Categorical proxy data fusion

- Alternatives
- Uncertainty analysis

Li-Chun Zhang

Statistics Norway

lcz@ssb.no

(University of Southampton from 1.9.2012)

An example: Education and election turnout data

	Turnout		
Education	No	Yes	Total
Low	(0, 0.104 = 182/1743)	(885)	1039
High	(28)	(676)	704
Total	182	1551	1743

	Y2		
Y1	0	1	Proportion
0	L = max(0, p+q-1) U = min(p, q)		p
1			1-p
Proportion	q	1-q	

Proxy data fusion (PDF) given { (Z1, Y1), (Z1, Y2) }

Proxy conditinoal independence assumption (CIA): $\lambda_{ij}^{h} = \lambda_{i}^{h}\lambda_{j}^{h}$

Target joint distribution: θ_{ij}

Surrogate joint distribution by proxy CIA-based PDF: $\tilde{\theta}_{\mathrm{ij}} = \lambda_{i}^{h}\phi_{h}\lambda_{j}^{h}$

$$\begin{pmatrix} \phi_{11} & \cdots & \phi_{1J} \\ \vdots & \ddots & \vdots \\ \phi_{H1} & \cdots & \phi_{HJ} \end{pmatrix} \leftarrow \begin{pmatrix} \phi_{1}(Z_1) \\ \vdots \\ \phi_{\mathrm{H}}(Z_1) \end{pmatrix} \Rightarrow \text{IPF/Raking} \Rightarrow \begin{pmatrix} \tilde{\theta}_{11} & \cdots & \tilde{\theta}_{1J} \\ \vdots & \ddots & \vdots \\ \tilde{\theta}_{H1} & \cdots & \tilde{\theta}_{HJ} \end{pmatrix} \leftarrow \begin{pmatrix} \theta_{1}(Y_1) \\ \vdots \\ \theta_{\mathrm{H}}(Y_1) \end{pmatrix}$$

$$\uparrow \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \uparrow$$

$$\begin{pmatrix} \theta_{1}(Y_2) & \cdots & \theta_{\mathrm{J}}(Y_2) \end{pmatrix} \qquad\qquad\qquad\qquad \begin{pmatrix} \theta_{1}(Y_2) & \cdots & \theta_{\mathrm{J}}(Y_2) \end{pmatrix}$$

Structure preserving estimation (SPREE): $\alpha_{\mathrm{ij}}^{Y_1Y_2} = \alpha_{\mathrm{hj}}^{Z_1Y_2}$

Minimum data requirement: $\{ Y_1, (Z_1, Y_2) \}$

Distribution calibration (DC) and Proxy DC

$$\mathbf{F} = \begin{pmatrix} f_{11} & \cdots & f_{1H} \\ \vdots & \ddots & \vdots \\ f_{H1} & \cdots & f_{HH} \end{pmatrix} \leftarrow \begin{pmatrix} \phi_1(Z) \\ \vdots \\ \phi_H(Z) \end{pmatrix}$$

$$\uparrow$$

$$(\theta_1(Y) \quad \cdots \quad \theta_H(Y))$$

Example: $\mathbf{Z} = (31, 28, 41)^T$ and $\mathbf{Y} = (30, 30, 40)^T$

$$\mathbf{f}_0 = \begin{pmatrix} 31 & 0 & 0 \\ 0 & 28 & 0 \\ 0 & 0 & 41 \end{pmatrix} \Rightarrow \tilde{\mathbf{F}} = \begin{pmatrix} 30 & 1 & 0 \\ 0 & 28 & 0 \\ 0 & 1 & 40 \end{pmatrix}$$

$$\mathbf{f} = \begin{pmatrix} f_{11} \\ \vdots \\ f_{1H} \\ \vdots \\ f_{H1} \\ \vdots \\ f_{HH} \end{pmatrix} \quad \& \quad \mathbf{b} = \begin{pmatrix} \phi_1(Z) \\ \vdots \\ \phi_H(Z) \\ \theta_1(Y) \\ \vdots \\ \theta_H(Y) \end{pmatrix} \quad \Rightarrow \quad \mathbf{Af} = \mathbf{b}$$

$$\text{Lagrangian} \quad L = \Delta + \lambda^T(\mathbf{Af} - \mathbf{b})$$

$$\Delta = \tfrac{1}{2}(\mathbf{f} - \mathbf{f}_0)^T(\mathbf{f} - \mathbf{f}_0)$$

$$\mathbf{f}_0 = \mathbf{f}(\text{Diag}\{\phi_h(Z); h=1,...,H\})$$

$$\Rightarrow \quad \tilde{\mathbf{f}} = \mathbf{f}_0 + \mathbf{A}^T(\mathbf{AA}^T)^{-1}(\mathbf{b} - \mathbf{Af}_0)$$

Surrogate joint distribution by Proxy DC: $\tilde{\theta}_{ij} = \xi_i^h \phi_h \lambda_j^h$ for ξ_i^h by DC $\mathbf{Z}_1 \rightarrow Y_1$

Minimum data requirement: $\{ Y_1, (Z_1, Y_2) \}$

Table 1: Proxy data fusion for target (Y_1, Y_2) given proxy (Z_1, Z_2).

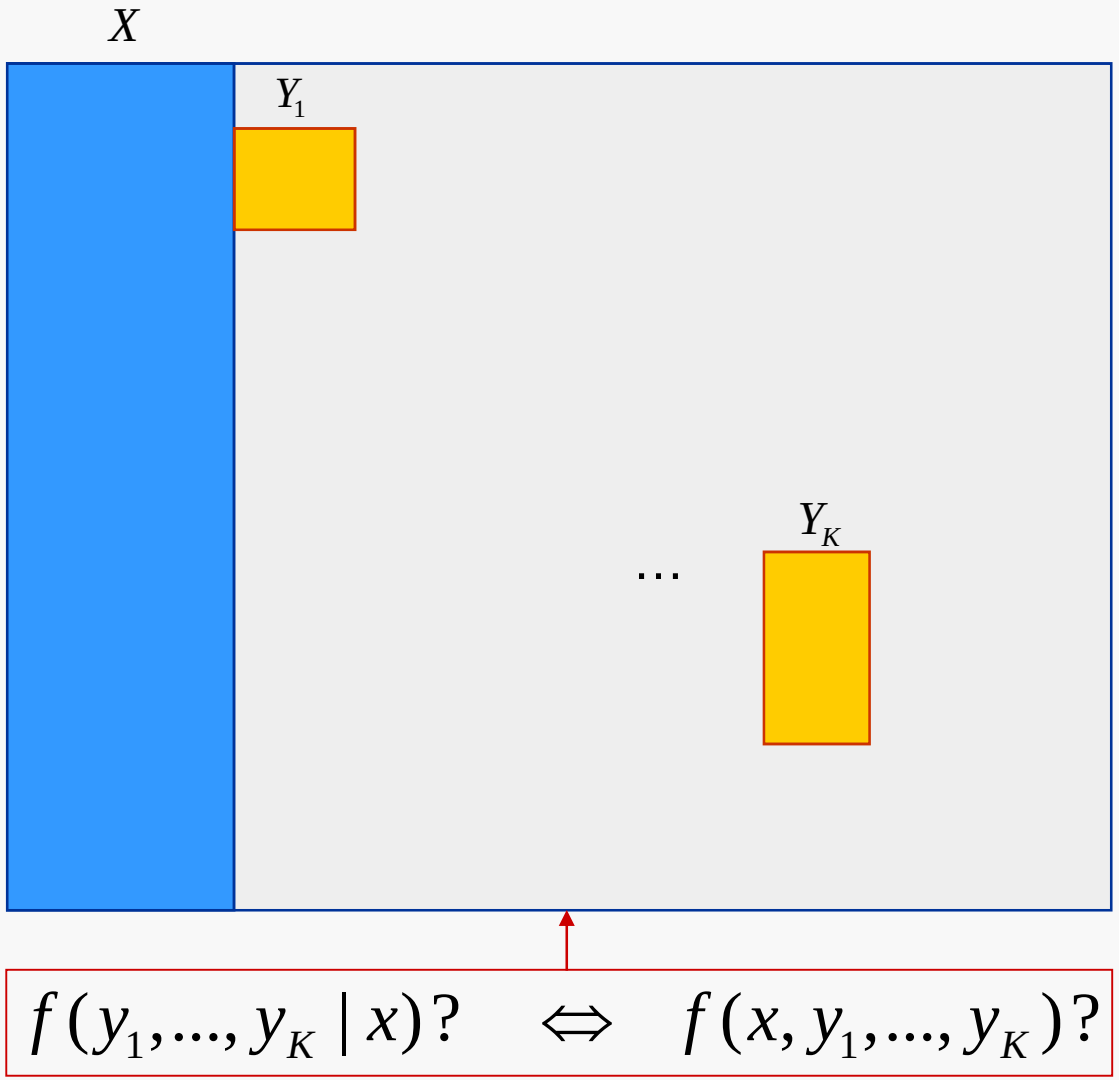
Minimum Data	Assumptions for PDF
$\{(Z_1, Y_1, Z_2),$	CIA: $\lambda_{ij}^{hk} = \lambda_i^{hk}\lambda_j^{hk}$
$(Z_1, Y_2, Z_2)\}$	CIA: $\lambda_{ij}^{hk} = \lambda_i^{h}\lambda_j^{k}$
	SPREE: $\alpha_{(ijk)}^{Y_1Y_2} = \alpha_{(hjk)}^{Z_1Y_2}$, $\alpha_{(ijk)}^{Y_1Y_2Z_2} = \alpha_{(hjk)}^{Z_1Y_2Z_2}$
	SPREE: $\alpha_{(hij)}^{Y_1Y_2} = \alpha_{(hik)}^{Y_1Z_2}$, $\alpha_{(hij)}^{Z_1Y_1Y_2} = \alpha_{(hik)}^{Z_1Y_1Z_2}$
$\{Y_1, (Z_1, Y_2, Z_2)\}$	SPREE: $\alpha_{(ijk)}^{Y_1Z_2} = \alpha_{(hjk)}^{Z_1Z_2}$, $\alpha_{(ijk)}^{Y_1Y_2} = \alpha_{(hjk)}^{Z_1Y_2}$, $\alpha_{(ijk)}^{Y_1Y_2Z_2} = \alpha_{(hjk)}^{Z_1Y_2Z_2}$
	SPREE: $\alpha_{(ijk)}^{Y_2Z_2} = \alpha_{(hjk)}^{Y_2Z_2}$, $\alpha_{(ijk)}^{Y_1Z_2} = \alpha_{(hjk)}^{Z_1Z_2}$, $\alpha_{(ijk)}^{Y_1Y_2} = \alpha_{(hjk)}^{Z_1Y_2}$, $\alpha_{(ijk)}^{Y_1Y_2Z_2} = \alpha_{(hjk)}^{Z_1Y_2Z_2}$
	CIA: $\lambda_{ik}^{hj} = \lambda_i^{hj}\lambda_k^{hj}$ and SPREE: $\alpha_{(ij)}^{Y_1Y_2} = \alpha_{(hj)}^{Z_1Y_2}$
$\{(Z_1, Y_1, Z_2), Y_2\}$	SPREE: $\alpha_{(hij)}^{Z_1Y_2} = \alpha_{(hik)}^{Z_1Z_2}$, $\alpha_{(hij)}^{Y_1Y_2} = \alpha_{(hik)}^{Y_1Z_2}$, $\alpha_{(hij)}^{Z_1Y_1Y_2} = \alpha_{(hik)}^{Z_1Y_1Z_2}$
	SPREE: $\alpha_{(hij)}^{Z_1Y_1} = \alpha_{(hik)}^{Z_1Y_1}$, $\alpha_{(hij)}^{Z_1Y_2} = \alpha_{(hik)}^{Z_1Z_2}$, $\alpha_{(hij)}^{Y_1Y_2} = \alpha_{(hik)}^{Y_1Z_2}$, $\alpha_{(hij)}^{Z_1Y_1Y_2} = \alpha_{(hik)}^{Z_1Y_1Z_2}$
	CIA: $\lambda_{hj}^{ik} = \lambda_h^{ik}\lambda_j^{ik}$
$\{Y_1, (Z_1, Z_2), Y_2\}$	SPREE: $\alpha_{(ij)}^{Y_1Y_2} = \alpha_{(hk)}^{Z_1Z_2}$
	Separate Proxy DC: from $\{\phi_h\}$ to $\{\phi_i\}$ and from $\{\phi_k\}$ to $\{\phi_j\}$

Integration of statistical data by Proxy Data Fusion

SM

X:
- Census/register
- Common, known

Y:
- Sample
- Selective/partial, target, expansive



Integration of statistical data by Proxy Data Fusion

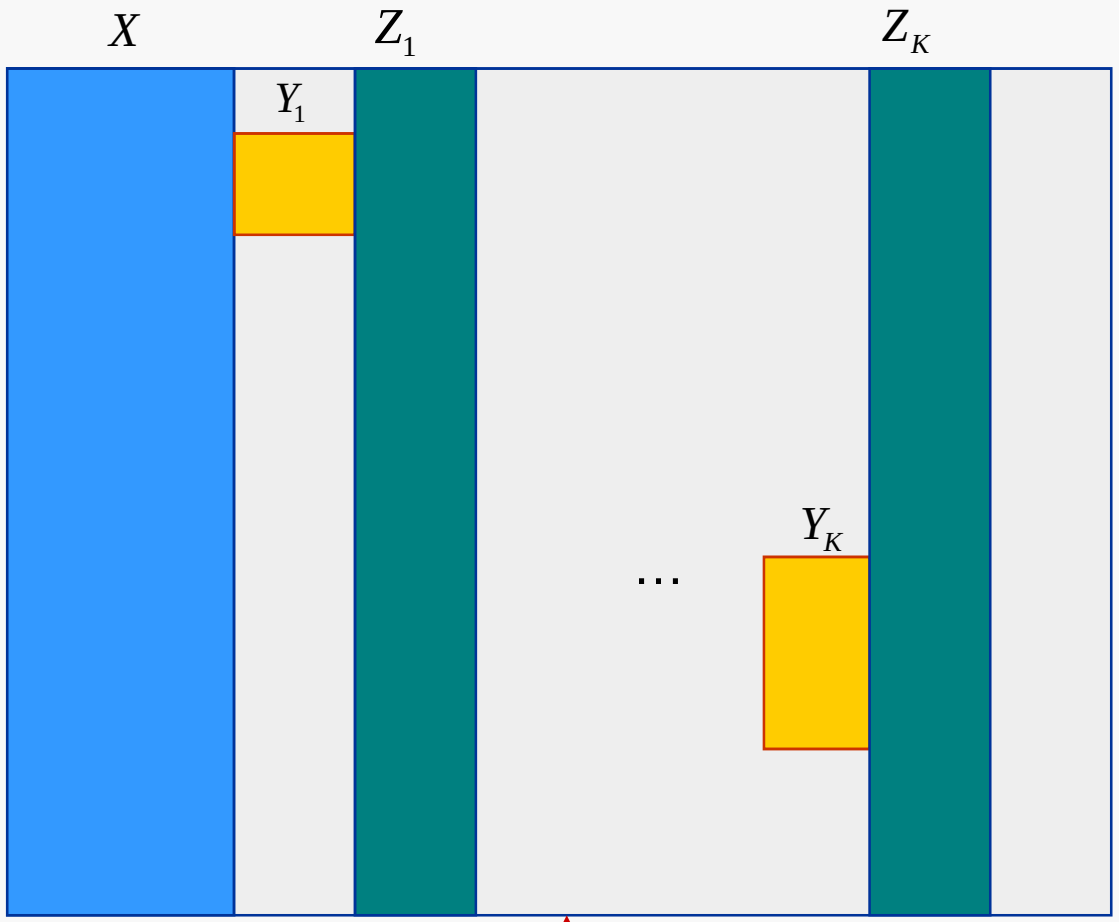
SM

X:
- Census/register
- Common, known

Y:
- Sample
- Selective/partial, target, expansive

Z:
- Census/register
- Common, proxy, cheap

PDF



$$f(x, z_k) \neq f(x, y_k), \quad \text{for } k = 1, 2, \ldots K$$

Integration of statistical data by Proxy Data Fusion

SM

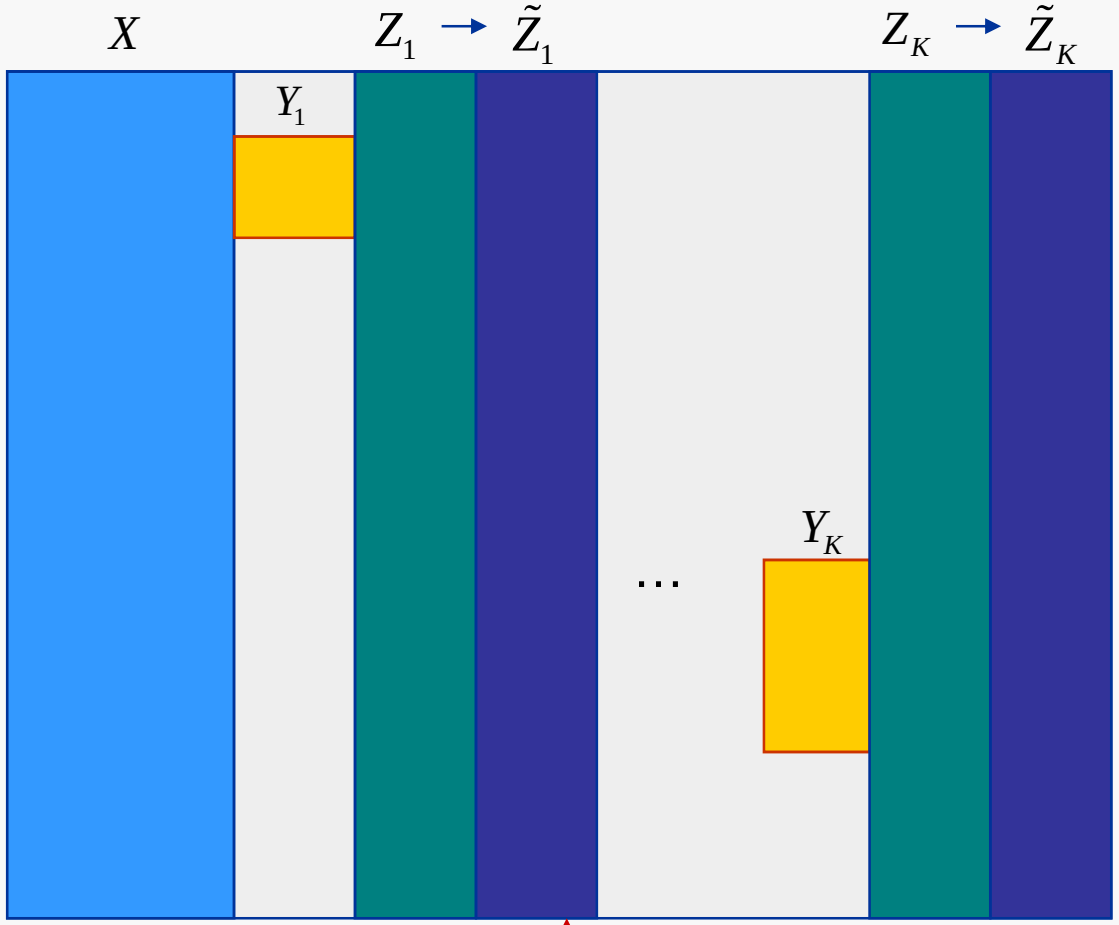
X:
- Census/register
- Common, known

Y:
- Sample
- Selective/partial, target, expansive

Z:
- Census/register
- Common, surrogate, cheap

Z□:
- Micro calibration
- Valid (Zhang, 2012), complete

PDF



$$f(x, \tilde{z}_k) = f(x, y_k), \quad \text{for } k = 1, 2, \ldots K$$

Uncertainty analysis:
Relative efficiency of proxy data

$$\Delta = w^{ij}\Delta_{ij} \qquad \text{where} \quad \Delta_{ij} = U_{ij} - L_{ij} \tag{7}$$

and w^{ij} is a (non-negative) weight function that sums to unity, i.e. $w^{ij}\mathbf{1}_{ij} = 1$, such as $w^{ij} = \phi_i\phi_j$.

$$\Delta_{ij}^{hk} = U_{ij}^{hk} - L_{ij}^{hk}$$

where

$$\max(0, \lambda_i^{hk} + \lambda_j^{hk} - 1) = L_{ij}^{hk} \leq \lambda_{ij}^{hk} = P_{ij|hk}^{Y_1Y_2|Z_1Z_2} \leq U_{ij}^{hk} = \min(\lambda_i^{hk}, \lambda_j^{hk})$$

Let $E_{Z_1Z_2}$ denote expectation over the distribution of (Z_1, Z_2). We have

$$\bar{L}_{ij} \overset{\text{def}}{=} L_{ij}^{hk}\phi_{hk} = E_{Z_1Z_2}(L_{ij}^{hk}) \leq \theta_{ij} = E_{Z_1Z_2}(\lambda_{ij}^{hk}) \leq E_{Z_1Z_2}(U_{ij}^{hk}) = U_{ij}^{hk}\phi_{hk} \overset{\text{def}}{=} \bar{U}_{ij}$$

Meanwhile, $\phi_i = E_{Z_1Z_2}(\lambda_i^{hk})$ and $\phi_j = E_{Z_1Z_2}(\lambda_j^{hk})$. It follows from the Jensen's inequality that

$$L_{ij} \leq \bar{L}_{ij} \qquad \text{and} \qquad U_{ij} \geq \bar{U}_{ij} \qquad \Longrightarrow \qquad \bar{\Delta}_{ij} = \bar{U}_{ij} - \bar{L}_{ij} \leq \Delta_{ij}$$

The uncertainty at θ_{ij} can thus be reduced in the presence of proxy data. So is the overall uncertainty, i.e. $\bar{\Delta} = w^{ij}\bar{\Delta}_{ij} \leq w^{ij}\Delta_{ij} = \Delta$. A general measure of the *relative efficiency* of the extra information due to the proxy data may be given as

$$\gamma(Z_1, Z_2; w) = \bar{\Delta}(w)/\Delta(w) = (w^{ij}\bar{\Delta}_{ij})/(w^{ij}\Delta_{ij}) \qquad \text{where} \quad w^{ij} = \phi_i\phi_j \tag{8}$$

Uncertainty analysis:
Uncertainty upper bound of (proxy) data fusion

Take the setting $\{(Z_1, Y_1, Z_2), (Z_1, Y_2, Z_2)\}$. PDF necessarily stipulates a particular value $\tilde{\lambda}_{ij}^{hk}$ between L_{ij}^{hk} and U_{ij}^{hk}, such as $\tilde{\lambda}_{ij}^{hk} = \lambda_i^{hk}\lambda_j^{hk}$ under the proxy CIA (3). Its *maximum* difference to the true but unidentifiable λ_{ij}^{hk} is given by

$$\Lambda_{ij}^{hk} = \max(\tilde{\lambda}_{ij}^{hk} - L_{ij}^{hk},\ U_{ij}^{hk} - \tilde{\lambda}_{ij}^{hk}) = \Delta_{ij}^{hk}/2 + \varepsilon_{ij}^{hk} \geq \Delta_{ij}^{hk}/2$$

where $\varepsilon_{ij}^{hk} = |\tilde{\lambda}_{ij}^{hk} - \mu_{ij}^{hk}|$ and $\mu_{ij}^{hk} = (L_{ij}^{hk} + U_{ij}^{hk})/2$. This gives us a point-wise uncertainty upper bound for $\tilde{\lambda}_{ij}^{hk}$. It has a minimum value $\Delta_{ij}^{hk}/2$, regardless of the choice of PDF. Next, an uncertainty bound of the corresponding $\theta_{ij}(\tilde{\lambda}) = \tilde{\lambda}_{ij}^{hk}\phi_{hk}$ can be given as

$$\bar{\Lambda}_{ij} = \Lambda_{ij}^{hk}\phi_{hk} = \bar{\Delta}_{ij}/2 + \bar{\varepsilon}_{ij} \geq \bar{\Delta}_{ij}/2$$

for $\bar{\varepsilon}_{ij} = \varepsilon_{ij}^{hk}\phi_{hk}$, and an overall uncertainty bound of PDF as

$$\bar{\Lambda} = w^{ij}\bar{\Lambda}_{ij} = \bar{\Delta}/2 + \bar{\varepsilon} \geq \bar{\Delta}/2 \qquad \text{where} \quad w^{ij} = \phi_i\phi_j \quad \text{and} \quad \bar{\varepsilon} = w^{ij}\bar{\varepsilon}_{ij}$$

It is now possible to define the *minimum uncertainty bound (MUB)* PDF as $\theta_{ij}(\eta)$, where

$$\{\eta_{ij}^{hk}\} = \arg\min_{\{\tilde{\lambda}_{ij}^{hk}\}\in\Omega} \bar{\varepsilon}(\tilde{\lambda}; w) \quad \text{and} \quad \bar{\varepsilon}(\tilde{\lambda}; w) = w^{ij}\varepsilon_{ij}^{hk}(\tilde{\lambda})\phi_{hk} = w^{ij}|\tilde{\lambda}_{ij}^{hk} - \mu_{ij}^{hk}|\phi_{hk}$$

The binary case

Table 2: Uncertainty analysis of binary data fusion. Notations: (L, U) for lower and upper bounds of cell probability, $\tilde{\theta} = \theta(\xi)$ for stipulated cell probability and Λ for its uncertainty upper bound. Conditions: $\theta_i \leq \theta_j$ and $\theta_i + \theta_j \leq 1$.

Cell	L	U	$U - L$	$\tilde{\theta}$	$\tilde{\theta} - L$	$U - \tilde{\theta}$	Λ
(i, j)	0	θ_i	θ_i	ξ	ξ	$\theta_i - \xi$	$\theta_i/2 + \lvert\xi - \theta_i/2\rvert$
(i, j^c)	0	θ_i	θ_i	$\theta_i - \xi$	$\theta_i - \xi$	ξ	$\theta_i/2 + \lvert\xi - \theta_i/2\rvert$
(i^c, j)	$\theta_j - \theta_i$	θ_j	θ_i	$\theta_j - \xi$	$\theta_i - \xi$	ξ	$\theta_i/2 + \lvert\xi - \theta_i/2\rvert$
(i^c, j^c)	$1 - (\theta_i + \theta_j)$	$1 - \theta_j$	θ_i	$\xi + 1 - (\theta_i + \theta_j)$	ξ	$\theta_i - \xi$	$\theta_i/2 + \lvert\xi - \theta_i/2\rvert$

$$\Delta = \min(\theta_i, \theta_{i^c}, \theta_j, \theta_{j^c}) = \theta_i \wedge \theta_{i^c} \wedge \theta_j \wedge \theta_{j^c} \tag{9}$$

The same simplification of analysis applies to the setting $\{(Z_1, Y_1, Z_2), (Z_1, Y_2, Z_2)\}$. We have

$$\Delta^{hk} \equiv \Delta^{hk}_{ij} = U^{hk}_{ij} - L^{hk}_{ij} = \lambda^{hk}_i \wedge \lambda^{hk}_{i^c} \wedge \lambda^{hk}_j \wedge \lambda^{hk}_{j^c} \qquad \text{and} \qquad \bar{\Delta} = \Delta^{hk}\phi_{hk} \tag{11}$$

where Δ^{hk} is the same for all (i, j), and $\bar{\Delta}$ does not depend on the weight function w^{ij}.

Conditions for absolute reduction of uncertainty space ⇔ Relative efficiency < 1

Table 3: Uncertainty analysis of binary PDF given $\{(Z_1, Y_1), (Z_1, Y_2)\}$.

$P(Y_1\|Z_1)$	Z_1	
Y_1	Low	High
Low	α	$1-\alpha$
High	$1-\alpha$	α

$P(Y_2\|Z_1)$	Z_1	
Y_2	Low	High
Yes	β_L	β_H
No	$1-\beta_L$	$1-\beta_H$

Conditions	Δ^L	Δ^H	RE (γ)
$1-\alpha < \alpha < \beta_L < \beta_H$	$1-\beta_L$	$1-\beta_H$	1
$1-\alpha < \beta_L < \alpha < \beta_H$	$1-\alpha$	$1-\beta_H$	< 1
$\beta_L < 1-\alpha < \alpha < \beta_H$	β_L	$1-\beta_H$	< 1
$\beta_L < 1-\alpha < \beta_H < \alpha$	β_L	$1-\alpha$	< 1
$\beta_L < \beta_H < 1-\alpha < \alpha$	β_L	β_H	1

Two conditions are sufficient and necessary for $\gamma < 1$, or absolute reduction of uncertainty:

$$\beta_L < \alpha \quad \text{and} \quad 1-\alpha < \beta_H \tag{12}$$

Table 4: 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	1551			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	

$\Delta = \frac{182}{1743} \propto 182$	RE = 1
$\bar{\Delta} \propto (127 + 24) = 151$	RE = 0.83
$\bar{\Delta} \propto (53 + 16) = 69$	RE = 0.38
$\bar{\Delta} \propto (15 + 14 + 6 + 2) = 37$	RE = 0.20

Table 5: Estimated lower and upper bounds for education and election turnout.

	Without (Z_1, Z_2)	With Z_1	With Z_2	With (Z_1, Z_2)
$P[(Y_1, Y_2) = (Low, No)]$	(0, 0.104)	(0.018, 0.104)	(0.065, 0.105)	(0.074, 0.095)

Estimates: True 0.0884 Proxy-CIA 0.0852 MUB 0.0845