Statistical data-fusion for cross-tabulation

Kamakura, Wagner A.; Wedel, Michel

STATISTICAL DATA-FUSION FOR CROSS-TABULATION

Wagner A. Kamakura
University of Pittsburgh

and

Michel Wedel
University of Groningen

SOM theme B: Marketing and Networks

March 12, 1996

ABSTRACT
The main purpose of this paper is to address the situation where the researcher wants to cross-tabulate two discrete variables collected at different occasions from independent samples. We propose a statistical data-fusion model for cross-tabulation, that allows for statistical tests of association via the technique of multiple imputations. Our approach is illustrated with an application in which we compare the cross-tabulation results from "fused" data with those obtained from complete data.

We thank Prof. Jose A. Mazzon from the University of Sao Paulo, Brazil for providing the data for our illustration.

INTRODUCTION

In this study, we address the following problem: a marketing researcher conducts a survey, and then needs to relate its results to those obtained in a previous study conducted with another independent sample. This problem is not uncommon in marketing research, and is found whenever one needs to consolidate results obtained from two independent samples. Consider, for example the situation of a telecommunications firm in the EC. Faced with the threat of EC competitors entering into its national market, this firm recently issued a large nationally representative segmentation study, to identify the benefits underlying the use of mobile telephones among segments of current and potential customers. The firm wants to use these segments in future research without having to re-assess the benefit variables on which the segments are based. In future research, these segments need to be cross-tabulated with other variables that were not measured in the segmentation study, but that are available from the new studies. The solution offered by the market research industry to this type of problem has been called data fusion.

A second situation arises for market research agencies that maintain consumer panels for consumer nondurables. The panel data typically contain information on brands purchased in a large variety of categories, as well as prices, sales promotions and so on. However, media exposure of the respondents is in general not available, because the collection of such data imposes too great a burden on the panel members and potentially biases buying behavior. Moreover, the costs of collecting single-source data on both purchase behavior and media exposure, are often prohibitive. However, media exposure is available from separate studies supplied by specialized research firms. Market research agencies wanting to provide their customers insight into advertising effectiveness need to cross tabulate product usage assessed in their panels, with media exposure as assessed by the media studies. This may be accomplished by data fusion of purchase and media data, and has become popular as a media planning tool, particularly in Europe (cf. Roberts 1994, Adamek 1994, Buck 1989, Baker, Harris and O'Brien 1989)

The main purpose in these and many other similar situations is to obtain cross tabulations of variables from two studies, with data that have been collected from two disjoint samples of consumers. The problem of combining data from two different sources has been solved by data-fusion; in the statistical literature this class of methods is also known asfile-concatenation.

In this paper, we propose a model-based classification procedure for data-fusion of discrete variables that is based on maximum likelihood. The advantage of such a model-based procedure is that it uses all the information available to impute the missing information in the two independent studies. This model-based procedure also forces the market researcher to state assumptions that might otherwise not be explicit in the data--fusion process. In order to assess the uncertainty associated with the data-fusion process, we use multiple imputations, where the missing values are multiply imputed to reflect the uncertainty about the missing data (Rubin 1986). We show how cross-tabulations of any two partially observed variables in the two data sets, and tests of their association can be computed. The proposed procedure is a multivariate approach to file concatenation, which simultaneously handles all missing variables, while taking into account all the available information. It allows us to directly cross classify variables that are unique to two different samples, and to calculate simple statistical tests of association, accounting for the uncertainty caused by the data-fusion process. We first review previous work on this area, and describe our data fusion model. We then apply our approach to a customer satisfaction study, and assess its validity empirically.

LITERATURE REVIEW

A number of authors have addressed the file concatenation problem in the statistics literature. Reviews are provided by Ford (1983), Rogers (1984) and Rubin (1986). A commonly used class of procedures for the concatenation of two files, say A and B, are the so-calledhot-deck procedures. A hot deck-procedure is essentially a procedure of data-duplication: when a value is missing from sample A (the recipient sample), a reported value is duplicated from sample B (the donor sample) to replace it

(Ford 1983). Thus, each recipient subject is linked to one (or more) donor subjects in file B, on the basis of variables that are observed in both files. This is accomplished using matching algorithms. The common variables on the basis of which subjects are matched are often demographics, but other variables may be included in the files for the specific purpose of data-fusion (Roberts 1994).

A large number of different matching algorithms can be used to search for the linkages of the subjects in the two files. First, when quantitative (metric) variables are assessed, a distance function between subjects in the two samples is defined on the basis of the common variables. The subjects in one file are then matched with subjects in the other file on the basis of a minimum distance function (Rubin 1976). The procedure used may involve regression functions estimated between the common variables and the unique variables in sample A, and subsequently used to predict the unobserved values of these variables in sample B. A distance function relates the predicted values for the subjects in that sample to the measured values in sample A (Rubin 1986).

Second, if the variables are qualitative (discrete), the common variables are grouped into categories, and an exact match between subjects in the two files is sought. For those subjects for which an exact match does not exist, the procedure searches for a match at a lower level of detail, by omitting some variables, or by collapsing the categories of one or more variables. Sometimes, a distinction is made between **critical** variables, for which the match between the donor and recipient must be exact, and matching variables, that are used to find the best possible matchings between donors and recipients within the classes defined by the critical variables. The hot-deck procedure based on discrete variables involve a complicated and arbitrary classification process, where the sample units are classified into disjoint, homogeneous groups, and missing values in the recipient sample are imputed from subjects in the same group in the donor sample (see for example Baker, Harris and O'Brien 1994, Roberts 1994, O'Brien 1991, Buck 1989, Antoine and Santini 1987 for applications to purchase- and media-data fusion)

A major disadvantage of regression-based and hot-deck procedures for data fusion is that they consider only the relationship between each variable to be imputed (i.e., from the unique sets) and the common variables, thus ignoring the information contained in the inter-relationships among variables within each unique set. Further disadvantages of the hot-deck matching procedures for data-fusion are caused by their heuristical nature, not grounded on statistical theory. The absence of theory clouds the subject with inconsistencies and ambiguities. The hot deck procedures involve a number of subjective decisions which may critically affect the quality of the complete data set obtained. The choice of the distance and matching measures, the definition of the levels of the matching procedure in terms of categories and variables, and the distinction between critical and matching variables affects the matches of the subjects in the two files. Moreover, the (statistical) properties of the fused data set are generally unknown. It may not be clear what the quality of the data-fusion is with respect to specific variables, and statistical properties of the fused data set, e.g. with respect to the significance levels of chi-square tests from the required cross-tabulations, are unknown.

The data-fusion approach proposed next is an attempt to overcome some of the limitations of the methods discussed above. Rather than concatenate two independent files by physically matching them according tdemographic criteria only we propose a probabilistic model that first identifies homogeneous groups on the basis all information available from the two samples. Our approach then uses multiple imputations to obtain the posterior joint density functions for each cell in the fused cross-table, thus providing an assessment of the uncertainty caused by the data-fusion process. As shown below, this approach takes into account the information contained in the inter-relationship among all variables within each set of variables, as well as the relationships between the unique sets and the common variables.

DEVELOPMENT OF THE DATA FUSION MODEL

We start by introducing the basic notation required for our approach. We have observed two samples, denoted by A and B, in which sets of categorical variables are measured. Let:

I = 1,...,N indicate subjects in sample A,

I = N+1,...,N+M indicate subjects in sample B,
j = 1,...,P indicate variables unique to sample A,
j = P+1,...,P+Q indicate variables common to samples A and B,
j = P+Q+1,...,P+Q+R indicate variables unique to sample B,
k_j = 1,...,K_j indicate categories of variable j,
t = 1,...,T denote homogeneous imputation groups,
s = 1,...,S denote multiple imputations.

The purpose is to fuse the two samples in order to enable the investigation of bivariate relationships of the discrete variables across them. We will reformulate the data fusion problem into an equivalent missing-data problem. To this end, we form a complete (N+M) by (P+Q+R) data matrix $\mathbf{X}$, in which the missing observations are indicated by $\mathbf{X}^m$, and the observed data by $\mathbf{X}^o$. For example, the variables common to sample A and B can be demographics while the variables unique to sample A and sample B can be choice behavior towards a set of brands and media exposure, respectively. Figure 1 displays the structure of the complete data matrix.

[INSERT FIGURE 1 HERE]

To simplify our presentation, we assume no additional missing data among the P+Q variables in sample A, nor among the Q+R variables in sample B. We also assume an extreme case where there are no subjects for which all variables are assessed. The sample design underlying both samples may be different, e.g. simple or stratified random samples, where the known weights associated with the sampling procedures are denoted by w_i. The stratification variables can be included among the Q variables common to samples A and B. As a consequence of these assumptions, the samples A and B can be considered repeated independent samples from the same population, which enables us to form the data matrix $\mathbf{X}$ displayed in Figure 1.

It is useful to note from Figure 1 that the missing observations have a special structure. The advantages of formulating the data fusion problem as a missing-data problem are that the corresponding missing-data problem has several attractive properties, which facilitate the statistical treatment of the problem. First, the missing-data generation mechanism is not stochastic as in many other missing-value problems, but deterministic because it is determined by the study designs underlying the two samples. Therefore, the missing-data mechanism is *ignorable* (Rubin 1976). A missing-data mechanism is said to be ignorable if (1) the missing-data generation mechanism is not relevant for inference on parameters describing the complete data; and (2) the missing-data are Missing at Random (MAR), which means that the mechanism generating the missing data only depends on the observed data and not on the missing data. Second, the missing data in our data fusion problem are *monotone*, where for any j, x_{ij} is missing implies that $x_{ij'}$ is missing for all j' < j.

Data-fusion depends on the relationship between the missing and non-missing observations. A model that describes the dependencies between these two sets of observations can be used to describe that relationship. For that purpose, we use a mixture model, because of its property of local independence, which is particularly attractive in this context, as shown later. The mixture model formulation theoretically extends the traditional classification hot-deck procedures described above, and is used to derive the homogeneous imputation groups for data-fusion *post-hoc* instead of *a-priori*. Mixture models are by now a very well established and frequently used tool in marketing research, with proven explanatory power (c.f. Dillon and Kumar 1994, Wedel and DeSarbo 1994). The idea here is that we approximate the multivariate distribution of the observations semi-parametrically by mixtures of T homogeneous imputation groups:

$$(1) \qquad f_i(X\,|\,\mathbf{\Omega}) \sim \sum_{t=1}^{T} \eta_t f_{i|t}(X_i\,|\,\boldsymbol{\theta}_t)$$

where $\mathbf{\Omega} = \{\eta_t, \boldsymbol{\theta}_t ; t=1,2,...,T\}$ contains the relevant parameters of the distribution, and X_i denotes the I-th row of *X*. For the purposes of data-imputation, the mixture formulation has the attractive property of local independence, within an imputation group t. This means that, conditional upon the imputation groups, we can write the joint distribution of the missing and non-missing observations as the product of their marginal distributions:

$$(2) \qquad f_{i|t}(X_i^o, X_i^m | \boldsymbol{\theta}_t) = f_{i|t}(X_i^o | \boldsymbol{\theta}_t) f_{i|t}(X_i^m | \boldsymbol{\theta}_t) \; .$$

We assume that the P+Q+R variables are binary or multinomial with K_j categories, and therefore can be described by the following probability distribution function:

$$(3) \qquad f_{ij|t}(x_{ij} | \boldsymbol{\theta}_t) = \prod_{k=1}^{K_j} \left[\frac{e^{\theta_{jkt}}}{\sum_{k'=1}^{K_j} e^{\theta_{jk't}}} \right]^{x_{ijk}}$$

Here, x_{ij} may denote both the non-missing and missing observations. Note that the above model is based upon the assumption that the same imputation groups can be used do describe the relations underlying the P, Q, and R variables in both samples. The set of Q variables is common to both samples and thus provides the link between the P and R variables, through the T components. The effect of a larger number Q of common variables is therefore to increase the strength of the link between these two sets of variables. Moreover, if Q increases, the relative amount of missing information decreases, thus improving the statistical properties of the imputed values. Note however, that the imputation groups are formed on the basis of all available information, thus using the information about the inter-relationships within each unique set of P and R variables. The basic assumption of the model is of local independence among the observed variables in the sample, given the imputation groups. In the model, given t, the variables j and j' are assumed independent, where j indicates the unique and common variables in sample A (j=1,...,P+Q), and j’ indicates the unique and common variables in sample B (j'=P+1,...,P+Q+R).

Model estimation

In order to estimate the model, we first define the likelihood function. The likelihood is by definition proportional to the marginal density of the non-missing values in X, which is obtained by integrating the missing observations X^m out of the joint density of X^m and X^o. Since the missing data mechanism is ignorable for the data-fusion problem, the likelihood can be written as (Rubin 1976):

$$(4) \qquad L(\mathbf{\Omega} \,|\, X^o) = \int \prod_{i=1}^{N+M} w_i \sum_{t=1}^{T} \eta_t f_{i|t}(X_i^o, X_i^m \,|\, \boldsymbol{\theta}_t) d\,X^m$$

Due to the monotonicity of the missing observations and the local independence property of the mixture, it can be shown that the log-likelihood can be factored into:

$$(5) \qquad l(\mathbf{\Omega} \,|\, X^o) = \sum_{i=1}^{N} \ln\{w_i \sum_{t=1}^{T} \eta_t \prod_{j=1}^{P+Q} f_{ij\,|t}(X_i^o \,|\, \boldsymbol{\theta}_t)\} +$$

$$+ \sum_{i=N+1}^{M} \ln\{w_i \sum_{t=1}^{T} \eta_t \prod_{j=P+1}^{P+Q+R} f_{ij\,|t}(X_i^o \,|\, \boldsymbol{\theta}_t)\}$$

Because of the monotone pattern of missing data, equation (5) does not involve the missing observations and therefore the imputation procedure described below is non-iterative. The above log-likelihood will be maximized using an iterative EM algorithm (Dempster, Laird and Rubin 1977). Notice that because the likelihood function is defined for all observed data, the imputation groups and their parameter estimates are obtained on the basis of all available information on the relationship among the unique variables within each sample, and between the unique and common variables.

Once estimates of the model parameters are obtained, estimates of the posterior probabilities, π_{it}, of observation I relative to imputation group t can be calculated using Bayes' Theorem. These posteriors are calculated on the basis of the subject-specific likelihoods, which are again obtained by integrating out the missing observations from the joint distribution of X^m and X^o. It can be shown that these posteriors in sample A and B equal:

$$(6) \qquad \pi_{it} = \frac{\eta_t \prod_{j=1}^{P+Q} f_{ij|t}(x_{ij} \mid \boldsymbol{\theta}_t)}{\sum_{t=1}^{T} \eta_t \prod_{j=1}^{P+Q} f_{ij|t}(x_{ij} \mid \boldsymbol{\theta}_t)}, \quad (i=1,\ldots,N)$$

for subjects in sample A and,

$$(7) \qquad \pi_{it} = \frac{\eta_t \prod_{j=P+1}^{P+Q+R} f_{ij|t}(x_{ij} \mid \boldsymbol{\theta}_t)}{\sum_{t=1}^{T} \eta_t \prod_{j=P+1}^{P+Q+R} f_{ij|t}(x_{ij} \mid \boldsymbol{\theta}_t)}, \quad (i=N+1,\ldots,N+M)$$

for subjects in sample B.

In the EM algorithm, the E- and M-steps are alternated until no significant improvement in the likelihood function is possible. We will not provide the derivation of the EM algorithm for this problem since it is straightforward and available elsewhere (Dempster, Laird and Rubin 1977) , but will schematically summarize its main steps:

1. At the first step of the iteration, h=0, the algorithm is initialized by fixing the number of imputation groups T, and generating a starting partition $\pi_{it}^{(0)}$.
2. Given $\pi_{it}^{(h)}$, M.L. estimates of $\eta_t^{(h)}$ are obtained from the closed form expression:

$$(8) \qquad \eta_t = \frac{\sum_{i=1}^{N+M} w_i \pi_{it}}{\sum_{i=1}^{N+M} w_i},$$

3. The $\theta_t^{(h)}$ are estimated in three steps for the two sets of variables idiosyncratic to

samples A and B and the common variables, by maximizing respectively:

(9)
$$\sum_{i=1}^{N} \ln\{w_i \sum_{t=1}^{T} \pi_{it} \prod_{j=1}^{P} f_{ij|t}(X_i^o | \boldsymbol{\theta}_t)\}$$

(10)
$$\sum_{i=N+1}^{M} \ln\{w_i \sum_{t=1}^{T} \pi_{it} \prod_{j=P+Q+1}^{P+Q+R} f_{ij|t}(X_i^o | \boldsymbol{\theta}_t)\}$$

(11)
$$\sum_{i=1}^{N} \ln\{w_i \sum_{t=1}^{T} \pi_{it} \prod_{j=P+1}^{P+Q} f_{ij|t}(X_i^o | \boldsymbol{\theta}_t)\} + \sum_{i=N+1}^{N+M} \ln\{w_i \sum_{t=1}^{T} \pi_{it} \prod_{j=P+1}^{P+Q} f_{ij|t}(X_i^o | \boldsymbol{\theta}_t)\}$$

4. Convergence test: stop if the change in the log-likelihood from iteration (h-1) to iteration (h) is sufficiently small.
5. $h \leftarrow h+1$, calculate new estimates of the posterior membership according to equations (9) and (10), goto step 2.

Under standard regularity conditions, the estimates of the parameters of the model are asymptotically normal, with covariance matrix Σ_{θ}, which is estimated by inverting the negative Hessian matrix of second derivatives. In order to determine the appropriate number of imputation groups T we will use the Consistent Akaike Information Criterion (CAIC, Bozdogan 1987). We select that number of groups T for which the CAIC reaches a minimum. One must also ensure that the posterior probabilities provide well separated imputation groups. This is particularly important in light of the imputation procedure described below. We use an entropy statistic E_T to investigate the degree of separation in the estimated posterior probabilities (cf. Wedel and DeSarbo 1994). E_T is a relative measure and is bounded between 0 and 1. Values close to 1 indicate that the posteriors are well separated. A value close to zero is of concern as it implies that the posteriors are not sufficiently well separated.

Multiple imputation of the missing data

Once the parameters of the model are obtained from the above procedure, the data-fusion process involves the imputation of the missing observations X^m. Due to the monotone pattern of missing observations, the imputation procedure is non-iterative, which offers obvious computational advantages. The imputed values are based on the predictive distribution of the missing values, given the observed values. There are important problems in imputing a single, “best prediction” value for one missing value. A single imputed value cannot represent the uncertainty involved in the imputation process, and therefore generally leads to an underestimation of uncertainty in subsequent analyses of the imputed data set. Several studies have shown that these effects can be quite serious and that multiple imputation is far superior to single imputation with regard to the estimates of confidence intervals and significance levels (cf. Little and Schenker 1995). In our application this is of important since we want to assess the significance level of the association of pairs of discrete variables. Multiple imputation alleviates this problem by replacing each missing observation with $S(\geq 2)$ values. From a Bayesian perspective, the data-imputation procedure involves drawing S sets of values from the posterior density of the X_i^m, which is obtained by integrating the density of X_i with respect to the distribution of the parameters:

$$(12) \qquad f_i(X_i^m \mid X_i^o) = \int f_i(X_i^m \mid X_i^o, \tilde{\mathbf{\Omega}})\, f(\tilde{\mathbf{\Omega}} \mid X^o) d\tilde{\mathbf{\Omega}}$$

By inserting the posterior conditional density of X_i^m in (12), we obtain:

$$(13) \qquad = \int \sum_{t=1}^{T} \tilde{\pi}_{it} f_{i|t}(X_i^m \mid X_i^o, \tilde{\mathbf{\Omega}})\, f(\tilde{\mathbf{\Omega}} \mid X^o) d\tilde{\mathbf{\Omega}}$$

Now, the purpose is to cross-classify two discrete variables from the sets $\{X_j$; j=1,...,P} and $\{X_{j'}$; j'=P+Q+1,...,P+Q+R), resulting in a K_j by $K_{j'}$ contingency table.

This is equivalent to estimating the joint probability distribution of X_j and $X_{j'}$: $\boldsymbol{\phi} = \{\phi_{kl};$ $k=1,...,K_j$, $l=1,...,K_{j'}$), where part of the observations on both variables are missing (for convenience of notation we will omit the subscripts for the variables j and j' and their categories k and l). The posterior density of these parameters can be written as:

$$(14) \qquad f(\boldsymbol{\phi}\,|\,X^o) = \int f(\boldsymbol{\phi}\,|\,X^o,X^m) \prod_{i=1}^{N+M} w_i[f_i(X_i^{\,m}\,|\,X_i^{\,o})]dX^m$$

By substitution of equation (13) we obtain:

(15)

$$f(\boldsymbol{\phi}\,|\,X^o) = \int f(\boldsymbol{\phi}\,|\,X^o,X^m) \prod_{i=1}^{N+M} w_i[\int \sum_{t=1}^{T} \tilde{\pi}_{it} f_{i\,|\,t}(X_i^{\,m}\,|\,X_i^{\,o},\tilde{\boldsymbol{\Omega}}) f_i(\tilde{\boldsymbol{\Omega}}\,|\,X_i^o) d\tilde{\boldsymbol{\Omega}}]dX^m$$

In using equation (15) for the imputation of the missing values, we first draw a value of $\boldsymbol{\Omega}$ from its posterior distribution $f_i(\boldsymbol{\Omega}|X_i^{\,o})$, which is asymptotically MV-Normal with covariance matrix $\Sigma_{\boldsymbol{\Omega}}$, thus accounting for the estimation error. Then we draw t from π_{it}, and finally we draw X^m conditional upon the value of those parameters from $f_{i|t}(X_i^{\,m}|X_i^{\,o})$ (using sampling weights w_i, whenever required). This is repeated S times, and the posterior density of the parameters is computed by averaging over the S repeated samples.

The imputation procedure can be summarized as follows. We start by imputing the R missing variables for I=1,...,N:

1. Initialize s =1;
2. Initialize I^*=1;
3. Sample $\boldsymbol{\Omega}$ from MVN($\boldsymbol{\Omega}$,$\boldsymbol{\Sigma}_{\Omega}$);
4. Calculate $\pi_{i*t}(\boldsymbol{\Omega})$;
5. Sample t from $\pi_{i*t}(\boldsymbol{\Omega})$;
6. Sample $X_{i*}^{\,m}$ from $f_{i|t}(X_i^{\,o}$, I=N+1,...,N+M; j=P+Q+1,...,P+Q+R; $\pi_{i*t}(\boldsymbol{\Omega})=\pi_{it}$);

7. $I^*=I^*+1$, repeat steps 3 to 6 if $I^* \leq N$;

8. $s=s+1$, repeat steps 2 to 7 if $s \leq S$.

The imputation procedure of the P variables for $I^*=N+1,...,N+M$ is the same. Thus, in multiple imputation, the S imputed data sets are analyzed and the results are combined. For moderate fractions of missing information, a relatively small number (e.g. S=3) of draws suffices. However, since the fraction of missing information in the data fusion problem is relatively large, a larger number (S) of imputations must be used.

Testing for association in the fused table

Our multiple imputation procedure produces one cross-tabulation $\boldsymbol{\phi}=\{\phi_{11s},...,\phi_{KjKj's}\}$ for each of the S imputations. The final estimates of the cell proportions in the cross-tabulation ϕ_{kl} are obtained by averaging the S estimates for each imputation:

$$(16) \qquad \bar{\boldsymbol{\phi}} = \frac{1}{S}\sum_{s=1}^{S} \hat{\boldsymbol{\phi}}_s,$$

Estimates of the covariance matrix of the probabilities can be obtained from (Meng and Rubin 1992):

$$(17) \qquad \hat{S}_{\boldsymbol{\phi}}^{2} = \bar{W}_{\boldsymbol{\phi}} + (1+\frac{1}{S})B_{\boldsymbol{\phi}},$$

where W is the average within-imputation covariance matrix with elements:

$$(18) \qquad \bar{W}_{\boldsymbol{\phi}} = \frac{1}{(NM-1)S}\sum_{s=1}^{S} (\mathrm{diag}(\hat{\phi}_{kls}) - \hat{\boldsymbol{\phi}}_s\hat{\boldsymbol{\phi}}_s^{\prime\prime}), \qquad (18)$$

and B is the between-imputations covariance matrix:

$$B_{\phi} = \frac{1}{S-1}\sum_{s=1}^{S}(\hat{\phi}_s - \bar{\phi})(\hat{\phi}_s - \bar{\phi})'. \qquad (19)$$

In order to test for association between variables j and j’ in the fused table, we use a modification of the likelihood-ratio (LR) test statistic proposed by Meng and Rubin (1992), which takes into account the uncertainty arising from the imputation of missing information. Meng and Rubin’s statistic D for testing the independence assumption of the variables j and j' after data-fusion is computed by:

$$(20 \qquad D = \frac{D_B}{H(1+\gamma)}$$

where

$$(21) \qquad \gamma = \frac{S+1}{H(S-1)}(D_W - D_B)$$

is the estimated average odds ratio of the fraction of missing information with respect to the parameters ϕ. In order to compute the statistic D, it suffices to have access to the LR test-statistics for each of the S fused data sets, and the estimates of the joint probabilities ϕ_s in each of these data sets. Both are routinely computes by standard packages. D_B is simply the LR test statistic calculated at the average of the S estimates of the joint probabilities, ϕ_{kl}, D_W is the average of the LR statistics, calculated at each estimate $\hat{\phi}_{kls}$.

Under the null hypothesis the test-statistic D follows an approximate asymptotic $F_{H,G}$-distribution, with the denominator degrees of freedom, G, calculated as:

$$(22) \qquad G = 4 + (H(S-1) - 4)\{1 + \frac{1}{\hat{\gamma}}(1 - \frac{1}{2}H(S-1))\}^2$$

EMPIRICAL ILLUSTRATION

In this section we will apply the proposed statistical data fusion method to data from a customer satisfaction study. The study pertains the measurement of customer satisfaction among the most valuable customers of a large multi-branch bank in Latin America. Three types of variables have been assessed, among many others:

1 Eight measures assessing the staff (perceived politeness, attentiveness, availability of clerks), binary coded ("a little" or "not at all" versus "a lot"), one measure of bank use and one measure of satisfaction (whether the respondent would recommend the bank to friends) on a 5-point scale (from "certainly not" to "certainly yes"),

2 Fifteen variables on demographics (including gender, age, education) and the usage of a number of the bank's products (savings, credit cards, insurances etc.),

3 Eight performance variables available from internal records, including number of transactions per month, contribution of the account, funds in the bank, marginal contribution of the account per $, etc.

The data for this illustration come from a single study, so that in fact, observations on all variables are obtained for all 2000 subjects randomly sampled for our tests. We will however simulate a situation with two different samples, by randomly assigning the subjects to two groups. We assume that the variables for the first group have been collected by a survey among the bank customers, in which only the customer satisfaction measurements and the demographics and product-usage variables have been assessed. For this group the performance variables from internal records are assumed missing, and are not used for estimation. For the second group, we assume that the data are obtained from internal records, from which only the demographics, product-usage and performance variables are available. For this group, the satisfaction measurements are assumed not to be known, and these are ignored in the estimation stage.

The main purpose of this treatment of the original data is to allow for an objective assessment of the data-fusion results. Using the complete data in this way allows us to assess the validity of the proposed imputation procedure. We compare the imputed cross tables from the incomplete data to the actual cross tables from the complete data. We also assess the correspondence between the association tests computed from the complete tables and those obtained from the imputed tables. The first test will show how well the imputed data fit to the complete data. The second test will indicate whether the fused tables would lead to the same conclusions regarding the association between the two discrete variables in each cross-tabulation.

Estimation results

The first step in the analysis is to estimate the mixture imputation model provided by equations (1) to (3). This model was applied to the data for T=1 to T=7 imputation groups. Table 1 provides the log-likelihood, the CAIC statistic, and the ρ^2 measure of explained variance (relative to the aggregate model) for these solutions. Each of the solutions reported is the best among 15 solutions obtained from different starts, in order to overcome problems of local optima. The CAIC statistic indicates that the T=5 solution provides the best representation of the data, and consequently, we will use that solution as the basis of our imputation procedure. Table 2 provides the parameter estimates of the T=5 solution. The sizes of the five imputation groups are: 22.3%, 21.8%, 15.7% 19.8% and 20.4%, respectively. The imputation groups are well-separated: the entropy criterion calculated from the posteriors equals 0.85. In this illustration, we are not interested in the actual interpretation of the estimates in Table 2, since we only want to use them as a vehicle for the imputation of the sets of the missing observations in each of the two groups.

[INSERT TABLE 1 HERE]

[INSERT TABLE2 HERE]

Imputation results

In order to assess the degree of fit between the tabulations of actual data and our imputed estimates, we processed all 80 possible pairs of the 10 survey variables and 8 internal measures, and compared the imputed cell frequencies on these tables against the actual frequencies. As expected, the fit between the cell frequencies from the "fused" tables and those from the complete data is not perfect, since the imputed cell frequencies were obtained from incomplete data. Nevertheless, our imputation procedure performs satisfactorily, showing a correlation of 0.80 with the complete data, across all 960 cells of 80 two-way tables. This comparison is shown in Figure 2, where each observation is labeled by the number of the column variable (from set III) in the cross-tabulation. Figure 2 clearly shows that the imputed frequencies depart the most from the actual ones when the tabulation involves variables III-2, III-3 and III-7, and to a lesser extent, III-8. The parameter estimates displayed in Table 2 provide an explanation for these results. These estimates (for set III) show quite extreme (i.e., zero) within-class probabilities for variables III-2, III-3, III-7 and III-8, which led to extreme cell frequencies in the imputed tables.

[FIGURE 2 ABOUT HERE]

The proposed data-fusion procedure can be further assessed by comparing the results from tests of association in the imputed cross-tabulations with those obtained from the actual (i.e., complete) ones. Table 3 summarizes the results of these tests across all 80 cross-tabulations, at the 0.05 and 0.10 significance levels. One can see that these tests agreed in 63 and 62 of the 80 tables at the 0.05 and 0.10 significance levels, respectively. In relative terms, the tests of the imputed tables (Meng & Rubin 1992) were more likely (7 out of 30 cases) to produce false positives (i.e., finding association when there was independence) than false negatives (10 out of 50 cases).

[TABLE 3 ABOUT HERE]

To illustrate our imputation procedure, we will focus on a selection of cross-classifications of the variables in Set I and in Set III that serves to highlight both the strengths and the weaknesses of the proposed procedure. First, we cross-classify the satisfaction survey variables I-5 “Clerks are fast/agile” and I-9 “Clerks are available when needed” from Set I, with the variables III-1 “Respondent’s number of transactions”, III-3 “Respondent’s funds in the bank”, III-4 “Number of customers in respondent’s branch”, and III-5 “Ratio of customers per employees in respondent’s branch”, from set III. The purposes of these cross-classifications are two-fold: a) to investigate how the evaluation of the bank’s clerks, obtained from the survey, relates to clients’ frequency and volume of service use, obtained from internal records, and b) to test whether certain characteristics of a branch (number of customers, and ratio of customers per employee) are related to customers’ assessment of its personnel.

Second, we cross-tabulate the survey variables (Set I): I-10 “Would recommend the bank to friends”, and I-1 “Percentage of funds the customer states to have in this bank”, with the variables from the internal records (Set III) III-1 “Respondent’s number of transactions”, III-2 “Contribution of the respondent’s account”, III-3 “Respondent’s funds in the bank”, III-7 “Respondent’s applications per transaction”, and III-8 “Marginal contribution of the respondent’s account (per $)”. The purpose is to relate overall satisfaction measures from the survey, to these specific performance variables.

[INSERT TABLE 4 HERE]

Table 4 displays the Likelihood-Ratio (LR) test statistic and the corresponding p-value calculated from the actual cross classification, as well as the results from the Meng and Rubin F-test (MR) and its corresponding p-value calculated from the imputed data. (The number of imputations used was 100.) The table also shows $\hat{\gamma}/(1+\hat{\gamma})$, the estimated effective fraction of missing information. This table shows that the estimated fractions of effective missing information vary across the cross-tables, but are in all cases greater than 0.5. In four of the cross-tabulations, the test-statistic on the actual data does not reject the nul-hypothesis of independence at p<0.05. The test statistic from the imputed data agrees with these results, leading to the conclusion that “Clerks are fast, agile” is not related to the number of customers in the respondent’s branch, but is related to the ratio of customers per employee in their main branch. On the other hand, “Clerks are available when needed” is not associated to the respondent’s total number of transactions, nor to the number of customers or ratio of customer per employee in the respondent’s main branch. Both tests on the actual and the imputed data support these findings, with a few differences in the significance levels of the tests, as discussed above. The indirect measures of satisfaction, i.e., whether the customer would recommend the bank to friends and the percentage of funds the customer allocates to this bank, are significantly associated with the amount of funds in the bank, the volume of applications per transaction, and the marginal contribution of the account per unit of deposits.

We inspect four of the cross classification tables in detail in order to further reveal strengths and weaknesses of the proposed procedure. We show two cross-tables for which the tests on the imputed data correspond to the tests on the actual data, and two tables for which this is not the case. Table 5 depicts the results of the cross-classifications of the survey variables I-10 “Would recommend bank to friends” and I-1 “%Of funds in the bank”, with the internal record variable III-1 “Number of transactions”. The estimates of the ϕ_{kl} are obtained by averaging the S estimates for each imputation (see equation 16). The table shows the estimates of the joint classification probabilities $\phi_{kl,}$ from the actual data, and from the imputed data. The joint classification probabilities are expressed as percentages of the grand total. Note from Table 4 that the association of I-10 and III-1 is significant for both the actual and the imputed data, while the association of I-1 and III-1 is significant for the actual table, but it is not for the imputed data. Apparently the variability across the 100 imputations in this case (left-hand panel of Table 5) causes the association not to be significant for the incomplete data. Table 5 shows that the customers that would recommend the bank to friends execute more transactions per month (i.e., are heavy users of the bank). The percentages calculated from the imputed and actual data (expressed as percentages of the grand total) are quite close in both the left-hand panel of Table 5; the U^2 measure of fit of the actual and imputed table is 94.2%. The right-hand panel of Table 5 shows a positive association of having more funds in the bank and the number of transactions per month. This holds for both the actual (in parenthesis) and the imputed values, and the percentages are relatively close; the U^2 measure of fit between the actual and imputed tables is 86.5%.

[INSERT TABLE 5 HERE]

Table 6 shows the cross-classifications of the survey variables (Set I) 5, “Clerks are fast/agile”, and 9, “Clerks are available when needed”, with the “Ratio of customers to employees” from the internal records (Set III variable 5). From Table 4 it can be seen that the association of I-5 and III-5 is significant for both the actual and the imputed data. The association of I-9 and III-5 is not significant for both the actual and imputed tables. The left panel of Table 6 shows some intuitive results: when the ratio of customers to employees increases, the percentage of customers who rate clerks as “fast and agile” decreases for both the imputed and the actual data. The percentages calculated from the imputed and actual data (expressed as percentages of the grand total) in the left-hand panel of Table 6 are relatively close; the U^2measure of fit between the actual and imputed tables is 79.5 %. The right-hand panel of Table 6 shows that for the actual data (in parenthesis) the percentage of customers indicating that clerks are available decreases with the customers/employees ratio. Note that this effect is not significant in either case. The degree of fit between the actual and imputed tables is also relatively weaker, with a U^2 of 67.5 %.

[INSERT TABLE 6 HERE]

CONCLUSIONS AND DISCUSSION

Data fusion, coming originally from the statistical sciences, is a recent innovation for market researchers, being applied and further developed over the last ten years. Problems that lend themselves to data fusion are relatively common in marketing research. The demands placed on the market research industry with respect to the collection of data are constantly increasing. Managers nowadays do not only require insights into the buying behavior of their potential customers, but also on their demographic and socio-economic profile, their life-styles, attitudes, media exposure, and so on. The burden on consumers and business by extensive mail and telephone interviews is constantly increasing, and the response rates and the quality of the data are declining in consequence. On the other hand, the costs of collecting complete single source data are often prohibitive. Therefore the need of combining data from different sources is increasing, and data fusion presents a potential solution to the problem. Fused databases already have already been successfully used in practice, in particular in applications that fuse product usage and TV-viewing (Adamek 1994). Nevertheless, fusion cannot be claimed to be superior to single source data, but it may be a viable alternative when single source data is unavailable, or too expensive to collect.

In this paper we have presented an approach to data fusion that presents the possibility of performing appropriate statistical tests on fused databases, specifically for the common situation where cross-tables of variables assessed in different samples are required. The procedure is model-based, and formulated within a mixture model framework. Due to the relatively simple computational procedure for imputation, a large number of imputations can be performed. Using complete data, we have provided evidence of the validity of the proposed procedure. The tests of association in each of the 80 cross-tabulations show the same conclusions for the complete and imputed tables in 77.5% (62/100) of the tables. A comparison of 80 imputed tables with those obtained from complete data showed a reasonable fit, with a correlation of 0.80 across all 980 cell frequencies. Furthermore, we found that the imputations departed the most from the actual table when that table involved a variable with extreme (near zero) estimates of within-class probabilities. In such situations it would be advisable to recode the discrete variables to a smaller number of categories in order to avoid extreme within-class probabilities, and thus improve the quality of the imputations.

An important determinant of the quality of a data fusion are the number of common variables, and the strength of their relationships with the variables in the two samples. The final quality of the data fusion is therefore strongly dependent on these common variables. This holds not only for previous data fusion methods, but also for the method presented here. We suggest that future research should address these problems: the choice of appropriate common variables, the optimal number of common variables, and the effects of sample sizes on the quality of data fusion. Furthermore, the procedure could be extended beyond the application of cross-tabulation, including for example models tests for continuous variables, including regression and analysis of variance.

Table 1

Log-likelihood, CAIC and ρ^2 for the T=1 to T=7 imputation models

T	Log-l	CAIC	ρ^2
1	-36484.63	73519.72	0.1298
2	-33338.07	67871.01	0.1850
3	-32500.19	66668.96	0.1970
4	-31954.07	66135.77	0.2038
5	-31589.97	65966.62 *	0.2076
6	-31367.50	66080.76	0.2087
7	-31121.15	66147.11	0.2110

* Indicates the minimum value of CAIC

Table 2a: Parameter estimates - Set I: satisfaction variables (survey)

Segment	1	2	3	4	5	
1. Would recommend bank to friends						
Certainly not	.012	.051	.519	.072	.237	
	.005	.128	.202	.044	.253	
	.131	.410	.203	.233	.347	
Certainly yes	.852	.411	.076	.651	.167	
2. Clerks are competent						
A bit	.005	.398	.829	.014	.964	
A lot	.995	.602	.171	.986	.036	
3. Clerks are attentive						
A bit	.016	.239	.832	.050	.969	
A lot	.984	.761	.168	.950	.031	
4. Clerks are polite						
A bit	.006	.091	.677	.046	.837	
A lot	.994	.909	.323	.954	.163	
5. Clerks are fast, agile						
A bit	.072	.768	.976	.157	.987	
A lot	.928	.232	.024	.843	.013	
6. Clerks answer questions						
A bit	.013	.659	.881	.034	.958	
A lot	.987	.341	.119	.966	.042	
7. Clerks are well informed						
A bit	.056	.823	.938	.099	.988	
A lot	.944	.177	.062	.901	.012	
8. Clerks are well groomed						
A bit	.033	.307	.673	.022	.785	
A lot	.967	.693	.327	.978	.215	
9. Clerks are available when needed						
A bit	.038	.478	.928	.077	.967	
A lot	.962	.522	.072	.923	.033	
10. % of funds in this bank						
<25%	.102	.174	.772	.357	.181	
25-50%		.065	.162	.103	.163	.242
50-75%		.132	.111	.038	.158	.096
75-100%		.701	.553	.087	.322	.481

Table 2b: Parameter estimates - Set II: demographics and status variables (common variables)

Segment	1	2	3	4	5
1. Gender					
Male	.195	.212	.355	.285	.209
Female	.805	.788	.645	.715	.791
2. Age					
<40	.149	.279	.355	.137	.278
40-50	.303	.428	.390	.218	.372
50-60	.289	.203	.150	.248	.246
>60	.259	.090	.105	.397	.104
3. Education					
Junior	.085	.021	.088	.303	.041
High	.191	.189	.131	.139	.148
College	.724	.790	.781	.558	.811
4. Savings usage					
Use	.760	.685	.304	.650	.710
Don’t	.240	.315	.696	.350	.290
5. Credit Card usage					
Use	.368	.304	.080	.062	.252
Don’t .632	.696	.920	.938	.748	
6. Bank Card usage					
Use	.857	.902	.631	.618	.871
Don’t	.143	.098	.369	.382	.129
7. Certificate of Deposit usage					
Use	.687	.434	.058	.347	.498
Don’t	.313	.566	.942	.653	.502
8. Special Checking usage					
Use	.936	.915	.447	.567	.897
Don’t	.064	.085	.553	.433	.103
9. Automatic Bill Payment usage					
Use	.773	.695	.240	.437	.687
Don’t	.227	.305	.760	.563	.313
10. Mutual fund usage					
Use	.168	.114	.011	.029	.126
Don’t	.832	.886	.989	.971	.874
11. Tax Deferred Fund usage					
Use	.698	.589	.114	.341	.619
Don’t	.302	.411	.886	.658	.381
12. Commodities Fund usage					
Use	.449	.273	.042	.172	.390
Don’t	.551	.727	.958	.828	.610
13. Fixed Rate fund usage					
Use	.237	.098	.008	.057	.169
Don’t	.763	.902	.992	.943	.831
14. Auto insurance usage					
Use	.279	.222	.009	.022	.161
Don’t	.721	.778	.991	.978	.839
15. Home insurance usage					
Use	.204	.171	.007	.019	.118
Don’t	.796	.828	.993	.981	.882

Table 2c: Parameter estimates - Set III: performance variables in quintiles (internal records)

Segment	1	2	3	4	5
1. Number of transactions in respondent's account					
1 bottom 20%	.088	.013	.355	.504	.008
2	.227	.070	.298	.323	.164
3	.257	.303	.189	.101	.233
4	.199	.279	.100	.023	.262
5 top 20%	.229	.334	.059	.048	.333
2. Contribution of respondent's account					
1 bottom 20%	.000	.660	.112	.000	.162
2	.000	.259	.493	.000	.378
3	.034	.053	.271	.154	.460
4	.492	.025	.088	.416	.000
5 top 20%	.475	.003	.037	.429	.000
3. Respondent's funds in bank					
1 bottom 20%	.000	.340	.833	.000	.000
2	.000	.623	.167	.054	.073
3	.027	.037	.000	.251	.634
4	.378	.000	.000	.337	.284
5 top 20%	.595	.000	.000	.358	.009
4. Number of customers in respondent's branch					
1 bottom 20%	.127	.241	.174	.084	.174
2	.294	.244	.225	.159	.225
3	.170	.193	.098	.192	.176
4	.236	.165	.238	.329	.255
5 top 20%	.173	.157	.265	.236	.170
5. Customers/employee in respondent's branch					
1 bottom 20%	.184	.114	.151	.165	.198
2	.193	.147	.076	.253	.140
3	.178	.194	.191	.225	.226
4	.306	.249	.376	.221	.213
5 top 20%	.139	.297	.206	.135	.223
6. Customers/managersratio in respondent's branch					
1 bottom 20%	.187	.075	.189	.143	.195
2	.172	.218	.155	.220	.169
3	.226	.232	.240	.198	.291
4	.207	.213	.247	.247	.137
5 top 20%	.207	.263	.169	.192	.208
7. Volume of applications per transaction in respondent's account					
1 bottom 20%	.000	.403	.758	.000	.000
2	.000	.597	.227	.005	.080
3	.080	.000	.014	.149	.750
4	.439	.000	.000	.280	.170
5 top 20%	.480	.000	.000	.566	.000
8. Marginal contribution (per $) of respondent's account					
1 bottom 20%	.000	.689	.359	.000	.011
2	.048	.215	.100	.028	.688
3	.434	.012	.010	.317	.157
4	.437	.036	.025	.204	.144
5 top 20%	.081	.047	.506	.451	.000

Table 3: Summary of tests of association for all 80 cross-tabulations

Conclusions at 5% significance level

Actual \ Imputed	Association	Independence	Total Actual
Association	40	10	50
Independence	7	23	30
Total Imputed	47	33	80

Conclusions at 10% significance level

Actual \ Imputed	Association	Independence	Total Actual
Association	43	11	54
Independence	7	19	26
Total Imputed	50	30	80

Table 4 : χ^2 tests on the actual data (LR) and the Meng and Rubin test (MR) on the fused data

Set I (Survey)	Set III (Internal Records)	Df	LR	p-value	(/(1+ ()	MR	p-value
Clerks are fast, agile	Respondent's number of transactions	4	18.3	0.00	0.80	2.77	0.03
Clerks are fast, agile	Respondent's funds in the bank	4	51.4	0.00	0.95	8.39	0.00
Clerks are fast, agile	Number of customers in respondent's branch	4	4.55	0.34	0.61	1.36	0.25
Clerks are fast, agile	Customers/employee in respondent's branch	4	56.4	0.00	0.54	3.23	0.01
Clerks are available when needed	Respondent's number of transactions	4	7.89	0.10	0.81	1.44	0.22
Clerks are available when needed	Respondent's funds in the bank	4	28.6	0.00	0.96	5.26	0.00
Clerks are available when needed	Number of customers in respondent's branch	4	1.79	0.77	0.59	0.84	0.47
Clerks are available when needed	Customers/employee in respondent's branch	4	6.36	0.17	0.57	2.16	0.08
Would recommend the bank	Respondent's number of transactions	12	65.8	0.00	0.57	0.93	0.51
Would recommend the bank	Contribution of respondent's account	12	105.6	0.00	0.75	8.98	0.00
Would recommend the bank	Respondent's funds in the bank	12	132.3	0.00	0.84	4.78	0.00
Would recommend the bank	Respondent's volume/transaction	12	110.4	0.00	0.79	6.86	0.00
Would recommend the bank	Respondent's unit contribution(per $)	12	34.3	0.00	0.80	3.32	0.00
% of funds in the bank	Respondent's number of transactions	12	165.8	0.00	0.41	3.66	0.00
% of funds in the bank	Contribution of respondent's account	12	113.9	0.00	0.73	1.52	0.12
% of funds in the bank	Respondent's funds in the bank	12	285.6	0.00	0.74	4.10	0.00
% of funds in the bank	Respondent's volume/transaction	12	274.8	0.00	0.66	4.22	0.00
% of funds in the bank	Respondent's unit contribution (per $)	12	115.3	0.00	0.71	2.29	0.01

Table 5

Cell Percentages of imputed and actual (in parenthesis) tables crossing column variables with “Number of transactions ” (III-1, in quintiles)

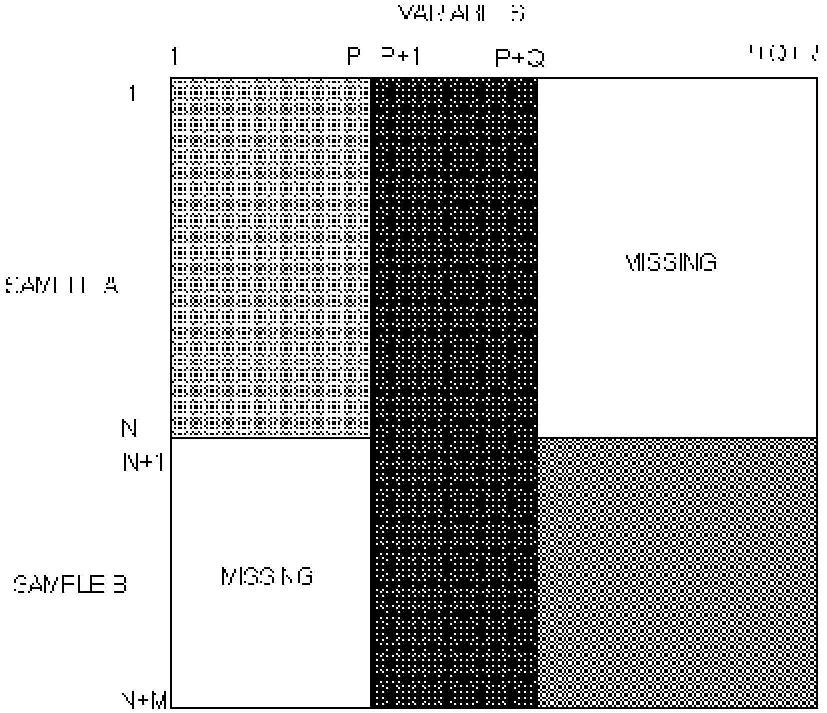
	I-1. “Recommend bank to friends ”				I-10. “% Of funds in bank ”			
III-1	1.No	2	3	4.Yes	0-25	25-50	50-75	75-100
1 bottom 20%	3.5 (4.0)	1.5 (2.1)	3.9 (4.7)	8.7 (7.8)	8.1 (9.1)	2.4 (3.6)	2.1 (1.9)	5.3 (4.4)
2	3.8 (3.6)	2.2 (3.2)	4.9 (4.0)	10.0 (9.2)	7.2 (6.8)	2.9 (2.9)	2.3 (2.8)	8.5 (7.7)
3	3.2 (3.1)	2.8 (3.7)	6.2 (6.5)	9.9 (8.7)	5.6 (4.8)	3.2 (3.3)	2.4 (2.6)	10.9 (11.1)
4	2.5 (2.2)	2.5 (2.8)	5.4 (4.7)	7.6 (8.4)	3.8 (4.6)	2.8 (2.3)	1.9 (1.7)	9.4 (9.9)
5 top 20%	2.6 (1.9)	2.9 (1.6)	6.5 (4.8)	9.2 (12.4)	4.1 (3.3)	3.4 (2.2)	2.3 (2.2)	11.3 (13.0)

Table 6

Cell percentages for imputed and actual (in parenthesis) tables crossing satisfaction variables with
“customer/employee ratio in respondent’s branch ” (III-5, in quintiles)

	I-5. “Clerks are fast ”		I-9 “Clerks are available ”	
III-5	A bit	A lot	A bit	A lot
1 bottom 20%	8.9 (9.3)	7.2 (7.1)	7.5 (7.1)	8.7 (9.6)
2	7.5 (7.4)	9.0 (9.3)	6.1 (5.2)	10.6 (11.6)
3	11.7 (10.9)	8.5 (11.1)	9.7 (8.5)	10.5 (13.6)
4	15.4 (15.8)	11.5 (8.6)	12.8 (13.5)	14.1 (10.8)
5 Top 20%	13.2 (14.0)	6.9 (6.4)	10.7 (11.0)	9.3 (9.2)

Figure 1
Schematic representation of the data fusion problem



REFERENCES

Adamek, James (1994), “Fusion: Combining Data from Separate Sources,” Marketing Research: A Magazine of Management and Applications , 6 (Summer), 48-50.

Antoine, Jacques and Gilles Santini (1987), “Fusion Techniques: Alternative to Single Source Methods?,” European Research, 15 (August), 178-187.

Baker, Ken, Paul Harris and John O’Brien (1994).”Data Fusion: An Appraisal and Experimental Evaluation,” Journal of The Market Research Society, 31 (2),152-212.

Bozdogan, Herman (1987), “ Model Selection and Akaike’s Information Criterion (AIC): The General Theory and its Analytical Extensions,” Psychometrika, 52, 345-370.

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

Dempster, A.P. N.M. Laird, and R.B. Rubin (1977), “Maximum Likelihood for Incomplete Data via the EM-Algorithm,” Journal of the Royal Statistical Society, B39, 1-38.

Ford, Barry (1983), “An Overview of Hot-Deck Procedures,” In: Incomplete Data In Sample Surveys, Academic Press, Volume II, Part 2, 185-207.

Little, Roderick J.A., and Nathaniel Schenker (1995), "Missing Data," in: Handbook of Statistical Modelling for the Social and

Behavioral Sciences , Gerard Arminger, Clifford C. Clogg and Michael E. Sobel (eds). Plenum Press, New York, 39-75.

Meng, Xiao-Li, and Donald B. Rubin (1992), " Performing Likelihood-ratio Tests with Multiply Imputed Data Sets," Biometrika , 79, 103-111.

O'Brien, Sarah (1991), "The Role of Data Fusion in Actionable Media Targeting in the 1990s," Marketing & Research Today, 19 (February), 15-22.

Roberts, Andrew (1994), "Media Exposure and Consumer Purchasing: An Improved Data Fusion Technique," Marketing and Research Today, 22 (August) 159-172.

Rogers, Willard L. (1984), "An Evaluation of Statistical Matching," Journal of Business and Economic Statistics, 2 (January), 91-105.

Rubin, Donald B. (1986), " Statistical Matching and File Concatenation with Adjusted Weights and Multiple Imputations," Journal of Business and Economic Statistics , 4, 87-94.

Rubin, Donald B. (1976), "Inference and Missing Data," Biometrika, 63, 581-592.

Rubin, Donald B. and Nathaniel Schenker (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," Journal of the American Statistical Association, 81, 366-374.

Wedel, Michel, and Wayne S. DeSarbo (1994), "A Review of

Recent Developments in Latent Class Regression Models," In: Advanced Methods of Marketing Research, Rick P. Bagozzi (ed.), Cambridge: Blackwell, 352-388.