AUSTRIAN JOURNAL OF STATISTICS
Volume 33 (2004), Number 1&2, 153-171

# Data Fusion: Identification Problems, Validity, and Multiple Imputation

Susanne Rässler[1]
Institute for Employment Research (IAB), Nürnberg, Germany

**Abstract:** Data fusion techniques typically aim to achieve a complete data file from different sources which do not contain the same units. Traditionally, data fusion, in the US also addressed by the term statistical matching, is done on the basis of variables common to all files. It is well known that those approaches establish conditional independence of the (specific) variables not jointly observed given the common variables, although they may be conditionally dependent in reality. However, if the common variables are (carefully) chosen in a way that already establishes conditional independence, then inference about the actually unobserved association is valid. In terms of regression analysis, this implies that the explanatory power of the common variables is high concerning the specific variables. Unfortunately, this assumption is not testable yet. Hence, we structure and discuss the objectives of statistical matching in the light of their feasibility. Four levels of validity a matching technique may achieve are introduced. By means of suitable multiple imputation (MI) techniques, the identification problem which is inherent in data fusion is reflected. In a simulation study it is also shown that MI allows to efficiently and easily use auxiliary information.

**Key words:** Data Fusion, Data Merging, Mass Imputation, File Concatenation, Multiple Imputation, Missing Data, Missing by Design, Observed-Data Posterior.

## 1 Introduction

Statistical matching techniques, as they are referred to in the U.S., typically aim to achieve a complete data file from different sources that do not contain the same units. On the contrary, if samples are exactly matched using identifiers such as social security numbers or name and address, this is called record linkage. Traditionally, statistical matching is done on the basis of variables common to all files. Statistical twins, i.e., donor and recipient units that are similar according to their common variables, are usually found by means of nearest neighbor or hot deck procedures. The specific variables of a donor unit which are observed only in one file are added to the record of the recipient unit to finally create the matched sample. We like to note that in our sense statistical matching is not restricted to the case of merging different samples without overlap. Also one single file may contain some records with observations on more variables than others, then, these records can be matched with those containing less information based on the variables common to all units. Basically, there are a couple of different situations when statistical

[1]Acknowledgment: The author wants to thank Friedrich Wendt, who was one of the first and most engaged persons to develop data fusion techniques in Europe. He gave me generously insight to all is procedures and material, unfortunately, he died at the 29th of June 2003 at the age of 82.

matching can be applied. Figure 1 gives an overview of these occasions. The white boxes represent the missing variables.

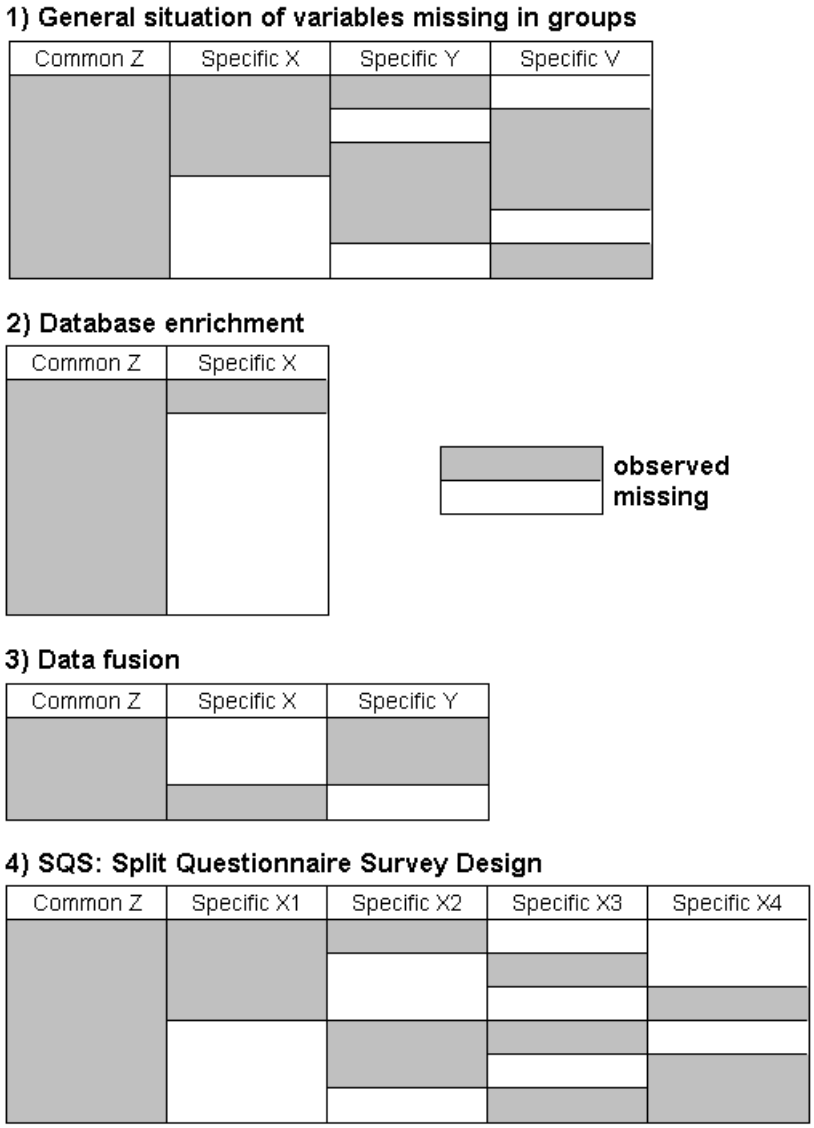


Figure 1: Different situations for statistical matching

In this paper we refer to the situation of picture no. 3 in Figure 1 which we call data fusion. This Figure illustrates that only in the case of data fusion there are groups of variables that are *never jointly observed*, say $X$ and $Y$. In all other cases we assume that, at least, every pair of variables has been jointly observed in one or the other data set. The fusion of data sets with the aim of analyzing the unobserved relationship of $X$ and $Y$ and addressing quality of data fusion is done, e.g., by National Statistical Institutes such as Statistics Canada or the Italian National Institute of Statistics, see, e.g., Liu and Kovacevic (1997) or D'Orazio et al. (2003). The focus often is on analyzing consumer's expenditures and income which are in detail only available from different surveys. In the U.S., e.g., data fusion is used for microsimulation modeling, where "what if" analyses

of alternative policy options are carried out using matched data sets, see Moriarity and Scheuren (2001, 2003). Especially in Europe and among marketing research companies, data fusion has become a powerful tool for media planning, see, e.g., Wendt (1986). Often surveys concerning the purchasing behavior of individuals or households are matched to those containing valuable information about print, radio and television consumption.

# 2 Data Fusion - Its Identification Problem

Data fusion initially is connected to an identification problem concerning the association of the variables not jointly observed. The conditional association (i.e., the association of the variables never jointly observed, $X$ and $Y$, given the variables common to both files, say $Z$) cannot be estimated from the observed data; for a detailed proof see Rubin (1974). However, depending on the explanatory power of the common variables $Z$ there is a smaller or wider range of admissible values of the unconditional association of $X$ and $Y$. By means of multiple imputation based on explicit models, imputations can be made for different values describing the conditional association. From these imputed data sets the unconditional association can then be estimated.

Consider, for example, a common variable $Z$ determining another variable $X$ only observed in one file. For demonstration purpose, we discuss linear dependencies; i.e., let the correlation $\rho_{ZX} = 1$, and thus $X = a + bZ$ for some $a, b \in \mathbb{R}^2, b \neq 0$. The correlation between this common variable $Z$ and a variable $Y$ in a second file may be $\rho_{ZY} = 0.8$. It is easy to see that the unconditional correlation of the two variables $X$ and $Y$ which are not jointly observed is determined by $Z$ with $\rho_{XY} = \rho_{a+bZ\,Y} = \rho_{ZY} = 0.8$. If the correlation between $X$ and $Z$ is less than one, say 0.9, we can easily calculate the possible range of the unconditional association between $X$ and $Y$ by means of the determinant of the covariance matrix which has to be positive semi-definite; i.e., the determinant of the covariance matrix $\text{cov}(Z, X, Y)$ should be positive or at least zero, see, e.g., Cox and Wermuth (1996).

Given the above values and setting the variances to one without loss of generality, the covariance matrix of $(Z, X, Y)$ is

$$\text{cov}(Z, X, Y) = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & \text{cov}(X, Y) \\ 0.8 & \text{cov}(X, Y) & 1 \end{pmatrix} \quad \text{with}$$

$$\det(\text{cov}(Z, X, Y)) = -\text{cov}(X, Y)^2 + 2 \cdot 0.72\,\text{cov}(X, Y) - 0.45\,.$$

Calculating the roots of $\det(\text{cov}(Z, X, Y))$ we get the two solutions $\text{cov}(X, Y) = 0.72 \pm \sqrt{0.0684}$. Hence we find the unconditional correlation bounded between [0.4585, 0.9815]; i.e., every value of the unknown covariance $\text{cov}(X, Y)$ greater than 0.4585 and less than 0.9815 leads to a valid and thus feasible covariance structure for $(Z, X, Y)$. By means of the observed data we are not able to decide which covariance matrix could have generated the data.

If the variables $X$ and $Y$ are conditionally independent or at least uncorrelated given $Z$, a correlation of $X$ and $Y$ of $\text{cov}(X, Y) = 0.72$ is computed which is exactly the middle of the interval [0.4585, 0.9815] yielding the maximum value for the determinant

$|\text{cov}(Z, X, Y)|$. Finally we have found an upper (0.9815) and a lower (0.4585) bound for $\text{cov}(X, Y)$. An estimation procedure of these bounds in the fusion context was first published by Kadane in 1978. The strength of the conditional independence assumption is also discussed in a similar example by Rodgers (1984). He shows that only an extremely high correlation narrows the range of the unconditional association considerably. Only few approaches, basically three different procedures, have been published to assess the effect of alternative assumptions of this inestimable value. One approach is due to Kadane (2001) (originally 1978, now reprinted), generalized by Moriarity and Scheuren (2001). The next approach dates back to Rubin and Thayer (1978), it is used to address data fusion explicitly by Rubin (1986), and generalizations are presented by Moriarity and Scheuren (2003). Both approaches use regression based procedures to produce synthetic data sets under various assumptions on this unknown association. Finally, a full Bayesian regression approach using multiple imputations is first given by Rubin (1987), p. 188, and then generalized by Rässler (2002).

We propose to discuss the explanatory power of the common variables and, thus, of the validity of the data fusion procedure based on these bounds. Since it is not possible to judge the quality of the matched data concerning this unobserved association whenever the variables are never jointly observed and no auxiliary data file is available, the range of these bounds may be used to evaluate any data fusion procedure. The less the bounds differ, the better is the explanatory power of the common variables and the more valid results traditional matching techniques will produce. However, if the data structure is complex and high-dimensional we are not able to calculate these bounds directly. Therefore, we propose multiple imputation to fix the conditional association using prior information. It provides a helpful and flexible tool for fusion in general and for the derivation of the bounds in particular. Moreover, we will illustrate by means of a simulation study how MI can make efficient use of auxiliary data.

# 3 Validity Levels of Data Fusion

The general benefit of data fusion is the creation of one complete data source containing information about all variables. Without loss of generality, let the $(X, Z)$ sample be the recipient sample $B$ of size $n_B$ and the $(Y, Z)$ sample the donor sample $A$ of size $n_A$. The traditional matching procedures determine for every unit $i$, $i = 1, 2, \ldots, n_B$, of the recipient sample with the observations $(x_i, z_i)$ a value $y$ from the observations of the donor sample. Thus, a composite data set $(x_1, \widetilde{y}_1, z_1), \ldots, (x_{n_B}, \widetilde{y}_{n_B}, z_{n_B})$ with $n_B$ elements of the recipient sample is constructed. The main idea is to search for a statistical match, i.e., for a donor unit $j$ with $(y_j, z_j) \in \{(y_1, z_1), (y_2, z_2), \ldots, (y_{n_A}, z_{n_A})\}$ whose observed data values of the common variables $z_j$ are identical to those $z_i$ of the recipient unit $i$ for $i = 1, 2, \ldots, n_B$. Notice that $\widetilde{y}_i$ is not the true $y$-value of the $i$th recipient unit but the $y$-value of the matched statistical twin. In the following, all density functions (joint, marginal, or conditional) and their parameters produced by the fusion algorithm are marked by the symbol $\widetilde{\ }$. Notice that $\widetilde{Y}$ is called fusion or imputed variable herein.

There are very sophisticated fusion techniques in practice; for a recent overview see Rässler (2002). Here we focus on the validity of the fusion process. Therefore, we suggest

to distinguish four levels of validity a fusion procedure may achieve. We use the term validity rather than efficiency, because efficiency usually refers to a minimum mean squared error criterion as it is common, for example, in survey sampling theory and not to different levels of reproduction and preservation of the original associations and distributions.

## 3.1 First Level: Preserving Individual Values

The individual values are preserved when the true but unknown values of the (multivariate) $Y$ variable of the recipient units are reproduced; i.e., $\widetilde{y}_i = y_i$ for $i = 1, 2, \ldots, n_A$. We call this situation a "hit" for any unit in the recipient sample and may calculate a "hit rate" therefrom. But some words of warning should be placed.

Within continuous distributions the probability of drawing a certain value $y$ is zero; counting the hits is meaningless then. In the case of discrete or classified variables $Y$ a hit rate may be calculated for the purpose of demonstration, counting a hit for the imputation of a $p$-dimensional variable $Y$ when the whole imputed vector equates the original vector; i.e.,

$$(\widetilde{y}_{1i}, \widetilde{y}_{2i}, \ldots, \widetilde{y}_{pi}) = (y_{1i}, y_{2i}, \ldots, y_{pi})$$

for $i = 1, 2, \ldots, n_A$. Notice that the calculation of a single hit rate for each variable may mislead the interpretation because it does not ensure that the joint distributions are well preserved. Moreover, any fusion technique that additionally assures the preservation of the marginal distributions will automatically lead to "high" hit rates, especially when the variables have only few categories, see Table 1. In this case, the "worst" fusion still yields 80% of hits. A simple random assignment produces $(0.9 \cdot 0.9 + 0.1 \cdot 0.1)100\%$ = 82% of hits.

| Before \ After Fusion | 0 | 1 | Total |
|---|---|---|---|
| 0 | 800 | 100 | 900 |
| 1 | 100 | 0 | 100 |
| Total | 900 | 100 | 1000 |

Table 1: Hit rate for a constraint match

Finally, consider the following example which was mentioned by Rubin in a talk at the DataClean2002 Conference in Finland:
Suppose we have 10,000 bernoulli trials with $p = 0.6$, and 9000 are missing completely at random. The observed 1000 trials show (more or less) 600 heads and 400 tails, and we are asked to impute the missing 9000 to draw inferences about the population probability of a head. Let us decide to evaluate different imputation methods by the hit rate, i.e., the number of agreements between the imputed values and the real data.

- Method 1: impute all heads, then we count in total 9600 heads and the inference for $p$ is that $p = 0.96$.
- Method 2: draw imputes with probability of head of 0.6, then the inference for $p$ will give that $p = 0.6$.

Which method is better by the hit rate? Method 1 gives agreement 60% of the time, whereas method 2 gives agreement $0.4 * 0.4 + 0.6 * 0.6 = 52$ % of the time. Thus the hit rate says method 1 is the one to use. One should always remember that imputations are not meant to reflect the real values and that their microdata interpretation is meaningless.

## 3.2 Second Level: Preserving Joint Distributions

The joint distribution is preserved after data fusion when the true joint distribution of all variables is reflected in the fused file; i.e., $\widetilde{f}_{X,Y,Z} = f_{X,Y,Z}$.

We usually assume that the units of both samples are drawn independently within and between the two samples and the fused file can be regarded as a random sample from the underlying fusion distribution $\widetilde{f}_{X,Y,Z}$. The most important objective of data fusion is the generation of a complete sample that can be used as a single-source sample drawn from the underlying distribution $f_{X,Y,Z}$. It is less the reconstruction of individual values but the possibility of making valid statistical inference based on the fused file. Sims (1972), Rodgers (1984), and Rässler (2002) show that this is only possible if the specific variables $Y$ and $X$ are conditionally independent given the common variables $Z = z$; i.e., $f_{X,Y|Z} = f_{X|Z} f_{Y|Z} = \widetilde{f}_{X,Y|Z}$ holds.

## 3.3 Third Level: Preserving Correlation Structures

The correlation structure and higher moments of the variables are preserved after data fusion with $\widetilde{\text{cov}}(X, Y, Z) = \text{cov}(X, Y, Z)$. Also the marginal distributions are reflected correctly with $\widetilde{f}_{Y,Z} = f_{Y,Z}$ and $\widetilde{f}_{X,Z} = f_{X,Z}$.

Sometimes the analyst's interests are more specific concerning, for instance, only the association of variables measured by their correlation structure. Then the fused file must be considered as randomly generated from an artificial population which has, at least, the same moments and correlation structure as the actual population of interest. Rässler and Fleischer (1998) show that the fusion covariance is given by $\widetilde{\text{cov}}(X, Y) = \text{cov}(\text{E}(X|Z), \text{E}(Y|Z))$. Because

$$\text{cov}(X, Y) = \text{E}(\text{cov}(X, Y|Z)) + \text{cov}(\text{E}(X|Z), \text{E}(Y|Z))$$

holds, the fusion covariance $\widetilde{\text{cov}}(X, Y)$ only equals the true $\text{cov}(X, Y)$, if $X$ and $Y$ are on the average conditionally uncorrelated given $Z = z$; i.e., $\text{E}(\text{cov}(X, Y|Z)) = 0$. Notice that variables which are conditionally independent are also conditionally uncorrelated and, of course, on the average conditionally uncorrelated, but not vice versa in general.

## 3.4 Fourth Level: Preserving Marginal Distributions

After data fusion, at least, the marginal and joint distributions of the variables in the donor sample are preserved in the fused file. Then $\widetilde{f}_Y = f_Y$ and $\widetilde{f}_{Y,Z} = f_{Y,Z}$ are expected to hold if $Y$ is imputed in the $(X, Z)$ sample.

In practice, the preservation of the distributions observed in the separate samples is usually required. Analysis concerning the marginal distributions based on the fused file should provide the same valid inference when based on the separate samples. Therefore,

the empirical distributions of the common variables $Z$ as well as the imputed variables $Y$ in the resulting fused file are compared with their empirical distributions in the donor sample to evaluate the similarity of both samples. The empirical distributions $\widehat{\widetilde{f}}_Y$ and $\widehat{\widetilde{f}}_{Y,Z}$ should not differ from $\widehat{f}_Y$ and $\widehat{f}_{Y,Z}$ more than two random samples drawn from the true underlying population. Notice that this implies the different samples being drawn according to the same sampling design.

In the typical fusion situation only the fourth level can be controlled. Therefore, often data fusion is said to be successful if the marginal and joint empirical distributions of $Z$ and $Y$, as they are observed in the donor sample, are "nearly" the same in the fused file.

In common approaches, first of all, averages for all common variables $Z$ between the donor and the recipient sample are compared. Then the average values between the imputed variables $\widetilde{Y}$ and the corresponding variables $Y$ in the donor sample are compared. Often the preservation of the relation between variables is measured by means of correlations. Therefore, for each common variable $Z$, the correlation with every original variable $Y$ and imputed variable $\widetilde{Y}$ is computed, both for the fused data set and the donor sample. The mean difference between common-fusion correlations in the donor versus the fused data set are calculated and empirically evaluated, see, e.g., van der Putten et al. (2002).

The German association for media analysis[2], for example, still postulates the following data controls after a match has been performed.

- First the empirical distributions of the common variables $Z$ in the recipient and the donor sample are compared to evaluate whether their marginal distributions are the same in both samples.

- Next the empirical distributions of the imputed variables $\widetilde{Y}$ in the recipient and $Y$ in the donor sample are compared.

- Finally the joint distribution $f_{Z,Y}$ as observed in the donor sample is compared to the joint distribution $\widetilde{f}_{Z,Y}$ as observed in the fused file.

All these comparisons are done using different tests such as $\chi^2$-tests or $t$-tests to compare empirical distributions or their moments. A successful match should lead to similar relationships between common and specific variables in the donor and the fused file; discrepancies should not be larger than expected between two independent random samples from the same underlying population. In particular, often each pair of variables $Y$ and $Z$ in the donor sample is tested at a significance level $\alpha$ for positive or negative association by, for example, a $\chi^2$-test or a $t$-test (depending on the scale of the variables). Then the same test of association between $\widetilde{Y}$ and $Z$ is performed for each pair in the fused file. If the results of the tests only differ in about $\alpha$ percent of the possible $(Y, Z)$ combinations, then the fusion procedure is regarded as successful, although this means accepting the Null hypotheses rather than discarding them. Among others, nonparametric tests and

[2] "Media Analysis Association" called in German Arbeitsgemeinschaft Media Analyse, for short, AG.MA. The AG.MA is a media association, i.e., publishing houses, radio and TV stations, and many advertising agencies, as well as a certain number of advertisers.

multiple regression models may be used in the same manner.[3]

Any discussion of validity of a data fusion technique can now be based on these four levels. Besides so-called split half or simulations studies, all tests actually applied in practice only indicate the fourth-level validity. Therefore, we like to go further to evaluate the predictive power of the common variables.

# 4 A Multiple Imputation Algorithm

In the cases pictured in Figure 1 (at least in nos. 2 to 4), it is assumed that the data are missing completely at random or, at least, missing at random because the missingness is induced by design. Thus, the fusion task can be viewed as a typical imputation problem. Because the amount of information to be imputed is large, we use the term mass imputation. In missing data problems the multiple imputation theory, initially introduced by Rubin (1977) and extensively described in Rubin (1987), provides very flexible procedures for imputation with good statistical properties from a Bayesian as well as a frequentist view. We follow this approach and suggest a non-iterative Bayesian multiple imputation procedure, called NIBAS, especially suited for data fusion.

Let us assume a multivariate normal data model for $(X, Y|Z = z) = (X_1, X_2, \ldots, X_q, Y_1, Y_2, \ldots, Y_p|Z = z)$ with expectation $\mu_{XY|Z}$ and covariance matrix $\Sigma_{XY|Z}$ is denoted by

$$\Sigma_{XY|Z} = \begin{pmatrix} \Sigma_{XX|Z} & \Sigma_{XY|Z} \\ \Sigma_{YX|Z} & \Sigma_{YY|Z} \end{pmatrix} .$$

Moreover, the general linear model for both data sets is applied with

$$\begin{aligned} (\text{file } A) \quad & Y = Z_A\beta_{YZ} + U_A\,, \qquad U_A \sim N_{pn_A}(0, \Sigma_{YY|Z} \otimes I_{n_A})\,, \\ (\text{file } B) \quad & X = Z_B\beta_{XZ} + U_B\,, \qquad U_B \sim N_{qn_B}(0, \Sigma_{XX|Z} \otimes I_{n_B})\,, \end{aligned}$$

with $Z_A$ and $Z_B$ denoting the corresponding parts of the common derivative matrix $Z$. This data model assumes that the units can be observed independently for $i = 1, 2, \ldots, n$. The correlation structure refers to the variables $X_{1i}, X_{2i}, \ldots, X_{qi}, Y_{1i}, Y_{2i}, \ldots, Y_{pi}$ for each unit $i = 1, 2, \ldots, n$. For abbreviation we use the Kronecker product $\otimes$ denoting that the variables $X_i$ and $Y_i$ of each unit $i$, $i = 1, 2, \ldots, n$, are correlated but no correlation of the variables is assumed between the units.

As a suitable noninformative prior we assume prior independence between $\beta$ and $\Sigma$ choosing

$$f_{\beta_{YZ},\beta_{XZ},\Sigma_{XX|Z},\Sigma_{YY|Z},R_{XY|Z}} \propto \Sigma_{XX|Z}^{-(q+1)/2}\Sigma_{YY|Z}^{-(p+1)/2} f_{R_{XY|Z}}\,.$$

The joint posterior distribution for the fusion case can be factored into the prior and likelihood derived by file $A$ and file $B$, respectively, see Rässler (2002). Then the joint

[3] If the samples have different structures, e.g., due to oversampling in one survey or differing sampling designs, weights can be applied accounting for differing selection probabilities of the units in the separate samples. Also, samples drawn according to different sampling designs could be made "equal" by using propensity scores according to an idea by Rubin (2002) before performing the final match. However, this is beyond the scope of this article.

posterior distribution can be written with

$$f_{\beta_{XZ},\beta_{YZ},\Sigma_{XX|Z},\Sigma_{YY|Z},R_{XY|Z}|X,Y} = c_X^{-1} L(\beta_{XZ}, \Sigma_{XX|Z}; x) f_{\Sigma_{XX|Z}|R_{XY|Z}} \\ c_Y^{-1} L(\beta_{YZ}, \Sigma_{YY|Z}; y) f_{\Sigma_{YY|Z}|R_{XY|Z}} f_{R_{XY|Z}} .$$

Thus, our problem of specifying the posterior distributions reduces to standard derivation tasks described, for example, by Box and Tiao (1992), p. 439. $\Sigma_{XX|Z}$ and $\Sigma_{YY|Z}$ given the observed data each are following an inverted-Wishart distribution. The conditional posterior distribution of $\beta_{XZ}$ ($\beta_{YZ}$) given $\Sigma_{XX|Z}$ ($\Sigma_{YY|Z}$) and the observed data is a multivariate normal distribution. The posterior distribution of $R_{XY|Z}$ equals its prior distribution. Having thus obtained the observed-data posteriors and the conditional predictive distributions a multiple imputation procedure for multivariate variables $X$ and $Y$ can be proposed with the following algorithm:

**Algorithm NIBAS**

- Compute the ordinary least squares estimates $\widehat{\beta}_{YZ} = (Z_A' Z_A)^{-1} Z_A' Y$ and $\widehat{\beta}_{XZ} = (Z_B' Z_B)^{-1} Z_B' X$ from the regression of each data set. Note that $\widehat{\beta}_{YZ}$ is a $k \times p$ matrix and $\widehat{\beta}_{XZ}$ is a $k \times q$ matrix of the OLS or ML estimates of the general linear model.

- Calculate the following matrices proportional to the sample covariances for each regression with

$$S_Y = (Y - Z_A \widehat{\beta}_{YZ})'(Y - Z_A \widehat{\beta}_{YZ}) , \\ S_X = (X - Z_B \widehat{\beta}_{XZ})'(X - Z_B \widehat{\beta}_{XZ}) .$$

- Choose a value for the correlation matrix $R_{XY|Z}$ or each $\rho_{X_i Y_j|Z}$ for $i = 1, 2, \ldots, q$, $j = 1, 2, \ldots, p$

  (a) from its prior according to some distributional assumptions, i.e., uniform over the $p + q$-dimensional ]-1,1[-space, or

  (b) several arbitrary levels, or

  (c) estimate a value from a small but completely observed data set.

  The latter should be the most realistic case in many practical situations.

- Perform random draws for the parameters from their observed data posterior distribution according to the following scheme.

$$\begin{array}{lll} \text{Step 1: } \Sigma_{YY|Z}|y & \sim W_p^{-1}(v_A, S_Y^{-1})\, v_A = n_A - (k+p) + 1 \\ \quad \Sigma_{XX|Z}|x & \sim W_q^{-1}(v_B, S_X^{-1})\, v_B = n_B - (k+q) + 1 \\ \text{Step 2: } \beta_{YZ}|\Sigma_{YY|Z}, y & \sim N_{pk}(\widehat{\beta}_{YZ}, \Sigma_{YY|Z} \otimes (Z_A' Z_A)^{-1}), \\ \quad \beta_{XZ}|\Sigma_{XX|Z}, y & \sim N_{qk}(\widehat{\beta}_{XZ}, \Sigma_{XX|Z} \otimes (Z_B' Z_B)^{-1}), \\ \text{Step 3: Set } \Sigma_{XY|Z} & = \{\sigma_{X_i Y_j|Z}\} \text{ with } \sigma_{X_i Y_j|Z} \\ & = \rho_{X_i Y_j|Z} \sqrt{\sigma^2_{X_i|Z} \sigma^2_{Y_j|Z}} \\ & \text{with } \sigma^2_{X_i|Z}, \sigma^2_{Y_j|Z} \text{ derived by Step 1} \\ & \text{for } \; i = 1, 2, \ldots, q, \;\; j = 1, 2, \ldots, p . \end{array}$$

$$\text{Step 4:}\quad X|y,\beta,\Sigma \sim N_{qn_A}\Big(Z_A\beta_{XZ} + (Y - Z_A\beta_{YZ})\Sigma^{-1}_{YY|Z}\Sigma_{YX|Z};\; (\Sigma_{XX|Z} - \Sigma_{XY|Z}\Sigma^{-1}_{YY|Z}\Sigma_{YX|Z}) \otimes I_{n_A}\Big)$$
$$Y|x,\beta,\Sigma \sim N_{pn_B}\Big(Z_B\beta_{YZ} + (X - Z_B\beta_{XZ})\Sigma^{-1}_{XX|Z}\Sigma_{XY|Z};\; (\Sigma_{YY|Z} - \Sigma_{YX|Z}\Sigma^{-1}_{XX|Z}\Sigma_{XY|Z}) \otimes I_{n_B}\Big)\,.$$

The predictive power of the common variables $Z$ especially affects the last step. Choosing the correlation matrix $R_{XY|Z}$ uniform over the $pq$-dimensional ]-1,1[-space may lead to invalid conditional variances in Step 4. To achieve imputations reflecting the bounds we propose to set $R_{XY|Z} = 0_{q\times p}$ and add some $\pm\epsilon$ iteratively until the variance matrices in Step 4 are no longer positive definite. A similar procedure to get admissible values of the covariance matrix is proposed by Moriarity and Scheuren (2001) for the multivariate case. For univariate $X$ and $Y$ variables $\rho_{XY|Z}$ can be chosen from $]-1,1[$. The bounds can be calculated directly then; see also Moriarity and Scheuren (2001).

# 5 Simulation Study

## 5.1 Data Model

Let $(Z_1, Z_2, X, Y_1, Y_2)$ each be univariate standard normally distributed variables with their joint distribution

$$(Z_1, Z_2, X, Y_1, Y_2) \sim N_5(0, \Sigma) \tag{1}$$

and

$$\Sigma = \left(\begin{array}{cc|c|cc} 1.0 & 0.2 & 0.5 & 0.8 & 0.5 \\ 0.2 & 1.0 & 0.5 & 0.6 & 0.6 \\ \hline 0.5 & 0.5 & 1.0 & \sigma_{XY_1} & \sigma_{XY_2} \\ \hline 0.8 & 0.6 & \sigma_{XY_1} & 1.0 & 0.4 \\ 0.5 & 0.6 & \sigma_{XY_2} & 0.4 & 1.0 \end{array}\right) = \left(\begin{array}{ccc} \Sigma_{ZZ} & \Sigma_{ZX} & \Sigma_{ZY} \\ \Sigma_{XZ} & \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YZ} & \Sigma_{YX} & \Sigma_{YY} \end{array}\right).$$

Assume that file $A$ contains $(Z_1, Z_2, Y_1, Y_2)$ and file $B$ $(Z_1, Z_2, X)$, thus $X$ and $Y = (Y_1, Y_2)'$ are never jointly observed. Thus, the partial correlations of $X$ and $Y_1$ and $X$ and $Y_2$, respectively, can not be estimated from the observed data. Also the unconditional covariance matrix $\Sigma_{XY}$ does not have an unique estimate, however, there is information in the data about their admissible values. As it is shown in Rässler (2002), the traditional nearest neighbor match leads to conditional independence of $X, Y|Z = z$ with the unconditional covariance after the fusion $\widetilde{\Sigma}_{XY} = \Sigma_{XZ}\Sigma^{-1}_{ZZ}\Sigma_{ZY} = (0.5833 \quad 0.4583)$ in this setting.

## 5.2 Calculation of the Range

Now let

$$\Sigma^{\star}_{XY} = \left(\begin{array}{ccc} 1.0 & \sigma_{XY_1} & \sigma_{XY_2} \\ \sigma_{XY_1} & 1.0 & 0.4 \\ \sigma_{XY_2} & 0.4 & 1.0 \end{array}\right) = \Sigma^{\star}_{XY|Z} + \left[\begin{array}{c} \Sigma_{XZ} \\ \Sigma_{YZ} \end{array}\right] \Sigma^{-1}_{ZZ} \left[\Sigma_{ZX} \;\; \Sigma_{ZY}\right] \tag{2}$$

with

$$\Sigma^{\star}_{XY|Z} = \left(\begin{array}{c|cc} \sigma_{XX|Z} & \sigma_{XY_1|Z} & \sigma_{XY_2|Z} \\ \hline \sigma_{XY_1|Z} & \sigma_{Y_1Y_1|Z} & \sigma_{Y_1Y_2|Z} \\ \sigma_{XY_2|Z} & \sigma_{Y_1Y_2|Z} & \sigma_{Y_2Y_2|Z} \end{array}\right) = \left(\begin{array}{cc} \Sigma_{XX|Z} & \Sigma_{XY|Z} \\ \Sigma_{YX|Z} & \Sigma_{YY|Z} \end{array}\right) ,$$

and

$$\Sigma_{XX|Z} = \Sigma_{XX} - \Sigma_{XZ}\Sigma^{-1}_{ZZ}\Sigma_{ZX} , \quad \text{and} \quad \Sigma_{YY|Z} = \Sigma_{YY} - \Sigma_{YZ}\Sigma^{-1}_{ZZ}\Sigma_{ZY} .$$

Finally, we can set

$$\sigma_{XY_1|Z} = \rho_{XY_1|Z}\sqrt{\sigma_{XX|Z}\sigma_{Y_1Y_1|Z}} \quad \text{and} \quad \sigma_{XY_2|Z} = \rho_{XY_2|Z}\sqrt{\sigma_{XX|Z}\sigma_{Y_2Y_2|Z}} ,$$

to get $\Sigma_{XY|Z} = \Sigma'_{YX|Z}$ therefrom.

(2) illustrates that the partial correlations between $X$ and $Y_1$ or $X$ and $Y_2$ depend on the partial correlations between $X$ and $Y_1$ or $Y_2$, respectively, but not on the partial correlations between any other pair of variables, see also Little and Rubin (2002), p. 159. Furthermore, they do not depend on the inestimable unconditional correlation of the respective alternative variable. Moreover, the relationship between partial and unconditional correlation is linear like Figure 2 shows.

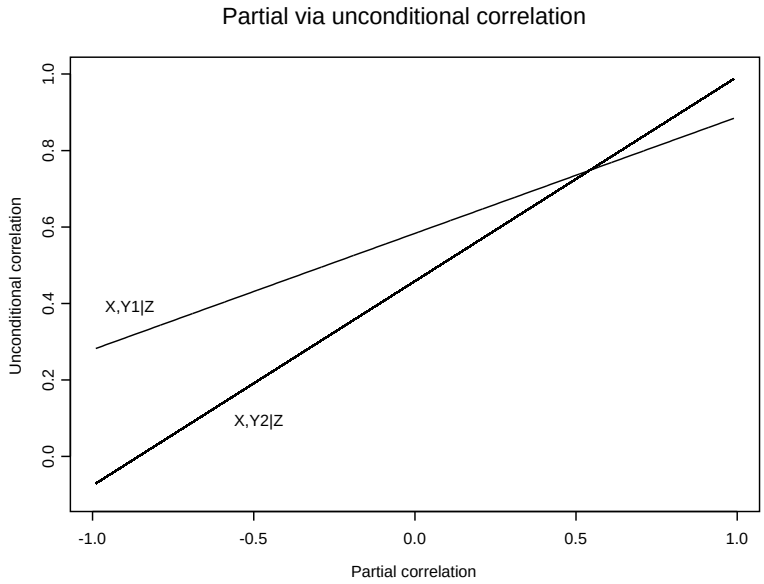


Figure 2: Linear relationship between conditional (partial) and unconditional correlations

However, these ranges of the partial correlations cannot be applied in a multivariate setting because of the restriction that the covariance matrices $\Sigma_{XY|Z}$ and $\Sigma$ have to be positive definite.

In a slight abuse of notation, we write for the general case

$$\Sigma_{XY|Z} = R_{XY|Z}\left(\sqrt{\text{diag}(\Sigma_{XX|Z})}\sqrt{\text{diag}(\Sigma_{YY|Z})'}\right)$$

which may depend on all possible values of $R_{XY|Z} \in \{-1, 1\}$ as long as $\Sigma_{XY|Z}$ and $\Sigma$ are positive definite.

For the above example, we can calculate the admissible values via searching the grid, i.e., we set $\rho_{XY_1|Z} \in ]-1,1[$ and $\rho_{XY_2|Z} \in ]-1,1[$ and store all combinations that yield a positive definite matrix $\Sigma$. We find the solutions shown in Figure 3.

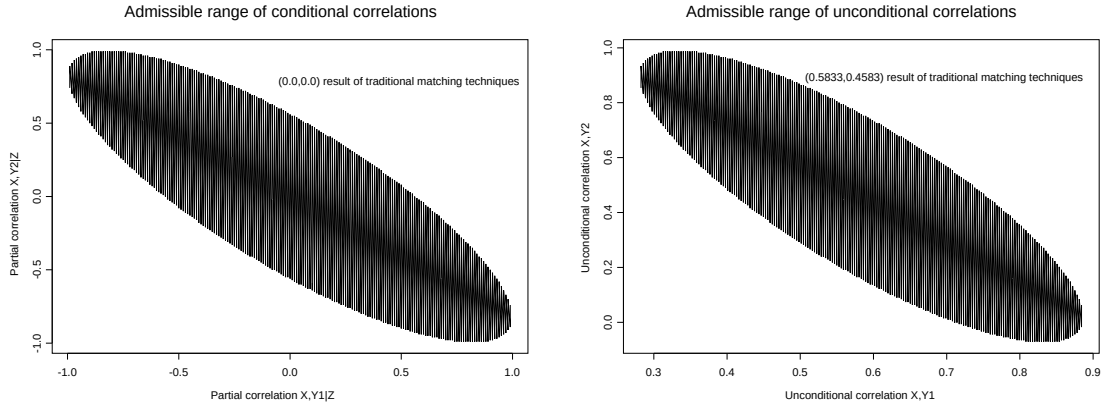


Figure 3: Admissible combinations of conditional/unconditional correlations

## 5.3 Simulation setup

From the data model of (1) we simulate $k = 200$ complete data sets of size $n_A + n_B = 5000$ and part them into two separate files $A$ with $n_A = 3000$ and $n_B = 2000$ observations. Then the $Y$ values of file $A$ are matched or imputed in file $B$ according to the following algorithms

- NN; i.e., a nearest neighbor match (always assuming conditional independence),
- RI; i.e., a regression imputation under different conditional correlations,
- RIEPS; i.e., a regression imputation with stochastic residual under different conditional correlations, and
- NIBAS; i.e., the proposed MI algorithm assuming different prior conditional correlations.

For details of the formulae used in RI and RIEPS see Rässler (2002). Notice that NN and RI are single imputation procedures whereas RIEPS and NIBAS create more than one imputed data set. However, imputations produced by RIEPS are expected to underestimate variability because they lack from additional random draws of the parameters. Finally, small 1% and 5% complete auxiliary files are created according to the data model and used with the multiple imputation algorithm NORM (standalone software NORM 2.03) that is provided by Schafer (1997). With NORM it is not possible to use a real informative prior for the unknown correlations, therefore, NORM is applied herein for the data fusion situation when some auxiliary data are available containing information about all variables $X$, $Y$, and $Z$.

This procedure of creating the data, dividing and matching them is carried out 200 times. Relevant point and interval estimates are stored and tabulated. For the MI procedures $m = 5$ imputations are used. The MI estimates are calculated according to $\widehat{\theta}_{MI} = m^{-1} \sum_{t=1}^{m} \widehat{\theta}^{(t)}$, as well as the within-imputation variance $W = m^{-1} \sum_{t=1}^{m} \widehat{\text{var}}(\widehat{\theta}^{(t)})$, and the between-imputation variance $B = (m-1)^{-1} \sum_{t=1}^{m} (\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2$. The 95% MI interval estimates are calculated with $\widehat{\theta}_{MI} \pm \sqrt{T}\, t_{0.975,\nu}$, $T = W + (1 + m^{-1})B$, and degrees of freedom $v = (m-1)\left(1 + W/((1+m^{-1})B)\right)^2$. According to the MI principle we assume that based on the complete data the point estimates $\widehat{\theta}$ are approximately normal with mean $\theta$ and variance $\widehat{\text{var}}(\widehat{\theta})$.[4] Therefore, some estimates should be transformed to a scale for which the normal approximation works well. For example, the sampling distribution of Pearson's correlation coefficient $\widehat{\rho}$ is known to be skewed, especially if the corresponding correlation coefficient of the population is large. Thus, usually the multiple imputation point and interval estimates of a correlation $\rho$ are calculated by means of the Fisher $z$-transformation $z(\widehat{\rho}) = 0.5 \ln((1+\widehat{\rho})/(1-\widehat{\rho}))$, which makes $z(\widehat{\rho})$ approximately normally distributed with mean $z(\rho)$ and constant variance $1/(n-3)$, see, e.g., Schafer (1997), p. 216. By back transforming the corresponding MI point and interval estimates of $z(\rho)$ via the inverse Fisher transformation the final estimates and confidence intervals for $\rho$ are achieved.

## 5.4 Results

The following Tables show the estimated expectations of some point estimates. In addition, the Tables give the simulated actual coverage, i.e., the number of times out of 200 that cover the true parameter value. To ease the reading we display the percentage. Also the average length of the confidence intervals is reported (ALCI). The following Tables 2, 3, and 4 concentrate on the most important results.

Table 2 shows the preservation of the prior values of the conditional correlation between $X$ and $Y_1$ or $Y_2$, respectively. As it was to be expected, the nearest neighbor match always establishes conditional independence. Thus, this matching procedure only works, when the conditional independence assumption is satisfied. Even with slight derivations from it, see block 4 in Table 2, the simulated actual coverage is far beyond its true nominal value. Also the single regression imputation does not reflect the correct coverage and typically leads to a strong overestimation of the true population correlation. The regression imputation with random residual performs quite well as long as the true unconditional correlation is not too high. Best in all cases is the new procedure NIBAS. In every setting it preserves the prior correlation with a higher nominal coverage than expected.[5]

Moreover, its average confidence intervals are only a little bit larger than those produced by RIEPS. Also Table 2 demonstrates that the multiple imputation procedure NORM very efficiently allows to use auxiliary data. With an additional file of size 5%, i.e., a file of only 250 observations completely observed in $X, Y$, and $Z$, the simulated actual coverage in most of the cases is higher than its nominal value. For NIBAS and RIEPS we

[4] Notice that Barnard and Rubin (1999) relax this assumption of a normal reference distribution to allow a $t$-distribution for the complete-data interval estimates and tests.

[5] Notice that according to classical and current formal definition of confidence intervals such conservative intervals are valid.

could also use auxiliary information to estimate the potential prior correlations therefrom, but other simulations have shown that NORM is more powerful here, see Rässler (2002). With NORM the confidence intervals are typically much larger than with NIBAS. Thus, when prior information has to be used, NIBAS is the best choice at hand. The preservation of the distributions of the matched or imputed variables $Y_1$ and $Y_2$ is displayed in Tables 3 and 4. Again the nearest neighbor match leads to similar results regardless of the true correlations between $X$ and $Y_1$ or $Y_2$. Always the coverage is too low and the variances are underestimated as it is typical for single imputation approaches. Regression imputation also typically underestimates the variances even more, the coverage often is 0. As before, adding a random residual improves the regression imputation considerably but not in all cases. The best preservation again provides NIBAS, throughout the coverage is higher than its nominal value. Using auxiliary data works fine for NORM also if only 1% (i.e., 50 observations) are completely observed.

# 6 Summary and Outlook

In this paper we structure the validity a data fusion procedure may achieve by four levels. It is shown that the first level is meaningless and only the last, fourth, level typically is controlled when traditional techniques of data fusion are applied. The preservation of the joint distribution and the correlation structure of the variables not jointly observed can be evaluated by using the non-iterative multiple imputation procedure NIBAS. Data fusion can be viewed as a problem of mass imputation and MI procedures are applicable in general. In a simulation study, we find the multiple imputation approaches superior to the traditional matching techniques. Auxiliary data can be easily and efficiently used by standard MI procedures such as NORM. To avoid the identification problem inherent in data fusion, we suggest to use split questionnaire surveys (SQS) as proposed by Raghunathan and Grizzle (1995). This situation is pictured in Figure 1 no. 4. For the SQS design identification problems are avoided by creating special patterns of missingness. The missing data can then be quite successfully multiply imputed, for an application in media planning see Rässler et al. (2002).

In general, a multiple imputation procedure seems to be the best alternative at hand even in the case of data fusion. It accounts for the missingness and exploits all valuable available information.

| | $\widehat{\mathrm{E}}(\widehat{\rho}_{XY_1})$ | ALCI | Cvg. | $\widehat{\mathrm{E}}(\widehat{\rho}_{XY_2})$ | ALCI | Cvg. |
|---|---|---|---|---|---|---|
| Procedure | $\rho_{XY_1\|Z} = -0.6032, \rho_{XY_1} = 0.4$ | | | $\rho_{XY_2\|Z} = 0.6393, \rho_{XY_2} = 0.8$ | | |
| NN | 0.5810 | 0.0581 | 0.000 | 0.4566 | 0.0694 | 0.000 |
| RI | 0.4204 | 0.0722 | 0.805 | 0.9480 | 0.0089 | 0.000 |
| RIEPS | 0.4054 | 0.0771 | 0.960 | 0.8496 | 0.0328 | 0.000 |
| NIBAS | 0.3984 | 0.0819 | 0.985 | 0.7989 | 0.0467 | 1.000 |
| NORM 1% | 0.4201 | 0.1256 | 0.890 | 0.7824 | 0.1330 | 0.785 |
| NORM 5% | 0.3948 | 0.0985 | 0.960 | 0.8010 | 0.1372 | 1.000 |
| Procedure | $\rho_{XY_1\|Z} = -0.2742, \rho_{XY_1} = 0.5$ | | | $\rho_{XY_2\|Z} = 0.2651, \rho_{XY_2} = 0.6$ | | |
| NN | 0.5828 | 0.0579 | 0.000 | 0.4584 | 0.0692 | 0.000 |
| RI | 0.5415 | 0.0620 | 0.265 | 0.8125 | 0.0298 | 0.000 |
| RIEPS | 0.5029 | 0.0730 | 0.960 | 0.6108 | 0.0764 | 0.980 |
| NIBAS | 0.5003 | 0.0753 | 0.995 | 0.6000 | 0.0815 | 1.000 |
| NORM 1% | 0.5331 | 0.1626 | 0.865 | 0.5604 | 0.2704 | 0.820 |
| NORM 5% | 0.4931 | 0.1084 | 0.960 | 0.6102 | 0.1963 | 0.970 |
| Procedure | $\rho_{XY_1\|Z} = 0, \rho_{XY_1} = 0.5833$ | | | $\rho_{XY_2\|Z} = 0, \rho_{XY_2} = 0.4583$ | | |
| NN | 0.5817 | 0.0580 | 0.970 | 0.4579 | 0.0693 | 0.940 |
| RI | 0.6354 | 0.0523 | 0.025 | 0.6410 | 0.0517 | 0.000 |
| RIEPS | 0.5828 | 0.0664 | 0.995 | 0.4589 | 0.0941 | 1.000 |
| NIBAS | 0.5830 | 0.0664 | 0.995 | 0.4581 | 0.0993 | 1.000 |
| NORM 1% | 0.6018 | 0.1741 | 0.920 | 0.4310 | 0.3184 | 0.920 |
| NORM 5% | 0.5732 | 0.1033 | 0.955 | 0.4702 | 0.2033 | 0.950 |
| Procedure | $\rho_{XY_1\|Z} = 0.0548, \rho_{XY_1} = 0.6$ | | | $\rho_{XY_2\|Z} = 0.078, \rho_{XY_2} = 0.5$ | | |
| NN | 0.5818 | 0.0580 | 0.765 | 0.4590 | 0.0692 | 0.345 |
| RI | 0.6540 | 0.0502 | 0.015 | 0.6981 | 0.0450 | 0.000 |
| RIEPS | 0.5999 | 0.0646 | 0.975 | 0.5014 | 0.0914 | 1.000 |
| NIBAS | 0.5998 | 0.0648 | 0.960 | 0.5006 | 0.0934 | 1.000 |
| NORM 1% | 0.6286 | 0.1891 | 0.940 | 0.4536 | 0.3275 | 0.905 |
| NORM 5% | 0.5921 | 0.0984 | 0.905 | 0.5063 | 0.1831 | 0.945 |
| Procedure | $\rho_{XY_1\|Z} = 0.7129, \rho_{XY_1} = 0.8$ | | | $\rho_{XY_2\|Z} = -0.6705, \rho_{XY_2} = 0.1$ | | |
| NN | 0.5821 | 0.0580 | 0.000 | 0.4578 | 0.0693 | 0.000 |
| RI | 0.8331 | 0.0268 | 0.030 | 0.1168 | 0.0864 | 0.850 |
| RIEPS | 0.8156 | 0.0316 | 0.535 | 0.1055 | 0.0965 | 0.975 |
| NIBAS | 0.7999 | 0.0385 | 0.965 | 0.1008 | 0.1202 | 0.995 |
| NORM 1% | 0.7993 | 0.0888 | 0.945 | 0.1012 | 0.1986 | 0.970 |
| NORM 5% | 0.7972 | 0.0559 | 0.990 | 0.1063 | 0.1339 | 0.980 |

Table 2: Results for preserving the correlation structure

| | $\widehat{\mathrm{E}}(\widehat{\mu}_{Y_1})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\mu}_{Y_2})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\sigma}^2_{Y_1})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\sigma}^2_{Y_2})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\rho}_{Y_1Y_2})$ | Cvg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Procedure | $\rho_{XY_1\|Z} = -0.6032, \rho_{XY_1} = 0.4$ and $\rho_{XY_2\|Z} = 0.6393, \rho_{XY_2} = 0.8$ | | | | | | | | | |
| NN | 0.0032 | 0.93 | -0.0025 | 0.880 | 0.9903 | 0.910 | 0.9972 | 0.855 | 0.3975 | 0.885 |
| RI | 0.0019 | 0.96 | -0.0009 | 0.935 | 0.8986 | 0.090 | 0.7116 | 0.000 | 0.6532 | 0.000 |
| RIEPS | 0.0019 | 0.98 | -0.0014 | 0.980 | 0.9668 | 0.820 | 0.8863 | 0.055 | 0.4994 | 0.000 |
| NIBAS | 0.0009 | 0.99 | 0.0001 | 1.000 | 0.9998 | 0.980 | 1.0025 | 0.985 | 0.3995 | 0.995 |
| NORM 1% | 0.0066 | 0.97 | -0.0023 | 0.995 | 0.9978 | 0.985 | 1.0052 | 1.000 | 0.4065 | 0.990 |
| NORM 5% | 0.0022 | 0.99 | -0.0002 | 1.000 | 0.9906 | 0.960 | 1.0187 | 0.995 | 0.3964 | 0.990 |
| Procedure | $\rho_{XY_1\|Z} = -0.2742, \rho_{XY_1} = 0.5$ and $\rho_{XY_2\|Z} = 0.2651, \rho_{XY_2} = 0.6$ | | | | | | | | | |
| NN | 0.0006 | 0.920 | -0.0016 | 0.890 | 0.9917 | 0.915 | 0.9956 | 0.875 | 0.3988 | 0.815 |
| RI | 0.0000 | 0.925 | -0.0005 | 0.880 | 0.8540 | 0.000 | 0.5464 | 0.000 | 0.8925 | 0.000 |
| RIEPS | 0.0003 | 0.965 | -0.0009 | 0.985 | 0.9890 | 0.965 | 0.9687 | 0.945 | 0.4288 | 0.825 |
| NIBAS | -0.0001 | 0.955 | -0.0006 | 1.000 | 1.0018 | 0.980 | 1.0006 | 1.000 | 0.3996 | 0.985 |
| NORM 1% | 0.0019 | 0.935 | 0.0040 | 0.980 | 0.9999 | 0.995 | 1.0014 | 1.000 | 0.4069 | 0.995 |
| NORM 5% | -0.0008 | 0.990 | 0.0026 | 1.000 | 0.9929 | 0.985 | 1.0180 | 0.985 | 0.3953 | 0.980 |
| Procedure | $\rho_{XY_1\|Z} = 0, \rho_{XY_1} = 0.5833$ and $\rho_{XY_2\|Z} = 0, \rho_{XY_2} = 0.4583$ | | | | | | | | | |
| NN | 0.0027 | 0.920 | -0.0017 | 0.930 | 0.9897 | 0.915 | 0.9925 | 0.870 | 0.3995 | 0.895 |
| RI | 0.0015 | 0.955 | 0.0012 | 0.920 | 0.8428 | 0.000 | 0.5112 | 0.000 | 0.9600 | 0.000 |
| RIEPS | 0.0013 | 0.975 | 0.0010 | 0.995 | 1.0008 | 0.980 | 1.0003 | 0.990 | 0.4012 | 1.000 |
| NIBAS | 0.0013 | 0.970 | 0.0011 | 1.000 | 1.0006 | 0.995 | 1.0013 | 0.995 | 0.4009 | 0.995 |
| NORM 1% | 0.0014 | 0.975 | 0.0092 | 0.985 | 1.0017 | 1.000 | 0.9983 | 1.000 | 0.4061 | 0.995 |
| NORM 5% | 0.0005 | 0.995 | 0.0046 | 1.000 | 0.9931 | 0.990 | 1.0179 | 0.990 | 0.3964 | 0.990 |

Table 3: Results for preserving the moments of the fused/imputed variable (1)

| | $\widehat{\mathrm{E}}(\widehat{\mu}_{Y_1})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\mu}_{Y_2})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\sigma}^2_{Y_1})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\sigma}^2_{Y_2})$ | Cvg. | $\widehat{\mathrm{E}}(\widehat{\rho}_{Y_1Y_2})$ | Cvg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Procedure | $\rho_{XY_1\|Z} = 0.0548, \rho_{XY_1} = 0.6$ and $\rho_{XY_2\|Z} = 0.078, \rho_{XY_2} = 0.5$ | | | | | | | | | |
| NN | 0.0028 | 0.910 | 0.0021 | 0.880 | 0.9914 | 0.935 | 0.9973 | 0.910 | 0.3969 | 0.885 |
| RI | 0.0017 | 0.925 | 0.0035 | 0.835 | 0.8414 | 0.000 | 0.5150 | 0.000 | 0.9582 | 0.000 |
| RIEPS | 0.0015 | 0.950 | 0.0041 | 0.985 | 0.9998 | 0.985 | 0.9994 | 0.990 | 0.3987 | 0.990 |
| NIBAS | 0.0021 | 0.970 | 0.0032 | 1.000 | 0.9990 | 0.995 | 1.0038 | 1.000 | 0.3996 | 0.995 |
| NORM 1% | 0.0014 | 0.950 | 0.0107 | 0.970 | 0.9994 | 0.995 | 1.0018 | 1.000 | 0.4058 | 1.000 |
| NORM 5% | 0.0007 | 0.980 | 0.0068 | 0.995 | 0.9916 | 0.985 | 1.0198 | 0.995 | 0.3948 | 0.990 |
| Procedure | $\rho_{XY_1\|Z} = 0.7129, \rho_{XY_1} = 0.8$ and $\rho_{XY_2\|Z} = -0.6705, \rho_{XY_2} = 0.1$ | | | | | | | | | |
| NN | -0.0033 | 0.935 | 0.0015 | 0.920 | 0.9891 | 0.880 | 0.9922 | 0.845 | 0.3970 | 0.890 |
| RI | -0.0017 | 0.935 | 0.0005 | 0.910 | 0.9205 | 0.235 | 0.7303 | 0.000 | 0.6033 | 0.000 |
| RIEPS | -0.0018 | 0.940 | 0.0009 | 0.965 | 0.9604 | 0.775 | 0.8963 | 0.115 | 0.4968 | 0.005 |
| NIBAS | -0.0017 | 0.960 | 0.0005 | 0.980 | 0.9975 | 0.985 | 1.0019 | 0.995 | 0.3993 | 0.995 |
| NORM 1% | -0.0014 | 0.955 | 0.0075 | 0.970 | 0.9944 | 0.990 | 1.0122 | 0.990 | 0.4025 | 1.000 |
| NORM 5% | -0.0003 | 0.965 | 0.0000 | 0.985 | 0.9867 | 0.960 | 1.0108 | 1.000 | 0.4005 | 0.995 |

Table 4: Results for preserving the moments of the fused/imputed variable (2)

# References

Barnard, J., Rubin, D.B. (1999). Small-Sample Degrees of Freedom with Multiple Imputation, *Biometrika*, **86**, 948-955.

Box, G.E.P., Tiao, G.C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley, New York.

Cox, D.R., Wermuth, N. (1996). *Multivariate Dependencies*. Chapman and Hall, London.

Kadane, J.B. (2001). Some Statistical Problems in Merging Data Files, *Journal of Official Statistics*, **17**, 423-433.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Liu, T.P., Kovacevic, M.S. (1997). An Empirical Study on Categorically Constrained Matching, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 167-178.

Moriarity, C., Scheuren, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, *Journal of Official Statistics*, **17**, 407-422.

Moriarity, C., Scheuren, F. (2003). A Note on Rubin's Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations, *Journal of Business & Educational Studies*, **21**, 65-73.

D'Orazio, M., Di Zio, M., Scanu, M. (2003). Statistical matching and the likelihood principle: uncertainty and logical constraints, *ISTAT Technical Report*.

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

Rässler, S., Fleischer, K. (1998). Aspects Concerning Data Fusion Techniques, *ZUMA Nachrichten Spezial*, **4**, 317-333.

Rässler, S., Koller, F., Mäenpää, C. (2002). A Split Questionnaire Survey Design applied to German Media and Consumer Surveys, *Proceedings of the International Conference on Improving Surveys*, ICIS 2002, Copenhagen.

Raghunathan, T.E., Grizzle, J.E. (1995). A Split Questionnaire Survey Design, *Journal of the American Statistical Association*, **90**, 54-63.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching, *Journal of Business and Econometric Statistics*, **2**, 91-102.

Rubin, D.B. (1974). Characterizing the Estimation of Parameters in Incomplete-Data Problems, *Journal of the American Statistical Association*, **69**, 467-474.

Rubin, D.B. (1977). Formalizing subjective notations about the effect of nonrespondents in sample surveys, *Journal of the American Statistical Association*, **72**, 538-543.

Rubin, D.B. (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations, *Journal of Business and Economic Statistics*, **4**, 87-95.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D.B. (2002). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation, *Health Services & Outcomes Research Methodology*, **2**, 178-186.

Rubin, D.B., Thayer, D. (1978). Relating Tests Given to Different Samples, *Psychometrika*, 43, 3-10.

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

Sims, C.A. (1972). Comments, *Annals of Economic and Social Measurement*, **1**, 343-345.

Van der Putten, P., Kok, J.N., Gupta, A. (2002). Data Fusion Through Statistical Matching, *MIT Sloan School of Management*, Working Paper 4342-02.

Wendt, F. (1986). Einige Gedanken zur Fusion, *Auf dem Wege zum Partnerschaftsmodell*, Arbeitsgemeinschaft Media-Analyse e.V., Media-Micro-Census GmbH, Frankfurt, 109-140.

Author's address:

Susanne Rässler
Institute for Employment Research of the Federal Employment Services
Competence Centre Empirical Methods
Regensburger Straße 104
D-90478 Nürnberg, Germany

E-mail: susanne.raessler@iab.de