DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 783–807, http://dx.doi.org/10.1515/JOS-2015-0045

On Proxy Variables and Categorical Data Fusion

Li-Chun Zhang[1]

The problem of inference about the joint distribution of two categorical variables based on knowledge or observations of their marginal distributions, to be referred to as *categorical data fusion* in this paper, is relevant in statistical matching, ecological inference, market research, and several other related fields. This article organizes the use of proxy variables, to be distinguished from other auxiliary variables, both in terms of their effects on the uncertainty of fusion and the techniques of fusion. A measure of the gains of efficiency is provided, which incorporates both the identification uncertainty associated with data fusion and the sampling uncertainty that arises when the theoretical bounds of the uncertainty space are unknown and need to be estimated. Several existing techniques for generating fusion distributions (or datasets) are described and some new ones proposed. Analysis of real-life data demonstrates empirically that proxy variables can make data fusion more precise and the constructed fusion distribution more plausible.

Key words: Identification problem; sampling uncertainty; uncertainty analysis; fusion distribution; fusion data; proxy variable; relative efficiency.

1. Introduction

Some statistical problems are characterized by a lack of observations of interest. A familiar example is incomplete data due to survey nonresponse. Examples of other 'censoring' mechanisms that have received attention in the social sciences can be found in Manski (1995). In all these cases, the lack of observations of interest induces an *identification uncertainty* about any stipulated model assumptions that is not a question of the sample size but one of the data structure, such that "inference even from an infinite number of observations is subject" (Koopmans 1949, 132).

The particular situation to be considered in this article is inference about the joint distribution of two *target* categorical variables of interest based on knowledge or observations of their marginal distributions, to be referred to as *categorical data fusion*. The setting is readily recognizable in statistical matching (e.g., D'Orazio et al. 2006b; Rässler 2002), ecological inference (e.g., Wakefield 2004; King 1997), and several other related fields.

The first topic of interest in data fusion is uncertainty analysis. The identification problem implies that there exist a set of probability distributions of the target two-way contingency table, denoted by Θ and referred to as the *uncertainty space*, whose elements can be constrained by knowledge or observations of the table margins.

[1] University of Southampton, S3RI/Social Statistics and Demography, Highfield Southampton SO17 1BJ, UK and Statistics Norway, P.O. Box 8131 Dep. 0033 Oslo, Norway. Email: L.Zhang@soton.ac.uk.
Acknowledgment: I would like to thank three anonymous referees for their insightful comments, and an Associate Editor for helpful suggestions regarding the presentation.

The conceptualization and measure of uncertainty space for statistical matching have been considered in Kadane (1978), Moriarity and Scheuren (2001), D'Orazio et al. (2006a), Rässler and Kiesel (2009) and Conti et al. (2012, 2013).

The second topic of interest is data fusion techniques. Each element of the uncertainty space corresponds to a specific joint distribution. Identification is only possible by stipulation. The thus-identified joint distribution will be referred to as the *fusion distribution*. A fusion distribution should be regarded as a *pseudo* estimate of the target distribution, since the underlying assumption is not empirically verifiable. Sometimes, as is often the case in statistical matching, the practical interest is to construct a *fusion dataset* that conforms to the fusion distribution. It is natural to treat the two as the dual aspects of each data fusion technique. Indeed, D'Orazio et al. (2006b) refer to the construction of fusion distribution as statistical matching at the macro level and to fusion data as that at the micro level.

In this article we organize for the first time the use of proxy variables for categorical data fusion. We define a *proxy* variable to be similar in concept to the target variable and have the same support. For example, having a registered job-seeker status or not can be considered a proxy variable of being unemployed or not in the Labor Force Survey (LFS), but not whether a person is male or female even though both are binary variables. On the other hand, having a registered job-seeker status or not is not a proxy variable of the there-category LFS status (employed, unemployed, not in the labor force), because of the different support. It is helpful to distinguish between proxy and other auxiliary variables in data fusion both with regards to uncertainty and technique.

The rest of the article is arranged as follows. In the first place, when available, the proxy variables are usually the covariates that have the strongest association with the target ones. To facilitate a precise statement of this, in Section 2 we propose a measure of the relative efficiency of fusion with and without the proxy (or other auxiliary) variables, which builds on the measure of uncertainty space proposed by Conti et al. (2012), but here incorporates additionally the sampling uncertainty when the relevant theoretical uncertainty bounds are unknown and need to be estimated.

Next, existing methods, including conditional independence model, middle-of-bounds estimation and iterative proportional fitting, are discussed in Section 3. Note is given whether a technique can be more readily motivated depending on the availability of proxy variables. We also introduce some new methods, including a recursive derivation of the middle-of-bounds estimates, and in particular a flexible technique of *distribution calibration* for making use of proxy variables.

Thirdly, using real-life data on education, election turnout, and labor force status, we demonstrate empirically in Section 4 that proxy variables can potentially yield not only huge reduction of the identification uncertainty of data fusion, but also more plausible pseudo estimates of the target joint distribution. Finally, a short summary is given in Section 5.

2. Uncertainty Analysis

2.1. The Identification Problem

There is a general identification problem in data fusion due to the lack of joint observations of the target data. The problem can be characterized by the breakdown of

likelihood-based inference of the uncertainty space Θ. Binary data can be used to provide an illustration.

Let $Y_1 = 0, 1$ and $Y_2 = 0, 1$ be the two target variables. Consider first the situation where Y_1 and Y_2 are separately observed in two independent and disjoint samples. This is a typical setting for statistical matching. Let n_1 and n_2 be the respective sample sizes, and let y_1 and y_2 be the respective numbers of $Y_1 = 1$ and $Y_2 = 1$. Let y_1 have the Binomial (n_1, ϕ_1) distribution where $\phi_1 = P(Y_1 = 1)$, and let y_2 have the Binomial (n_2, ϕ_2) distribution where $\phi_2 = P(Y_2 = 1)$. Note that the two outcomes y_1 and y_2 are independent of each other because they are observed in two independent samples of Y_1 and Y_2, respectively. The likelihood is then given by

$$\begin{aligned} L(\phi_1, \phi_2; y_1, y_2) &\propto \phi_1^{y_1}(1-\phi_1)^{n_1-y_1}\phi_2^{y_2}(1-\phi_2)^{n_2-y_2} \\ &= (\theta_{10}+\theta_{11})^{y_1}(\theta_{00}+\theta_{01})^{n_1-y_1}(\theta_{01}+\theta_{11})^{y_2}(\theta_{00}+\theta_{10})^{n_2-y_2} \propto L(\theta; y_1, y_2) \end{aligned}$$

where $\theta = (\theta_{ij})_{i,j=0,1}$, and $\theta_{ij} = P(Y_1 = i, Y_2 = j)$ for $i, j = 0, 1$. The maximum-likelihood estimate (MLE) of (ϕ_1, ϕ_2) is $(\hat{\phi}_1, \hat{\phi}_2) = (y_1/n_1, y_2/n_2)$. However, the MLE $\hat{\boldsymbol{\theta}}$ can not be uniquely identified, but is constrained to a region called the likelihood ridge (D'Orazio et al. 2006a) defined by

$$\hat{\theta}_{10} + \hat{\theta}_{11} = y_1/n_1 \qquad \text{and} \qquad \hat{\theta}_{01} + \hat{\theta}_{11} = y_2/n_2 \qquad \text{and} \qquad \sum_{ij} \hat{\theta}_{ij} = 1$$

Next, suppose a single sample of the target data of size n, where joint observations of Y_1 and Y_2 are unavailable due to some censoring mechanism. Instead, only the marginal totals y_1 of $Y_1 = 1$ and y_2 of $Y_2 = 1$ are observed. Let n_{ij} be the number of units with $(Y_1, Y_2) = (i, j)$, for $i, j = 0, 1$, where $y_1 = n_{11} + n_{10}$ and $y_2 = n_{01} + n_{11}$. Suppose the joint cell counts follow the multinomial distribution with parameters θ as defined above. The likelihood is then the sum of the probabilities of all possible joint cell counts subjected to the marginal constraints, that is,

$$\begin{aligned} L(\boldsymbol{\theta}; y_1, y_2) &\propto P(y_1, y_2) \\ &= \sum_{m=L_{11}}^{U_{11}} P(n_{11} = m, n_{10} = y_1 - m, n_{01} = y_2 - m, n_{00} = n - y_1 - y_2 + m) \\ &= \sum_{m=L_{11}}^{U_{11}} b_m \theta_{11}^m \theta_{10}^{y_1-m} \theta_{01}^{y_2-m} \theta_{00}^{n-y_1-y_2+m} \end{aligned}$$

where $L_{11} = \max(y_1 + y_2 - n, 0)$, and $U_{11} = \min(y_1, y_2)$, and the coefficient b_m is given by

$$b_m = \frac{n!}{m!(y_1 - m)!(y_2 - m)!(n - y_1 - y_2 + m)!}$$

A variation of the setting is when one of the margins is known, as is usual in ecological inference. Suppose the marginal distribution of Y_1, that is $\phi_1 = P(Y_1 = 1)$, is known. Conditional on y_1, n_{11} and n_{01} are now modelled as two independent binomial

distributions, that is Binomial $(y_1, \theta_{11}/\phi_1)$ for n_{11}, and Binomial $(n_1 - y_1, \theta_{01}/(1-\phi_1))$ for n_{01}. The likelihood is then given by

$$L(\theta; y_1, y_2) \propto P(y_2|y_1) = \sum_{m=L_{11}}^{U_{11}} P(n_{11} = m, n_{01} = y_2 - m|y_1)$$
$$= \sum_{m=L_{11}}^{U_{11}} P(n_{11} = m|y_1)P(n_{01} = y_2 - m|n - y_1)$$

This is the same likelihood as above, except that the coefficient b_m is replaced by

$$b_m^c = \left(\frac{y_1!}{m!(y_1 - m)!}\right)\left(\frac{(n - y_1)!}{(y_2 - m)!(n - y_1 - y_2 + m)!}\right) = b_m / \left(\frac{n!}{y_1!(n - y_1)!}\right)$$

that is $b_m^c \propto b_m$ for fixed (y_1, y_2, n). Plackett (1977) demonstrates that the MLE of the log odds ratio of this 2×2 table is either ∞ or $-\infty$. Equivalently, the MLE of either $P(Y_2 = 1|Y_1 = 1)$ or $P(Y_2 = 1|Y_1 = 0)$ is 0 or1, which are all on the boundary of the likelihood ridge.

The reason for the breakdown of likelihood-based inference above is not the sample size. The number of observations might as well be infinite in any of the settings, the problem would still remain. Identification of a particular θ is only possible by stipulation, which is thus associated with an identification uncertainty that is distinct from the sampling uncertainty. The former is due to the structure of the available data, whereas the latter is basically a function of the sample size. While the sampling uncertainty will become negligible as the sample size tends to infinity, the identification uncertainty could remain fundamentally unchanged. Therefore, for proper inference in data fusion, it is necessary to quantify the identification uncertainty.

2.2. *Measure of Identification Uncertainty*

A natural approach is to construct a measure of the uncertainty space Θ, in the sense that larger Θ would imply greater identification uncertainty and *vice versa*. Denote by $Y_1 = 1, \ldots, H$ and $Y_2 = 1, \ldots, J$ the target variables of interest. Let $\phi_i = P[Y_1 = i]$ and $\phi_j = P(Y_2 = j)$, where the simplified notation requires that one observe the notational correspondence between i and Y_1 and between j and Y_2. Let $\theta_{ij} = P[(Y_1, Y_2) = (i, j)]$ be the target joint distribution. The Fréchet inequalities for θ_{ij} are given as

$$\max(\phi_i + \phi_j - 1, 0) = L_{ij} \leq \theta_{ij} \leq U_{ij} = \min(\phi_i, \phi_j)$$

It should be noted that logical constraints among the variables may invalidate these bounds. Such situations of incoherence are excluded from the general discussion below (see e.g., Lindley et al. 1979, Vantaggi 2008 and Brozzi et al. 2012 for discussions).

The Fréchet inequalities can also be given for any subtable as follows. Let $R_1 \subseteq \{1, \ldots, H\}$ be a subset of categories of Y_1, and let $R_2 \subseteq \{1, \ldots, J\}$ be that of Y_2. Let $\theta_R = \sum_{i \in R_1} \sum_{j \in R_2} \theta_{ij}$ be the total measure of the subtable corresponding to $R_1 \times R_2$. Let ϕ_{Rj} and ϕ_{Rj} be the respective marginal probabilities of the subtable, satisfying

$\theta_R = \sum_{i\in R_1} \phi_{Ri} = \sum_{j\in R_2} \phi_{Rj}$, given which the Fréchet inequalities for θ_{ij}, where $i \in R_1$ and $j \in R_2$, are given as

$$\max\,(\phi_{Ri} + \phi_{Rj} - \theta_R, 0) = L_{Rij} \le \theta_{ij}|\theta_R \le U_{Rij} = \min\,(\phi_{Ri}, \phi_{Rj}) \tag{1}$$

The full-table bounds thus correspond to the case of $\theta_R = 1$, $R_1 = \{1, \ldots, H\}$ and $R_2 = \{1, \ldots, J\}$.

Conti et al. (2012) propose using the interval width as a point-wise measure of Θ at θ_{ij}, that is,

$$\Delta_{ij} \stackrel{\text{def}}{=} U_{ij} - L_{ij} \tag{2}$$

Below we derive two results Lemma 1 and Corollary 1 in the case of categorical (Y_1, Y_2).

Lemma 1 The point-wise measure Δ_{ij} given by (2) can be directly calculated as

$$\Delta_{ij} = \min\,(\phi_i, 1 - \phi_i, \phi_j, 1 - \phi_j) \tag{3}$$

Proof. First, it is only necessary to consider the situation where $\phi_i \le \phi_j$, since one can handle the situation where $\phi_i \ge \phi_j$ by exchanging the generic denotation of Y_1 and Y_2. Next, provided $\phi_i \le \phi_j$, one only needs to distinguish between two situations: $\phi_i + \phi_j \le 1$ or $\phi_i + \phi_j > 1$. The result (3) follows then from observing:

$$\phi_i \le \phi_j \text{ and } \phi_i + \phi_j \le 1 \;\Rightarrow\; \Delta_{ij} = \phi_i \text{ and } \phi_i \le \min\,(\phi_j, 1 - \phi_j) \le 1/2 \le 1 - \phi_i$$

$$\phi_i \le \phi_j \text{ and } \phi_i + \phi_j > 1 \;\Rightarrow\; \Delta_{ij} = 1 - \phi_j \text{ and } 1 - \phi_j < \phi_i \le \phi_j \text{ and } 1 - \phi_j \le 1 - \phi_i \;\blacksquare$$

Corollary 1 The identification uncertainty (2) is the same everywhere for binary Y_1 and Y_2.

Proof. The binary outcome space can be specified as (i, i^c) and (j, j^c), respectively, such that $\phi_{i^c} = 1 - \phi_i$ and $\phi_{j^c} = 1 - \phi_j$. It follows from (1) that Δ_{ij} is the same for any (i, j). ■

Next, suppose there are additional categorical auxiliary variables X, and let $k = 1, \ldots, K$ be the levels arising from cross classifying all the variables in X. The joint distributions $\phi_{ik} = P(Y_1 = i, X = k)$ and $\phi_{jk} = P(Y_2 = j, X = k)$ are assumed to be observable or known, but not the target conditional distribution $\lambda_{ij}^k = P(Y_1 = i, Y_2 = j|X = k)$. Note that, in this paper, ϕ can designate any unconditional probability while θ will be reserved for that of (Y_1, Y_2). Note also the special tensor (or Einstein) notation for conditional probability λ_{ij}^k, which facilitates the summation convention *whenever an index appears both as superscript and subscript*. An index that appears only as subscript, or only as superscript, remains constant. Thus, for example, we have $\lambda_i^k = P(Y_1 = i|X = k)$, and $\phi_i = \lambda_i^k \phi_k = \sum_k P(Y_1 = i|X = k)P(X = k) = E_X(\lambda_i^k)$, where E_X denotes expectation over X.

As a measure of the conditional identification uncertainty given $X = k$, Conti et al. (2012) use

$$\Delta_{ij}^k \stackrel{\text{def}}{=} U_{ij}^k - L_{ij}^k \tag{4}$$

where $L_{ij}^k = \max\left(\lambda_i^k + \lambda_j^k - 1, 0\right) \le \lambda_{ij}^k \le \min\left(\lambda_i^k, \lambda_j^k\right) = U_{ij}^k$. It follows from Lemma 1 that

$$\Delta_{ij}^k = \min\left(\lambda_i^k, 1 - \lambda_i^k, \lambda_j^k, 1 - \lambda_j^k\right)$$

Note that sharper bounds are available when Y_1 and Y_2 are *ordered* categorical variables (Conti et al., 2012, 2013). Note also that it is sometimes possible to achieve point-wise identifiability due to logical constraints between the target and auxiliary variables. For instance, let Y_1 be the employment status and let X contain the payroll records at the tax authority, then the presence of wage payment in X would imply null probability of Y_1 being other than employed.

To assess the contribution of the auxiliary information $\{\phi_{ik}\}$ and $\{\phi_{jk}\}$ on θ_{ij}, put

$$\bar{\Delta}_{ij} \stackrel{\text{def}}{=} E_X\left(\Delta_{ij}^k\right) = \phi_k \Delta_{ij}^k \tag{5}$$

$$\bar{L}_{ij} \stackrel{\text{def}}{=} \phi_k L_{ij}^k = E_X\left(L_{ij}^k\right) \leq \theta_{ij} = E_X\left(\lambda_{ij}^k\right) \leq E_X\left(U_{ij}^k\right) = \phi_k U_{ij}^k \stackrel{\text{def}}{=} \bar{U}_{ij} \tag{6}$$

One observes that $\phi_i = \phi_k \lambda_i^k = E_X\left(\lambda_i^k\right)$ and $\phi_j = \phi_k \lambda_j^k = E_X\left(\lambda_j^k\right)$. It follows from Jensen's inequality that $L_{ij} \leq \bar{L}_{ij}$ and $\bar{U}_{ij} \leq U_{ij}$ (Conti et al., 2009), such that

$$\bar{\Delta}_{ij} = \bar{U}_{ij} - \bar{L}_{ij} \leq \Delta_{ij} \tag{7}$$

The result (7) means that the bounds $(\bar{L}_{ij}, \bar{U}_{ij})$ are never wider but can only be narrower than $\left(L_{ij}, U_{ij}\right)$ due to the additional information $\{\phi_{ik}\}$ and $\{\phi_{jk}\}$. A measure of the *relative efficiency (RE)* of this additional information for θ_{ij} can thus be given as

$$\gamma_{ij} = \bar{\Delta}_{ij} / \Delta_{ij} \tag{8}$$

In particular, powerful auxiliary information is often the case when proxy values for the target ones are available, which can greatly reduce the identification uncertainty, as will be illustrated in Section 4. Moreover, the scope of data fusion techniques is widened by the proxy variables (Section 3).

Conti et al. (2012) propose combining the point-wise measure (4) to yield an overall measure of the identification uncertainty through a set of normalising weights, that is,

$$\bar{\Delta} = w_k^{ij} \Delta_{ij}^k \qquad \text{where} \quad w_k^{ij} / \phi_k = \lambda_i^k \lambda_j^k$$

and $w_k^{ij} = \tilde{\phi}_{ijk} = P[(Y_1, Y_2, X) = (i, j, k)]$ when Y_1 and Y_2 are independent conditional on X. But other choices may be possible. In particular, setting $w_k^{ij} = w^{ij} \phi_k$, where $1_{ij} w^{ij} = 1$, yields

$$\Delta = w^{ij} \Delta_{ij} \qquad \text{and} \qquad \bar{\Delta} = w^{ij} \bar{\Delta}_{ij} \qquad \text{and} \qquad \gamma = \bar{\Delta} / \Delta = \bar{w}^{ij} \gamma_{ij} \tag{9}$$

where $\bar{w}^{ij} / w^{ij} = \Delta_{ij} / \Delta$. The choice (9) expresses the overall RE $\gamma = \bar{\Delta} / \Delta$ as a weighted average of the point-wise RE γ_{ij}s. The weights may be set as $w^{ij} = \phi_i \phi_j$. Or they may be chosen to reflect the relative 'importance' of θ_{ij}, for example, both $\Delta = \max \Delta_{ij}$ and $\Delta = \min \Delta_{ij}$ can be accommodated by (9). Note that, in the special case of binary data without auxiliary data, Δ_{ij} is a constant of (i, j), so that the overall measure Δ does not depend on the choice of the weights.

2.3. Estimation of Uncertainty Bound

The uncertainty bounds (L_{ij}, U_{ij}) for the target θ_{ij} depend on the marginal probabilities ϕ_i and ϕ_j. In reality these may be unknown and need to be estimated. Consequently, in uncertainty analysis one *also* needs to take into consideration the sampling uncertainty.

Take first the case where observations of Y_1 and Y_2 are available in separate and independent samples. Assume asymptotic normal distributions of $\hat{\phi}_i$ and $\hat{\phi}_j$. The distribution of the max and min of bivariate normal random variables has been studied in the literature (e.g., Nadarajah and Kotz 2008; Cain 1994). These results apply directly to $\hat{U}_{ij}$, but further derivation is needed for $\hat{L}_{ij}$. An alternative is to directly evaluate the expectations and variances by Monte Carlo calculation.

Take next the situation with a single sample, where $\hat{\phi}_i$ and $\hat{\phi}_j$ are not independent. Without losing generality, it suffices to consider $\left(\hat{L}_{11}, \hat{U}_{11}\right)$ for cell $(1, 1)$ in a 2×2 table. Denote the true cell counts by $(n_{11}, n_{10}, n_{01}, n_{00})$ where n_{11} is the cell of concern. Let $n = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij}$. The estimates $\hat{L}_{11}$, $\hat{U}_{11}$ and $\hat{\Delta} = \hat{\Delta}_{11} = \hat{U}_{11} - \hat{L}_{11}$ are, respectively, given as

$$\begin{aligned}
\hat{L}_{11} &= n^{-1} \max\,(n_{11} - n_{00}, 0) \\
\hat{U}_{11} &= n^{-1}(n_{11} + \min\,(n_{10}, n_{01})) \\
\hat{\Delta} = \hat{\Delta}_{11} &= n^{-1}(\min\,(n_{10}, n_{01}) + \min\,(n_{11}, n_{00}))
\end{aligned}$$

The expectation and variance of $\hat{L}_{11}$ can be evaluated *via* conditioning on $m = n_{11} + n_{00}$, for $m = 1, \ldots, n$. For convenience, denote by $\wp_{b;m,\psi}$ the generic binomial probability function, that is, $\wp_{b;m,\psi} = P(B = b)$ for $B \sim \text{Binomial}(m, \psi)$. Then,

$$E(\hat{L}_{11}) = n^{-1}\sum_{m=1}^{n} \mu^{+}_{m;\psi}\wp_{m;n,\xi}$$

$$V(\hat{L}_{11}) = n^{-2}\left(\sum_{m=1}^{n} \tau^{+}_{m;\psi}\wp_{m;n,\xi} - \left(\sum_{m=1}^{n} \mu^{+}_{m;\psi}\wp_{m;n,\xi}\right)^2\right)$$

where $\xi = \theta_{11} + \theta_{00}$, and $\mu^{+}_{m;\psi} = \sum_{b=k+1}^{m}(2b - m)\wp_{b;m,\psi}$ and $\tau^{+}_{m;\psi} = \sum_{b=k+1}^{m}(2b - m)^2\wp_{b;m,\psi}$, and $\psi = \theta_{11}/(\theta_{11} + \theta_{00})$, and $k = \lfloor m/2 \rfloor$ is the largest integer less or equal to $m/2$. An alternative, closed expression for $\mu^{+}_{m;\psi}$ can be given as $\mu^{+}_{m;\psi} = m(2\psi - 1)P(B \geq k + 1) + 2(k + 1)(1 - \psi)\wp_{k+1;m,\psi}$, where $B \sim \text{Binomial}(m, \psi)$, on noting the following result (Patel et al. 1976, 201):

$$\sum_{b=k}^{m} b\binom{m}{b}\psi^b(1 - \psi)^{m-b} = m\psi P(B \geq k) + k(1 - \psi)P(B = k)$$

Similarly for $\hat{U}_{11}$. Let $m = n_{10} + n_{01}$ and $\xi = \theta_{10} + \theta_{01}$. One obtains

$$E(\hat{U}_{11}) = \theta_{11} + n^{-1}\sum_{m=1}^{n} \mu_{m;\psi}\wp_{m;n,\xi}$$

$$V(\hat{U}_{11}) = n^{-1}\theta_{11}(1 - \theta_{11}) + n^{-2}\left(\sum_{m=1}^{n} \tau_{m;\psi}\wp_{m;n,\xi} - \left(\sum_{m=1}^{n} \mu_{m;\psi}\wp_{m;n,\xi}\right)^2\right)$$

$$+2n^{-2}\left(\sum_{m=1}^{n} \eta_m\mu_{m;\psi}\wp_{m;n,\xi} - n\theta_{11}\left(\sum_{m=1}^{n} \mu_{m;\psi}\wp_{m;n,\xi}\right)\right)$$

where $\eta_m = E(n_{11}|m) = (n-m)\theta_{11}/(\theta_{11}+\theta_{00})$, and

$$\mu_{m;\psi}=E(\min(A,B)|A+B=m,B\sim\text{Binomial}(m,\psi))=\sum_{b=1}^{k} b\wp_{b;m,\psi}+\sum_{b=1}^{m-k-1} b\wp_{b;m,1-\psi}$$

$$\tau_{m;\psi}=E\big(\min(A,B)^2|A+B=m,B\sim\text{Binomial}(m,\psi)\big)=\sum_{b=1}^{k} b^2\wp_{b;m,\psi}+\sum_{b=1}^{m-k-1} b^2\wp_{b;m,1-\psi}$$

Again, a closed expression can be given for $\mu_{m;\psi}=m\psi P(B\leq k)-(k+1)(1-\psi)\wp_{k+1;m,\psi}+m(1-\psi)P(B\geq k+1)-(m-k)\psi_{\wp k;m,\psi}$. Finally, *via* the same conditioning on $m=n_{10}+n_{01}$, one obtains

$$E(\hat{\Delta})=n^{-1}\sum_{m=1}^{n}(\mu_{m;\psi_1}+\mu_{n-m;\psi_2})\wp_{m;n,\xi}$$

$$\begin{aligned}V(\hat{\Delta})=&n^{-2}\left(\sum_{m=1}^{n}\tau_{m;\psi_1}\wp_{m;n,\xi}-\left(\sum_{m=1}^{n}\mu_{m;\psi_1}\wp_{m;n,\xi}\right)^2\right)\\&+n^{-2}\left(\sum_{m=1}^{n}\tau_{n-m;\psi_2}\wp_{m;n,\xi}-\left(\sum_{m=1}^{n}\mu_{n-m;\psi_2}\wp_{m;n,\xi}\right)^2\right)\\&+2n^{-2}\left(\sum_{m=1}^{n}\mu_{m;\psi_1}\mu_{n-m;\psi_2}\wp_{m;n,\xi}-\left(\sum_{m=1}^{n}\mu_{m;\psi_1}\wp_{m;n,\xi}\right)\left(\sum_{m=1}^{n}\mu_{n-m;\psi_2}\wp_{m;n,\xi}\right)\right)\end{aligned}$$

where $\psi_1=\theta_{10}/(\theta_{10}+\theta_{01})$ and $\psi_2=\theta_{11}/(\theta_{11}+\theta_{00})$.

Now that the true target distribution $\boldsymbol{\theta}$ is not identifiable, one needs to *stipulate* a particular element in the uncertainty space $\tilde{\boldsymbol{\theta}}\in\Theta$, in order to evaluate the expectations and variances above. Various fusion distributions described in Section 3 can be used. As it will be illustrated in Section 4, the choice seems to matter little to the results. In other words, the identification uncertainty of the sampling uncertainty is usually small compared to the sampling uncertainty itself.

3. Fusion Techniques

Data fusion techniques depend not only on whether auxiliary data are available, but also the nature of the auxiliary data that are available. Note will be given whether a technique requires proxy variables or not. To focus on the identification that results from the underlying assumptions, the techniques will be described in terms of the relevant theoretical distributions. It is understood that some of these may be known while some may require estimation in a particular application.

3.1. Conditional Independence Assumption

Denote by $\{(X,Y_1),(X,Y_2)\}$ the setup where each target variable is separately observed with the auxiliary ones. The *conditional independence assumption (CIA)* is given by

$$\tilde{\lambda}_{ij}^{k}=\lambda_i^k\lambda_j^k \tag{10}$$

The corresponding fusion distribution can be given in several expressions:

$$\tilde{\theta}_{ij} = \lambda_i^k \lambda_j^k \phi_k = \lambda_i^k \phi_{jk} = \lambda_j^k \phi_{ik} = \phi_{ik}\phi_{jk}/\phi_k$$

A schematic denotation of data fusion by the CIA is $Y_1 \coprod Y_2 | X$. The auxiliary data may or may not include proxy variables. However, the possibility of including a good proxy variable for at least one of the variables can be beneficial (Rässler 2002; D'Orazio et al. 2006b). The independence assumption (IA), that is, $Y_1 \coprod Y_2$ or $\tilde{\theta}_{ij} = \phi_i \phi_j$, can be considered as a special case of the CIA in the absence of auxiliary information.

To obtain categorical fusion data, some variant of the hot-deck imputation can be used (see e.g., Singh et al. 1993). Constraints of hot-deck imputation may easily be imposed when generating synthetic fusion data. For instance, starting from the dataset $\{(x_s, y_{1s}); s = 1, \ldots, n\}$, synthetic $\tilde{y}_{2s}$ can be generated randomly for each $s = 1, \ldots, n$ from the conditional distribution λ_j^k given $x_s = k$. However, one may wish to constrain the synthetic dataset $\{(x_s, y_{1s}, \tilde{y}_{2s}); s = 1, \ldots, n\}$ such that $\tilde{n}_{jk} = \sum_{s=1}^n I_{x_s=k} I_{\tilde{y}_{2s}=j} = n_k \lambda_j^k = n_k \phi_{jk}/\phi_k$ for all (j,k) and $n_k = \left(\sum_{s=1}^n I_{x_s=k}\right)$. This can be accomplished as follows: first, construct a vector of n_k components where $\tilde{n}_{jk}$ of them have value j, for $j = 1, \ldots, J$; then, assign any permutation of this vector to the units that have $x_s = k$. The difference between the unconstrained and constrained hot decks here is an example of the matching noise (see e.g., Conti et al. 2008 and Marella et al. 2008 for discussions).

It is convenient to merge separate datasets under the CIA. Okner (1972) is often cited as an early reference. But the CIA is understandably avoided in ecological inference, where it would have defeated its own purpose. It is interesting to note that the same assumption may be popular for generating fusion data, but disreputable when it comes to the construction of fusion distribution.

3.2. *Middle of Bounds*

To start with, consider the situation with no auxiliary data. The difference between the true θ_{ij} and any admissible $\tilde{\theta}_{ij}$, or the 'loss' of $\tilde{\theta}_{ij}$ as measured by $|\tilde{\theta}_{ij} - \theta_{ij}|$, has an upperbound

$$\Lambda_{ij} = \max\left(\tilde{\theta}_{ij} - L_{ij},\ U_{ij} - \tilde{\theta}_{ij}\right) = \Delta_{ij}/2 + |\tilde{\theta}_{ij} - \mu_{ij}|$$

where $\mu_{ij} = (L_{ij} + U_{ij})/2$ and $\Delta_{ij} = U_{ij} - L_{ij}$. In other words, Λ_{ij} is the *upper bound* of the identification error of $\tilde{\theta}_{ij}$. It attains the minimum value $\Delta_{ij}/2$ at

$$\tilde{\theta}_{ij} = \mu_{ij} = (L_{ij} + U_{ij})/2 \tag{11}$$

which is the *middle-of-bounds* (MoB) value that minimizes the maximum potential loss. Note that D'Orazio et al. (2006a, 2006b) define the 'middle-of-bounds' as the expectation of θ_{ij} with respect to a Bayesian distribution of the parameter. Theirs differs from the definition (11) and its minimax interpretation, except in the special case of binary Y_1 and Y_2.

The MoB fusion distribution $\tilde{\boldsymbol{\theta}}$ should be well defined and preserve all the margins of Y_1 and Y_2. Take first the binary base, and let Y_1 and Y_2 take values (i, i^c) and (j, j^c), respectively. Then,

$$
\begin{aligned}
2(\theta_{\not ij} + \theta_{\not ij^c}) &= (U_{ij} + L_{ij}) + (U_{ij^c} + L_{ij^c}) \\
&= \theta_{ij} + \min(\theta_{ij^c}, \theta_{i^c j}) + \max(\theta_{ij} - \theta_{i^c j^c}, 0) \\
&+ \theta_{ij^c} + \min(\theta_{ij}, \theta_{i^c j^c}) + \max(\theta_{ij^c} - \theta_{i^c j}, 0) = 2\phi_i
\end{aligned}
$$

since $\min(a,b) + \max(a-b,0) \equiv a$ for any a and b. An MoB fusion distribution in the nonbinary case can be constructed recursively, by repeatedly referring to the basic binary case and the subtable bounds (1). Example 1 below suffices to illustrate the idea.

Example 1. Consider the target 3×3 table to the left in Table 1. The marginal ϕ_i and ϕ_j are as given, as well as the MoB values directly derived from them. Clearly, since these do not sum to the total measure $\theta_R = 1$, they do not yield a well-defined fusion distribution. However, starting from any of them, which is by definition an admissible value of the corresponding $\tilde{\theta}_{ij}$, one can construct the corresponding MoB fusion distribution *rooted* in the chosen cell. The choice of cell (1,1) is illustrated here. The initial $\tilde{\theta}_{11} = 1/8$ partitions the remaining MoB $\tilde{\theta}_{ij}$s into three groups:

1. Cell (1,2) and (1,3). The implied row margin is $\phi_{Ri} = \phi_i - \tilde{\theta}_{i1} = 3/8$ for $i = 1$. The column margins are $\phi_{Rj} = \phi_j = 1/8$ for $j = 2$ and $5/8$ for $j = 3$. The relevant subtable is given by deleting the initial first column since, whatever the values $(\tilde{\theta}_{21}, \tilde{\theta}_{31})$, they have no effect on $(\tilde{\theta}_{12}, \tilde{\theta}_{13})$ given ϕ_{Ri} and ϕ_{Rj}. Thus the total measure of the relevant subtable is $\theta_R = 1 - 1/4 = 3/4$. The MoB $(\tilde{\theta}_{12}, \tilde{\theta}_{13}) = (1/16, 5/16)$ follow from the subtable bounds (1).
2. Similarly for cell (2,1) and (3,1). The implied column and row margins are as given. The relevant subtable is given by deleting the initial first row, yield the corresponding subtable total measure $\theta_R = 1 - 1/2 = 1/2$. The MoB $(\tilde{\theta}_{21}, \tilde{\theta}_{31})$ follow from (1).
3. The remaining cells on deleting the initial row and column occupied by the root cell (1,1). The implied row margins are $\phi_{Ri} = \phi_i - \theta_{i1}$ and $\phi_{Rj} = \phi_j - \theta_{1j}$. The implied subtable total measure is $\theta_R = 1 - \sum_j \tilde{\theta}_{1j} - \sum_i \tilde{\theta}_{i1} + \tilde{\theta}_{11}$, which is $3/8$ in this case. Clearly, the initial problem is thus reduced to the smaller, remaining 2×2 table, which can be solved recursively.

Table 1. Illustration of MoB fusion distribution rooted in cell (1,1).

	$(\theta_R = 1)$	Y_2		
		1/4	1/8	5/8
Y_1	1/2	1/8	1/8	5/16
	1/3	1/8	1/16	1/6
	1/6	1/12	1/16	1/12

$\tilde{\theta}_{11}=\frac{1}{8}$ $\longrightarrow$

Cell (1,2) & (1,3)		
$(\theta_R = 3/4)$	1/8	5/8
3/8	1/16	5/16

Cell (2,1) & (3,1)	
$(\theta_R = 1/2)$	1/8
1/3	1/16
1/6	1/16

Cell (2,2), (2,3), (3,2) & (3,3)		
$(\theta_R = 3/8)$	1/16	5/16
13/48	1/32	23/96
5/48	1/32	7/96

The resulting MoB distribution is well defined and preserves all the margins of Y_1 and Y_2. ■

In the setting $\{(X, Y_1), (X, Y_2)\}$, the *conditional* binary MoB fusion distribution is given by

$$\tilde{\lambda}_{ij}^{k} = \mu_{ij}^{k} = \frac{1}{2}\left(\max\left(\lambda_i^k + \lambda_j^k - 1, 0\right) + \min\left(\lambda_i^k, \lambda_j^k\right)\right) \tag{12}$$

such that $\tilde{\theta}_{ij} = \phi_k \tilde{\lambda}_{ij}^{k}$, denoted by $\mu_{ij}|X$. The conditional nonbinary MoB fusion distribution can be constructed recursively as described above, separately for each $X = k$. Again, the auxiliary data may or may not include proxy variables, although the plausibility of the MoB distribution can be quite different with or without the latter.

The use of binary MoB fusion distribution has been considered, for example, by Chambers and Steel (2001) in the context of ecological inference, but rarely in statistical matching. The discussion above shows that the MoB fusion distribution is more complicated to handle than CIA when merging data files containing nonbinary and/or multiple target variables.

3.3. *Structure-Preserving Estimation*

Consider the setting $\{(X', Y_1), (X', Y_2)\}$, and suppose now the auxiliary data are $X' = (X, Z_1)$, where Z_1 is a proxy variable for Y_1 and X contains the rest of the nonproxy variables. Data fusion is yielded by turning Z_1 into $\tilde{Y}_1$, under certain distributional constraints derived from the knowledge or observations available. Denote by ϕ_{ijk} the joint distribution of (Y_1, Y_2, X), and by ϕ_{hjk} that of (Z_1, Y_2, X) where the proxy Z_1 is indexed by h, and by $\tilde{\phi}_{ijk}$ the fusion distribution $(\tilde{Y}_1, Y_2, X)$ where $\tilde{Y}_1$ has the same index as Y_1 but a distinction is made between ϕ and $\tilde{\phi}$.

Structure-preserving estimation (SPREE) operates by raking (or iterative proportional fitting) of the initial table $\{\phi_{hjk}\}$ towards certain sufficient margins that are available. To identify the constraints that may be imposed, one only needs to inspect, in a 'descending' order, the log-linear representation of the fusion distribution, that is,

$$\log \tilde{\phi}_{ijk} = \tilde{\alpha}_0 + \tilde{\alpha}_i + \tilde{\alpha}_j + \tilde{\alpha}_k + \tilde{\alpha}_{ij} + \tilde{\alpha}_{ik} + \tilde{\alpha}_{jk} + \tilde{\alpha}_{ijk}$$

Take first $\tilde{\alpha}_{ijk}$, which corresponds to the sufficient margin $\tilde{\phi}_{ijk}$. Since ϕ_{ijk} is unavailable, no constraint can be imposed on $\tilde{\alpha}_{ijk}$. Next, take $\tilde{\alpha}_{jk}$, for which ϕ_{jk} can be derived from $\{(X, Z_1, Y_2)\}$ and imposed through raking. The case similar for $\tilde{\alpha}_{ik}$, where ϕ_{ik} can be derived from $\{(X, Z_1, Y_1)\}$, but not $\tilde{\alpha}_{ij}$, which requires the knowledge of ϕ_{ij}. There is no need to go through the lower-order terms as these will be fixed through the constraints already included: $\{\phi_{ik}\}$ and $\{\phi_{jk}\}$. Note that one needs to ensure that these two distributions are consistent with each other if they are estimated from separate data sources. The fusion distribution by SPREE can be characterized by the proxy interactions, derived from (Z_1, Y_2, X), which are preserved by raking

$$(\tilde{\alpha}_{ij}, \tilde{\alpha}_{ijk}) = (\alpha_{hj}, \alpha_{hjk}) \qquad \text{and} \qquad \tilde{\lambda}_i^{jk} = \tilde{\phi}_{ijk}/\phi_{jk} \tag{13}$$

A schematic representation of SPREE (13) is $(Z_1, Y_2, X) \rightarrow (\tilde{Y}_1, Y_2, X)|(Y_2, X)\&(Y_1, X)$.

Singh et al. (1993) consider a similar approach of exploring proxy data through log-linear constraints in the setting of merging three data files. The term SPREE, however, is taken directly from the small-area estimation literature that dates further back (e.g., Purcell and Kish 1980).

Two other generic settings for SPREE are worth noting. First, consider $\{(X', Y_1), (X', Y_2)\}$ where $X' = (X, Z_1, Z_2)$, that is, proxy variables are available for both Y_1 and Y_2. The SPREE can either turn (Z_1, Z_2, X, Y_2) into $(\tilde{Y}_1, Z_2, X, Y_2)$, or (Z_1, Z_2, X, Y_1) into $(Z_1, \tilde{Y}_2, X, Y_1)$. Afterwards, the 'redundant' proxy variable can be integrated out to obtain the fusion distribution $\tilde{\phi}_{ijk}$, that is, Z_2 out of the distribution of $(\tilde{Y}_1, Z_2, X, Y_2)$ or Z_1 out of the distribution of $(Z_1, \tilde{Y}_2, X, Y_1)$. For instance, the SPREE turns Z_2 (indexed by g) into $\tilde{Y}_2$ by raking of $\{\phi_{higk}\}$ towards $\{\phi_{hik}\}$ and $\{\phi_{hjk}\}$, that is, $(Z_1, Z_2, X, Y_1) \rightarrow (Z_1, \tilde{Y}_2, X, Y_1) | (Z_1, Y_1, X) \& (Z_1, Y_2, X)$, which is characterized by

$$(\tilde{\alpha}_{hij}, \tilde{\alpha}_{ijk}, \tilde{\alpha}_{hijk}) = (\alpha_{hig}, \alpha_{hgk}, \alpha_{higk}) \qquad \text{and} \qquad \tilde{\lambda}_j^{hik} = \tilde{\phi}_{hijk} / \phi_{hjk} \tag{14}$$

In the second case, consider $\{Y_1, Y_2, (X, Z_1, Z_2)\}$, where there are no joint observations of the target and auxiliary variables of any kind, but there do exist joint proxy variables among the auxiliaries. The SPREE remains operative by raking of $\{\phi_{hgk}\}$ towards $\{\phi_i\}$, $\{\phi_j\}$ and $\{\phi_k\}$, that is, $(Z_1, Z_2, X) \rightarrow (\tilde{Y}_1, \tilde{Y}_2, X) | Y_1 \& Y_2 \& X$, under which

$$(\tilde{\alpha}_{ij}, \tilde{\alpha}_{ik}, \tilde{\alpha}_{jk}, \tilde{\alpha}_{ijk}) = (\alpha_{hg}, \alpha_{hk}, \alpha_{gk}, \alpha_{hgk}) \tag{15}$$

It is instructive to note that neither the CIA (10) nor the MoB (12) is able to utilize the auxiliary data (X, Z_1, Z_2) in this setting.

3.4. *Distribution Calibration*

To start with, observe the setting $\{Y_1, Z_1\}$, where the target Y_1 and proxy Z_1 are separately available. To turn Z_1 into $\tilde{Y}_1$ that has the same distribution as Y_1, one only needs to identify an $H \times H$ matrix $\xi = \{\xi_i^h; i, h = 1, \ldots, H\}$, where $1^i \xi_i^h = 1$, such that

$$\tilde{\phi}_i = \xi_i^h \phi_h = \phi_i$$

Morever, being a gross-flow matrix from Z_1 to $\tilde{Y}_1$, ξ tells one how to generate a set of values $\{\tilde{Y}_{1s}; s = 1, \ldots, n\}$ from the initial proxy values $\{z_{1s}; s = 1, \ldots, n\}$ by constrained hot deck. Subjected to rounding, Tr(nξ) initial proxy values will then remain the same, where n is the diagonal matrix of $(n\phi_h)_{h=1,\ldots,H}$, while the rest $n - \text{Tr}(\text{n}\xi)$ will be changed. By contrast, with $\delta = \{\delta_i^h\}$ where $\delta_i^h = 1$ if $i = h$ and 0 otherwise, no proxy values will be changed at all. This suggests as a well defined approach to obtain some minimum-change $\tilde{\xi}$ by solving the following optimization problem:

$$\min_{\xi} D(\xi, \delta) \qquad \text{subject to} \quad \phi_i = \xi_i^h \phi_h \quad \text{and} \quad 1^i \xi_i^h = 1 \quad \text{and} \quad \xi_i^h \geq 0 \tag{16}$$

where $D(\xi, \delta)$ is the distance function of choice. For instance, to minimize the number of changes of the initial proxy values, one can put $D = \text{Tr}(\mathbf{n}\delta) - \text{Tr}(\mathbf{n}\xi) = n - \text{Tr}(\mathbf{n}\xi)$. Or, a squared Euclidean distance function between ξ and δ is given by $D = \sum_{i,h} (\xi_i^h - \delta_i^h)^2$.

Provided additional nonproxy auxiliary data, *distribution calibration (DC)* defined by (16) can be applied conditionally. Suppose the setting $\{(X, Y_1), (X, Z_1)\}$.

Conditional distribution calibration (CDC) from Z_1 to $\tilde{Y}_1$ for each $X = k$ yields $\tilde{\xi}_k = \{\tilde{\xi}_i^{hk}; i, h = 1, \ldots, H\}$, such that

$$\tilde{\lambda}_i^k = \lambda_h^k \tilde{\xi}_i^{hk} = \lambda_i^k \quad \text{and} \quad \tilde{\phi}_i = \phi_k \tilde{\lambda}_i^k = \phi_k \lambda_i^k = \phi_i \quad \text{and}$$

$$\tilde{\phi}_{ik} = \phi_k(\phi_{ik}/\phi_k) = \phi_{ik}$$

Note that a different distribution $\tilde{\phi}_{ik}$ of $(\tilde{Y}_1, X)$ would be generated by unconditional DC, that is, $\tilde{\phi}_i = \tilde{\xi}_i^h \phi_h$, since $\tilde{Y}_1$ is then independent of X given Z_1, such that $\tilde{\phi}_{ik} = \phi_{hk} \tilde{\xi}_i^h \neq \phi_{ik}$.

Given the relevant proxy variables, DC and CDC can be used to generate a fusion distribution, whether or not there are joint observations of the target and proxy variables. Consider again the setting $\{Y_1, Y_2, (X, Z_1, Z_2)\}$. A scheme of DC can be as follows:

$$\left.\begin{array}{lll} Z_1 \xrightarrow{\text{DC}} \tilde{Y}_1 & \Rightarrow & \tilde{\phi}_i = \phi_h \xi_i^h = \phi_i \\ Z_2 \xrightarrow{\text{DC}} \tilde{Y}_2 & \Rightarrow & \tilde{\phi}_j = \phi_g \xi_j^g = \phi_j \end{array}\right\} \Rightarrow \quad \tilde{\phi}_{ijk} = 1^{gh} \tilde{\phi}_{ghijk}^{Z_2 Z_1 \tilde{Y}_1 \tilde{Y}_2 X} = \lambda_k^{gh} \phi_{gh} \xi_i^h \xi_j^g$$

where the last expression follows since $\tilde{Y}_1$ is independent of the other variables given Z_1 and similarly for $\tilde{Y}_2$ given Z_2. This is a different fusion distribution than that by SPREE (15).

It is worth noting that DC and CDC can be useful for generating fusion data prescribed by another fusion technique. Take the SPREE (15) under the setting $\{Y_1, Y_2, (X, Z_1, Z_2)\}$. It is not immediately clear how to generate the fusion data it implies. However, let $\tilde{\lambda}_p^k$ be the fusion conditional probability of $p = (i, j)$ given $X = k$. Let $q = (h, g)$ index (Z_1, Z_2) in accordance. Then, CDC satisfying $\tilde{\lambda}_p^k = \lambda_q^k \xi_p^{qk}$ yields the gross-flowmatrix that can turn (Z_1, Z_2) into the SPREE $(\tilde{Y}_1, \tilde{Y}_2)$ with minimum changes given $X = k$. As another example, consider CDC under the setting $\{(X, Y_1), (X, Y_2), (X, Z_1, Z_2)\}$:

$$\left.\begin{array}{lllll} Z_1 \xrightarrow{\text{CDC}} \tilde{Y}_1|X & \Rightarrow & \tilde{\lambda}_i^k = \lambda_i^k & \Rightarrow & \tilde{\lambda}_{ik} = \phi_{ik} \\ Z_2 \xrightarrow{\text{CDC}} \tilde{Y}_2|X & \Rightarrow & \tilde{\lambda}_j^k = \lambda_j^k & \Rightarrow & \tilde{\lambda}_{jk} = \phi_{jk} \end{array}\right\} \Rightarrow \quad \tilde{\phi}_{ijk} = \frac{\tilde{\phi}_{ik}\tilde{\phi}_{jk}}{\phi_k} = \frac{\phi_{ik}\phi_{jk}}{\phi_k}$$

that is, exactly the same fusion distribution as that of the CIA in the setting $\{(X, Y_1), (X, Y_2)\}$. But CDC can yield different fusion data. For instance, suppose $\{(X, Y_1), (X, Y_2)\}$ represent two separate sample datasets, while $\{(X, Z_1, Z_2)\}$ is a population register dataset. On the one hand, a population fusion dataset can be generated by CDC; on the other hand, a synthetic CIA population fusion dataset can be obtained by randomly and separately generating $\tilde{Y}_1$ and $\tilde{Y}_2$ conditional on X in the population. Both datasets will have the same fusion distribution, but the CDC data will resemble the real population much more than the CIA data.

4. Two Cases

Two real-life datasets involving education, election turnout, and labor force status variables are used to illustrate the approach to uncertainty analysis and the fusion

techniques described above, and to empirically evaluate the relative efficiency of the available proxy data.

4.1. Education and Election Turnout: Binary Data

Both the highest level of education and election turnout are collected in the Norwegian Election Survey 2005, to be treated as Y_1 and Y_2, respectively. A level of education can also be compiled based on the register information available at Statistics Norway, denoted as Z_1, while the true head count can be obtained from the local electoral offices, denoted by Z_2. Both Z_1 and Z_2 can be linked to the survey at the individual level, and the observed four-way table for the respondents in Election Survey 2005 provides all the data for this illustration. For ease of exposition, only two categories "Low" and "High" are coded for the education variable.

Various settings of the data are given in Table 2. All the cross counts of Y_1 and Y_2 are given in parentheses and assumed to be unobserved. In the top block, the overall unconditional counts of (Y_1, Y_2) are given to the left, and those of (Z_1, Z_2) to the right. Together they provide the setting $\{Y_1, Y_2, (Z_1, Z_2)\}$. The next block gives the setting $\{(Z_1, Y_1), (Z_1, Y_2)\}$, where Z_1 is the only auxiliary data. The case is similar for $\{(Z_2, Y_1), (Z_2, Y_2)\}$ in the third block. Lastly, the bottom block provides the setting $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$.

Table 3 illustrates the results of uncertainty analysis for $P[(Y_1, Y_2) = (Low, No)]$. The first row corresponds to the setting $\{Y_1, Y_2, (Z_1, Z_2)\}$. The estimated lower and upper bounds are $(0.0, 0.104)$. The estimated width of the uncertainty space at this point is 0.104. Since Θ measures the same everywhere in the case of binary data, as previously noted for (2), 0.104 is also the estimated overall measure of the uncertainty space. The relative efficiency is unity by definition. The associated sampling uncertainty is evaluated as described in Subsection 2.3, for which it is necessary to stipulate a joint distribution. Three alternatives are illustrated in Table 3. The first one is the true sample distribution of (Y_1, Y_2) given in Table 2; the second one is the CIA fusion distribution; and the last one is the MoB fusion distribution. It is seen that the estimated standard errors (SEs) are virtually the same using any of the three alternatives.

In a similar manner, the other rows of Table 3 provide the results under different settings of jointly available auxiliary data. It is seen that with only Z_1 available, the identification uncertainty is reduced by 17% (that is, RE $= 0.83$), whereas the reduction is 62% (that is, RE $= 0.38$) with Z_2, so that it is much more informative than Z_1. With both proxy variables available, the estimated uncertainty bounds are $(0.074, 0.095)$, strictly narrower than the initial $(0.0, 0.104)$ on both sides. The width of the interval is 0.021, which is about one fifth of that without (Z_1, Z_2). Taking into account the sampling uncertainty, an approximate 95% confidence interval of the width of the identification uncertainty interval is $(0.014, 0.028)$. In comparison, had the joint sample of (Y_1, Y_2) been available, the width of the approximate 95% confidence interval of $P[(Y_1, Y_2) = (Low, No)]$ would have been 0.027. Thus, *in this respect*, there is at least as much information about $P[(Y_1, Y_2) = (Low, No)]$ in $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$ as that in $\{(Y_1, Y_2)\}$.

Table 4 illustrates a number of (pseudo) estimates of $P[(Y_1, Y_2) = (Low, No)]$ together with their respective identification assumptions. The first one (from the top) is based on the true data of (Y_1, Y_2). The next five are derived under the setting

$\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$. Note the difference between the two CIAs. The two situations of single proxy variable follow next. In the last setting where (Z_1, Z_2) are not jointly observed with any of the target variables, only SPREE and DC can make use of them. A few general impressions can be noted.

- All the different SPREE estimates appear reasonable here; the best ones (that is, 0.0877 and 0.0876) yield an estimated cell count 153 after rounding, which is almost identical to the true observation 154. Adjusting Z_1 towards Y_1 gives better results than adjusting Z_2 towards Y_2. But at this stage of knowledge one is unable to *deduce* this from the higher association between Z_2 and Y_2 compared to that between Z_1 and Y_1.

Table 2. Education and election turnout data.

	Y_2				Z_2		
Y_1	No	Yes		Z_1	No	Yes	
Low	(154)	(885)	1039	Low	210	920	1130
High	(28)	(676)	704	High	44	569	613
	182	1561			254	1489	
	$Z_1 =$ Low Y_2				$Z_1 =$ High Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(149)	(854)	1003	Low	(5)	(31)	36
High	(9)	(118)	127	High	(19)	(558)	577
	158	972			24	589	
	$Z_2 =$ No Y_2				$Z_2 =$ Yes Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(140)	(61)	201	Low	(14)	(824)	838
High	(26)	(27)	53	High	(2)	(649)	651
	166	88			16	1473	
	$(Z_1, Z_2) =$ (Low, No) Y_2				$(Z_1, Z_2) =$ (Low, Yes) Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(136)	(59)	195	Low	(13)	(795)	808
High	(8)	(7)	15	High	(1)	(111)	112
	144	66			14	906	
	$(Z_1, Z_2) =$ (High, No) Y_2				$(Z_1, Z_2) =$ (High, Yes) Y_2		
Y_1	No	Yes		Y_1	No	Yes	
Low	(4)	(2)	6	Low	(1)	(29)	30
High	(18)	(20)	38	High	(1)	(538)	539
	22	22			2	567	

Table 3. Estimated lower and upper bounds for $\boldsymbol{P}[(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = (\boldsymbol{Low}, \boldsymbol{No})]$ and associated standard error (SE) using true data, CIA or MoB fusion distribution as basis of evaluation, estimated width of uncertainty space and true SE in parentheses, relative efficiency (RE) of proxy data.

Joint proxy variable	Bound (Lower, Upper)	1,000 × SE of Bound (Lower, Upper)			Width	RE
		True	CIA	MoB		
-	(0.000, 0.104)	(0.0, 7.3)	(0.0, 7.3)	(0.0, 7.3)	0.104 (0.0073)	1
Z_1	(0.018, 0.104)	(9.1, 7.2)	(8.9, 7.2)	(7.1, 7.2)	0.086 (0.0065)	0.83
Z_2	(0.065, 0.104)	(6.2, 4.9)	(5.7, 4.9)	(6.2, 4.9)	0.039 (0.0044)	0.38
(Z_1, Z_2)	(0.074, 0.095)	(4.6, 4.7)	(4.4, 4.7)	(4.6, 4.7)	0.021 (0.0034)	0.20

- The CIA results are worse than SPREE in every setting for this dataset. The advantage of SPREE is particularly useful in cases without any joint observations between the proxy and target variables, where it makes much better use of the auxiliary information.
- The MoB estimates are quite reasonable as long as Z_2 is available, and Z_1 appears to bring little improvement either on its own or in addition to Z_2. The effect of the proxy data is evident if 0.0846 given (Z_1, Z_2) is compared to 0.0521 in the absence of (Z_1, Z_2).
- The Euclidean distance is used to generate the DC. The result is worse than the SPREE, but better than CIA and MOB, which are unable to make use of the proxy variables in this setting.

Table 4. Illustrated (pseudo) estimates of $\boldsymbol{P}\big[(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = (\boldsymbol{Low}, \boldsymbol{No})\big]$.

Setting	Estimate	Identification assumptions
$\{(Z_1, Z_2, Y_1, Y_2)\}$	0.0884	True sample
$\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$	0.0856	CIA: $Y_1 \coprod Y_2 \mid (Z_1, Z_2)$
	0.0761	CIA: $Y_1 \coprod (Y_2, Z_2) \mid Z_1$ and $Y_2 \coprod (Y_1, Z_1) \mid Z_2$
	0.0846	MoB: $\mu_{ij} \mid (Z_1, Z_2)$
	0.0876	SPREE: $(Z_1, Y_2, Z_2) \rightarrow (\tilde{Y}_1, Y_2, Z_2) \mid (Y_1, Z_2) \& (Y_2, Z_2)$
	0.0863	SPREE: $(Z_1, Y_1, Z_2) \rightarrow (Z_1, Y_1, \tilde{Y}_2) \mid (Y_1, Z_1) \& (Y_2, Z_1)$
$\{(Z_1, Y_1), (Z_1, Y_2)\}$	0.0813	CIA: $Y_1 \coprod Y_2 \mid Z_1$
	0.0592	MoB: $\mu_{ij} \mid Z_1$
	0.0877	SPREE: $(Z_1, Y_2) \rightarrow (\tilde{Y}_1, Y_2) \mid Y_1 \& Y_2$
$\{(Z_2, Y_1), (Z_2, Y_2)\}$	0.0805	CIA: $Y_1 \coprod Y_2 \mid Z_2$
	0.0845	MoB: $\mu_{ij} \mid Z_2$
	0.0833	SPREE: $(Y_1, Z_2) \rightarrow (Y_1, \tilde{Y}_2) \mid Y_1 \& Y_2$
$\{Y_1, Y_2, (Z_1, Z_2)\}$	0.0622	IA: $Y_1 \coprod Y_2$
	0.0521	MoB: μ_{ij}
	0.0833	SPREE: $(Z_1, Z_2) \rightarrow (\tilde{Y}_1, \tilde{Y}_2) \mid Y_1 \& Y_2$
	0.0794	DC: $Z_1 \xrightarrow{\text{DC}} \tilde{Y}_1$ and $Z_2 \xrightarrow{\text{DC}} \tilde{Y}_2$

Finally, it may be reiterated that the choice of a particular fusion distribution is empirically unverifiable within the identification uncertainty bounds. Indeed, under each of the four settings considered in Table 4, the *same* uncertainty analysis, as given in Table 3 for the corresponding datasetting, should be reported for all the different pseudo estimates.

4.2. Labor Force Gross Flows

Labor force gross flows are of concern for both policy makers and researchers. Let the labor force status be classified as "employed (E)", "unemployed (U)" and "not in the labor force (N)" for each eligible person in some given age range. Let Y_1 be the status at time point t_1 and Y_2 that at t_2, then gross flow (i,j) refers here to the probability $\theta_{ij} = P[Y_1 = i, Y_2 = j]$. Together these form the 3×3 matrix, where the row margins $\phi_i = \sum_j \theta_{ij}$, for $i = 1, 2, 3$, form the marginal distribution of Y_1 and the column margins $\phi_j = \sum_i \theta_{ij}$, for $j = 1, 2, 3$, that of Y_2. Further classification by region, age, and so on may be of practical interest, but will not be considered here.

Countries that conduct the LFS typically apply some form of rotating panel design, so that joint observation (or panel data) of Y_1 and Y_2 are available for various combinations of t_1 and t_2. However, concerns for response burden and cost of following the same person over time will place a practical limit on the length of LFS participation, so that joint observations are not available if the difference between t_1 and t_2 is beyond that limit. For instance, in the Norwegian LFS (NLFS), each sample person participates in eight successive quarters, such that panel data are available for any two time points within a two-year span but not otherwise.

Two questions are considered below. Subsection 4.2.1 studies the efficiency of proxy data for labor force gross flows. To this end, proxy labor force status, denoted by Z_1 and Z_2 respectively, are compiled based on the various administrative data available to Statistics Norway (SN) and linked to the NLFS yearly panel between 2011 and 2012. The sources include employer/employee and self-employer registration, administration of job seekers, related health and welfare, payroll tax records, military services, and so on. Essentially the same proxy labour force status is used for the register-based census 2011. At the same time, it is acknowledged that at the individual level the proxy values will not always coincide with those that could be collected in the NLFS.

The second question to be considered is data fusion of (Y_1, Y_2), for which no joint observations are available. In particular, there is then an issue of how to make use of the data that are available for the time period between t_1 and t_2. For instance, although one does not have panel data between 2011 and 2013, one does have data between 2011 and 2012 and between 2012 and 2013, respectively. Various fusion methods can be used. For instance, under the CIA between (Y_1, Z_1) in 2011 and (Y_2, Z_2) in 2013 conditional on (Y_t, Z_t) in 2012, it becomes possible both to generate the fusion distribution of (Y_1, Y_2) and to assess the associated sampling uncertainty. However, this would not be appropriate if the identification uncertainty surrounding the CIA is ignored (Subsection 4.2.2).

4.2.1. Relative Efficiency of Proxy Labor Force Status

The data between 2011 and 2012 are given in Table 5. All joint observations of (Y_1, Y_2) are given in parentheses and assumed to be unobservable. The proxy register variables

(Z_1, Z_2) are jointly available with either of the target status, that is, the generic setting $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$.

The target NLFS sample gross flows of (Y_1, Y_2) and the proxy flows of (Z_1, Z_2) are given in Table 6, together with four fusion distributions by the CIA, MoB and two SPREE methods, respectively. Comparisons between the target and proxy joint distribution show that the register flow is higher for the stable employed persons (E, E), but lower for the stable unemployed persons (U, U) and 'inactive' ones (N, N). The largest relative deviations among the off-diagonal flows occur for (U, E) and (N, U). The causes for these differences are complex. For instance, persons who are on the way back into the labor force from N may be classified as U or E if interviewed in the NLFS, but they may well remain as N in the register sources until they first become E (possibly lagging behind the NLFS), which can be a cause for register underestimation of (N, U).

Focusing on the results of data fusion, it may be noted that all the techniques adjust the proxy flows (E, E) and (N, N) downwards. The adjustment of the proxy flow (U, U) differs across the method. In particular, the off-diagonal proxy flows are all adjusted upwards, and the flows (U, E) and (N, U) are no longer the ones that relatively deviate most from the target flows. Overall, the CIA results are worse than the others, especially for the diagonal flows, whereas the MoB results may seem slightly better than the two SPREE. Indeed, compared to the average of the two SPREE results, the MoB fusion distribution is closer to the target distribution for five out of nine flows.

Still, regardless of how plausible the fusion distributions may seem compared to the direct register-based proxy distribution, they can only be treated as potentially useful pseudo estimates. Proper inference is only facilitated by uncertainty analysis. Table 7 provides the estimated identification uncertainty bounds and the associated SE with and without the proxy variables as auxiliary data. The SEs are evaluated here on the basis of the true sample distribution, but any of the fusion distributions would have yielded virtually the same results. Again, the identification uncertainty matters little to the assessment of the sampling uncertainty.

It can be seen that the sampling uncertainty is relatively small compared to the identification uncertainty, especially in terms of the width of the identification uncertainty interval. The proxy variables are most effective for reducing the identification uncertainty of the 'corner' flows (E, E), (E, N), (N, E) and (N, N). As these four measure over 95% of the outcome space, the overall measure of the uncertainty space is greatly reduced in the presence of the proxy variables. Depending on the choice of w^{ij} in the calculation of $\hat{\Delta} = w^{ij}\hat{\Delta}_{ij}$ and $\hat{\tilde{\Delta}} = w^{ij}\hat{\tilde{\Delta}}_{ij}$, one obtains $\hat{\Delta} = 0.266, 0.263$ or 0.269 when w^{ij} is set to the true θ_{ij}, the CIA or MoB $\tilde{\theta}_{ij}$, and $\hat{\tilde{\Delta}} = 0.069, 0.069$ or 0.070 in correspondence. The overall relative efficiency of the proxy variables is 0.26 by all means.

4.2.2. Making Use of Available Data in Data Fusion

Where observations are unavailable for gross flows (Y_1, Y_2) over t_1 and t_2, various fusion distributions can be generated based on the intermediate observable target and proxy data. For instance, the gross flows between 2011 and 2013 can be derived from the observable flows between 2011 and 2012 and that between 2012 and 2013, under assumption that the labor force status in 2011 is independent of that in 2013 conditional on the status in 2012.

Table 5. Proxy and survey labor force status 2011 – 2012. Source: NLFS and registers at SN.

	$(Z_1, Z_2) = (E, E)$ Y_2					$(Z_1, Z_2) = (E, U)$ Y_2					$(Z_1, Z_2) = (E, N)$ Y_2			
Y_1	E	U	N		Y_1	E	U	N		Y_1	E	U	N	
E	(5289)	(15)	(124)	5428	E	(23)	(1)	(3)	27	E	(97)	(2)	(140)	239
U	(5)	(15)	(13)	33	U	(0)	(2)	(0)	2	U	(2)	(1)	(3)	6
N	(84)	(6)	(138)	228	N	(3)	(1)	(3)	7	N	(12)	(2)	(92)	106
	5378	36	275			26	4	6			111	5	235	
	$(Z_1, Z_2) = (U, E)$ Y_2					$(Z_1, Z_2) = (U, U)$ Y_2					$(Z_1, Z_2) = (U, N)$ Y_2			
Y_1	E	U	N		Y_1	E	U	N		Y_1	E	U	N	
E	(8)	(0)	(0)	8	E	(10)	(0)	(1)	11	E	(6)	(0)	(0)	6
U	(0)	(0)	(0)	0	U	(2)	(5)	(4)	11	U	(1)	(2)	(3)	6
N	(0)	(1)	(2)	3	N	(2)	(6)	(4)	12	N	(0)	(0)	(16)	16
	8	1	2			14	11	9			7	2	19	
	$(Z_1, Z_2) = (N, E)$ Y_2					$(Z_1, Z_2) = (N, U)$ Y_2					$(Z_1, Z_2) = (N, N)$ Y_2			
Y_1	E	U	N		Y_1	E	U	N		Y_1	E	U	N	
E	(146)	(4)	(10)	160	E	(23)	(1)	(1)	25	E	(140)	(8)	(66)	214
U	(3)	(3)	(3)	9	U	(1)	(0)	(1)	2	U	(2)	(1)	(3)	6
N	(67)	(3)	(44)	114	N	(5)	(9)	(26)	40	N	(73)	(38)	(1487)	1598
	216	10	57			28	11	28			215	47	1556	

However, analysis of the register-based status overtime suggests that such a CIA is unattainable. Moreover, even if the CIA had seemed reasonable for the proxy gross flows, it would only have yielded plausible pseudo estimates of the target gross flows, due to the fact that identification is not verifiable empirically but is only achieved on the strength of stipulation.

To illustrate data fusion under alternative settings in this situation, a synthetic dataset has been constructed as follows. Denote by $(Y_1, Z_1) = (i, h)$ the data in 2011 and by $(Y_t, Z_t) = (k, l)$ the data in 2012, with the joint sample distribution ϕ_{hikl}. Assume the CIA and the same conditional transition probabilities from 2012 to 2013 as from 2011 to 2012, that is, put $\lambda_{jg}^{lk} = \phi_{lkjg}/\phi_{lk}$ equal to the corresponding $\lambda_{kl}^{hi} = \phi_{hikl}/\phi_{hi}$, for $j, g = 0, 1$. The synthetic joint distribution over 2011, 2012, and 2013 is then given by $\phi_{hikljg} = \phi_{hikl}\phi_{lkjg}/\phi_{lk}$, from which the synthetic joint distribution of (Z_1, Y_1, Y_2, Z_2) can be obtained by integrating out (Y_t, Z_t), that is, $\phi_{hijg} = 1^{kl}\phi_{hikljg}$, and so on.

Consider three settings: (i) ignore (Y_t, Z_t) and assume the setting $\{(Z_1, Z_2, Y_1), (Z_1, Z_2, Y_2)\}$, that is, with joint auxiliary data (Z_1, Z_2), (ii) assume the setting $\{(Z_1, Z_2, Z_t, Y_t, Y_1), (Z_1, Z_2, Z_t, Y_t, Y_2)\}$, that is, with joint auxiliary data (Z_1, Z_2, Z_t, Y_t), and (iii) ignore (Z_1, Z_2, Z_t) and assume the setting $\{(Y_t, Y_1), (Y_t, Y_2)\}$, where Y_t may be considered a proxy for Y_1 as well as for Y_2.

The respective theoretical uncertainty bounds and width of the nine gross flows between Y_1 and Y_2 are given in Table 8. It is clear that using all the available joint auxiliary data, that is (Z_1, Z_2, Z_t, Y_t) here, provides the narrowest uncertainty bounds. There is more

Table 6. Target, proxy and fusion labor force gross flows by CIA, MoB and SPREE

	Target gross flows Y_2				Proxy gross flows Y_2		
Y_1	E	U	N	Y_1	E	U	N
E	0.6736	0.0057	0.0402	E	0.6846	0.0049	0.0410
U	0.0083	0.0030	0.0052	U	0.0030	0.0020	0.0037
N	0.0400	0.0065	0.2176	N	0.0480	0.0020	0.2107
	CIA: $Y_1 \coprod Y_2 \mid (Z_1, Z_2)$ Y_2				MoB: $\mu_{ij} \mid (Z_1, Z_2)$ Y_2		
Y_1	E	U	N	Y_1	E	U	N
E	0.6460	0.0059	0.0675	E	0.6628	0.0075	0.0501
U	0.0078	0.0013	0.0074	U	0.0078	0.0058	0.0076
N	0.0681	0.0080	0.1880	N	0.0510	0.0068	0.2058
	SPREE: $Z_1 \rightarrow \tilde{Y}_1 \mid (Y_1, Z_2) \& (Y_2, Z_2)$ Y_2				SPREE: $Z_2 \rightarrow \tilde{Y}_2 \mid (Y_1, Z_1) \& (Y_2, Z_1)$ Y_2		
Y_1	E	U	N	Y_1	E	U	N
E	0.6530	0.0053	0.0612	E	0.6567	0.0084	0.0543
U	0.0081	0.0022	0.0062	U	0.0077	0.0022	0.0065
N	0.0609	0.0077	0.1956	N	0.0575	0.0045	0.2022

Table 7. Estimated identification bounds and width (SE in parentheses; all numbers in 10^4)

Proxy variable	$(Y_1, Y_2) = (E, E)$ Lower	Upper	Width	$(Y_1, Y_2) = (E, U)$ Lower	Upper	Width	$(Y_1, Y_2) = (E, N)$ Lower	Upper	Width
-	4413	7194	2781	0	151	151	0	2630	2630
	(92)	(48)	(48)	(0)	(13)	(13)	(0)	(48)	(48)
(Z_1, Z_2)	6276	6980	703	12	138	126	155	857	702
	(38)	(28)	(26)	(5)	(12)	(12)	(18)	(27)	(25)
Proxy Variable	$(Y_1, Y_2) = (U, E)$ Lower	Upper	Width	$(Y_1, Y_2) = (U, U)$ Lower	Upper	Width	$(Y_1, Y_2) = (U, N)$ Lower	Upper	Width
-	0	165	165	0	151	151	0	165	165
	(0)	(14)	(14)	(0)	(12)	(12)	(0)	(14)	(14)
(Z_1, Z_2)	12	143	131	1	115	114	5	148	143
	(3)	(13)	(12)	(3)	(11)	(11)	(3)	(13)	(13)
Proxy variable	$(Y_1, Y_2) = (N, E)$ Lower	Upper	Width	$(Y_1, Y_2) = (N, U)$ Lower	Upper	Width	$(Y_1, Y_2) = (N, N)$ Lower	Upper	Width
-	0	2642	2642	0	151	151	0	2630	2630
	(0)	(48)	(48)	(0)	(13)	(13)	(0)	(47)	(47)
(Z_1, Z_2)	155	866	711	0	136	136	1695	2422	726
	(14)	(27)	(26)	(0)	(12)	(12)	(27)	(28)	(26)

Table 8. Theoretical identification bounds and width given auxiliary data (all numbers in 10^4)

Joint auxiliary data	$P(E,E)=6,411$ Lower	Upper	Width	$P(E,U)=71$ Lower	Upper	Width	$P(E,N)=685$ Lower	Upper	Width
(Z_1,Z_2)	6,038	6,738	702	19	149	130	357	1,065	708
(Z_1,Z_2,Z_t,Y_t)	6,287	6,718	431	26	137	111	376	811	435
Y_t	6,241	7,164	923	0	151	151	0	857	857
Joint auxiliary data	$P(U,E)=98$ Lower	Upper	Width	$P(U,U)=9$ Lower	Upper	Width	$P(U,N)=60$ Lower	Upper	Width
(Z_1,Z_2)	19	151	132	0	94	94	9	146	137
(Z_1,Z_2,Z_t,Y_t)	40	148	108	2	86	84	11	120	109
Y_t	0	167	167	0	138	138	0	167	167
Joint auxiliary data	$P(N,E)=710$ Lower	Upper	Width	$P(N,U)=71$ Lower	Upper	Width	$P(N,N)=1,885$ Lower	Upper	Width
(Z_1,Z_2)	381	1,086	705	0	131	131	1,505	2,216	711
(Z_1,Z_2,Z_t,Y_t)	400	835	435	2	119	117	1,754	2,199	445
Y_t	0	880	880	0	151	151	1,721	2,630	909

information about the target distribution of (Y_1, Y_2) in the register proxy (Z_1, Z_2) than in the survey proxy Y_t, as witnessed by the widths of the uncertainty bounds. In other words, there is more information in the concurrent proxy variables that are of a different definition than in the proxy variable that has the same definition but is from a different reference time point. Although the actual figures in Table 3 are obtained on a synthetic dataset, the basic results appear to reinforce the message that in data fusion one should strive to make use of all available auxiliary data.

5. Summary

The usefulness of proxy variables for categorical data fusion is considered above. A measure of the relative efficiency with and without proxy (or other auxiliary) variables is proposed. In practice, the uncertainty analysis must also take into account the sampling uncertainty in cases where the identification uncertainty bounds are unknown and need to be estimated. A flexible technique of distribution calibration is introduced for making use of proxy variables, which can be useful for constructing the fusion distribution as well as the fusion dataset. Empirical results demonstrate that proxy variables can play two beneficial roles at the same time: not only do they provide a general means for reducing the uncertainty associated with data fusion, they also widen the scope of plausible pseudo estimates of the target joint distribution.

6. References

Brozzi, A., A. Capotorti, and B. Vantaggi. 2012. "Incoherence Correction Strategies in Statistical Matching." *International Journal of Approximate Reasoning* 53: 1124–1136. Doi: http://dx.doi.org/10.1016/j.ijar.2012.06.009.

Conti, P.L., D. Marella, and M. Scanu. 2008. "Evaluation of Matching Noise for Imputation Techniques Based on Nonparametric Local Linear Regression Estimators." *Computational Statistics & Data Analysis* 53: 354–365. Doi: http://dx.doi.org/10.1016/j.csda.2008.07.041.

Conti, P.L., M. Di Zio, D. Marella, and M. Scanu. 2009. "Uncertainty Analysis in Statistical Matching." *Paper given at the First Italian Conference on Survey Methodology (ITACOSM09), June 10–12, 2009, Siena*

Conti, P.L., D. Marella, and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28: 69–88.

Conti, P.L., D. Marella, and M. Scanu. 2013. "Uncertainty Analysis for Statistical Matching of Ordered Categorical Variables." *Computational Statistics & Data Analysis* 68: 311–325. Doi: http://dx.doi.org/10.1016/j.csda.2013.07.004.

Cain, M. 1994. "The Moment-generating Function of the Minimum of Bivariate Normal Random Variables." *The American Statistician* 48: 124–125. Doi: http://dx.doi.org/10.1080/00031305.1994.10476039.

Chambers, R.L. and R.G. Steel. 2001. "Simple Methods for Ecological Inference in 2 x 2 Tables." *Journal of the Royal Statistical Society Series A* 164: 175–192. Doi: http://dx.doi.org/10.1111/1467-985X.00195.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006a. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints." *Journal of Official Statistics* 22: 137–157.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006b. *Statistical Matching: Theory and Practice*. Chichester: Wiley.

Kadane, J.B. 1978. "Some Statistical Problems in Merging Data Files." *In 1978 Compendium of Tax Research*, (pp. 159–171). Washington, D.C. Department of Treasury. (Reprinted in Journal of Official Statistics 17: 423–433.).

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

Koopmans, T. 1949. "Identification Problems in Economic Model Construction." *Econometrica* 17: 125–144. Doi: http://dx.doi.org/10.2307/1905689.

Lindley, D.V., A. Tversky, and R.V. Brown. 1979. "On the Reconciliation of Probability Assessments (incl. discussions)." *Journal of the Royal Statistical Society Series A* 142: 146–180. Doi: http://dx.doi.org/10.2307/2345078.

Manski, C.F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Marella, D., P.L. Conti, and M. Scanu. 2008. "On the Matching Noise of Some Nonparametric Imputation Procedures." *Statistics and Probability Letters* 78: 1593–1600. Doi: http://dx.doi.org/10.1016/j.spl.2008.01.020.

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

Nadarajah, S. and S. Kotz. 2008. "Exact Distribution of the Max/Min of Two Gaussian Random Variables." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 16: 210–212. Doi: http://dx.doi.org/10.1109/TVLSI.2007.912191.

Okner, B.A. 1972. "Constructing a New Microdata Base From Existing Microdatasets: the 1966 Merge File." *Annals of Economic and Social Measurement* 1: 325–342.

Patel, J.K., C.H. Kapadia, and D.B. Owen. 1976. *Handbook of Statistical Distributions*. New York: Marcel Dekker.

Plackett, R.L. 1977. "The Marginal Totals of a 2 x 2 Table." *Biometrika* 64: 37–42. Doi: http://dx.doi.org/10.1093/biomet/64.1.37.

Purcell, N.J. and L. Kish. 1980. "Postcensal Estimates for Local Areas (or Domains)." *International Statistical Review* 48: 3–18. Doi: http://dx.doi.org/10.2307/1402400.

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

Rässler, S. and H. Kiesl. 2009. "How Useful Are Uncertainty Bounds? Some Recent Theory With an Application to Rubin's Causal Model." *In Proceedings of the 57th Sessions of the International Statistical Institute.* (2009) CD-ROM. Durban, South Africa.

Singh, A.C., H. Mantel, M. Kinack, and G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption." *Survey Methodology* 19: 57–79.

Vantaggi, B. 2008. "Statistical Matching of Multiple Sources: A Look Through Coherence." *International Journal of Approximate Reasoning* 49: 701–711. Doi: http://dx.doi.org/10.1016/j.ijar.2008.07.005.

Wakefield, J. 2004. "Ecological Inference for 2 x 2 Tables (incl. discussions)." *Journal of the Royal Statistical Society Series A* 167: 385–445. Doi: http://dx.doi.org/10.1111/j.1467-985x.2004.02046.x.

Received July 2013
Revised August 2015
Accepted September 2015