Why the Information Explosion Can Be Bad for Data Mining, and How Data Fusion Provides a Way Out

Peter van der Putten[1], Joost N. Kok[2] and Amar Gupta[3]

1 Introduction and motivation

One may claim that the exponential growth in the amount of data provides great opportunities for data mining. Reality can be different though. In many real world applications, the number of sources over which this information is fragmented grows at an even faster rate, resulting in barriers to widespread application of data mining and missed business opportunities. Let us illustrate this paradox with a motivating example from database marketing.

In marketing, direct forms of communication are becoming increasingly popular. Instead of broadcasting a single message to all customers through traditional mass media such as television and print, the most promising potential customers receive personalized offers through the most appropriate channels. So it becomes more important to gather information about media consumption, attitudes, product propensity etc. at an individual level. Basic, company specific customer information resides in customer databases, but market survey data depicting a richer view of the customer are only available for a small sample or a disjoint set of reference customers. Collecting all this information for the whole customer database in a single source survey would certainly be valuable but is usually a very expensive proposition. The common alternative within business to

[1] Leiden Institute of Advanced Computer Science. Niels Bohrweg 1, 2333 CA Leiden, The Netherlands. putten@liacs.nl / pvdutten@hotmail.com

[2] Leiden Institute of Advanced Computer Science. Niels Bohrweg 1, 2333 CA Leiden, The Netherlands. joost@liacs.nl

3 MIT Sloan School of Management. Room E60-309, 30 Memorial Drive, Cambridge, MA 02139, USA. agupta@mit.edu

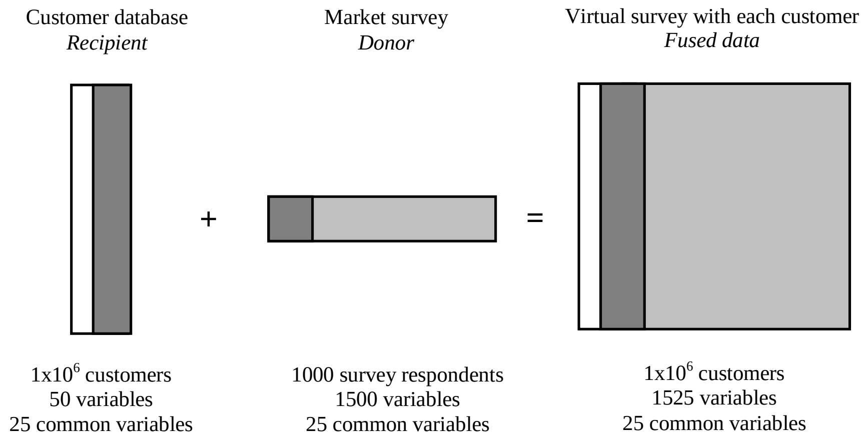

Fig. 1. ***Data Fusion Example***

consumer marketing is to buy syndicated socio-demographic data that have been aggregated at a geographical level. All customers living in the same geographic region, for instance in the same zip code area, receive the same characteristics. In reality, customers from the same area will behave differently. Furthermore, regional information may be absent in surveys because of privacy concerns.

The zip code based data enrichment procedure provides a crude example of data fusion: the combination of information from different sources. But one needs more generalized and powerful fusion procedures that cater to any number and kind of variables. Data mining algorithms can help to carry out such generalized fusions and create rich data sets for marketing and other applications [14].

In this paper we position data fusion as both a key enabling technology and an interesting research topic for data mining. A fair amount of work has been done on data fusion over the past 30 years, but primarily outside the knowledge discovery community. We would like to share and summarize the main approaches taken so far from a data mining perspective (section 2). A case study from database marketing serves as a clarifying example (section 3). We conclude with a discussion of some the interesting opportunities for future research (section 4).

2 Data Fusion

Valuable work has been done on data fusion in areas other than data mining. From the 1970s through the 1990s, the subject was quite popular and controversial, with a number of initial applications in economic statistics in the US and Germany ([2,4,8,12,17,18,19]; [15] provides an overview] and later in the field of media research in the Europe and Australia ([3,6]; [1] provides an overview). It is also known as micro data set merging, statistical record linkage, multi-source imputation and ascription. Until today, data fusion is used to reduce the required number of respondents or questions in a survey. For instance, for the Belgian National Readership survey questions regarding media and questions regarding products are collected in 2 separate groups of 10,000 respondents each, and then fused into a single survey, thereby reducing costs and the required time for each respondent to complete a survey.

2.1 Data Fusion Concepts

We assume that we start from two data sets. These can be seen as two tables in a database that may refer to disjoint data sets. The data set that is to be extended is called the recipient set *A* and the data set from which this extra information has to come is called the donor set *B*. We assume that the data sets share a number of variables. These variables are called the common variables *X*. The data fusion procedure will add a number of variables to the recipient set. These added variables are called the fusion variables *Z*. Unique variables are variables that only occur in one of the two sets: *Y* for *A* and *Z* for *B*. See Figure 1 for a marketing example. In general, we will learn a model for the fusion using the donor *B* with the common variables *X* as input and the fusion variables *Z* as output and then apply it to the recipient *A*.

2.2 Core Data Fusion Algorithms

In nearly all studies, statistical matching is used as he core fusion algorithm. The statistical matching approach can be compared to *k*-nearest neighbor prediction with the donor as training set and the recipient as a test set. The procedure consists of two steps. First, given some element from the recipient set, the set of *k* best matching donor elements is selected. The matching distance is calculated over some subset of the common variables. Standard distance measures such as Euclidian distance can be used, but often more complex measures are designed to tune the fusion process. For instance, it may be desirable that men are never matched with women, to prevent that 'female' characteristics like 'pregnant last year' are predicted. In this case, the gender variable will become a so-called cell or critical variable; the match between recipient and donor must be 100% on the cell variable; otherwise they will not be matched at all. Another enhancement is called constrained matching. Experiments with statistical matching have shown that even if the donor and recipient are large samples of the same population, some donors are used more than others, which can result in a fusion that is not representative. By taking into account how many times an element of the donor set has been used when calculating the distance, we can counter this effect [13]. In the second step, the prediction for the fusion variables can be constructed using the set of found nearest neighbors, e.g. by calculating averages (numerical), modes (categorical) or distributions (categorical or numerical).

A number of constraints have to be satisfied by any fusion algorithm in order to produce valid results. First, the donor must be representative for the recipient. This does not necessarily mean that the donor and recipient set need to be samples of the same population, although this would be preferable. For instance, in the case of statistical matching only the donors that were actually used need to be representative of the recipient set. For example, the recipients could be buyers of a specific product and the donor set could be very large sample of the general population. Second, the donor variables must be good predictors for the fusion variables. In addition, the Conditional Independence Assumption must be satisfied: the commons *X* must explain all the relations that exist between unique variables *Y* and *Z*. In other words, we assume that $P(Y|X)$ is independent of $P(Z|X)$. This could be measured by the partial correlation $r_{ZY.X}$, however there is generally no data available on *X*, *Y* and *Z* to compute this. In most of our fusion projects we start with a small-scale fusion exercise to test the validity of the assumptions and to produce ballpark estimates of fusion performance.

There have been some exceptions to the standard statistical matching approach. In [2], constrained fusion is modeled as a large-scale linear programming transportation model. The main idea was to minimize the match distance under the constraint that all donors should be used only once, given recipients and donors of equal size. Various methods derived from solutions to the well-known stable marriage problem [7] are briefly mentioned in [1]. In statistics extensive work has been done on handling missing data [9], including likelihood based methods based on explicit models such as linear and logistic regression. Some researchers have proposed to impute values for the fusion variables using multiple models to reflect the uncertainty in the correct values to impute [18]. In [8] a statistical clustering approach to fusion is described based on mixture models and the EM algorithm.

2.3 Data Fusion Evaluation and Deployment

An important issue in data fusion is to measure the quality of the fusion; this is not a trivial problem [6].

We can distinguish between internal evaluation and external evaluation. This refers to the different stages in the data mining process. If one considers data fusion to be part of the data enrichment step and evaluates the quality of the fused data set only within this step then this is an internal evaluation. However, if the quality is measured using the results within the other steps in the data mining process, then we call this an external evaluation.

Of course, in practice the external evaluation is the bottom line evaluation. Assume for instance that one wants to improve the response on mailings for a certain set of products, this being the reason why the fusion variables were added in the first place. In this case, one way to evaluate the external quality is to check whether an improved mail response prediction model can be built when fused data is included in the input. Ideally, the fusion algorithm is tuned towards the kinds of analysis that is expected to be performed on the enriched data set.

3. Case Study: Cross Selling Credit Cards

We assume the following (hypothetical) business case. A bank wants to learn more about its credit card customers and expand the market for this product. Unfortunately, there is no survey data available that includes credit card ownership; this variable is only known for customers in the customer base. Data fusion is used to enrich a customer database with survey data. The resulting data set serves as a starting point for further data mining.

To simulate the bank case we do not use separate donors; instead we draw a sample from an existing real world market survey and split the sample into a donor set and a recipient set. The original survey contains over one thousand variables and over 5000 possible variable values and covers a wide variety of consumer products and media.

The recipient set contains 2000 records with a cell variable for gender, common variables for age, marital status, region, number of persons in the household and income. Furthermore, the recipient set contains a unique variable for credit card ownership. One of the goals is to predict this variable for future customers. The donor set contains the remaining 4880 records, with 36 variables for which we expect that there may be a relationship to the credit card ownership: general household demographics, holiday and leisure activities, financial product usage and personal attitudes. These 36 variables are either numerical or Boolean.

First we discuss the specific kind of matching between the databases and then the way the matching is transformed into values of the fusion variables. The matching is done on all common variables. Given an element of the recipient set we search for elements in the donor set that are similar. Elements of the donor set need to agree on the cell variable gender. All the common variables are transformed to numerical values. Next we take as distance on the vectors of values of common variables the root mean squared differences. We select the k best matching elements from the donor. For the values of the fusion variables, we take the average of the corresponding values of the *k* best matching elements of the donor set.

3.1 Internal evaluation

As a baseline analysis we first compared averages for all common variables between the donor and the recipient. As could be expected from the donor and recipient sizes and the fact that the split was done randomly, there were not many significant differences between donor set and recipient set for the common variables. Within the recipient 'not married' was over represented (30.0% instead of 26.6%), 'married and living together' was under represented (56.1% versus 60.0%) and the countryside and larger families were slightly over represented.

Then we compared the average values between the values of the fusion variables and the corresponding average values in the donor. Only the averages of "Way Of Spending The Night during Summer Holiday" and "Number Of Savings Accounts" differed significantly, respectively by 2.6% and 1.5%. Compared to the differences between the common variables, which were entirely due to sampling errors, this was a good result.

Next, we evaluated the preservation of relations between variables, for which we used the following measures. For each common variable, we listed the correlation with all fusion variables, both for the fused data set and for the donor. The mean difference between common-fusion correlations in the donor versus the fused data set was 0.12 ± 0.028. In other words, these correlations were well preserved. A similar procedure was carried out for correlations between the fusion variables with a similar result.

3.2 External evaluation

The external evaluation concerns the value of data fusion for further analysis. Typically only the enriched recipient database is available for this purpose.

We first performed some descriptive data mining to discover relations between the target variable, credit card ownership, and the fusion variables using straightforward univariate techniques. We selected the top 10 fusion variables with the highest absolute correlations with the target (see Table 1).

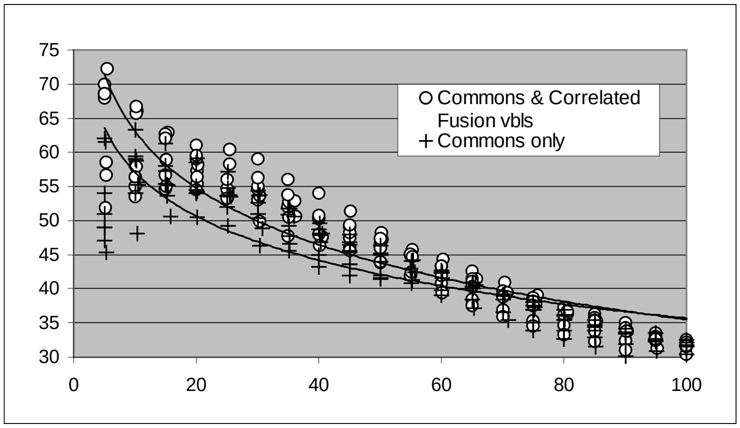

Fig. 2. ***Lift chart linear regression models for predicting credit card ownership (7 randomly selected runs)***

Welfare class	Eating out (privately): money per person
Income household above average	Frequency usage credit card
Is a manager	Frequency usage regular customer card
Manages which number of people	Statement current income
Time per day of watching television	Spend more money on investments

Table 1. ***Fusion variables in recipient strongly correlated with credit card ownership***

	SCG Neural Network	**Linear Regression**	**Naive Bayes Gaussian**	**Naive Bayes Multinomial**	***k*-NN**
Only common variables	*c*=0.692 ± 0.012	*c*=0.692 ± 0.014	*c*=0.701 ± 0.015	*c*=0.707 ± 0.015	*c*=0.702± 0.012
Common and correlated fusion variables	*c*=0.703 ± 0.015 *p*=0.041	*c*=0.724± 0.012 *p*=0.000	*c*=0.720 ± 0.012 p=0.003	*c*=0.720 ± 0.011 *p*=0.200	*c*=0.716 ± 0.013 *p*=0.0093
Common and all fusion variables	*c*=0.694 ± 0.019 *p*=0.38	*c*=0.713 ± 0.013 *p*=0.002	*c*=0.719 ± 0.012 p=0.005	*c*=0.704 ± 0.009 *p* not relevant	*c*=0.720 ± 0.012 *p*=0.0023

Table 2. ***External evaluation results: using enriched data generally leads to improved performance***

Note that this analysis was not possible before the fusion, because the credit card ownership variable was only available in the recipient. If other new variables become available for the recipient customer base in future, e.g. product ownership of some new product, their estimated relationships with the donor survey variables can directly be calculated, without the need to carry out a new survey.

Next we investigated whether different predictive modeling methods would be able to exploit the added information in the fusion variables. The specific goal of the models was to predict a response score for credit card ownership for each recipient, so that they could be ranked from top prospects to suspects. We compared models trained only on values of common variables to models trained on values of common variables plus either all or a selection of correlated fusion variables. We used feed forward neural networks, linear regression, k nearest neighbor search and naive Bayes classification. We provide the details of the algorithms in Appendix A.

We report results over ten runs with train and test sets of equal size. Error criteria such as the root mean squared error (rmse) do not always suffice to evaluate a ranking task. Take for instance a situation where there are few positive cases, say people that own a credit card. A model that predicts that there is no credit card holder has a low rmse, but is useless for the ranking and the selection of prospects. In fact, one has to take the costs and gains per mail piece into account. If we do not have this information, we can consider rank based tests that measure the concordance between the ordered lists of real and predicted cardholders. We use a measure we call the *c*-index, which is a test related to Kendall's Tau [18]. See Appendix B for details about the *c* -index.

The results of our experiments are in Table 2 (c=0.5 corresponds to random prediction and c =1 corresponds to perfect prediction). The results show that for the data set under consideration the models that are allowed to take the fusion variables into account outperform the models without the fusion variables. For linear regression these differences were most significant (one tailed two sample T test; the *p*-value intuitively relates to the probability that the improvement gained by using fusion is coincidental).

In Figure 2, the cumulative response curves are shown for the linear regression models. The elements of the recipient database that belong to the test set are ordered from high score to low score on the x-axis. The data points correspond to the actual proportion of cardholders up to that percentile. Random selection of customers results in an average proportion of 32.5% cardholders. Credit card ownership can be predicted quite well: the top 10% of cardholder prospects according to the model contains around 50-65% cardholders. The added logarithmic trend lines indicate that the models which include fusion variables are better in 'creaming the crop', i.e. selecting the top prospects.

4 Discussion and Future Research

Data fusion can be a valuable, practical tool. For descriptive data mining tasks, the additional fusion variables and the derived patterns may be more understandable and easier to interpret. For predictive data mining, enriching a data set using fusion may make sense, notwithstanding the fact that the fusion variables are derived from information already contained in the donor variables. In general, including derived variables can often improve prediction quality. Fusion may make it easier for algorithms such as linear regression to discover complex non-linear relations between commons and target variables by exploiting the information in the fusion variables. Of course, it is recommended to use appropriate variable selection techniques to remove the noise that is added by 'irrelevant' fusion variables and counter the 'curse of dimensionality', as demonstrated by the experiments.

The fusion algorithms provide a great opportunity for further research and improvement. There is no fundamental reason why the fusion algorithm should be based on k-nearest neighbor prediction instead of clustering methods, decision trees, regression, the expectation-maximization (EM) algorithm or other data mining algorithms. It is to be expected that future applications will require massive scalability. For instance, in the past the focus on fusion for marketing was on fusing surveys with each other, each containing up to tens of thousands of respondents. There exists an important future opportunity to start fusing such surveys with customer databases containing millions of customers.

It goes without saying that evaluating the quality of data fusion is also crucial. We hope to have demonstrated that this is not straightforward and that it ultimately depends on the type of data mining that will be performed on the enriched data set.

Overall, data fusion projects can be quite complex. We have started to describe the phases, steps and choices in a data fusion process model [16], inspired by the CRISP_DM model for data mining [5].

5 Conclusion

We started by discussing how the information explosion provides barriers to the application of data mining and positioned data fusion as a possible solution to the data availability problem. We presented an overview of the main approaches adopted by researchers from outside the data mining community and described a marketing case.

The application of data fusion increases the value of data mining, because there is more integrated data to mine. Data mining algorithms can also be used to perform fusions. Therefore we think that data fusion is an interesting research topic for knowledge discovery and data mining research.

Appendix A: Algorithms Used in the External Evaluation

To repeat, the goal of the external evaluation was to assess the added value of the fusion variables for the prediction of credit card ownership. Whereas the core fusion was solely based on statistical matching, we used a variety of algorithms to build the prediction models for the external evaluation: feedforward neural networks, linear regression, *k* nearest neighbor and naive Bayes models [18].

The feedforward neural networks had a fixed architecture of one hidden layer with 20 hidden nodes using a *tanh* activation function and an output layer with linear activation functions. The weights were initialized by Nguyen-Widrow initialization [9] to enforce that the active regions of the layer's neurons were distributed roughly evenly over the input space. The inputs were linearly scaled between -1 and 1. The networks were trained using scaled conjugate gradient learning [8] as provided within Matlab. The training was stopped after the error on the validation set increased during five consecutive iterations.

For the regression models we used standard least squares linear regression modeling.

For the *k* nearest neighbor algorithm, we used the same simple approach as in the fusion procedure, so without normalization and variable weighting, with k=75.

The naive Bayes algorithm is a well known algorithm based on Bayes rule, using the assumption that the input variables are mutually independent. Let *D* be the training set, *c* a binary target class variable and *x* an input vector to be classified. The a posteriori probability for *c*=1 given *x* can be calculated as follows:

$$
\begin{aligned}
&P(c=1\,|\,x)\\
&=\frac{P(c=1)P(x\,|\,c=1)}{P(x)}\\
&=\frac{P(c=1)P(x\,|\,c=1)}{P(c=0)P(x\,|\,c=0)+P(c=1)P(x\,|\,c=1)}\\
&=\frac{P(c=1)\prod_{i=1}^{n}P(x_i\,|\,c=1)}{P(c=0)\prod_{i=1}^{n}P(x_i\,|\,c=0)+P(c=1)\prod_{i=1}^{n}P(x_i\,|\,c=1)}
\end{aligned}
\qquad \mathbf{(1)}
$$

In the last step we assume conditional independence on the variables given the class. The probabilities in the last part of formulae are then estimated from the training set *D* as follows. The probabilities *P*(*c*=1), *P*(*c*=0) are just the fractions of examples in the training set with class 1 and class 0, respectively. We also have to estimate $P(x_i \mid c=0)$ and $P(x_i \mid c=1)$ for each *i*. For these estimations we take into account whether the data is categorical or numerical.

- Categorical: we assume that each x_i has a multinomial distribution within each class.
- Numerical: we assume that each x_i has a normal distribution within each class.

Hence, for each class and for each element of *x*, we estimate the parameters of the (either multinomial or normal) distribution of x_i from the training set *D* and these we use to estimate $P(x_i \mid \mathrm{c}=0)$ and $P(x_i \mid c=1)$ for each *i*.

Appendix B: c-index

The c-index is a rank based test statistic that can be used to measure how concordant two series of values are, assuming that series is real valued and the other series is binary valued.

Assume that all records are sorted ascending on rank scores. Records can be positive or negative (for example, if they are credit card holders or not). We assign points to all positive records: in fact we give *k*-0.5 points to the *k*-th ranked positive record and records with equal scores share their points. These points are summed and scaled to obtain the *c*-index , so that an optimal predictor results in a *c*-index of 1and a random predictor results in a *c*-index of 0.5. Under these assumptions, the *c*-index is equivalent (but not equal) to Kendall’s Tau; see [18] for details.

The scaling works as follows. Assume that l is the total number of points that we have assigned, and that we have a total of *n* records with *s* positive records. If the *s* positives all have a score higher than the other *n-s* records, then the ranking is perfect and $l = s * (n - s / 2)$. If the *s* positives all have a score that is lower than the *n-s* others, then we have used a worst case model and $l = s^2 / 2$. The *c*-index is thus calculated by:

$$\frac{l - \frac{s^2}{2}}{s(n - \frac{s}{2}) - \frac{s^2}{2}} = \frac{l - \frac{s^2}{2}}{s(n - s)} \qquad \textbf{(2)}$$

Take as an example a score list of (0.1, 0.2, 0.3, 0.4, 0.5) for the targets (0, 0, 0, 1, 1) is optimal: the c-index is 1/6*((3.5 + 4.5)-2)=1. A sub-optimal score list of (0.1,0.2,0.4,0.3,0.5) results in a *c*-index of 1/6*((2.5+4.5)-2)=5/6. A score list of (0.1,0.2,0.4,0.4,0.5) results in a *c*-index of 1/6*((3+4.5)-2)=11/12.

References

[1] BAKER, K, HARRIS, P. AND O'BRIEN, J. *Data Fusion: An Appraisal and Experimental Evaluation.* Journal of the Market Research Society 31 (2) (1989) pp 152-212.

[2] BARR, R.S. AND TURNER, J.S. *A new, linear programming approach to microdata file merging.* In 1978 Compendium of Tax Research, Office of Tax Analysis, Washington D.C. (1978)

[3] O'BRIEN, S. *The Role of Data Fusion in Actionable Media Targeting in the 1990's.* Marketing and Research Today 19 (1991) pp 15-22.

[4] BUDD, E.C. *The creation of a microdata file for estimating the size distribution of income.* Review of Income and Wealth (December) 17, (1971) pp 317-333

[5] CHAPMAN, P., CLINTON, J. KHABAZA, T. REINARTZ., T. WIRTH, R. *The CRISP-DM Process Model.* Draft Discussion paper Crisp Consortium (1999). http://www.crisp-dm.org/.

[6] JEPHCOTT, J. AND BOCK, T.. *The application and validation of data fusion.* Journal of the Market Research Society, vol 40, nr 3 July (1998) pp. 185-205.

[7] GUSFIELD, D. AND ROBERT W. IRVING. *The stable marriage problem: structure and algorithms.* MIT Press, Cambridge.

[8] KAMAKURA, W. AND WEDEL, M., *Statistical data fusion for cross-tabulation.* JMR, Journal of Marketing Research, Chicago; Vol. 34, Iss. 4 (1997) pp. 485-498

[9] LITTLE, R.J. AND DONALD B. RUBIN. *Statistical analysis with missing data.* New York, John Wiley & Sons (1986).

[10] MOLLER, M.F. *A scaled conjugate gradient algorithm for fast supervised learning.* Neural Networks, 6 (1993) pp 525-533.

[11] NGUYEN, D.H. AND WIDROW, B. *Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights.* IJCNN International Joint Conference on Neural Networks. San Diego, CA, USA 17-21 June. (1990) p. 21-6 vol.3

[12] PAASS, G. *Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information.* In: Microanalytic Simulation Models to Support Social and Financial Policy. Orcutt, G.H. and Merz, K, (eds). Elsevier Science Publishers BV, North Holland (1986)

[13] VAN PELT, X. *The Fusion Factory: A Constrained Data Fusion Approach.* MSc. Thesis, Leiden Institute of Advanced Computer Science (2001).

[14] VAN DER PUTTEN, P.. *Data Fusion: A Way to Provide More Data to Mine in?* Proceedings 12th Belgian-Dutch Artificial Intelligence. Conference BNAIC’2000, De Efteling, Kaatsheuvel, The Netherlands (2000)

[15] RADNER D.B, RICH, A. GONZALEZ, M.E. JABINE T.B. AND MULLER, H.J.. *Report on Exact and Statistical Matching Techniques.* Statistical Working Paper 5, Office of Federal Statistical Policy and Standards US DoC (1980). See http://www.fcsm.gov/working-papers/wp5.html

[16] RAMAEKERS, M. *Procesmodel The Fusion Factory.* Sentient Machine Research, Amsterdam, (2000).

[17] RODGERS, W.L. *An Evaluation of Statistical Matching.* Journal of Business & Economics Statistics. Vol 2, nr 1, January (1984)

[18] RUBIN, D.B. *Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations.* Journal of Business and Economic Statistic, January, Vol 4., No 1, (1986).

[19] RUGGLES, N. AND RUGGLES, R. *A strategy for merging and matching microdata sets.* Annals Of Social And Economic Measurement 3 (2) (1974) pp 353-371

[20] DE RUITER, M. *Bayesian classification in data mining: theory and practice.* MSc. Thesis, BWI, Free University of Amsterdam, The Netherlands (1999).