21].

The energy supplier knows that each individual has a different attitude towards energy and issues related to energy. Because of this, a motivational research study, called Brand Strategy Research (BSR)[22,24,25], is conducted among 1,751 customers (data set **B** in Figure 1). The 1,751 customers are a fraction (simply using the whole customer database is too expensive) of the supplier's customer database. From the motivational study it is known that there are actually five groups (or clusters) of customers who have more of less the same attitude towards energy and issues related to energy. Short descriptions (see www.smartagent.nl for detailed descriptions) of these five motivational clusters are:

- cluster 1: energy stands for creating a cozy and warm atmosphere. Customers in this cluster try to find a balance between their own comfort and the comfort for persons in their neighborhood. The usage of energy is a well-considered choice;

- cluster 2: for customers in this cluster, energy is self-evident; the goal of the energy supplier must be to deliver as much energy as needed. Customers in this cluster are followers; the usage of customers from this cluster is mainly oriented on their peer group and the rules and values of this group;

- cluster 3: customers in this cluster use as much energy as needed for their own well-being and their own comfort; they do not conform to rules and values in society. Energy is an uncomplicated and single product. As such, the energy supplier must deliver energy with a price as low as possible, and with as least contact moments as possible;

- cluster 4: customers in this cluster feel guilty towards nature, when using energy. The usage of energy is a well-considered choice. Customers from this cluster are looking for a energy supplier that is active in the field of energy saving technique;

- cluster 5: customers in this cluster think they are smarter than their energy supplier. They live according to their own (superior) rules and values. Customers in this cluster are looking for an energy supplier that acknowledge the customer's expertise in the field of energy. The usage of energy is a smart and well-considered choice. This results in all kind of energy saving products and services.

These five motivational clusters provide a basis for developing a company's vision and/or a company's marketing directions on the strategic, tactical and operational levels, aligning the total marketing mix around the consumers needs in the domain energy. Table 1 displays the frequencies of the resulting motivational clustering.

Using the descriptions of the five motivational clusters, for each cluster, a separate questionnaire is made by a specialized communication agency. The content of the questionnaires, that are, the questions about the customer's interests in energy products and services, is the same for each questionnaire. Only the lay out (colors and pictures used in the questionnaire) and the tone-of-voice of the invitation letters are different for the cluster specific questionnaires.

Because the energy supplier wants to send a differentiated written questionnaire to all their customers, data fusion is used. Using the fraction of the supplier's customer database (data set **B** in Figure 1), for which the motivational clusters are known, data fusion methods are used to fuse the motivational clusters to the rest of the supplier's customer database. In order to do this, ten items (data set **A** in Figure 1), that are, gender, age, education, position in household, type of work, occupation, number of persons in household, household stage, social economic status and income, which are known for all the 1,133,405 customers, are used.

2.3 Methods and algorithms used

In literature, data fusion problems have been solved by several, more traditional methods, like for example: regression techniques, discriminant analysis, nearest neighbor algorithms, network approaches, etc.. More recently, Kamakura and Wedel[2] proposed a mixture model based methodology for data fusion. Each method with its own advantages and disadvantages.

But despite of all these advantages and disadvantages, not all the above mentioned methods are appropriate for data fusion as described in Section 2.2. In the following subsections two of above mentioned traditional methods for data fusion, that are nearest neighbor algorithms and regression techniques, are adjusted and described for the purpose of this paper. Furthermore, two new methods, based on model based clustering, are introduced and described.

In order to describe the data fusion methods, the schematic representation of data fusion in Figure 1 is used with the following notation: data set $\mathbf{A}$ is a customer database and contains information from $i = 1, \ldots, N$ customers about $j = 1, \ldots, J$ items ($\mathbf{X}$), Where $\mathbf{X} = (x_{i1}, \ldots, x_{iJ})$, for all customer i. Data set $\mathbf{B}$ comes from a market research study, and contains information from M customers about $J + 1$ items. The description of the first J items, is equal for both data sets. The goal of this paper is to fuse the extra information in data set $\mathbf{B}$, that is, item $J + 1$ ($\mathbf{Y}$), to data set $\mathbf{A}$. Where $\mathbf{Y} = y_i$, for all customer i. As a result of this data fusion, the information about item $J + 1$ becomes 'known' for the N customers in data set $\mathbf{A}$, that is $\widehat{\mathbf{Y}}$ in data set $\mathbf{A}^+$.

2.3.1 Nearest neighbor method

In practice, the most commonly used algorithms for data fusion are based on nearest neighbor methods[8–11]. In other words, values that are missing in one data set, usually called the recipient data set, are duplicated from another data set, usually called donor data set. The choice of the duplication record from the donor data set, is based on a certain (distance) measure, calculated on the common items in both data sets.

Translated to the application about the Dutch energy supplier, data set $\mathbf{A}$, the total customer database, is the recipient data set and data set $\mathbf{B}$, the fraction of the customer database containing data set $\mathbf{Y}$, is the donor data set. As is illustrated (for simplification only the first three items are used) in Figure 2, the motivational clusters in the recipient data set (column 'bsr; bsr represents the resulting motivational clusters as described in Section 2.2.) are missing, and needs to be fused using the records from the donor data set. As can be seen from this figure, customer record 5 in the recipient data set is exactly the same as customer record 2 in the donor data set. Consequently, the value of the motivational cluster in customer record 2 (bsr=1) in the donor data set is duplicated (or fused) to customer record 5 in the recipient data set. Likewise, the value of the motivational cluster in customer record 4 (bsr=2) in the donor data set is duplicated (or fused) to customer record 6 in the recipient data set.

An important aspect in nearest neighbor algorithms, is the choice of the (distance) measure, calculated on the common items in both data sets. Different measures have been used for data fusion, for example: Euclidean distance, City-block distance, Mahalanobis distance, etc.. See D'Orazio et al.[3] for calculation of several distances. Besides selecting the appropriate distance measure, the duplication of records can also be restricted by all kind of constraints. For example, girls less than 12 years old can't be pregnant, etc.. See D'Orazio et al.[3] and Rässler[6] for an overview of different (un)constrained measures that can be used in nearest neighbor algorithms.

Despite the popularity of nearest neighbor methods in data fusion practice, the major disadvantage of these algorithms is the heuristic rule where the duplication of data from the donor data is based on. Kamakura and Wedel[2] state that the choice of the type of distance measure is subjective, and can critically affect the quality of the data fusion. Also D'Orazio et al.[3] warn for these disadvantages, when using nearest neighbor methods.

However, in practice, nearest neighbor methods are still the most commonly used in data fusion problems[5,7]. In Germany, it is common practise to use Euclidean or City-Block distances[6], where D'Orazio et al.[3] state, that the Mahalanobis distance is the most popular distance in data fusion practise. This paper uses a nearest neighbor method with a Euclidean distance measure.

2.3.2 Polytomeous logistic regression

Regression methods have become an important aspect of any data analysis. These methods are used to describe the relationship between an outcome item and some explanatory items. When the outcome item is categorical, logistic regression has become the standard method of analysis in this situation[12]. The best known usage of logistic regression, is the case in which the categorical outcome item has only two categories. In literature, this is often called binary logistic regression. For an overview of binary logistic regression, see Hosmer and Lemeshow[12]. However and Ratner[26], when the categorical outcome item has three or more levels, this is called polytomeous logistic regression. For an overview of polytomeous logistic regression, see Hosmer and Lemeshow[12] and Ratner[26].

Above description of logistic regression can also be used in data fusion[27−30]. This application of logistic regression techniques in data fusion problems is in fact a single imputation in a missing data problem[30,31], that is, using the estimates of the logistic regression model, fitted on the complete data set, the missing value, in this case the fusion value, is imputed for the incomplete data set.

Translated to the application about the Dutch energy supplier, a polytomeous logistic regression model is fitted, in order to describe the relationship between the motivational clusters in data set $\mathbf{Y}$ and the ten socio economical items in data set $\mathbf{X}$ (in data set $\mathbf{B}$). Using the fitted regression coefficients, for each customer $i = 1, \ldots, N$ in data set $\mathbf{A}$, it is now possible to calculated the probabilities $P(y_i = 1 \mid x_{i1}, \ldots, x_{i10}), \ldots, P(y_i = 5 \mid x_{i1}, \ldots, x_{i10})$. Where, customer i, is classified to the motivational cluster value with the highest probability.

For all the technical details of fitting a logistic model, the reader is referred to Hosmer and Lemeshow[12]. This paper uses the statistical software program SPSS 15.0 to fit the polytomeous logistic model.

2.3.3 Fusion value specific probabilities model

This new method is based on latent class analysis, where the role of the latent classes is taken by the fusion values and the explanatory variables are the items. This is illustrated in two steps, using a simple example (for simplification only the first three items, are used) in the context of the application about the Dutch energy supplier. As can be seen from Table 2 the items x_1 (gender), x_2 (age) and x_3 (education) in data set $\mathbf{B}$ are used to fit a fusion value specific probabilities model in order to predict (or fuse) item y (motivational cluster) to data set $\mathbf{A}$.

Step 1: Using data set $\mathbf{B}$, the fusion value sizes (displayed in the row 'ω_y' in Table 2) and the fusion value specific probabilities (displayed in column '1', '2', '3', '4' and '5' in Table 2) are estimated. For example, there are 263 customers classified to motivational cluster 1 (=fusion value 1), that is 0.30 of the total number of customers in the data set. From these 263 customers, classified to motivational cluster 1, there are 105 customers for which $x_1 = 1$ and 158 customers for $x_1 = 2$. This is 0.40 and 0.60, respectively. Likewise, the other model parameters are calculated and displayed in Table 2).

Step 2: Using the classification rule of latent class analysis[32], the model parameters in Table 2 are used for fusing the motivational clusters to the customers in data set A. This is done in the following way: suppose a customer in data set A has the following answer pattern: $x_1 = 1$ (gender=male), $x_2 = 2$ (age=old) and $x_3 = 2$ (education=high). Using the estimated fusion value specific probabilities (column '1', '2', '3', '4' and '5') and the estimated fusion value sizes (row 'ω_y') the probabilities of fusing the motivational clusters to this customer, with answer pattern $x_1 = 1$ (gender=male), $x_2 = 2$ (age=old) and $x_3 = 2$ (education=high), are calculated, that are, $P(y = 1 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.17$, $P(y = 2 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.05$, $P(y = 3 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.16$, $P(y = 4 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.13$ and $P(y = 5 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.49$. Because the probability of fusing motivational cluster 5 is the highest of the five probabilities, motivational cluster 5 is fused to this customer in data set **A**, with this particular answer pattern.

2.3.4 Model based clustering approach

In recent years, model based clustering has become a popular technique. Also in marketing, model based clustering has become an established tool[2,7,33]. An important difference between traditional clustering[34] and model based clustering[2,32,33,35−39] is that in the latter it is assumed that the data is generated by a certain mixture of underlying probability distributions. Kamakura and Wedel[2], Vermunt and Magidson[32] and Moustaki and Papageorgiou[40] describe some advantages of a probabilistic clustering approach.

A model based clustering approach has been developed, that can be used for data fusion in the context of this paper. The goal of this model based clustering approach is to 'unmix' the mixture of underlying probability distributions. Translated to this paper, the goal of the proposed model based clustering approach is to 'unmix' the fusion value specific probabilities from the previous subsection. As a result of the model based clustering approach, there will be a fusion value specific probabilities model for each latent cluster found. Translated to the application about the energy supplier, the number of latent clusters found is 16; for each of the 16 latent clusters a fusion value specific probabilities model is estimated.

This paper does not go into detail about the model based clustering approach. The interested reader is referred to Hoijtink and Notenboom[13] for all the technical details about the proposed model based clustering approach.

2.4 Data sets: real or simulated?

The purpose of this paper is to evaluate the data fusion methods, introduced in the previous section. One of the most important evaluation criteria in comparing the four methods is the quality, or reconstruction, of the individual (missing) values.

In order to show the number of mismatches for each fusion method, we need two thing. First of all, we need a training data set to which each of the four fusion models can be fitted. Secondly, we need a test data set with the true individual fusion values known, for which the predicted values are obtained using the fitted models. Comparing the true fusion values with the predicted values for each of the four fusion methods, gives us insight in the performances of the fusion methods. However, the problem in reality is the lack of test data sets with known fusion values.

A solution to above problem is simulating the training and the test data set. A major disadvantage of simulating data sets, is that you can choose the simulation model (e.g. nearest neighbor, regression, model based clustering, et cetera) and the simulation parameters (e.g. regression parameters, number of clusters, within cluster parameters, et cetera), such that it favors one of the four data fusion methods. Another disadvantage of simulating data sets is, that it is almost impossible to choose the simulation model and the simulation parameters, such that the simulated data set is a good representation of reality. And more important, with simulated data sets it is impossible to validate the results of the data fusion externally (this is further described in the next subsection).

A good alternative for simulating data sets, without the disadvantages of simulation, is cross-validation[41,42]. In cross-validation, the data set, that is used for fitting the data fusion models, and where the true individual values are known, is randomly split in a training data set and test data set. The training data set is used for training (or calibrating) the data fusion models and, since the true individual values are known, the test data set is used for evaluating the fusion models. The use of cross-validation in the validation of the fusion models is further described in the next subsection.

In this paper, cross validation on a real data set is used in both marketing applications. Not favoring simulation models or simulation parameters in simulated data sets; the experiments described in this paper are performed in their most realistic context.

2.5 Validation

After fusing two data sets, the big question is how good (or bad) is the data fusion. In her book, Rässler[6] describes four levels of data fusion validation. Rässler[6] states that the first level of validation, that is the preservation of individual values or the reconstruction of the individual values, is the most challenging level of the data fusion validation. Furthermore, Rässler[6] states that this first level is very difficult to achieve and in many case not practical. However, this is not the case in the context of this paper.

In this paper, the goal is to fuse a data set, containing an item with information about a customer's reaction on a certain marketing mix strategy, to another data set. Taking this goal into consideration, it is undesirable that the reconstruction of customer's individual values are ill-performed. Or, translated to the application about the Dutch energy supplier, it is undesirable that a customer, that belongs to motivational cluster 1, is fused to motivational cluster 2. In order to show the realistic number of such mismatches, this paper concentrates on Rässler's first level of data fusion validation. More specific, this paper concentrates on both a validation step within the data set, and on a validation step in the actual market, after a real marketing strategy has taken place. In this paper, the first validation step is called the internal validation, and is described in Section 2.5.1. As described in the previous subsection, the internal validation step uses cross-validation for validation of the results. The second validation is called the external validation, and is described in Section 2.5.2.

2.5.1 Internal validation

One of the most important goals of this paper is to minimize the number of mismatches, or to maximize the number of correct matches, in the reconstruction of customer's individual values. Since the customer's true fusion values are not known, Rässler[6] only validates by means of simulation studies. However, this paper makes use of real data sets, and, in order to get an idea of the number of correct matches, data set $\mathbf{B}$ is randomly split according to a $2:1:1$ proportion. Which means that roughly $\frac{2}{4}$th of data set $\mathbf{B}$, or data set $\mathbf{B}_{train}$, is used for training (or calibrating) the data fusion model, roughly $\frac{1}{4}$th of data set $\mathbf{B}$, or data set $\mathbf{B}^1_{test}$, is used for the first validation of the data fusion model and roughly $\frac{1}{4}$th of data set $\mathbf{B}$, or data set $\mathbf{B}^2_{test}$, is used for the second validation of the data fusion model. In the case of the application about the Dutch energy supplier the number of customers in $\mathbf{B}_{train} = 875$, in $\mathbf{B}^1_{test} = 411$ and in $\mathbf{B}^2_{test} = 409$. Because the customer's true fusion values are known in the test data sets, the number of correct matches can easily be determined[41].

The advantage of splitting the data set $\mathbf{B}$ into a train data set and test data sets, is the prevention against model overfitting. Overfitting refers to the phenomenon in which a data fusion model may well describe the relationship between explanatory items and an outcome item in the data set used to develop the model, but may subsequently fail to provide valid predictions, when cross validating a new data set. The model shows an adequate fit in the data set under study, but does not cross validate, that is, does not provide accurate predictions for observations from a new data set. In the remainder of this subsection some examples of model overfitting are shown. However, this paper does not go into detail about this topic. The interested reader is referred to Verstraeten[42] for more details of model overfitting. Using two test data sets, is done because of the dependency of the validation results of one particular split of the data set used[42].

In order to draw conclusions about the quality of the data fusion, Ratner[26] introduces the statistics model lift and total correct classification rate (TCCR). These statistics are explained and described using the application about the Dutch energy supplier. Table 3 displays the classification table after the data fusion method 'logistic regression' is applied on $\mathbf{B}_{train}$. As is described in Section 2.2, the fusing item is the motivational cluster about the domain energy. Customers are classified to either motivational cluster 1, 2, 3, 4 or 5. The row totals of Table 3 show the actual counts in data set $\mathbf{B}_{train}$. The column totals show how the predicted classification counts are after applying the data fusion method 'logistic regression' on data set $\mathbf{B}_{train}$. The percentages under the total counts (between brackets) are with respect to the total number of customers in data set $\mathbf{B}_{train}$. For example, in data set $\mathbf{B}_{train}$ the actual percentage of customers classified to motivational cluster 2 is 19.8%. However, the predicted percentage is 21.5%.

The diagonal in Table 3 displays the numbers of correct matches for each motivational cluster. For example, 90 customers, which is 47.9% ($=\frac{90}{188}$), are correct classified to motivational cluster 2. In this paper this is called the total correct classification rate for motivational cluster 2 ($TCCR(2)$), which is derived from Ratner's total correct classification rate for the overall model ($TCCR(model)$). However, using the actual percentages one would expect to find 19.8% of the customers to be classified to motivational cluster 2, or, in other words, based on a random chance model one would expect to find 19.8% of the customers to be classified to motivational cluster 2. In this paper this is called the total correct classification rate for motivational cluster 2 that can be obtained by a random chance model ($TCCR_{chance}(2)$). Using these

TCCR's the model lift for motivational cluster 2 is $\frac{TCCR(2)}{TCCR_{chance}(2)}$=242, which means that the data fusion method 'logistic regression' provides 142% more correct matches for motivational cluster 2, than obtained by chance.

These statistics can also be calculated in order to draw conclusions about the overall quality of the data fusion. From Table 3 it is clear that 436 (=167+90+61+59+59) customers, are correctly classified to one of the five motivational clusters. This results in a total correct classification rate for the overall model ($TCCR(total)$) of 49.8% (=$\frac{436}{875}$). To calculate the model lift for the overall model, the $TCCR(total)$ is compared with the $TCCR_{chance}(total)$, that is the total correct classification rate for the overall model that can be obtained by a random chance model. The $TCCR_{chance}(total)$ is defined as the sum of the square actual value percentages. For Table 3 the $TCCR_{chance}(total)$ is 21.4% (=30.1%^{2}+19.8%^{2}+15.9%^{2}+17.7%^{2}+16.6%^{2}). Using these TCCR's the overall model lift is $\frac{TCCR(total)}{TCCR_{chance}(total)}$=233, which means that the data fusion model provides 133% more correct matches for all the motivational clusters than obtained by chance.

Above described classification table are made for each data fusion method, applied to one of the tree data sets. However, the figures that are the most important from these tables, are the actual and predicted frequencies and the information necessary to calculate the statistics TCCRs and model lifts. Table 4, 5 and 6 summarize this important information. Table 4 displays the actual and predicted frequencies, after applying the data fusion methods to the train and two data sets. This table also shows how many customers are in each data set. Table 5 displays the total correct classification rates (TCCRs) for each motivational cluster and for the total model. This table also shows what the percentage of correct matches would be, in each data set, when obtained by chance. Table 6 displays the model lifts for each motivational cluster and for the total model. Note that in all three tables the figures in the nearest neighbor method for the train data set are missing. This because the train data set is defined as the donor data set (see Section 2.3.1), and the motivational clusters are duplicated from this data set for the two test data sets (the recipients files).

In order to determine which data fusion method performs the best, several considerations need to be made. First of all, for each data fusion method, the predicted frequencies of the motivational clusters are compared with the actual frequencies. Table 4 shows that, for each data fusion method, the predicted frequencies for motivational cluster 2, 3 and 5, are closer to the actual frequencies. The predicted frequencies for motivational cluster 1 and 4 are more different.

Secondly, the TCCRs and the model lifts are examined for each data fusion method. Corresponding with this second consideration, a third consideration, the degree of model overfitting plays a part in the determination of the best performing data fusion method. From Table 5 and 6 it is clear that, both the TCCRs and the model lifts are the lowest for data fusion method 'nearest neighbor'. From these two tables it is also clear that, the data fusion method 'model based clustering approach', applied on the train data set, has the highest TCCRs and model lifts. However, these statistics drop, when applying the same data fusion model on the two test sets. This is the model overfitting phenomenon, as described above. This model overfitting can be seen in all the data fusion method used. However, it seems that for the data fusion method 'fusion value specific probability approach', this is the least. For the method 'fusion value specific probabilities approach', both the TCCRs and the model lifts are among the highest, and the difference between the train data set and the test data sets is not as large as the other fusion methods.

Taking into account the three considerations, the fusion value specific probabilities approach, turns out to be the best performing data fusion method. Consequently, this method is used to fuse the motivational clusters to the company's customer database. As a result of this data fusion, the motivational clusters become 'known' for all the customers in the database. This is the starting point for differentiated marketing strategies as described in the next subsection.

2.5.2 External validation

Despite the internal validation described in the previous subsection, the final validation is in the real world. Before a data fusion is proposed, the marketing company has a certain goal to achieve. This can be for example, improving the response on a certain questionnaire, increasing sales, etc.. External validation is done in order to draw conclusions about how this goal is achieved. It is clear that each marketing company has a different goal to achieve with data fusion, that's why there are no unified statistics for external validation. For each external validation, tailor-made criteria need to be made.

Unfortunately external validation is not common practice for most marketing companies[43]. It is expensive and time-consuming. However, these type of cross-validation experiments are highly recommended. In the end, it is not important what the statistics are in the internal validation step, but what the effect is of the differentiated marketing strategy in the real world. This is best described by the proverb '*the proof of the pudding is in the eating*'.

Keeping this proverb into consideration, the following external validation is performed for the application about the Dutch energy supplier:
In the case of the supplier, the initial goal was to improve the response on the written questionnaire. From past research experiences, the supplier knows that the response percentage on regular questionnaires is 19.9%. The first goal with the differentiated questionnaire approach is to improve this response percentage.

The second goal is to improve the number of sales leads. The energy supplier defines the sales leads as the number of products or services, customers are interested in. In the questionnaire, customers are asked about their interests in ten energy products and services. From past experiences the supplier knows that the average number of sales leads is 2.25 per customer. The second goal with the differentiated questionnaire approach is to increase the average number of sales leads per customer.

As a results of the data fusion, the total customer database, with $1,133,405$ customers, is classified. The columns 'Frequency customers' and 'Percentage customers' in Table 7 show the resulting motivational cluster frequencies of the fused data set $\widehat{\mathbf{Y}}$. For 120,843 (=10.7%) customers there are no or insufficient common items available in order to classify to one of the five motivational clusters.

Using the descriptions of the five motivational clusters, for each cluster a separate questionnaire is made by a specialized communication agency. For the group of customers with no motivational cluster, the regular questionnaire is used. The content of the questionnaires, that are, the questions about the customer's interests in energy products and services, is the same for each questionnaire. Only the lay out (colors and pictures used in the questionnaire) and the tone-of-voice of the invitation letters are different for the cluster specific questionnaires. The focus of the questionnaire for motivational cluster 1 is on the balance between comfort and nature. The questionnaire for motivational cluster 2 emphasizes that the interests, wishes, desires, complaints, etc. from society, are taken seriously. For motivational cluster 3 the focus of the questionnaire is on the supplier's differentiated approach in order to increase the customer's comfort and to decrease the energy prices. The focus of the questionnaire for motivational cluster 4 is on, the saver the customer is with energy, the better it is for nature. And, finally, the focus of the questionnaire for motivational cluster 5 is on the question: 'Would you like to help us to improve our service for you?'.

Eventually in four batches 1,133,405 (un)differentiated questionnaires were sent to all the customers. Table 8 shows when and how many questionnaires were sent to the customers in each batch. This table also shows how many customers responded on the questionnaires. Table 9 further splits these responses into the motivational clusters. From this table it is also clear that the first goal is attained. The total response percentage is 25.2%, which is higher than the target percentage of 19.9%. The difference in response percentage equals almost 60,000 extra customers, which is of course, valuable for the supplier.

Although the total response percentage in Table 9 displays 25.2% it is interesting to see what the response behavior is for each motivational cluster. From Table 9 it can be seen that the response percentages for the customers, classified to motivational cluster 3 and 4, are relatively low. From past experiences with the motivational clusters, it is known that customers classified to motivational cluster 3 and 4 are, in general, less willing to fill out questionnaires.

The second goal to attain, is increasing the number of sales leads. Table 10 shows the average number of sales leads per motivational cluster. From this table it is clear that also the second goal is attained; the average number of sales leads is 2.63, whereas an average of 2.25 sales leads was the target. Also from Table 10 it is interesting to see what the average number of sales leads is for each of the motivational clusters. The results in the table are completely consistent with the description of these five motivational cluster. Cluster 1 with a higher interest in energy products and services, in order to get a good balance between own comfort and nature. Cluster 3 with a higher interest in energy products and services, in order to get a differentiated approach for more comfort and lower prices. Cluster 5 with a higher interest in energy products, in order to stay in control with their own thoughts about energy. And, cluster 2 and 4 with a lower interest in energy products and services, because they totally rely on the expertise of the energy supplier. However, there is no logical explanation for the fact that customers, not classified to one of the five motivational clusters, have a relative high average number of sales leads.

Although the responses and sales leads can be determined before and after the marketing strategy, it is impossible to conclude that the increase (or decrease) in responses and sales leads can be fully dedicated to the differentiated marketing strategy[43,44]. When sending the questionnaires it was impossible to control for all kind of side effects that may be associated with response behavior and interests. However, for this applications both goals are attained: almost 60,000 customers more responded to the differentiated questionnaires and, on average, the total responding customers were more interested in energy products and services. Furthermore, instead of conducting a motivational research study among all 1,133,405 customers, only a small, representative number of these customers (1,751) where used. Which is, in terms of dollars, a huge saving in marketing research costs.

3 Discussion

In this paper, data sets were fused (or integrated) to each other. In order to be as realistic as possible, this paper used only real data sets. No simulated data sets were used, where, inevitably, one could favor a simulation model and simulation parameters. The experiments described in this paper, were performed in their most realistic context.

In the marketing application the customer database of a energy supplier was fused to a motivational research study about energy. One of the most important goals was the reconstruction of customer's individual fusion values. Or, translated to the marketing application, it was undesirable that a customer, that belongs to motivational cluster 1, was fused to motivational cluster 2. In order to show the realistic number of such mismatches, this paper concentrated on two very important validation steps, that were, the internal validation step and the external validation step.

3.1 Internal validation

The most important thing in the internal validation step was the prevention against model overfitting. The application showed that model overfitting was a serious problem. For example, in the case of the model based clustering approach, the method showed the best statistics on the train data set, but subsequently failed to preserve these good statistics on the test data sets.

In order to prevent against model overfitting, this paper used a train data set and two test data sets. The latter was done because of the dependency of the validation results of one particular split of the data set used. The train data set was used for training (or calibrating) the data fusion models and the two test data sets were used for validating the data fusion models.

The lesson that can be learnt from this is that, one should never trust a data fusion company that uses only one data set to train and test data fusion models. Because, you have to take into account model overfitting, as we have shown using the train and test data sets.

In order to draw conclusions about the quality of the data fusion, this paper used the statistics model lift and total correct classification rate (TCCR). The latter was calculated for both the random chance model and the data fusion model under study. In the application the fusion value specific probabilities approach was found to be the 'best' method. This is not only the case in the application described in this paper, but also for past marketing applications in domains like care, insurance, gardening, financial services, et cetera (see track record on www.smartagent.nl). The problems and the goals of these marketing applications were similar to the application described in this paper. In these past marketing applications, the data fusion methods, as described in Section 2.3, were also used and compared. In each application the fusion value specific probabilities approach turned out to be (one of) the best methods in the internal validation, which makes this data fusion method, a method with stable results.

For the marketing application in this paper the TCCR for the overall model was around 40%, whereas the TCCR with a random chance mode was around 20%. The model lift was around the 200%, which means that the fusion value specific probabilities approach provided around 100% more correct matches than would be obtained by chance. Of course, the goal of the data fusion was to get a TCCR that was as close to 100% as possible, but when analyzing the TCCRs, we had to take into account the type of the fusion item and the type of the explanatory items. The fusion items in the application were motivational clusters, that came from a motivational research study. The explanatory items were socio demographical and socio economical items. When it was possible to predict (almost) perfectly the motivational clusters with these explanatory items, the initial motivational research study would loose their uniqueness.

3.2 External validation

As a result of the data fusion, the motivational clusters were estimated for all the customers in the database. In the real world application, this was the starting point for differentiated marketing strategies. In the application about the energy supplier, differentiated written questionnaires were made.

As *the proof of the pudding is in the eating*, the external validation step was even more important, than the internal validation step. In the end, the external validation step determined, whether the data fusion was profitable or not.

Using a cross validation experiment, different marketing goals were tested and attained. In the case of the energy supplier almost 60.000 more customers responded on the differentiated questionnaires. Also the average number of sales leads per customer increased.

Given the large number of customers involved in the application, the increases in responses and sales leads, gave the company a tremendous amount of extra information and sales opportunities. Furthermore, by using only a small proportion of the customers for a domain study, a lot of dollars were saved on marketing research costs. In the application the data fusion project was profitable, and, consequently, was successful.

References

1. Craig, C.S. and McCann,J.M. (1978). Item Nonresponse in Mail surveys: Extent and correlates. *Journal of Marketing Research*, 15(2), 285-289.

2. Kamakura, W.A. and Wedel, M. (1997). Statistical Data Fusion for Cross-Tabulation. *Journal of Marketing Research*, 34(4), 485-498.

3. D'Orazio, M., Di Zio, M. and Scanu, M., (2006). *Statistical Matching: theory and practice.* Chichester: John Wiley & Sons Ltd..

4. Buck, S. (1989). Single Source Data-The Theory and the Practice. *Journal of the Market Research Society*, 31(1), 489-500.

5. Van der Putten, P., Kok, J.N. and Gupta, A. (2002). Data Fusion Through Statistical Matching. *Paper 185.* MIT Sloan School of Management.

6. Rässler, S., (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* Lecture Notes in Statistics, 168, New York: Springer.

7. Wedel, M. and Kamakura, W.A. (2000), *Market Segmentation: Conceptual and Methodological Foundations.* Norwell: Kluwer Academic Publishers.

8. Gilula, Z., McCulloch, R.E and Rossi, P.E. (2006). A Direct Approach to Data Fusion. *Journal of Marketing Research*, 43(1), 73-83.

9. Dillon, W.R., Goldstein, M. and Schiffman, L.G. (1978). Appropriateness of Linear Discriminant and Multinomial Classification Analysis in Marketing Research. *Journal of Marketing Research*, 15(1), 103-112.

10. Rogers, W.L. (1984). An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics*, 2, 91-105.

11. Wind, J. and Mahajan, V. (1997). Editorial: Issues and Opportunities in New Product Development: An Introduction to the Special Issue. *Journal of Marketing Research*, 34(1), 1-12.

12. Hosmer, D.W. and Lemeshow, S (2000), *Applied Logistic Regression.* Hoboken: John Wiley & Sons.

13. Hoijtink, H. and Notenboom, A. (2004), Model Based Clustering of Large Data Sets: Tracing the Development of Spelling Ability. *Psychometrika*, 69, 481-498.

14. Bronner, A.E. (1988), Einde fusie-fobie in Nederland?. *Jaarboek van de Nederlandse vereniging van Marktonderzoekers*, 1988, 9-18.

15. Baker, K., Harris, P. and O'Brien, J. (1989). Data Fusion: An Appraisal and Experimental Evaluation. *Journal of the Market Research Society*, 31, 153-212.

16. Bull, N.H. and Passewitz, G.R. (1994). Finding Customers: Market Segmentation. *Publication CDFS-1253-94.* Ohio State University.

17. Cui, G. and Choudhury, P. (2002). Marketplace diversity and cost-effective marketing strategies. *Journal of Consumer Marketing*, 19(1), 54-73.

18. Cui, G. and Choudhury, P. (2003). Consumer Interests and the Ethical Implications of Marketing: A Contingency Framework. *Journal of Consumer Affairs*, 37, 364-387.

19. Buckinx, W. (2005). Using Predictive Modeling for Targeted Marketing in a Non-Contractual Retail Setting. PhD. thesis, Marketing, Gent University, Belgium (2005).

20. Lattin, J.M. and Bucklin, R.E. (1989). Reference Effects of Price and Promotion on Brand Choice Behavior. *Journal of Marketing Research*, 26(3), 299-310.

21. Feinberg, F.F., Krishna, A. and Zhang, J.Z. (2002). Do We Care What Others Get? A Behaviorist Approach to Targeted Promotions. *Journal of Marketing Research*, 39(3), 277291.

22. BSR is based on Adler's social-psychology theory[23] and provides a framework for understanding customers at the 'deepest' level. This motivational level gives knowledge of consumer's fears, beliefs and values, thus providing an understanding of the fundamental motivations

that drive (future) purchase decisions of customers. The interested reader is referred to www.smartagent.nl for more information about BSR.

23. Callebaut, J., Janssens, M., Op de Beeck, D., Lorré, D. and Hendrickx, H. (1999). *Motivational Marketing Research Revisited.* Leuven: Garant Publishers.

24. Brethouwer, W., Lamme, A., Rodenburg, J., Du Chatinier, H. and Smit, M. (1995). *Quality Planning toegepast (Dutch).* Amsterdam: Janssen Offset.

25. Oppenhuisen, J. (2000). *Een schaap in de bus? Een onderzoek naar waarden van de Nederlander (Dutch).* Amsterdam: Grafische Producties.

26. Ratner, B. (2003). *Statistical Modelling and Analysis for Database Marketing: Effective Techniques for Mining Big Data.* Florida: Chapman&Hall/CRC.

27. Jones, J.M. and Zufryden, F.S. (1980). Adding Explanatory Variables to Consumer Purchase Behavior model: An Exploratory Study. *Journal of Marketing Research*, 17(3), 323-334.

28. Bucklin, R.E. and Gupta, S. (1992). Brand Choice, Purchase Incidence and Segmentation: An Integrated Modelling Approach. *Journal of Marketing Research*, 29(2), 201-215.

29. Mela, C.F., Gupta, S. and Lehmann, D.R. (1997). The Long-term Impact of Promotion and Advertising on Consumer Brand Choice. *Journal of Marketing Research*, 34(2), 248-261.

30. Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

31. Little, R.J. and Rubin, D.B. (2002), *Statistical analysis with missing data.* New York: John Wiley.

32. Vermunt, J.K. and Magidson J. (2000). *Latent Gold.* Belmont: Statistical Innovations Inc.

33. Varki, S. and Chintagunta, P.K. (2004). The Augmented Latent Class Model: Incorporating Additional Heterogeneity in Latent Class Model for Panel Data. *Journal of Marketing Research*, 41(2), 226-233.

34. Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1984), *Multivariate Data Analysis.* London: Prentice-Hall International, Inc.

35. Newcomb, S. (1886), A Generalized Theory of the Combination of Observations So As To Obtain the Best Result. *American Journal of Mathematics*, 8, 343-366.

36. Pearson, K. (1894), Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions*, A185, 71-110.

37. Banfield, J.D. and Raftery, A.E. (1993), Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803-821.

38. Bensmail, H., Celeux, G., Raftery, A.E. and Robert, C.P. (1997), Inference in Model Based Clustering. *Statistics and Computing*, 7, 1-10.

39. Fraley, C. and Raftery, A.E. (1998), How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *Technical Report No. 329*, Department of Statistics, University of Washington.

40. Moustaki, I. and Papageorgiou, I. (2005), Latent Class Models for Mixed Variables with Applications in Archaeometry. *Computational Statistics and Data Analysis*, 48, 659-675.

41. Kamakura, W.A., Wedel, M., De Rosa, F. Mazzon, J.A. (2003). Cross-selling through Database Marketing: a Mixed Data Factor Analyzer for Data Augmentation and Prediction. *International Journal of Research in Marketing*, 20, 45-65.

42. Verstraeten, G. (2005). Issues in Predictive Modelling of Individual Customer Behavior: Applications in Targeted Marketing and Consumer Credit Scoring. PhD. thesis, Marketing, Gent University, Belgium (2005).

43. Bell, M.L. and Vincze, J.W. (1988). *Managerial Marketing: Strategy and Cases.* New York: Elsevier Science Publishing Co..

44. Kooiker, R. (1997). *Marktonderzoek. (Dutch).* Groningen: Wolters-Noordhoff bv.

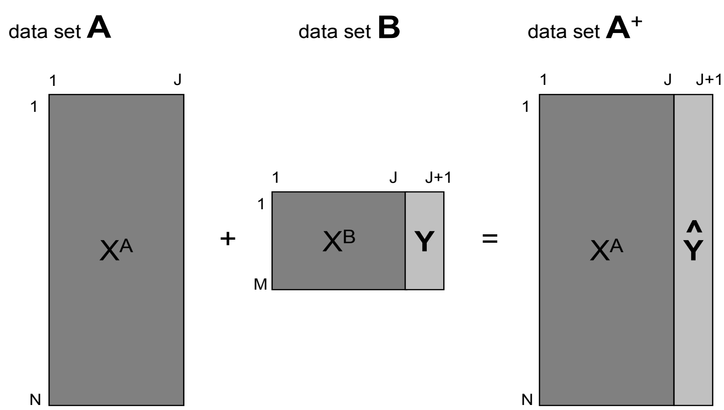

Figure 1: Schematic representation of data fusion in marketing (derived and adjusted from Van der Putten et al.[5])

Data set A: recipient file

customer	gender	age	eduction	bsr
5	male	old	low	
6	female	old	low	

Data set B: donor file

customer	gender	age	eduction	bsr
1	male	young	high	4
2	male	old	low	1
3	female	young	high	3
4	female	old	low	2

Figure 2: Nearest neighbor method

Table 1: Frequency respondents in motivational research study energy (between brackets are the percentages based on the total number of respondents classified to one of five motivational clusters)

Cluster	Frequency respondents	Percentage respondents
Cluster 1	537	30.7% (31.7%)
Cluster 2	337	19.2% (19.9%)
Cluster 3	265	15.1% (15.6%)
Cluster 4	305	17.4% (18.0%)
Cluster 5	251	14.3% (14.8%)
No cluster	56	3.2%
Total	1,751	100.0%

Table 2: Model parameters for the fusion value specific probabilities approach. Note: these counts and probabilities are fictive figures.

y		1	2	3	4	5	total
ω_y		0.30	0.20	0.16	0.18	0.17	1.00
		(263)	(173)	(139)	(155)	(145)	(875)
x_1	1	0.40	0.20	0.44	0.52	0.52	0.42
		(105)	(34)	(61)	(80)	(75)	(365)
	2	0.60	0.80	0.56	0.48	0.48	0.58
		(158)	(139)	(78)	(75)	(70)	(510)
x_2	1	0.43	0.80	0.46	0.74	0.32	0.46
		(114)	(139)	(64)	(115)	(46)	(399)
	2	0.57	0.20	0.54	0.26	0.68	0.54
		(149)	(34)	(75)	(40)	(99)	(476)
x_3	1	0.73	0.28	0.56	0.45	0.13	0.42
		(193)	(48)	(78)	(69)	(19)	(365)
	2	0.27	0.72	0.44	0.55	0.87	0.58
		(70)	(125)	(61)	(86)	(126)	(510)

Table 3: Classification table

		PREDICTED					
		cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total
ACTUAL	cluster 1	167 (50.9%)	43	18	21	14	263 (30.1%)
	cluster 2	61	90 (47.9%)	8	12	19	173 (19.8%)
	cluster 3	27	20	61 (48.4%)	12	19	139 (15.9%)
	cluster 4	35	25	14	59 (50.4%)	22	155 (17.7%)
	cluster 5	38	10	25	13	59 (50.9%)	145 (16.6%)
	total	328 (37.5%)	188 (21.5%)	126 (14.4%)	117 (13.4%)	116 (13.3%)	875 (100.0%)

Table 4: Frequencies after applying data fusion methods for domain energy

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
	actual	train	875	30.1%	19.8%	15.9%	17.7%	16.6%
		test1	411	33.8%	20.7%	15.6%	16.5%	13.4%
		test2	409	33.0%	19.3%	15.2%	20.0%	12.5%
nearest neighbor	predicted	train	875					
		test1	411	31.9%	19.7%	16.1%	18.0%	14.4%
		test2	409	32.2%	20.8%	17.4%	16.1%	13.4%
logistic regression	predicted	train	875	37.5%	21.5%	14.4%	13.4%	13.3%
		test1	411	39.4%	20.7%	15.6%	12.2%	12.2%
		test2	409	39.9%	17.1%	17.6%	16.4%	9.0%
fusion value specific approach	predicted	train	875	32.8%	22.7%	13.0%	12.3%	19.1%
		test1	411	39.2%	23.4%	13.4%	9.7%	14.4%
		test2	409	35.2%	22.7%	13.0%	14.2%	14.9%
model based clustering approach	predicted	train	875	33.1%	25.1%	13.3%	15.5%	12.9%
		test1	411	38.7%	25.5%	13.1%	11.4%	11.2%
		test2	409	35.5%	21.5%	12.7%	17.6%	12.7%

Table 5: Total correct classification rates after applying data fusion methods for domain energy

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total	chance
nearest		train	875							21.4%
neighbor	predicted	test1	411	40.5%	34.6%	24.2%	20.3%	22.0%	30.4%	22.7%
		test2	409	37.9%	25.9%	31.0%	25.8%	18.2%	29.6%	22.5%
logistic		train	875	50.9%	47.9%	48.4%	50.4%	50.9%	49.8%	21.4%
regression	predicted	test1	411	48.4%	40.0%	32.8%	24.0%	26.0%	38.7%	22.7%
		test2	409	46.6%	40.0%	33.3%	43.3%	32.4%	41.3%	22.5%
fusion value		train	875	47.7%	42.7%	46.5%	46.3%	43.1%	45.4%	21.4%
specific	predicted	test1	411	51.6%	44.8%	43.6%	37.5%	25.4%	43.8%	22.7%
approach		test2	409	54.2%	43.0%	49.1%	41.4%	34.4%	46.2%	22.5%
model based		train	875	56.2%	51.0%	56.0%	50.7%	54.9%	54.1%	21.4%
clustering	predicted	test1	411	45.9%	42.9%	38.9%	23.4%	23.9%	39.2%	22.7%
approach		test2	409	44.8%	37.5%	32.7%	29.2%	17.3%	35.5%	22.5%

Table 6: Model lifts after applying data fusion methods for domain energy

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total
nearest neighbor	predicted	train	875						
		test1	411	120	167	156	123	165	134
		test2	409	115	134	204	128	146	131
logistic regression	predicted	train	875	169	242	305	285	307	233
		test1	411	144	193	211	145	194	171
		test2	409	141	207	220	216	260	184
fusion value specific approach	predicted	train	875	159	216	293	261	260	213
		test1	411	152	217	280	227	190	193
		test2	409	164	223	324	206	276	205
model based clustering approach	predicted	train	875	187	262	353	286	331	253
		test1	411	136	207	250	141	179	173
		test2	409	136	194	216	146	139	158

Table 7: Frequency clusters in customer database for application energy

Cluster	Frequency customers	Percentage customers
Cluster 1	334,083	29.5% (33.0%)
Cluster 2	204,774	18.1% (20.2%)
Cluster 3	165,416	14.6% (16.3%)
Cluster 4	176,319	15.6% (17.4%)
Cluster 5	131,970	11.6% (13.0%)
No cluster	120,843	10.7%
Total	1,133,405	100.0%

Table 8: Responses per batch

Batch	Date sent	Questionnaires sent (#)	Response (#)	Response (%)
1	October 2002	549,818 (48.6%)	109,754 (38.5%)	20.0%
2	October 2002	260,151 (23.0%)	85,118 (29.8%)	32.7%
3	November 2002	217,928 (19.3%)	48,145 (16.9%)	22.1%
4	November 2002	103,508 (9.1%)	42,260 (14.8%)	40,8%
Total		1,133,405 (100.0%)	285,453 (100.0%)	25.2%

Table 9: Responses per motivational cluster

Cluster	Questionnaires sent (#)	Response (#)	Response (%)
Cluster 1	334,083 (29.5%)	99,961 (35.0%)	29.9%
Cluster 2	204,774 (18.1%)	55,064 (19.3%)	26.9%
Cluster 3	165,416 (14.6%)	30,638 (10.7%)	18.5%
Cluster 4	176,319 (15.6%)	34,264 (12.8%)	19.4%
Cluster 5	131,970 (11.6%)	41,210 (14.4%)	31.2%
No cluster	120,843 (10.7%)	24,316 (8.5%)	20,1%
Total	1,133,405 (100.0%)	285,453 (100.0%)	25.2%

Table 10: Sales leads per motivational cluster

Cluster	Response (#)	Sales leads
Cluster 1	99,961 (35.0%)	2.69
Cluster 2	55,064 (19.3%)	2.26
Cluster 3	30,638 (10.7%)	2.75
Cluster 4	34,264 (12.8%)	2.22
Cluster 5	41,210 (14.4%)	3.14
No cluster	24,316 (8.5%)	2.73
Total	285,453 (100.0%)	2.63