Aspects concerning data fusion techniques

Rässler, Susanne; Fleischer, Karlheinz

Aspects Concerning Data Fusion Techniques

SUSANNE RAESSLER AND KARLHEINZ FLEISCHER

Abstract: *Data fusion techniques merge data sets of different survey samples by means of statistical matching on the basis of common variables. As a result a virtual sample of complete, but artificial nature is generated. Because being completely unobserved, the missing information of an individual in one sample is imputed using the observed data values of some individual which is found to be most similar in the other sample. The power of data fusion techniques is analysed and the parameters of the distributions of all variables in the artificial sample are formulated. The correlation between variables not jointly observed, which can only be estimated by means of the matched file, is of main interest herein. Furthermore the influences of nearest neighbour matches, several so-called marriage processes, and small sample sizes are the focus of simulation studies.*

Key words: *missing information, imputation, merging data sets, statistical matching.*

1 Introduction

Empirical studies concerning the association between individual television viewing and purchasing behaviour, for instance, occure to be difficult in the majority of cases. The ideal medium would be a very large consumer panel, where each individual's purchasing and television viewing behaviour would both be measured. However, the costs of running a large single source panel of this kind are prohibitively high. Furthermore, a high percentage of nonresponses or poor quality of data are to be expected. A powerfully attractive alternative is to make use of a data fusion technique to link together, for example the viewing information available from a television measurement panel with the purchasing data available from an existing large market tracking panel.

Especially in the area of media analyses, data fusions have been performed in France and the UK with a reasonable degree of accuracy as published by Antoine

(1987), Baker (1990) and Roberts (1994). Further descriptions of data fusions done in practice could be found by Okner (1972a and b), Okner (1974), Ruggles and Ruggles (1974) or Scheler and Wiegand (1987). On the other hand there is scepticism among theoretical and practical statisticians about the power of fusion techniques (see Sims (1972a and b), Bennike (1987) or Gabler (1997)). Only a few publications, for instance, Woodbury (1983), Sims (1972a and b), Kovacevic and Liu (1994) or Wiedenbeck (1995), are known to mathematically study fusion algorithms and investigate their efficiency under certain circumstances.

This paper analyses the power of some data fusion techniques. Their influence on the estimated joint distribution of the variables not jointly observed will be shown herein with the help of mathematical methods as well as simulation studies.

2 Fusion algorithm

Data fusion is initiated by two 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, say Z, of the interesting individual's characteristics can be observed in both samples; they are called common variables. Other variables, Y, appear only in the larger sample while others, X, are observed exclusively in the smaller sample. (For generalization purpose X, Y, Z can be treated as vectors of variables.) 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 matching is performed at an individual level by means of statistical matching; this is often called the marriage process.

Without a loss of generality, let the smaller (X, Z) sample be the so-called recipient sample and the larger (Y, Z) sample the donor sample. For every unit i of the recipient sample with the observations (x_i, z_i) a value y from the observations of the donor sample is determined and a data set $(x_1, y_1, z_1), \ldots, (x_{n_E}, y_{n_E}, z_{n_E})$ is constructed with n_E elements of the recipient sample. The main idea is to search for a statistical match, i.e. a donor unit j whose observed data values of the common variables are identical to those of the recipient unit i.

As long as the overlap is poor, there is little chance to find a perfect match for each individual, especially if (some) common variables are continuous. Described by Baker (1990), Roberts (1994) or Antoine (1987), the marriage process is carried out using an algorithm based on nearest neighbour techniques calculated by means of a distance measure $d(.,.)$. The marriage algorithm may use all or some of the common variables, weighted or not, to find for each recipient unit i one (or more) donor unit(s) j whose distance $d(z_i, z_j)$ is minimal. By restricting the multiple choice of a donor for different recipients, further variations on the algorithm can be created. To limit the number of times a donor is taken, a penalty weight may be placed on donors already used, as the multiple choice will otherwise reduce the effective sample size and lead to underestimation of the true variance. Another modification is to take the next three (or any other number) donors and impute their (weighted) mean. If the multiple use of donors is restricted or combined with a penalty function, the resulting artificial sample, i.e. the fusion sample, may vary depending on the order the donor units are taken. Other algorithms are known to limit this problem by, for instance, cross-checking all matches after fusion and sometimes abandoning certain matches in order to find a better donator-recipient combination afterwards. Antoine (1987) gives a short description of such algorithms.

2.1 Distributions computed by fusion

In the following all density functions (joint, marginal or conditional) and their parameters produced by the fusion algorithm will be marked by the symbol $\tilde{}$.

Let X, Y, Z be multivariate random variables with joint discrete or continuous density function $f_{X,Y,Z}$. Thus, for discrete variables, $f_{X,Y,Z}(x_i, y_i, z_i)$ describes the probability to draw a certain unit i with observation (x_i, y_i, z_i) and for continuous variables it is the value of the joint density function at the point (x_i, y_i, z_i). To keep things simple, only the expression "probability" will be used hereafter. In case of continuous variables, f as the density function instead of the probability function may be taken.

If the units of the two samples are drawn independently from each other, the distribution of the donor sample of size n_S is $\prod_{i=1}^{n_S} f_{Y,Z}(y_i, z_i)$ and likewise the recipient sample is distributed with probability function $\prod_{i=1}^{n_E} f_{X,Z}(x_i, z_i)$.

Furthermore, let the fusion algorithm be one of multiple choice of the donor units without any penalty function. Thus the units of the artificial sample can be treated as being drawn independently with probability $\tilde{f}_{X,Y,Z}(x, y, z)$ each. The fusion algorithm therefore induces the probability distribution $\prod_{i=1}^{n_E} \tilde{f}_{X,Y,Z}(x_i, y_i, z_i)$ on the set of all possible fusion samples. They can be handled as simple random samples drawn from an artificial population with distribution $\tilde{f}_{X,Y,Z}(x, y, z)$, which may be called the "fusion distribution".

Often the fusion sample is used to estimate parameters (such as means, variances, covariances or higher moments) of the "initial" population following $f_{X,Y,Z}(x,y,z)$ with traditional methods. To judge the quality of such estimates, which in fact means the power of the fusion, the relation between $\tilde{f}_{X,Y,Z}(x,y,z)$ and $f_{X,Y,Z}(x,y,z)$ has to be examined.

2.1.1 Distribution of the artificial sample

As already specified, let the probability to get a particular unit i after the fusion with observation (x_i, y_i, z_i) be $\tilde{f}_{X,Y,Z}(x_i, y_i, z_i)$. This is equivalent to the probability of drawing a particular unit i of the recipient sample with observation (x_i, z_i) and merging this unit with a unit j from the donor sample with observed values (y_j, z_j), where $z_i = z_j$. The probability for a donor unit j with observed value $z_j = z_i$ from Z to have the observation y_j from Y is obviously $f_{Y|Z}(y_j|z_j)$.

Hence the probability to observe (x, y, z) for any unit of the fusion sample is

$$\tilde{f}_{X,Y,Z}(x, y, z) = f_{X,Z}(x, z) f_{Y|Z}(y|z) \tag{1}$$

provided that donor and recipient sample have been drawn independently from the same population. Thus,

$$\begin{aligned} \tilde{f}_{X,Y,Z}(x, y, z) &= f_{X,Z}(x, z) f_{Y|Z}(y|z) = f_{X|Z}(x|z) f_Z(z) f_{Y|Z}(y|z) \\ &= f_{X|Z}(x|z) f_{Y,Z}(y, z) \end{aligned} \tag{2}$$

and the conditional distribution is given by

$$\tilde{f}_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z)\, f_{Y|Z}(y|z). \tag{3}$$

It should be noted that this derivation is only admissible if, for every recipient unit, there is a donor unit with the same observed value z for Z. Especially in case of continuous distributions, this will not happen often, and a nearest neighbour unit in z has to be merged. The influence of such a nearest neighbour match on the computed distribution after the fusion will be discussed on the basis of simulation studies as well hereinafter.

2.1.2 Marginal distributions after the fusion

The marginal distributions of $\widetilde{f}_{X,Y,Z}$ are now easily obtained with the help of (2):

$$\begin{aligned}\widetilde{f}_X(x) &= \int\int \widetilde{f}_{X,Y,Z}(x,y,z)dydz = \int f_{X,Z}(x,z)\int f_{Y|Z}(y|z)dydz = \\ &= f_X(x) \end{aligned} \tag{4}$$

as is $\widetilde{f}_Y(y) = f_Y(y)$ and $\widetilde{f}_Z(z) = f_Z(z)$. Furthermore

$$\begin{aligned}\widetilde{f}_{X,Z}(x,z) &= \int \widetilde{f}_{X,Y,Z}(x,y,z)dy = f_{X,Z}(x,z)\int f_{Y|Z}(y|z)dy = \\ &= f_{X,Z}(x,z) \end{aligned} \tag{5}$$

and also $\widetilde{f}_{Y,Z}(y,z) = f_{Y,Z}(y,z)$. Different from their initial distributions are

$$\widetilde{f}_{X,Y}(x,y) = \int \widetilde{f}_{X,Y,Z}(x,y,z)dz = \int f_{X|Z}(x|z)f_Z(z)f_{Y|Z}(y|z)dz, \tag{6}$$

$$\begin{aligned}\widetilde{f}_{X,Y,Z}(x,y,z) &= f_{X,Z}(x,z)f_{Y|Z}(y|z) = \\ &= f_{X,Z}(x,z)f_{Y|X,Z}(y|x,z)\ \frac{f_{Y|Z}(y|z)}{f_{Y|X,Z}(y|x,z)} \\ &= f_{X,Y,Z}(x,y,z)\ \frac{f_{Y|Z}(y|z)}{f_{Y|X,Z}(y|x,z)} = \\ &= f_{X,Y,Z}(x,y,z)\ \frac{f_{X|Z}(x|z)}{f_{X|Y,Z}(x|y,z)} \end{aligned} \tag{7}$$

Thus, the distribution of X, Y, Z after the fusion is equal to the initial distribution if X and Y are independent, conditional on every possible value z of Z, i.e.

$$f_{Y|X,Z}(y|x,z) = f_{Y|Z}(y|z) \quad \text{or equivalent} \quad f_{X|Y,Z}(x|y,z) = f_{X|Z}(x|z).$$

Especially Sims (1972a and b) called for conditional independence as a main assumption for a reliable fusion.

Moreover, all marginal distributions, which could have been estimated already by the two separate samples, are identical before and after the fusion. Only the joint distributions of variables not jointly observed are different. Note that all derivations above are valid for vectors of random variables X, Y, Z as well. Accordingly all moments concerning variables of one or the other sample are identical for the fusion distribution and the initial distribution. See for instance $\widetilde{\mu}_X = \mu_X$,

$\widetilde{\mathrm{E}}(X^i) = \mathrm{E}(X^i)$, $\widetilde{\sigma}^2_X = \sigma^2_X$ and so on. Thus testing the accuracy of the fusion by properties of the marginal distributions of variables observed in one of the two samples is by no means useful for validation of fusion results.

Consider now the moments of the joint distribution of variables from different samples. The correlation between X and Y generated by the fusion, measured as covariance $\widetilde{\mathrm{Cov}}(X,Y)$, is

$$\widetilde{\mathrm{Cov}}(X,Y) = \mathrm{Cov}(X,Y) - \mathrm{E}(\mathrm{Cov}(X,Y|Z)). \tag{8}$$

This result, however, can easily be obtained by first calculating $\mathrm{E}(\mathrm{Cov}(X,Y|Z))$:

$$\begin{aligned}
\mathrm{E}(\mathrm{Cov}(X,Y|Z)) &= \int [\mathrm{E}(X\cdot Y|Z=z) - \mathrm{E}(X|Z=z)\mathrm{E}(Y|Z=z)]\, f_Z(z)dz \\
&= \int \Big[\int\int xy f_{X,Y|Z}(x,y|z)dxdy \\
&\quad - \int x f_{X|Z}(x|z)dx \int y f_{Y|Z}(y|z)dy\Big]\, f_Z(z)dz \\
&= \int\int\int xy f_{X,Y|Z}(x,y|z) f_Z(z)dxdydz \\
&\quad -\int\int\int xy f_{X|Z}(x|z) f_{Y|Z}(y|z) f_Z(z)dxdydz \\
&= \int\int\int xy f_{X,Y,Z}(x,y,z)dxdydz \\
&\quad -\int\int\int xy \widetilde{f}_{X,Y,Z}(x,y,z)dxdydz \\
&= \mathrm{E}(XY) - \widetilde{\mathrm{E}}(XY). \qquad (9) \\
\text{Thus}\quad \widetilde{\mathrm{E}}(XY) &= \mathrm{E}(XY) - \mathrm{E}(\mathrm{Cov}(X,Y|Z)) \quad \text{and} \\
\widetilde{\mathrm{E}}(XY) - \mu_X\mu_Y &= \mathrm{E}(XY) - \mu_X\mu_Y - \mathrm{E}(\mathrm{Cov}(X,Y|Z)) \text{ and because of } \mu = \widetilde{\mu}. \\
\widetilde{\mathrm{Cov}}(X,Y) &= \mathrm{Cov}(X,Y) - \mathrm{E}(\mathrm{Cov}(X,Y|Z)) \quad q.e.d.
\end{aligned}$$

This leads to

$$\mathrm{E}(\mathrm{Cov}(X,Y|Z)) = \mathrm{Cov}(X,Y) - \widetilde{\mathrm{Cov}}(X,Y), \tag{10}$$

i.e. the average covariance of X and Y is just the difference of the covariances from the initial, the real distribution and the fusion distribution. It may be used as a quality measurement of the fusion. The closer this value gets to zero, the better the true correlation is reproduced by the fusion.

In analogy to (9) it is quite simple to show, that

$$\begin{aligned} \mathrm{E}(\mathrm{Cov}(X^i, Y^j|Z)) &= \mathrm{E}(X^iY^j) - \widetilde{\mathrm{E}}(X^iY^j) \quad \text{and} && (11) \\ \widetilde{\mathrm{Cov}}(X^i, Y^j) &= \mathrm{Cov}(X^i, Y^j) - \mathrm{E}(\mathrm{Cov}(X^i, Y^j|Z)) \quad (i, j \in \mathbb{N}). && (12) \end{aligned}$$

Accordingly the fusion can produce "good results" concerning the true correlation between the variables X and Y never jointly observed only if they are **on the average conditionally uncorrelated**, i.e. $\mathrm{E}(\mathrm{Cov}(X, Y|Z)) = 0$. The same applies to higher moments. Therefore the independence of X und Y conditional on Z, as postulated by Sims (1972a and 1972b), is sufficient but not necessary.

2.2 Application on certain distributions

Under the assumption of a multivariate normal distribution for the joint distribution of X, Y, Z Wiedenbeck (1995), has shown the following results after the fusion, independent of the real correlation $\mathrm{Cov}(X, Y)$:

$$\widetilde{\mathrm{Cov}}(X, Y) = \tilde{\sigma}_{X,Y} = \frac{\sigma_{X,Z}\sigma_{Y,Z}}{\sigma_Z^2}. \qquad (13)$$

Using the expression (10) this leads to

$$\mathrm{E}(\mathrm{Cov}(X, Y|Z)) = \mathrm{Cov}(X, Y) - \widetilde{\mathrm{Cov}}(X, Y) = \sigma_{X,Y} - \frac{\sigma_{X,Z}\ \sigma_{Y,Z}}{\sigma_Z^2}.$$

Thus, after the fusion process the variables X and Y are computed uncorrelated without respect to their initial correlation, if X, Z or Y, Z are uncorrelated. Otherwise, if X, Z and Y, Z are correlated, then X, Y are computed correlated as well, although they may be uncorrelated initially.

Consider now X, Y, Z as being transformed via (e^X, e^Y, e^Z) to lognormally distributed random variables whose means, variances and covariance are specified by

$$\begin{aligned} \mu_X^\star &= e^{\mu_X + 0.5\sigma_X^2} \\ \sigma_X^{2\star} &= e^{2\mu_X + \sigma_X^2}(e^{\sigma_X^2} - 1) = \mu_X^{\star\,2}(e^{\sigma_X^2} - 1) \\ \sigma_{X,Y}^\star &= e^{\mu_X + 0.5\sigma_X^2 + \mu_Y + 0.5\sigma_Y^2}(e^{\rho_{X,Y}\sigma_X\sigma_Y} - 1) = \mu_X^\star\mu_Y^\star(e^{\sigma_{X,Y}} - 1) \end{aligned}$$

and so on. Let the parameters of the distribution of transformed variables based on a normal distribution be marked by the symbol *. Finally the reproduced correlation of X and Y is given by

$$\begin{aligned}\widetilde{\mathrm{Cov}}(e^X, e^Y) &= \widetilde{\sigma}^\star_{X,Y} = \mu^\star_X \mu^\star_Y \left(e^{\sigma_X \sigma_Y \frac{\sigma_{X,Z}\sigma_{Y,Z}}{\sigma_Z^2}} - 1 \right) = \\ &= \mu^\star_X \mu^\star_Y \left(e^{\sigma_X^2 \sigma_Y^2 \rho_{X,Z} \rho_{Y,Z}} - 1 \right) \qquad (14)\end{aligned}$$

In general, it could be difficult to calculate the exact formulas of the covariance reproduced by the fusion algorithm. Therefore, the investigation hereinafter will be if and to what degree of accuracy the presented results can be computed by simulation.

3 An experimental design

As mentioned before the fusion distribution was derived assuming the existence of a donor unit with identical z–values for every recipient unit. Since it is common practice to use several sociodemographical variables often combined with other continuous variables as common Z variables, the above assumption is most unlikely.

Hence the simulation study in the following is performed to consider the accuracy of the fusion distribution and estimators such as means, variances and covariances derived from it and affected by different continuous variables. Likewise, the influence of nearest neighbour matches, different marriage processes, and varying sample sizes will be discussed.

To keep it simple, the simulation study is limited to trivariate normal and lognormal distributions.

3.1 Random number generation

To generate random numbers considered as realizations of a standard normal distribution, a simple random number generator *randn()* of the MATLAB 4.0 program is used. All the programs needed for the simulation study have been created in the matrix programming language MATLAB 4.0 which is a product and trademark of The Math Works, Inc.

Based on standard normally distributed variables U, V, W produced by *randn()*, the multivariate normally distributed variables X, Y, Z are given by

$$\begin{aligned}(U,V,W) &\sim N(0,I)\\ Z &= U\cdot\sigma_Z+\mu_Z\\ X &= V\cdot\sigma_{X|Z}+\mu_{X|Z}\\ Y &= W\cdot\sigma_{Y|Z,X}+\mu_{Y|Z,X}\end{aligned}$$

assisted by their conditional distributions, for further notes see Johnson (1987), p. 50. Realizations of lognormally distributed random variables are easy to obtain by calculating (e^X, e^Y, e^Z).

3.2 Marriage processes

The statistical match or marriage process is carried out by using a simple nearest neighbour algorithm first. For each recipient unit i, the missing information Y is imputed taken from the donor unit j whose distance $|z_i - z_j|$ is minimal. The donor sample size is twice the recipient sample size with $n_S = 10000$ and $n_E = 5000$. A donor unit can be used many times without restrictions for other marriages; this process may be called "polygamy".

Moreover, the sample sizes are reduced considerably. Further variations of the algorithm are dealt with by restricting the multiple choice of the donor units. Thus the following experimental design results:

(1) "Polygamy", i.e. any multiple use of donor units is allowed with

- $n_S = 1000, n_E = 500$ and
- $n_S = 500, n_E = 500$.

(2) "Bigamy", i.e. any donor unit can be used twice only with

- $n_S = 1000, n_E = 500$ and
- $n_S = 500, n_E = 500$.

(3) "Monogamy", i.e. any donor unit can be used once only with

- $n_S = 1000, n_E = 500$ and
- $n_S = 500, n_E = 500$.

If the data sets are sorted, the order of sampling donor units influences the resulting fusion sample. Since the sample units are ordered at random, no special effect is to be expected by sampling one unit after the other.

(4) "Free triple", i.e. imputing the observations' mean of the next three donor units allowing multiple choice with

- $n_S = 1000, n_E = 500$ and
- $n_S = 500, n_E = 500$.

Since the variance of the mean of n observations is no longer identical with the variance of the population, it is not possible to reproduce even the true variance of Y by the fusion. In case of normal distributed Y and Z variables, the reproduced variance of Y is now given by

$$\begin{aligned}
\widetilde{\mathrm{Var}}(Y) &= \mathrm{E}(\widetilde{\mathrm{Var}}(Y|Z)) + \mathrm{Var}(\widetilde{\mathrm{E}}(Y|Z)) = \mathrm{E}(\mathrm{Var}(\overline{Y}|Z)) + \mathrm{Var}(\mathrm{E}(\overline{Y}|Z)) \\
&= \mathrm{E}\left(\frac{\sigma_Y^2}{n}\,(1-\rho_{Y,Z}^2)\right) + \mathrm{Var}\left(\mu_Y + \rho_{Y,Z}\,\frac{Z-\mu_Z}{\sigma_Z}\,\sigma_Y\right) \\
&= \frac{\sigma_Y^2}{n}\,(1-\rho_{Y,Z}^2) + \rho_{Y,Z}^2\,\frac{\sigma_Z^2}{\sigma_Z^2}\,\sigma_Y^2 = \sigma_Y^2\left(\frac{1}{n}\,(1-\rho_{Y,Z}^2)+\rho_{Y,Z}^2\right) \\
&= \sigma_Y^2\left(\frac{1}{n} + \rho_{Y,Z}^2\left(1-\frac{1}{n}\right)\right) = \frac{\sigma_Y^2}{n}\left(1+(n-1)\rho_{Y,Z}^2\right)
\end{aligned}$$

Thus, with the assumption of normal distributions for the free triple the following is true:

$$\widetilde{\mathrm{Var}}(Y) = \frac{\sigma_Y^2}{3}(1+2\rho_{Y,Z}^2), \tag{15}$$

which does not match the variance σ_Y^2 of the initial population.

3.3 Simulation of the reproduced covariance

Now n_E random variables (X_i, Z_i) and n_S random variables (Y_j, Z_j) are generated independently due to a given trivariate normal distribution or its transformations with mean vector μ and covariance structure Σ. This could be done either in accordance with their marginal distributions or by generating a sample of the trivariate distribution and splitting it at random. Then the two samples are merged in accordance with the specified algorithms. The empirical covariance, i.e. the estimate of the reproduced covariance, is calculated by

$$\widehat{\widetilde{\sigma}}_{X,Y} = \frac{1}{n_E - 1} \sum_{i=1}^{n_E} (x_i - \overline{x})(y_i - \overline{y})$$

from the fusion sample. This procedure is repeated k times. In this manner the estimated mean and variance of the empirical covariance are obtained from the simulated distribution. In particular, for every considered distribution

$$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y}) = \frac{1}{k} \sum_{i=1}^{k} \widehat{\widetilde{\sigma}}_{X,Y\,i}, \quad s^2(\widehat{\widetilde{\sigma}}_{X,Y}) = \frac{1}{k-1} \sum_{i=1}^{k} \left(\widehat{\widetilde{\sigma}}_{X,Y\,i} - \widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y}) \right)^2 \qquad (16)$$

and $\widetilde{\sigma}_{X,Y} - \widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$ are calculated. To assure the accuracy of the simulation, to what extend $\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$ and $\widetilde{\sigma}_{X,Y}$ agree is checked since the fusion only reproduces this covariance and not $\sigma_{X,Y}$. Furthermore, a t–statistic like value is computed with $t = \frac{\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y}) - \widetilde{\sigma}_{X,Y}}{s(\widehat{\widetilde{\sigma}}_{X,Y})} \sqrt{k}$ to ease interpretation.

As the true variance of the imputed variable Y is changed by use of the free triple, the value of $\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}^2_Y)$ is tabulated as well when using this algorithm.

To get results in reasonable time, the simulation has to be restricted to $k = 100$. On a 100 MHz pentium computer it takes about 4.2 hours to generate one empirical distribution with $n_S = 10000$.

4 Results of the simulation

4.1 Reproduced covariances

The simulation study is done with the following parameter set for the normal distribution

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{und} \quad \Sigma = \begin{pmatrix} 1 & \sigma_{X,Y} & \sigma_{X,Z} \\ \sigma_{Y,X} & 1 & \sigma_{Y,Z} \\ \sigma_{Z,X} & \sigma_{Z,Y} & 1 \end{pmatrix} \qquad (17)$$

That means $\sigma_{\cdot,\cdot} = \rho_{\cdot,\cdot}$; the parameters $\rho_{X,Z} = \rho_{Y,Z} = 0$ and 0.5 together with $\rho_{X,Y} = 0.1, 0.5$ and 0.9 each are assumed.

Using the multivariate lognormal distribution, it is possible to specify mean vectors and covariance structures such as shown by Johnson (1987), p. 83. As mentioned before, the parameters of the lognormal distribution are marked by the symbol *.

As shown in detail below, the results are quite stable despite the rather small simulation size of $k = 100$. The sample sizes are $n_S = 2n_E = 10000$ allowing multiple choice of donor units, i.e. polygamy. The different distributions used and the real correlations between X and Y have no influence on the reproduced covariances. Likewise the reproduced covariances are uninfluenced by the need to merge nearest neighbours instead of donor units identical in Z.

Table 1: Normal distribution with $n_S = 2n_E = 10000$ using polygamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\rho_{X,Y}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0.1	0	0.0012	0.0147	0.836
0	0.5	0	−0.0031	0.0149	−2.104
0	0.9	0	0.0026	0.0143	1.818
0.5	0.1	0.25	0.2469	0.0144	−2.161
0.5	0.5	0.25	0.2524	0.0174	1.365
0.5	0.9	0.25	0.2526	0.0148	1.739

Table 2: Lognormal distribution with $n_S = 2n_E = 10000$ using polygamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\rho_{X,Y}$	$\widetilde{\sigma}^\star_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}^\star_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}^\star_{X,Y})}$	t
0	0.1	0	0.0090	0.0747	1.199
0	0.5	0	−0.0231	0.0594	−3.880
0	0.9	0	0.0074	0.0629	1.181
0.5	0.1	0.7721	0.7553	0.1204	−1.388
0.5	0.5	0.7721	0.7903	0.1331	1.369
0.5	0.9	0.7721	0.7760	0.1352	0.294

As mentioned before, the true correlation $\rho_{X,Y}$ has no influence on the correlation generated by the fusion.

4.2 Influences of the marriage processes and the sample sizes

Even reducing the rather large sample sizes of recipient and donor samples to only $n_S = 1000$ and $n_E = 500$ does not affect the results. The same holds when several marriage processes are considered as is reported in the following tables. Since the true correlation $\rho_{X,Y}$ is of no influence, the simulation is done via k recipient and k donor samples generated for different $\rho_{X,Z}$ and $\rho_{Y,Z}$ values.

Table 3: Normal distribution with $n_S = 2n_E = 1000$ using polygamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0	−0.0094	0.0409	−2.293
0.2	0.04	0.0315	0.0417	−2.042
0.4	0.16	0.1573	0.0509	−0.528
0.6	0.36	0.3669	0.0497	1.379
0.8	0.64	0.6430	0.0601	0.501

Table 4: Normal distribution with $n_S = n_E = 500$ using polygamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0	0.0041	0.0487	0.839
0.2	0.04	0.0468	0.0434	1.574
0.4	0.16	0.1600	0.0533	0.001
0.6	0.36	0.3624	0.0535	0.443
0.8	0.64	0.6397	0.0587	−0.044

Table 5: Normal distribution with $n_S = 2n_E = 1000$ using bigamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0	−0.0029	0.0414	−0.694
0.2	0.04	0.0325	0.0458	−1.628
0.4	0.16	0.1587	0.0457	−0.283
0.6	0.36	0.3553	0.0540	−0.876
0.8	0.64	0.6387	0.0584	−0.220

Table 6: Normal distribution with $n_S = n_E = 500$ using bigamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0	−0.0140	0.0505	−2.764
0.2	0.04	0.0372	0.0429	−0.647
0.4	0.16	0.1568	0.0498	−0.641
0.6	0.36	0.3604	0.0501	0.085
0.8	0.64	0.6354	0.0433	−1.058

Table 7: Normal distribution with $n_S = 2n_E = 1000$ using monogamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0	−0.0001	0.0416	−0.028
0.2	0.04	0.0354	0.0454	−1.013
0.4	0.16	0.1573	0.0435	−0.613
0.6	0.36	0.3579	0.0500	−0.419
0.8	0.64	0.6338	0.0495	−1.263

Table 8: Normal distribution with $n_S = n_E = 500$ using monogamy

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t
0	0	0.0024	0.0510	0.470
0.2	0.04	0.0439	0.0504	0.770
0.4	0.16	0.1510	0.0486	−1.856
0.6	0.36	0.3414	0.0482	−3.858
0.8	0.64	0.6007	0.0467	−8.422

Table 10: Normal distribution with $n_S = 2n_E = 1000$ using free triple

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_Y^2)$	$\widetilde{\sigma}_Y^2$
0	0	−0.0011	0.0238	−0.465	0.3342	0.3333
0.2	0.04	0.0363	0.0261	−1.409	0.3542	0.3599
0.4	0.16	0.1589	0.0343	−0.334	0.4360	0.4400
0.6	0.36	0.3662	0.0394	1.566	0.5729	0.5733
0.8	0.64	0.6376	0.0546	−0.446	0.7549	0.7600

Table 11: Normal distribution with $n_S = n_E = 500$ using free triple

$\rho_{X,Z} = \rho_{Y,Z}$	$\widetilde{\sigma}_{X,Y}$	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_{X,Y})$	$\sqrt{s^2(\widehat{\widetilde{\sigma}}_{X,Y})}$	t	$\widehat{\mathrm{E}}(\widehat{\widetilde{\sigma}}_Y^2)$	$\widetilde{\sigma}_Y^2$
0	0	0.0047	0.0285	1.657	0.3321	0.3333
0.2	0.04	0.0442	0.0307	1.381	0.3615	0.3599
0.4	0.16	0.1594	0.0380	−0.163	0.4373	0.4400
0.6	0.36	0.3545	0.0432	−1.269	0.5743	0.5733
0.8	0.64	0.6297	0.0523	−1.976	0.7461	0.7600

Again the simulation results turn out very stable and match the theoretical results quite well. Therefore, it seems not necessary to examine more combinations such as $\rho_{X,Z} \neq \rho_{Y,Z}$ or positive/negative ρ.

Few absolute t–values are greater than 2. Neglecting the case of monogamy, this seems not to be systematic but a matter of small k. The different marriage processes show no further influence on the reproduced covariances as long as the donor sample is twice the recipient sample (or just larger). Even the very small sample sizes do not much affect the results. Only when using monogamy and identical sample sizes, meaning that every donor unit is used once, are bigger differences reported. If recipient and donor sample sizes are similar, this algorithm is of no practical use, of course. Furthermore, using the free triple the reproduced covariance is simulated just as well as the reproduced variance; for evidence see (15). Even this algorithm is not able to reproduce the true correlation between X and Y nor the true variance of Y.

5 Conclusions

The results are obvious. Fusion of data sets using such rather simple algorithms can reproduce the true correlation between variables X and Y not jointly observed if and only if they are uncorrelated on the average conditional on the common variable Z, i.e. if $\mathrm{E}(\mathrm{Cov}(X, Y|Z)) = 0$.

The stronger demand for conditional independence is not necessary if the interest is focused on the correlation (or higher moments) between X and Y only.

In general, the parameters reproduced by the fusion are not affected by merging nearest neighbour units instead of statistical twins. The influence of several marriage processes on the reproduced parameters is likewise low. The free triple (or any mean of n observations) should not be used if inference is done without correcting the variance of Y reproduced by the fusion. Finally, the sample sizes are not important, but the donor sample should be of larger size than the recipient sample if multiple use of donor units is restricted anyway.

References

Antoine, J. (1987). A Case Study Illustrating the Objectives and Perspectives of Fusion Techniques. In: Readership Research: Theory and Practice, ed. Henry, H., Amsterdam: Elsevier Science Publishers, 336–351.

Baker, K. (1990). The BARB/TGI Fusion. Technical Report on the Fusion Conducted in February/March 1990, Ickenham, Middlesex: Ken Baker Associates.

Bennike, S. (1987). Fusion — An Overview by an Outside Observer. In: Readership Research: Theory and Practice, ed. Henry, H., Amsterdam: Elsevier Science Publishers, 334–335.

Gabler, S. (1997). Datenfusion. Mannheim: ZUMA-Nachrichten, 40, 81–92.

Johnson, M.E. (1987). Multivariate Statistical Simulation. New York: John Wiley and Sons.

Kovacevic, M.S. and Liu, T. (1994). Statistical Matching of Survey Datafiles: A Simulation Study. Proceedings of the Section on Survey Research Methods, American Statistical Association, 497–484.

Okner, B.A. (1972a). Constructing a New Data Base from Merging Microdata Sets: The 1966 Merge File. Annals of Economic and Social Measurement, 1, 325–341.

Okner, B.A. (1972b). Reply and Comments. Annals of Economic and Social Measurement, 1, 359–362.

Okner, B.A. (1974). Data Matching and Merging: An Overview. Annals of Economic and Social Measurement, 3, 347–352.

Roberts, A. (1994). Media Exposure and Consumer Purchasing: An Improved Data Fusion Technique. Marketing and Research Today, 150–172.

Ruggles, N. and Ruggles, R. (1974). A Strategy for Merging and Matching Microdata Sets. Annals of Economic and Social Measurement, 3, 353–371.

Scheler, H.-E. and Wiegand, J. (1987). A Report on Experiments in Fusion in the 'Official' German Media Research (AG.MA). In: Readership Research: Theory and Practice, ed. Henry, H., Amsterdam: Elsevier Science Publishers, 352–360.

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

Sims, C.A. (1972b). Rejoinder. Annals of Economic and Social Measurement, 1, 355–357.

Wiedenbeck, M. (1995). Datenfusion in normalverteilten Populationen. Mannheim: Unpublished paper.

Woodbury, M.A. (1983). Statistical Record Matching for Files. In: Incomplete Data in Sample Surveys, 3, Proceedings of the Symposium, eds. Madow, W.G. and Olkin, I., New York: Academic Press, 173–181.