https://doi.org/10.20378/irbo-55154

Statistica Neerlandica (2003) Vol. 57, nr. 1, pp. 58–74

A Non-Iterative Bayesian Approach to Statistical Matching

Susanne Rässler*

University of Erlangen-Nürnberg, Department of Statistics and Econometrics, Lange Gasse 20, D-90403 Nürnberg, Germany

Data fusion or statistical matching techniques merge datasets from different survey samples to achieve a complete but artificial data file which contains all variables of interest. The merging of datasets is usually done on the basis of variables common to all files, but traditional methods implicitly assume conditional independence between the variables never jointly observed given the common variables. Therefore we suggest using model based approaches tackling the data fusion task by more flexible procedures. By means of suitable multiple imputation techniques, the identification problem which is inherent in statistical matching is reflected. Here a non-iterative Bayesian version of Rubin's implicit regression model is presented and compared in a simulation study with imputations from a data augmentation algorithm as well as an iterative approach using chained equations.

Key Words and Phrases: data fusion, data merging, mass imputation, file concatenation, multiple imputation, missing data, missing by design, observed-data posterior.

1 Data fusion – problems and perspectives

It seems that there is an ongoing controversy about statistical matching between statisticians. Statistical matching is blamed and repudiated by sceptical theoretical and practical statisticians about the power of matching techniques. On the other hand, well reputed statistical offices like Statistics Canada, as well as market research companies especially in Europe have done or are still doing statistical matching. In Europe this is typically called data fusion. However, from time to time there are reports published stating that data from different sources have been matched successfully.

Historically data fusion has been "invented" and extended in the early 1960s and 1970s in Germany and France to match print media information and television viewing behavior with purchasing information due to media planning needs. In the US and Canada surveys are matched since the 1970s by federal offices to achieve better comprehensive household income information or for means of microsimulation modeling.

*susanne.raessler@wiso.uni-erlangen.de.

Data fusion is initiated by two (or more) samples, one usually of larger size than the other, with the number of individuals appearing in both samples (i.e., the overlap) clearly negligible. Only certain variables, let us denote them by Z, of the interesting individual's characteristics can be observed in both samples; they are called common variables. Some variables, Y, appear only in one sample, while other variables, X, are observed exclusively in the second sample; X and Y are called specific variables. For the purpose of generalization X, Y and Z can be regarded as vectors of variables. In media practice, social class, housing conditions, marital status, terminal age of education, education, and many other variables as well as gender and age would be used as Z variables for a linking mechanism. Figure 1 illustrates the principle of statistical matching on a simplified example.

Since no single sample exists with information on X, Y and Z together, an artificial sample has to be generated by matching the observations of both samples according to Z. The objective of data fusion is the creation of a complete microdata file where every unit provides observations of all X, Y and Z variables. Once the data are matched the analysis proceeds as if the artificial fusion sample is a real sample representative for the true population of interest. Often the data are designed to be used by many analysts for many different purposes or, finally, become a public resource.

In the last decades, papers have been published showing that traditional data fusion techniques establish the so-called conditional independence, see especially RODGERS (1984) or for a more detailed discussion RÄSSLER (2002). Under conditional independence the variables never jointly observed are independent given the variables observed in both files after the fusion is performed. Referring to Figure 1 in the artificial fusion sample we find the TV viewing and the purchasing behavior being more or less (conditionally) independent given the demographic and socioeconomic information. So the gain of statistical matching is known a priori. What is the controversy about? From an information-theoretic point of view it is

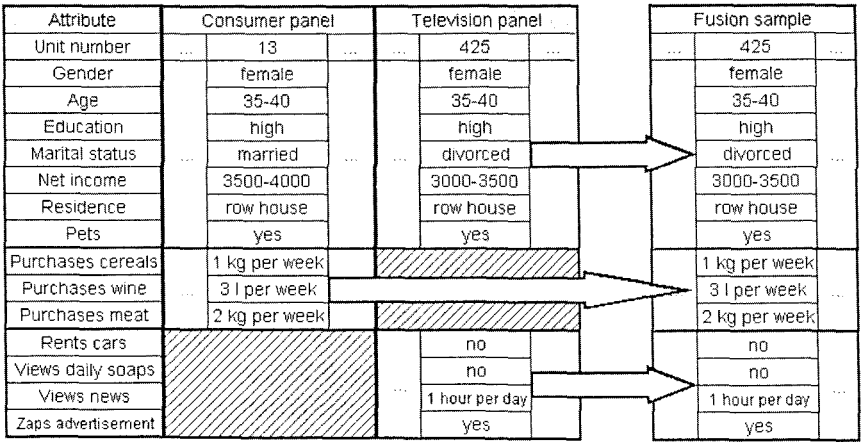

Fig. 1. Illustration of statistical matching.

easy to accept that the association of variables never jointly observed cannot be estimated from the observed data. RUBIN (1974) shows that whenever two variables are never jointly observed, the parameters of conditional association between them given the other variables are inestimable by means of likelihood inference. Nevertheless, many fusion techniques mainly based on nearest neighbor matches have been applied over years. But these traditional approaches to statistical matching establish (conditional) independence. Hence critical voices argue that any data fusion appears to be unnecessary because the outcome is already known. Moreover, conditional independence is produced for the variables not jointly observed although they may be conditionally dependent in reality. The critics are right so far. On the contrary advocates of data fusion argue, if the common variables are (carefully) chosen in a way that establishes more or less conditional independence among the variables not jointly observed given these common variables, 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.

To derive alternative procedures for matching we treat the data fusion task as a problem of nonresponse. More precisely, the missing information is regarded as missing at random because the missingness is induced by the study design of the separate samples. The missing data are due to unasked questions and the missingness mechanism is regarded as ignorable which in principle makes the application of conventional multiple imputation techniques obvious. Contrary to the usual missingness patterns, data fusion is characterized by its identification problem. The association of the variables never jointly observed is unidentifiable and cannot be estimated by means of likelihood inference. 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. Only a few approaches have been published to assess the effect of alternative assumptions of this inestimable value. KADANE (2001) (originally 1978, now reprinted), MORIARITY and SCHEUREN (2001), and RUBIN (1986) describe regression based procedures to produce synthetic datasets under various assumptions on this unknown association. A full Bayesian regression approach is given by RUBIN (1987), p. 188. We follow the latter to propose the use of multiple imputation (MI) techniques that are either based on informative prior distributions in the Bayesian context to overcome the conditional independence assumption or efficiently exploiting auxiliary data.

If no prior information about the unconditional association is available, we follow RUBIN's advice of investigating the sensitivity of the association between X and Y, rather than assuming one prior value for it. Such a prior specification can take a parametric form, e.g., as the partial correlation between X and Y given Z, with the advantage that it is relatively easy to manipulate. It also allows one to illustrate the explanatory power of the common variables. Another, but related, possibility is that additional data might be found on X, Y, and Z. Matching methodology should be able to take such auxiliary information into account, whenever it's available.

2 Stochastic regression imputation

2.1 *Introduction*

Many intuitively appealing approaches to imputation in general are based on hot deck, nearest neighbour or regression techniques; e.g., see RUBIN and SCHENKER (1998) or the recent discussions in GROVES *et al.* (2002). Instead of imputing predicted or observed values from suitable donor units, sometimes a random residual is added to account for the tendency of single imputation techniques to reduce variability. This so-called stochastic regression imputation, for example, can be used to produce multiple imputations. However, such procedures not derived within the Bayesian framework are often not proper in the sense defined by RUBIN (1987), because additional uncertainty due to random draws from the model parameters is missing. Proper MI methods reflect the sampling variability correctly; i.e., the resulting multiple imputation inference is valid also from a frequentist's view. Roughly speaking, if we get randomization-valid inference with the complete data then a MI method is proper if we get randomization-valid inference with the (theoretically infinitely) multiply imputed data.

In the context of statistical matching, RUBIN (1986) proposed an implicit (i.e., not Bayesian based) regression model concatenating the separate samples and using multiple imputations. Assuming different conditional associations of the specific variables given the common variables multiple imputations are created by means of a predictive mean matching process which was named and extended later by LITTLE (1988). We discuss a similar regression imputation procedure first using random residuals to create multiple imputations for a given conditional association. A full Bayesian MI model is derived easily then in the following section. Notice that other regression based procedures for assessing the effect of alternative assumptions of the inestimable unconditional association of the specific variables has been published by KADANE (2001) and MORIARITY and SCHEUREN (2001). The latter also use random residuals in the regression but prior to the matching process. Their matching procedure is performed using the Mahalonobis distance and leads to synthetic datasets which are imputed only once.

2.2 *Imputation procedure*

Let us consider data fusion as a problem of file concatenation; for illustration see Figure 2 which also introduces the notation used herein.

A regression imputation procedure for the fusion case can be constructed as follows. Assume the general linear model for both datasets with

$$(\text{file } A) \quad Y = Z_A\beta_{YZ} + U_A, \quad \text{and} \quad (\text{file } B) \quad X = Z_B\beta_{XZ} + U_B, \tag{1}$$

with Z_A $n_A \times k$ and Z_B $n_B \times k$ matrices of known values each with rank k. Usually we treat Z as the common derivative matrix including the constant $(1,1,\ldots,1)'$, thus, we let z_1 describe the constant. Y and X correspond to any multivariate variables as pictured in Figure 2. X denotes a $n_B \times q$ matrix and Y a $n_A \times p$ matrix according to

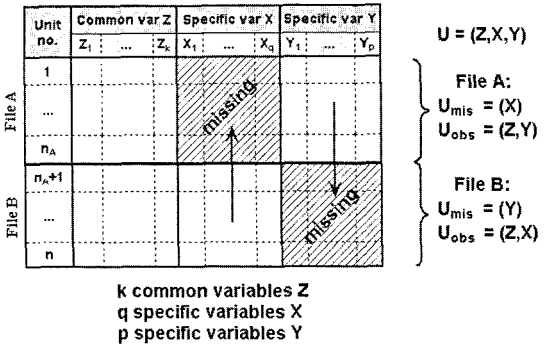

Fig. 2. Data fusion pictured as file concatenation.

the general multivariate normal model, see Box and TIAO (1992), pp. 423–425. Notice that although the common variables Z are usually regarded as being fixed, in a slight abuse of notation, we write $\sigma^2_{Y|Z}$ instead of σ^2_Y, and so on, just to correspond to the distinction between the unconditional association ρ_{XY} of X and Y and the conditional association $\rho_{XY|Z}$ of X and Y given $Z = z$.

We 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 a given common variable $Z = z$ and the parameters are also given. The expectation is $\mu_{XY|Z}$ and the covariance matrix $\Sigma_{XY|Z}$ is denoted by

$$\Sigma_{XY|Z} = \left(\begin{array}{ccc|ccc} \sigma_{X_1X_1|Z} & \cdots & \sigma_{X_1X_q|Z} & \sigma_{X_1Y_1|Z} & \cdots & \sigma_{X_1Y_p|Z} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{X_qX_1|Z} & \cdots & \sigma_{X_qX_q|Z} & \sigma_{X_qY_1|Z} & \cdots & \sigma_{X_qY_p|Z} \\ \hline \sigma_{Y_1X_1|Z} & \cdots & \sigma_{Y_1X_q|Z} & \sigma_{Y_1Y_1|Z} & \cdots & \sigma_{Y_1Y_p|Z} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{Y_pX_1|Z} & \cdots & \sigma_{Y_pX_q|Z} & \sigma_{Y_pY_1|Z} & \cdots & \sigma_{Y_pY_p|Z} \end{array}\right) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad (2)$$

then the residual $U_A \sim N_{pn_A}(0, \Sigma_{22} \otimes I_{n_A})$ and $U_B \sim N_{qn_B}(0, \Sigma_{11} \otimes I_{n_B})$. Sometimes we set $\Sigma_{11} = \Sigma_{XZ}$, $\Sigma_{22} = \Sigma_{YZ}$. 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.

The maximum likelihood (or ordinary least squares) estimates derived from the data model are $\hat{\beta}_{YZ} = (Z'_AZ_A)^{-1}Z'_AY$ and $\hat{\beta}_{XZ} = (Z'_BZ_B)^{-1}Z'_BX$. With the multivariate X and Y we get a $k \times p$ matrix of parameters β_{YZ} as well as a $k \times q$ matrix of parameters β_{XZ}. Using the estimates $\hat{\beta}_{YZ}$ and $\hat{\beta}_{XZ}$ the residual matrices Σ_{11} and Σ_{22} can be estimated for each regression with

$$
\begin{aligned}
S_Y/(n_A-k) &= (Y-Z_A\hat{\beta}_{YZ})'(Y-Z_A\hat{\beta}_{YZ})/(n_A-k)\\
&= \hat{\Sigma}_{22} = \hat{\Sigma}_{Y|Z} = \{\hat{\sigma}_{Y_iY_j|Z}\}, \quad i,j=1,2,\ldots,p\\
S_X/(n_B-k) &= (X-Z_B\hat{\beta}_{XZ})'(X-Z_B\hat{\beta}_{XZ})/(n_B-k)\\
&= \hat{\Sigma}_{11} = \hat{\Sigma}_{X|Z} = \{\hat{\sigma}_{X_iX_j|Z}\}, \quad i,j=1,2,\ldots,q. \qquad (3)
\end{aligned}
$$

Now we refer to the linear regression of X on Z and Y for file A and Y on Z and X for file B, respectively. Thus we model

$$
\begin{aligned}
&\text{(file } A\text{)} \quad X = Z_A\beta_{XZ.Y} + Y\beta_{XY.Z} + V_A, \quad V_A \sim N_{n_Aq}(0, \Sigma_{X|ZY}\otimes I_{n_A}),\\
&\text{(file } B\text{)} \quad Y = Z_B\beta_{YZ.X} + X\beta_{YX.Z} + V_B, \quad V_B \sim N_{n_Bp}(0, \Sigma_{Y|ZX}\otimes I_{n_B}). \qquad (4)
\end{aligned}
$$

To estimate the parameters of (4) we first calculate the conditional covariance matrix $\hat{\Sigma}_{XY|Z}$. Therefore any 'prior' information may be used to fix the conditional correlation matrix $R_{XY|Z}$,

$$
R_{XY|Z} = \begin{pmatrix} \rho_{X_1Y_1|Z} & \cdots & \rho_{X_1Y_p|Z} \\ \cdots & \cdots & \cdots \\ \rho_{X_qY_1|Z} & \cdots & \rho_{X_qY_p|Z} \end{pmatrix} = \{\rho_{X_iY_j|Z}\}, \qquad (5)
$$

for $i = 1, 2,\ldots,q$, $j = 1, 2,\ldots,p$ denoting a matrix of size $q \times p$. Notice that this regression imputation method is not based on Bayesian inference. Thus, prior information refers to any arbitrarily chosen values for the conditional correlation of X and Y given $Z = z$. The estimate of Σ_{12} is calculated by means of (3) with

$$
\hat{\Sigma}_{12} = \left\{\rho_{X_iY_j|Z}\sqrt{\hat{\sigma}_{X_iX_i|Z}\hat{\sigma}_{Y_jY_j|Z}}\right\}, \quad i=1,2,\ldots,q,\ j=1,2,\ldots,p, \text{ and} \qquad (6)
$$

$$
\hat{\Sigma}_{XY|Z} = \begin{pmatrix} \hat{\Sigma}_{X|Z} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{Y|Z} \end{pmatrix}, \quad \hat{\Sigma}_{21} = \hat{\Sigma}_{12}'. \qquad (7)
$$

The regression parameters for (4) are derived according to COX and WERMUTH (1996), p. 69, with

$$
\begin{aligned}
&\text{(file } A\text{)} \quad \hat{\beta}_{XZ.Y} = \hat{\beta}_{XZ} - \hat{\beta}_{YZ}\hat{\beta}_{XY.Z} \quad \text{with} \quad \hat{\beta}_{XY.Z} = \hat{\Sigma}_{Y|Z}^{-1}\hat{\Sigma}_{21}, \text{ and}\\
&\text{(file } B\text{)} \quad \hat{\beta}_{YZ.X} = \hat{\beta}_{YZ} - \hat{\beta}_{XZ}\hat{\beta}_{YX.Z} \quad \text{with} \quad \hat{\beta}_{YX.Z} = \hat{\Sigma}_{X|Z}^{-1}\hat{\Sigma}_{12}. \qquad (8)
\end{aligned}
$$

The predicted regression values of all X and Y variables are given by

$$
\begin{aligned}
&\text{(file } A\text{)} \quad \hat{X} = \hat{X}_A = Z_A\hat{\beta}_{XZ.Y} + Y\hat{\beta}_{XY.Z},\\
&\qquad\qquad \hat{Y} = \hat{Y}_A = Z_A\hat{\beta}_{YZ.X} + \hat{X}_A\hat{\beta}_{YX.Z},\\
&\text{(file } B\text{)} \quad \hat{Y} = \hat{Y}_B = Z_B\hat{\beta}_{YZ.X} + X\hat{\beta}_{YX.Z}, \text{ and}\\
&\qquad\qquad \hat{X} = \hat{X}_B = Z_B\hat{\beta}_{XZ.Y} + \hat{Y}_B\hat{\beta}_{XY.Z}. \qquad (9)
\end{aligned}
$$

The residual variances $\Sigma_{X|ZY}$ and $\Sigma_{Y|ZX}$ of the regression (4) can be estimated using (9)

$$\hat{\Sigma}_{Y|ZX} = (Y - \hat{Y})'(Y - \hat{Y})/(n_A - (k+q)) \quad \text{with } Y,\ \hat{Y} = \hat{Y}_A \text{ taken from file } A$$
$$\hat{\Sigma}_{X|ZY} = (X - \hat{X})'(X - \hat{X})/(n_B - (k+p)) \quad \text{with } X,\ \hat{X} = \hat{X}_B \text{ taken from file } B.$$

Random residuals can be generated according to $\hat{V}_{A,i} \sim N_q(0, \hat{\Sigma}_{X|ZY})$ for each single row $i = 1,2,\ldots,n_A$ and $\hat{V}_{B,i} \sim N_p(0, \hat{\Sigma}_{Y|ZX})$ for each single row $i = 1,2,\ldots,n_B$. More generally, we write $\hat{V}_A \sim N_{qn_A}(0, \hat{\Sigma}_{X|ZY} \otimes I_{nA})$ and $\hat{V}_B \sim N_{pn_B}(0, \hat{\Sigma}_{Y|ZX} \otimes I_{nB})$. These randomly generated values are added to the regression output. Finally, the imputed values are

$$\begin{aligned} &\text{(file } A\text{)} \quad \hat{X} = Z_A\hat{\beta}_{XZ.Y} + Y\hat{\beta}_{XY.Z} + \hat{V}_A, \\ &\text{(file } B\text{)} \quad \hat{Y} = Z_B\hat{\beta}_{YZ.X} + X\hat{\beta}_{YX.Z} + \hat{V}_B. \end{aligned} \tag{10}$$

The calculation of the imputed values according to (10) is equivalent to drawing the missing values from their conditional predictive distribution $f_{U_{mis}|U_{obs},\Theta}(u_{mis} \mid u_{obs}, \hat{\theta})$ given the observed data and some actual parameter values $\hat{\theta}$, i.e.,

$$\begin{aligned} &\text{(file } A\text{)} \quad \hat{X}|y \sim N_{qn_A}\left(\hat{\mu}_{X|ZY}, \hat{\Sigma}_{X|ZY} \otimes I_{nA}\right) \quad \text{and} \\ &\text{(file } B\text{)} \quad \hat{Y}|x \sim N_{pn_B}\left(\hat{\mu}_{Y|ZX}, \hat{\Sigma}_{Y|ZX} \otimes I_{nB}\right) \quad \text{with} \\ &\hat{\mu}_{X|ZY} = Z_A\hat{\beta}_{XZ.Y} + Y\hat{\beta}_{XY.Z} = Z_A\hat{\beta}_{XZ} + (Y - Z_A\hat{\beta}_{YZ})\hat{\Sigma}_{Y|Z}^{-1}\hat{\Sigma}_{21}, \\ &\hat{\mu}_{Y|ZX} = Z_B\hat{\beta}_{YZ.X} + X\hat{\beta}_{YX.Z} = Z_B\hat{\beta}_{YZ} + (X - Z_B\hat{\beta}_{XZ})\hat{\Sigma}_{X|Z}^{-1}\hat{\Sigma}_{12}. \end{aligned} \tag{11}$$

In this frequentist regression imputation with random residual, called RIEPS hereinafter, the missing values are imputed according to (11). The parameters are estimated from the observed data as described above.

2.3. *Discussion*

Regression imputation is a computationally interesting approach, because neither random draws for the parameters nor iterations to achieve any stationary distribution are necessary. Despite its computational advantages and its general acceptance among practitioners, this imputation procedure is not proper because the parameters are not randomly drawn according to their observed-data posterior. Roughly speaking, regression imputation often yields to biased estimates due to its lack of asymptotic properties. In the following section we extend this approach to its Bayesian version, performing random draws for the parameters instead of estimating them.

3 Non-iterative multivariate imputation procedure

3.1 *Introduction*

Extending the regression imputation procedure already proposed we now introduce a new non-iterative Bayesian based imputation procedure which is Bayesianly proper by definition. For short we call it **NIBAS** hereinafter. The elements of each column

of the common matrix Z can either be an appropriate sequence of -1's, 0's or 1's corresponding to a design matrix, or contain 0's and 1's describing some qualitative variables, or, finally, may simply contain continuous values. Being quite general, we only require Z to be suitable to serve as predictor matrix in a linear regression model. Concerning the specific variables X and Y our assumptions are more restrictive. Here we require variables that may be regarded as, at least, univariate normally distributed. Variables concerning media and consuming behavior may fulfill this demand after applying some useful transformations.

3.2 *Imputation procedure*

Again we assume the general linear model for both datasets with

$$\begin{aligned} &\text{(file } A\text{)} \quad Y = Z_A\beta_{YZ} + U_A, \qquad U_A \sim N_{pn_A}(0, \Sigma_{22} \otimes I_{n_A}), \\ &\text{(file } B\text{)} \quad X = Z_B\beta_{XZ} + U_B, \qquad U_B \sim N_{qn_B}(0, \Sigma_{11} \otimes I_{n_B}), \end{aligned} \tag{12}$$

with Z_A and Z_B denoting the corresponding parts of the common derivative matrix Z. Again we assume a multivariate normal data model for $(X,Y|Z = z) = (X_1,X_2,\ldots,X_q, Y_1, Y_2,\ldots,Y_p \mid Z = z)$ with expectation $\mu_{XY|Z}$ and covariance matrix $\Sigma_{XY|Z}$ like (2). As a suitable noninformative prior we assume independence between β and Σ choosing

$$f_{\beta_{YZ},\beta_{XZ},\Sigma_{X|Z},\Sigma_{Y|Z},R_{XY|Z}} \propto \Sigma_{X|Z}^{-\left(\frac{q+1}{2}\right)} \Sigma_{Y|Z}^{-\left(\frac{p+1}{2}\right)} f_{R_{XY|Z}}. \tag{13}$$

RÄSSLER (2002) shows that the joint posterior distribution for the fusion case can be factored into the prior and likelihood derived by file A and file B, respectively. Then the joint posterior distribution can be written with $f_{\beta_{XZ},\beta_{YZ},\Sigma_{X|Z},\Sigma_{Y|Z},R_{XY|Z}|X,Y} = c_X^{-1}$ $L(\beta_{XZ}, \Sigma_{X|Z}; x)\, f_{\Sigma_{X|Z}}\, f_{R_{XY|Z}}\, c_Y^{-1}\, L(\beta_{YZ}, \Sigma_{Y|Z}; y)\, f_{\Sigma_{Y|Z}} \mid R_{XY|Z} f_{R_{XY|Z}}$.

Thus, our problem of specifying the posterior distributions reduces to standard derivation tasks described, for example, by BOX and TIAO (1992), p. 439. $\Sigma_{X|Z}$ and $\Sigma_{Y|Z}$ given the observed data each is following an inverted-Wishart distribution. The conditional posterior distribution of β_{XZ} (β_{YZ}) given $\Sigma_{X|Z}$ ($\Sigma_{Y|Z}$) 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 algorithm NIBAS.

Algorithm 'NIBAS'

- Compute the ordinary least squares estimates

$$\hat{\beta}_{YZ} = (Z_A'Z_A)^{-1}Z_A'Y, \quad \text{and} \quad \hat{\beta}_{XZ} = (Z_B'Z_B)^{-1}Z_B'X$$

from the regression of each dataset. Note that $\hat{\beta}_{YZ}$ is a $k \times p$ matrix and $\hat{\beta}_{XZ}$ 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

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

- Choose a value for the correlation matrix $R_{XY|Z}$ or each $\rho_{X_iY_j|Z}$ for $i = 1, 2,\ldots,q, j = 1, 2,\ldots,p$ either
  (a) from its prior according to some distributional assumptions, e.g., 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 dataset.
- Perform random draws for the parameters from their observed-data posterior distribution according to the following scheme:

Step 1:

$$\Sigma_{22}|y \sim W_p^{-1}(v_A, S_Y^{-1}) \text{ with } v_A = n_A - (k+p) + 1,$$
$$\Sigma_{11}|x \sim W_q^{-1}(v_B, S_X^{-1}) \text{ with } v_B = n_B - (k+q) + 1.$$

Step 2:

$$\beta_{YZ}|\Sigma_{22}, y \sim N_{pk}(\hat{\beta}_{YZ}, \Sigma_{22} \otimes (Z_A'Z_A)^{-1}),$$
$$\beta_{XZ}|\Sigma_{11}, x \sim N_{qk}(\hat{\beta}_{XZ}, \Sigma_{11} \otimes (Z_B'Z_B)^{-1}).$$

Step 3:

$$\text{Set } \Sigma_{12} = \{\sigma_{X_iY_j|Z}\} \text{ with } \sigma_{X_iY_j|Z} = \rho_{X_iY_j|Z}\sqrt{\sigma^2_{X_i|Z}\sigma^2_{Y_j|Z}}$$
$$\text{with } \sigma^2_{X_i|Z}, \sigma^2_{Y_j|Z} \text{ derived by step 1}$$
$$\text{for } i = 1, 2, \ldots, q,\ j = 1, 2, \ldots, p.$$

Step 4:

$$X|y, \beta, \Sigma \sim N_{qn_A}\big(Z_A\beta_{XZ} + (Y - Z_A\beta_{YZ})\Sigma_{22}^{-1}\Sigma_{21};$$
$$(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \otimes I_{n_A}\big).$$
$$Y|x, \beta, \Sigma \sim N_{pn_B}\big(Z_B\beta_{YZ} + (X - Z_B\beta_{XZ})\Sigma_{11}^{-1}\Sigma_{12};$$
$$(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \otimes I_{n_B}\big).$$

Repeating this procedure m times yields m imputed datasets which can be analyzed by standard complete data inference. The results are combined then according to the MI paradigm. As proposed by RUBIN to create suitable imputations (e.g., see RUBIN, 1987 or RUBIN and SCHENKER, 1998), we obtain draws for the missing data from their posterior predictive distribution by first drawing values for the parameters from their observed-data posterior distribution and then drawing values for the missing

data from their predictive distribution conditional on the drawn parameter values. Thus, our imputations are repetitions from a Bayesian posterior predictive distribution for the missing data. Under the posited response mechanism and when the model for the data is appropriate, then in large samples such an imputation method should be proper for a wide range of standard statistics; for details see RUBIN (1987).

The similarity of this new multiple imputation procedure with the stochastic regression imputation method according to (11) is obvious. Instead of estimating the model parameters they are drawn from their observed-data posterior distribution. More uncertainty to account for the missing data is incorporated than by simply adding a random residual. We are also able to influence the resulting imputations by the prior choice of R_{XYZ}. The range of admissible values of R_{XY} can finally be estimated from the imputed datasets. Thus, it is possible to display the predictive power of the common variables Z by this procedure. Notice that in the multivariate case choosing the correlation matrix R_{XYZ} uniform over the $p + q$-dimensional $]-1,1[$-space may lead to invalid conditional variances in step 4. To achieve imputations reflecting the bounds of the possible range of the unconditional association between X and Y we propose to set $R_{XYZ} = 0_{q \times p}$ 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 ρ_{XYZ} may be chosen from $]-1,1[$. The bounds can be calculated directly then, see also MORIARITY and SCHEUREN (2001).

3.3 *Discussion*

We have shown that it is possible for the fusion problem to formulate a suitable data model and prior distribution and to derive the observed-data posterior therefrom. This is due to the special missingness pattern induced by the fusion. We find this approach rather encouraging because it allows a quick and controlled data fusion generating suitable multiple imputations. Further criteria and advantages of NIBAS are discussed in the following simulation study as well as in more detail by RÄSSLER (2002).

4 Simulation study

The simulation study described below has three objectives. The first objective is to explore the efficiency of estimating the unconditional association of the variables never jointly observed based on the imputed dataset when different prior information about their conditional association is used. The second objective is to investigate to what extent a third data source of a complete nature may improve the estimation of the unconditional association. Finally, we want to demonstrate the simplicity of application of the proposed fusion techniques and highlight their benefits. For an application of the following Bayesian procedures to real world media data see RÄSSLER (2002).

4.1 *Data model*

Let (Z_1, Z_2, X, Y) each be univariate standard normally distributed variables with their joint distribution $(Z_1, Z_2, X, Y) \sim N_4(0, \Sigma)$ and let

$$\Sigma = \left(\begin{array}{cc|c|c} 1.0 & 0.2 & 0.5 & 0.8 \\ 0.2 & 1.0 & 0.5 & 0.6 \\ \hline 0.5 & 0.5 & 1.0 & \sigma_{XY} \\ \hline 0.8 & 0.6 & \sigma_{YX} & 1.0 \end{array}\right) = \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZX} & \Sigma_{ZY} \\ \Sigma_{XZ} & \sigma_{XX} & \sigma_{XY} \\ \Sigma_{YZ} & \sigma_{YX} & \sigma_{YY} \end{pmatrix}, \quad \Sigma_{XY} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{pmatrix}. \quad (14)$$

Throughout the study we assume that the true covariance is given with $\sigma_{XY} = \sigma_{YX} = 0.8$. Furthermore, let file $A = (Z_1, Z_2, Y)$ and file $B = (Z_1, Z_2, X)$, thus X and Y are never jointly observed. As it is shown in RÄSSLER (2002) the simple nearest neighbor match leads to conditional independence of $X, Y \mid Z = z$ with the unconditional covariance after the fusion $\tilde{\sigma}_{XY} = \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} = 0.5833$.

To calculate the unconditional covariance, when a particular conditional correlation is given, the following well-known formula is used:

$$\Sigma_{XY} = \Sigma_{XY|Z} + \begin{bmatrix} \Sigma_{XZ} \\ \Sigma_{YZ} \end{bmatrix} \Sigma_{ZZ}^{-1} [\Sigma_{ZX}\Sigma_{ZY}] \text{ with } \Sigma_{XY|Z} = \begin{pmatrix} \sigma_{XX|Z} & \sigma_{XY|Z} \\ \sigma_{YX|Z} & \sigma_{YY|Z} \end{pmatrix}. \quad (15)$$

According to the Cauchy–Schwarz inequality the conditional covariance is bounded by $|\sigma_{XY|Z}| \leqslant \sqrt{\sigma_{XX|Z}\sigma_{YY|Z}} = \sqrt{0.5833 \cdot 0.1583} = 0.3039$. We calculate some associations according to (15) with $\sigma_{XY} = \rho_{XY|Z} \cdot 0.3039 + 0.5833$ and the corresponding determinant values $|\Sigma|$ as listed in Table 1. Due to $\sigma_{XX} = \sigma_{YY} = 1$ our setting is $\rho_{XY} = \sigma_{XY}$. Hence, only a range of the unconditional correlation of the two variables never jointly observed with $\rho_{XY} \in [0.2794, 0.8872]$ yields a positive definite matrix Σ. Notice that the "conditional independence value" of $\tilde{\rho}_{XY} = 0.5833$ is the midpoint of the range of admissible values of ρ_{XY} and maximizes Wilks' generalized variance; i.e., the determinant $|\Sigma|$ of the covariance matrix as given above with all other parameters fixed.

Table 1: Values of the determinant $|\Sigma|$ as function of $\rho_{XY|Z}$.

$\rho_{XY\|Z}$	$\sigma_{XY\|Z}$	σ_{XY}	$\lvert\Sigma\rvert$
−1.0	−0.3039	0.2794	0.0000
−0.8	−0.2431	0.3402	0.0319
−0.6	−0.1823	0.4010	0.0567
−0.4	−0.1216	0.4618	0.0745
−0.2	−0.0608	0.5226	0.0851
0.0	0.0000	0.5833	0.0887
0.2	0.0608	0.6441	0.0851
0.4	0.1216	0.7049	0.0745
0.6	0.1823	0.7657	0.0567
0.8	0.2431	0.8265	0.0319
1.0	0.3039	0.8872	0.0000

4.2 *Design of the study*

We draw $n = 5000$ random numbers for (z, x, y) according to $(Z, X, Y) \sim N_4(0, \Sigma)$. This generated dataset is divided into two parts, each file of size 2500 and all x (file A) or y (file B) values are eliminated. Then we either assume different conditional correlations of X and Y given $Z = z$ as prior information or take another random draw to generate a small but complete data source.

The multiple imputation procedures implemented for this study are as follows.

- NORM, the data augmentation algorithm assuming the normal model as proposed by SCHAFER (1997, 1999) is applied based on $m = 5$ multiple imputations. Using the S-PLUS library NORM we run a burn in period of 100 iterations, then impute the missing data from every further 50th iteration.
- NIBAS, the non-iterative multivariate Bayesian regression model based on $m = 5$ multiple imputations is used.
- An iterative univariate imputation method proposed by VAN BUUREN and OUDSHOORN (1999, 2000) implemented as S-PLUS library MICE is applied. Note that MICE does not allow the use of informative (parametric) priors as it does not rely on a parametric prior distribution for the parameters. This is typical for many of the actually available MI routines. The default settings of the S-PLUS function 'mice()' are taken here.
- RIEPS is the regression imputation technique discussed in section 2. For the fusion task we are able to introduce prior information. Moreover, regression imputation is fairly widespread among practitioners, thus, RIEPS may serve as the baseline here.

All computations are basically performed with S-PLUS 2000, copyrighted by MathSoft, Inc. Within NIBAS the prior conditional correlation $\rho_{XY|Z}^{prior}$ is used directly in step 3 of its algorithm. For NORM the unconditional covariance σ_{XY}^{prior} is calculated according to $\sigma_{XY}^{prior} = \sigma_{XY|Z}^{prior} + \hat{\Sigma}_{XZ}\hat{\Sigma}_{ZZ}^{-1}\hat{\Sigma}_{ZY}$ with $\sigma_{XY|Z}^{prior} = \rho_{XY|Z}^{prior}\sqrt{\hat{\sigma}_{XX|Z}\hat{\sigma}_{YY|Z}}$ and the covariances being estimated from all available data; i.e., $\hat{\sigma}_{XX|Z}$ and $\hat{\Sigma}_{XZ}$ from file A, $\hat{\sigma}_{YY|Z}$ and $\hat{\Sigma}_{ZY}$ from file B and $\hat{\Sigma}_{ZZ}$ using all values of Z from the available datasets. Finally, σ_{XY}^{prior} is taken as the starting value for the data augmentation algorithm itself to fix this parameter. Within RIEPS, the prior setting of $\sigma_{XY|Z}^{prior}$ is used to retain the correct regression coefficient for the regression of X on Z and Y in file A and Y on Z and Y in file B according to (6).

The whole procedure of generating and discarding data, predicting the missing values as described above $m = 5$ times for each generated dataset, calculating the usual multiple imputation estimates $\hat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}^{(t)}$ of the parameters $\theta = (\mu_X, \mu_Y, \sigma_{XX}, \sigma_{YY}, \rho_{XY})$ for each imputed dataset, is repeated $k = 50$ times. Note that NORM and NIBAS are computationally rather speedy algorithms, but MICE is not; thus we decided to restrict the repetitions to $k = 50$. The within-imputation variance $W = \frac{1}{m}\sum_{t=1}^{m}\hat{var}(\hat{\theta}^{(t)})$ and the between-imputation variance $B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2$ are computed $k = 50$ times. The 95% MI interval estimates are calculated with $\hat{\theta}_{MI} \pm \sqrt{T}t_{0.975,\nu}$, $T = W + (1 + m^{-1})B$, and degrees

of freedom $v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2$. This enables us to count the coverage; i.e., the number of times out of k that cover the true population parameter θ. To ease the reading we display the percentage. According to the MI principle we must assume that based on the complete data the point estimates $\hat{\theta}$ are approximately normal with mean θ and variance $\hat{var}(\hat{\theta})$. Therefore some estimates should be transformed to a scale for which the normal approximation works well. For example, the sampling distribution of the correlation coefficient $\hat{\rho}_{XY} = \hat{\sigma}_{XY}/\sqrt{\hat{\sigma}_{XX}\hat{\sigma}_{YY}}$ 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 ρ are calculated by means of the Fisher z-transformation $z(\hat{\rho}) = 0.5 \ln(\frac{1+\hat{\rho}}{1-\hat{\rho}})$, which makes $z(\hat{\rho})$ approximately normally distributed with mean $z(\rho)$ and constant variance $1/(n-3)$, see, e.g., SCHAFER (1997), p. 216, or BRAND (1999), p. 116. By back transforming the corresponding MI point and interval estimates of z via the inverse Fisher transformation the final estimates and confidence intervals for ρ are achieved.

4.3 *Results based on prior information*

The results are presented in detail in RÄSSLER (2002). Here we focus on the MI estimate $\hat{\rho}_{MI} = \hat{\rho}_{XY}$ of the unknown association of X and Y. The estimated expectation $\hat{E}(\hat{\rho}_{XY})$, the standard error $s(\hat{\rho}_{XY})$, denoted by

$$s(\hat{\rho}_{XY}) = \sqrt{\hat{var}(\hat{\rho}_{XY})} = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(\hat{\rho}_{XY}^{(j)} - \hat{E}(\hat{\rho}_{XY}))^2},$$

and the coverage from our $k = 50$ repetitions of the (back transformed) unconditional correlation for each procedure are listed in Table 2.

When calculating a t-statistic according to $t = \sqrt{k}(\hat{E}(\hat{\rho}_{XY}) - \rho_{XY}^{prior})/s(\hat{\rho}_{XY})$ to ease interpretation, we realize that with NORM and NIBAS only $\rho_{XY|Z}^{prior} = -0.9$ of the settings above yields an absolute value of t greater than three, whereas RIEPS only once under conditional independence has a t-value of less than three. The prior specified value of $\rho_{XY|Z}$, and, thus, also the value of ρ_{XY}, is well maintained by these procedures. Therefore, the entire range of admissible values of the unknown

Table 2: Simulation study using prior information.

	NORM			NIBAS			RIEPS		
$\rho_{XY\cdot Z}^{prior}/\rho_{XY}^{prior}$	$\hat{E}(\hat{\rho}_{XY})$	$s(\hat{\rho}_{XY})$	Cvg.	$\hat{E}(\hat{\rho}_{XY})$	$s(\hat{\rho}_{XY})$	Cvg.	$\hat{E}(\hat{\rho}_{XY})$	$s(\hat{\rho}_{XY})$	Cvg.
−0.9/0.3098	0.3007	0.0160	0.96	0.3034	0.0139	0.98	0.3186	0.0182	0.76
−0.8 0.3402	0.3339	0.0240	0.90	0.3405	0.0153	0.94	0.3547	0.0179	0.72
−0.4 0.4618	0.4588	0.0524	0.80	0.4616	0.0143	0.98	0.4720	0.0135	0.84
0.0 0.5833	0.5925	0.0780	0.84	0.5825	0.0088	1.00	0.5835	0.0106	0.98
0.4 0.7049	0.6954	0.0609	0.86	0.7048	0.0098	0.98	0.7218	0.0096	0.52
0.8 0.8265	0.8330	0.0230	0.80	0.8250	0.0100	0.92	0.8637	0.0076	0.00
0.9 0.8569	0.8563	0.0137	0.88	0.8551	0.0089	0.90	0.8833	0.0081	0.00

association is reproduced quite well by the non-iterative multivariate Bayesian regression and the data augmentation procedures. The reproduction is a little bit better and the between-imputation variance is a bit smaller with NIBAS. In the absence of any prior information MICE will assume conditional independence of X and Y given Z. See Table 3.

It seems to be a challenging area for further research to make the MICE library flexible to informative prior distributions.

4.4. *Results based on an auxiliary data file*

Now we make use of a third data source. While setting the true $\rho_{XY} = 0.8$ again we draw a small but complete dataset for (z, x, y) according to $(Z, X, Y) \sim N_4(0, \Sigma)$. For imputations via NORM and MICE the complete data are simply added to the incomplete data and their imputation procedures are performed based on the usual improper priors. For MICE the number of iterations is set to 250 (150) for the $n = 50$ (250) auxiliary file leading to a runtime of about 3 hours for each simulation run on an AMD Duron 750 MHz computer. With NIBAS and RIEPS the conditional correlation $\rho_{XY|Z}$ is estimated by means of the small sample first. We estimate σ_{XY} from this third data source and calculate $\hat{\sigma}_{XY|Z}$ according to $\hat{\sigma}_{XY|Z} = \hat{\sigma}_{XY} - \hat{\Sigma}_{XZ}\hat{\Sigma}_{ZZ}^{-1}\hat{\Sigma}_{ZY}$. Then $\hat{\rho}_{XY|Z}$ is derived by $\hat{\rho}_{XY|Z} = \hat{\sigma}_{XY|Z}/\sqrt{\hat{\sigma}_{XX|Z}\hat{\sigma}_{YY|Z}}$ and used as prior in step 3 of algorithm NIBAS or to calculate the regression coefficient for RIEPS according to (6).

Table 4 displays the mean estimate of the unconditional correlation and further statistics when samples of 1% and 5% of the actual sample are available to improve

Table 3: MICE: Simulation study assuming conditional independence.

$\rho_{XY\|Z}^{prior}$	ρ_{XY}^{prior}	$\hat{E}(\hat{\rho}_{XY})$	$s(\hat{\rho}_{XY})$	Cvg.
0.0	0.5833	0.5751	0.0146	0.96

Table 4: Simulation study using a third data source.

n	Procedure	Sample	NORM	NIBAS	MICE	RIEPS
50	$\hat{E}(\hat{\rho}_{XY})$	0.8014	0.7958	0.7985	0.7984	0.8359
	$s(\hat{\rho}_{XY})$	0.0583	0.0272	0.0255	0.0215	0.0278
	$\min(\hat{\rho}_{XY})$	0.6157	0.7092	0.7252	0.7426	0.7463
	$\max(\hat{\rho}_{XY})$	0.8995	0.8389	0.8403	0.8372	0.8726
	$\hat{E}(B)$	–	0.0016	0.0002	0.0027	0.0001
	Cvg.	0.90	0.90	0.56	0.88	0.10
250	$\hat{E}(\hat{\rho}_{XY})$	0.8001	0.8007	0.8001	0.8000	0.8386
	$s(\hat{\rho}_{XY})$	0.0231	0.0096	0.0109	0.0092	0.0106
	$\min(\hat{\rho}_{XY})$	0.7166	0.7784	0.7669	0.7777	0.8031
	$\max(\hat{\rho}_{XY})$	0.8571	0.8178	0.8233	0.8177	0.8595
	$\hat{E}(B)$	–	0.0006	0.0003	0.0006	0.0001
	Cvg.	0.96	0.94	0.88	0.92	0.02

the imputation procedure. We see from Table 4 that even a very small sample of size $n = 50$ is suitable to substitute the arbitrary prior used before and, thus, improves the imputation procedure. It is worth taking all the information provided by the two files into account then basing the estimation on the small third file alone. The range of the correlation estimate derived from the small 1% sample is considerably narrowed from $\hat{\rho}_{XY}^{sample} \in [0.6157, 0.8995]$ to the smaller intervals of $\hat{\rho}_{XY}^{NORM} \in [0.7092, 0.8389]$, $\hat{\rho}_{XY}^{NIBAS} \in [0.7252, 0.8403]$ and $\hat{\rho}_{XY}^{MICE} \in [0.7426, 0.8372]$ by using the proposed procedures. RIEPS usually overestimates the true correlation with a negligible between-imputation variance. Now the coverage is best with NORM. NIBAS produces rather small MI interval estimates due to its rather small between-imputation variance and therefore does not yield the best coverage. This might be due to the fact that estimating the prior value from auxiliary data is an ad hoc specification rather than a correct empirical Bayes procedure; this is left to future research.

4.5 *Summary*

Since all estimates used here are at least asymptotically unbiased and the sample size is quite large, the coverage of the MI interval estimate gives a good hint as to whether the implemented procedure may be proper. If the multiple imputations are proper the actual interval coverage should be equal to the nominal interval coverage. Concerning the marginal distributions NORM, NIBAS and MICE are apparently proper but only NORM (the S-PLUS library) and NIBAS are capable of including prior information in a parametric form; for details also concerning the preservation of other properties see RÄSSLER (2002). RIEPS provides the lowest coverage due to the fact that variances are often under- and correlations are over-estimated. Whether the parameters of the model are sampled from their complete or observed-data posterior distribution, the extra amount of uncertainty induced thereby improves the validity of the imputation techniques considerably. A third data source is best exploited by NORM and MICE here because the between-imputation variance based on NIBAS is small throughout. By means of the simulation study we have realized that RIEPS is not a proper imputation method even if the data follow simplifying assumptions; i.e., for example, if the data are generated according to the data model assumed. MICE has its disadvantages concerning speed and utilizing (parametric) prior information. If the normal model fits to the data or prior assumptions about the association of the variables never jointly observed others than conditional independence are suitable, we propose to use NORM or NIBAS for the imputation process, otherwise MICE is a rather flexible alternative at hand.

5 Concluding remarks

In the fusion task which can be viewed as a very special missing data pattern, the observed-data posterior distributions is derived under the assumption of a normal

data model. Thus, a particular model specially suited for the fusion task is postulated. The assumption of normality seems to be a great limitation at first glance but an application to real media data shows rather encouraging results, see RÄSSLER (2002). The iterative univariate imputation procedure MICE tries to reduce the problem of dimensionality to multivariate regressions with univariate responses. It is a very flexible procedure allowing different scales for the variable of interest. (Parametric) prior information cannot be used efficiently here. Another great advantage of the alternative approaches proposed herein we find in the property of multiple imputations to reflect the uncertainty due to the missing data. By means of MI it is possible to estimate the bounds of the unconditional association of the variables never jointly observed by using different prior settings. Furthermore, we have seen that prior information is most easily used by NIBAS whereas RIEPS does not impute enough variability. MICE does not allow the use of (parametric) prior information yet. With NORM prior information can only be applied via the hyperparameter when the standalone MS-Windows™ version is used. In the simulation study auxiliary data are most efficiently used by NORM and MICE and third by NIBAS.

References

BOX, G.E.P. and G.C. TIAO (1992), *Bayesian inference in statistical analysis*, John Wiley and Sons, New York.

BRAND, J.P.L. (1999), Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, *Thesis Erasmus University Rotterdam*. Print Partners Ispkamp, Enschede.

COX, D.R. and N. WERMUTH (1996), *Multivariate dependencies*, Chapman and Hall, London.

GROVES, R.M., D.A. DILLMAN, J.L. ELTINGE and R.J.A. LITTLE (2002), *Survey nonresponse*, Wiley, New York.

KADANE, J.B. (2001), Some statistical problems in merging data files, *Journal of Official Statistics* **17**, 423–433.

LITTLE, R.J.A. (1988), Missing-data adjustments in large surveys, *Journal of Business and Economic Statistics* **6**, 287–296.

MORIARITY, C. and F. SCHEUREN (2001), Statistical matching: a paradigm for assessing the uncertainty in the procedure, *Journal of Official Statistics* **17**, 407–422.

RÄSSLER, S. (2002), *Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches*. Lecture Notes in Statistics **168**, Springer, New York.

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. (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*, John Wiley and Sons, New York.

RUBIN, D.B. and N. SCHENKER (1998), Imputation, in S. KOTZ, C.B. READ, and D.L. BANKS (eds.) *Encyclopedia of Statistical Sciences*, Update, Volume 2, John Wiley and Sons, New York, 336–342.

SCHAFER, J.L. (1997), *Analysis of incomplete multivariate data*, Chapman and Hall, London.
SCHAFER, J.L. (1999), Multiple imputation under a normal model, Version 2, software for Windows 95/98/NT, available from http://www.stat.psu.edu/jls/misoftwa.html.
VAN BUUREN, S. and K. OUDSHOORN (1999), Flexible multivariate imputation by MICE, *TNO Report PG/VGZ/99.054*, Leiden.
VAN BUUREN, S. and C.G.M. OUDSHOORN (2000), Multivariate imputation by chained equations, *TNO Report PG/VGZ/00.038*, Leiden.

Received: February 2002. Revised: August 2002.