Current work on automatic multisource editing at Statistics Netherlands

Sander Scholtus, Arnout van Delden, Rob Willems, Frank Aelen

UNECE Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna

Overview

- New integrated uniform production system for business statistics
- Automatic editing: Review
- Automatic multisource editing
- Incorporating subject-matter knowledge
- Evaluating quality of multisource editing

New integrated uniform production system

Presented at last expert meeting (Vaasen-Otten et al., 2022)

Two particular aims:

- Reduce manual editing where possible
- Improve consistency between statistics

Approaches:

- Top-down editing to prioritize manual work
- Automatic editing

Automatic editing: Review

Observed variables for unit i: $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{iJ}\right)'$

Edit rules:

$$\text{IF } (\mathbf{a}_1'\mathbf{x}_i \leq b_1 \text{ AND } \mathbf{a}_2'\mathbf{x}_i \leq b_2 \cdots \text{AND} \cdots \mathbf{a}_{K-1}'\mathbf{x}_i \leq b_{K-1}) \text{ THEN } (\mathbf{a}_K'\mathbf{x}_i \leq b_K)$$

(Instead of $\leq$, also allowed: $\geq$, $<$, $>$ or $=$.)

Examples:

$$x_{i1} \geq 0$$

$$x_{i1} + x_{i2} + x_{i3} = x_{i4}$$

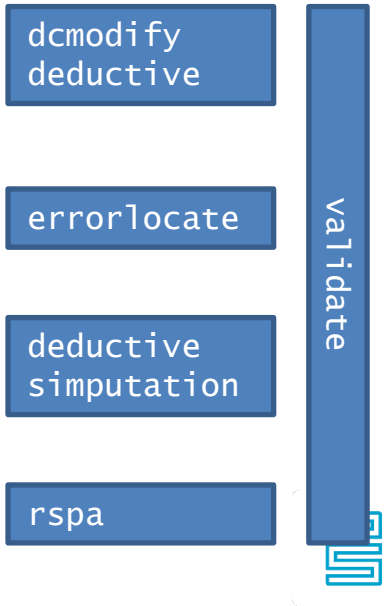
$$\text{IF } (x_{i1} > 0) \text{ THEN } (x_{i2} + x_{i3} > 0)$$

Automatic editing: Review

Typical automatic editing process:

1. Deductive correction (systematic errors)
   (IF-THEN rules, simple algorithms)
2. Error localization (other errors)
   (Fellegi-Holt paradigm)
3. Imputation of missing values

4. Adjustment of imputed values to satisfy all edit rules



Automatic editing: Review

Paradigm of Fellegi & Holt (1976):

The data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields). This we believe to be in agreement with the idea of keeping the maximum amount of original data unchanged, subject to the constraints of the edits, and so manufacturing as little data as possible. At the same time, if errors are comparatively rare, it seems more likely that we will identify the truly erroneous fields.

Generalized paradigm:

"fewest possible items" → "minimal sum of reliability weights"

Automatic editing: Review

Generalizations of Fellegi-Holt paradigm:

- including soft edit rules (Scholtus, 2015)
- including general edit operations
  (Scholtus, 2016; Daalmans & Scholtus, 2018)

Other approaches:

- model-based approaches
  (Little & Smith, 1987; Ghosh-Dastidar & Schafer, 2006; Kim et al., 2015)
- Nearest-neighbour Imputation Methodology (Bankier, 2006)

Automatic multisource editing

Source 1

observed variables

$\mathbf{x}_i^{(1)} = \left(x_{i1}^{(1)}, \ldots, x_{iJ_1}^{(1)}\right)'$

internal edit rules

common variables

$\mathbf{y}_i^{(1)}$

derivations / edit rules

$y_{il}^{(1)} = \left(\mathbf{c}_l^{(1)}\right)' \mathbf{x}_i^{(1)} + d_l^{(1)}$

Source 2

observed variables

$\mathbf{x}_i^{(2)} = \left(x_{i1}^{(2)}, \ldots, x_{iJ_2}^{(2)}\right)'$

internal edit rules

common variables

$\mathbf{y}_i^{(2)}$

derivations / edit rules

$y_{il}^{(2)} = \left(\mathbf{c}_l^{(2)}\right)' \mathbf{x}_i^{(2)} + d_l^{(2)}$

...

Source P

observed variables

$\mathbf{x}_i^{(P)} = \left(x_{i1}^{(P)}, \ldots, x_{iJ_P}^{(P)}\right)'$

internal edit rules

common variables

$\mathbf{y}_i^{(P)}$

derivations / edit rules

$y_{il}^{(P)} = \left(\mathbf{c}_l^{(P)}\right)' \mathbf{x}_i^{(P)} + d_l^{(P)}$

...

'true' values

$\mathbf{z}_i = (z_{i1}, \ldots, z_{iL})'$

edit rules

$\left|y_{il}^{(p)} - z_{il}\right| \leq \epsilon_l^* |z_{il}|$

other edit rules on $\mathbf{y}_i, \mathbf{z}_i$

Task: adjust the data $\left(\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(P)}, \mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ as necessary so that all edit rules are satisfied

Automatic multisource editing

Proposal: Use a three-step approach

(Scholtus et al., 2022)

1. Automatic editing of common variables across data sources
   - Identify errors in $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ using edit rules across data sources
2. Impute 'true' values and derive additional edit rules
   - Impute 'true' values $\mathbf{z}_i$ consistently with edit rules
   - Derive additional edit rules for $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}\right)$ from $\left|y_{il}^{(p)} - z_{il}\right| \leq \epsilon_l^* |z_{il}|$
3. Automatic editing within each individual data source
   - Identify errors and impute values in $\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)$ so all edit rules are satisfied

Example

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	.	.
2	.	.

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	0	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	800	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables:

$0.95 \times \text{COSTS\_GOODS} \leq \text{COSTS\_GOODS}^{(p)} \leq 1.05 \times \text{COSTS\_GOODS}$

$0.95 \times \text{COSTS\_OTHER} \leq \text{COSTS\_OTHER}^{(p)} \leq 1.05 \times \text{COSTS\_OTHER}$

Example: Step 1

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	.	.
2	.	.

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID			data source 1		
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	0	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID			data source 2		
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	800	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables:

$0.95 \times \text{COSTS\_GOODS} \leq \text{COSTS\_GOODS}^{(p)} \leq 1.05 \times \text{COSTS\_GOODS}$

$0.95 \times \text{COSTS\_OTHER} \leq \text{COSTS\_OTHER}^{(p)} \leq 1.05 \times \text{COSTS\_OTHER}$

Example: Step 1

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	.	.
2	.	.

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID			data source 1		
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	.	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID			data source 2		
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables:

$0.95 \times \text{COSTS\_GOODS} \leq \text{COSTS\_GOODS}^{(p)} \leq 1.05 \times \text{COSTS\_GOODS}$

$0.95 \times \text{COSTS\_OTHER} \leq \text{COSTS\_OTHER}^{(p)} \leq 1.05 \times \text{COSTS\_OTHER}$

Example: Step 2

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	.	.
2	.	.

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID			data source 1		
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	.	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID			data source 2		
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables:

$0.95 \times \text{COSTS\_GOODS} \leq \text{COSTS\_GOODS}^{(p)} \leq 1.05 \times \text{COSTS\_GOODS}$

$0.95 \times \text{COSTS\_OTHER} \leq \text{COSTS\_OTHER}^{(p)} \leq 1.05 \times \text{COSTS\_OTHER}$

Example: Step 2

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID			data source 1		
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	.	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID			data source 2		
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables:

$0.95 \times \text{COSTS\_GOODS} \leq \text{COSTS\_GOODS}^{(p)} \leq 1.05 \times \text{COSTS\_GOODS}$

$0.95 \times \text{COSTS\_OTHER} \leq \text{COSTS\_OTHER}^{(p)} \leq 1.05 \times \text{COSTS\_OTHER}$

Example: Step 2

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID			data source 1		
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	.	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID			data source 2		
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$

$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$

Edit rules common variables (ID 2):

$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$

$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$

Example: Step 3

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	.	100	0	0
2	60	80	60	30	50

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$$

$$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$$

Edit rules common variables (ID 2):

$$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$$

$$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$$

Example: Step 3

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	.	100	.	0
2	60	80	60	.	.

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$$

$$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$$

Edit rules common variables (ID 2):

$$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$$

$$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$$

Example: Step 3

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	**160**	100	**160**	0
2	60	80	60	**50**	**30**

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$$

$$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$$

Edit rules common variables (ID 2):

$$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$$

$$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$$

Example: Step 3

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	160	100	160	0
2	60	80	60	50	30

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	800	160	960
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$$

$$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$$

Edit rules common variables (ID 2):

$$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$$

$$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$$

Example: Step 3

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	160	100	160	0
2	60	80	60	50	30

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	.	160	.	160	.
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$$

$$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$$

Edit rules common variables (ID 2):

$$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$$

$$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$$

Example: Step 3

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	160	100	160	0
2	60	80	60	50	30

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	100	160	100	160	260
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables (ID 1):

$$95 \leq \text{COSTS\_GOODS}^{(p)} \leq 105$$

$$152 \leq \text{COSTS\_OTHER}^{(p)} \leq 168$$

Edit rules common variables (ID 2):

$$59.40 \leq \text{COSTS\_GOODS}^{(p)} \leq 65.66$$

$$76 \leq \text{COSTS\_OTHER}^{(p)} \leq 84$$

Example: Final result

ID	common variables	
	COSTS_GOODS	COSTS_OTHER
1	100	160
2	62.53	80

Edit rules common variables:

$\text{COSTS\_GOODS} \geq 0$

$\text{COSTS\_OTHER} \geq 0$

$\text{IF } (\text{COSTS\_GOODS} > 0) \text{ THEN } (\text{COSTS\_OTHER} > 0)$

ID	data source 1				
	$\text{COSTS\_GOODS}^{(1)}$	$\text{COSTS\_OTHER}^{(1)}$	$\text{costs\_goods}^{(1)}$	$\text{costs\_staff}^{(1)}$	$\text{costs\_misc}^{(1)}$
1	100	160	100	160	0
2	60	80	60	50	30

Edit rules source 1:

$\text{costs\_goods}^{(1)} \geq 0$

$\text{costs\_misc}^{(1)} \geq 0$

$\text{costs\_staff}^{(1)} \geq \text{costs\_misc}^{(1)}$

ID	data source 2				
	$\text{COSTS\_GOODS}^{(2)}$	$\text{COSTS\_OTHER}^{(2)}$	$\text{costs\_goods}^{(2)}$	$\text{costs\_other}^{(2)}$	$\text{costs}^{(2)}$
1	100	160	100	160	260
2	65	80	65	80	145

Edit rules source 2:

$\text{costs\_goods}^{(2)} \geq 0$

$\text{costs\_other}^{(2)} \geq 0$

$\text{costs}^{(2)} = \text{costs\_goods}^{(2)} + \text{costs\_other}^{(2)}$

Edit rules common variables:

$0.95 \times \text{COSTS\_GOODS} \leq \text{COSTS\_GOODS}^{(p)} \leq 1.05 \times \text{COSTS\_GOODS}$

$0.95 \times \text{COSTS\_OTHER} \leq \text{COSTS\_OTHER}^{(p)} \leq 1.05 \times \text{COSTS\_OTHER}$

Automatic multisource editing

Pilot study (2021–2022):

- Prototype implementation in R
- 7 data sources, 13 common variables, 100+ variables in total
- Main findings:
  - Technically feasible within current IT environment (unlike one-step approach)
  - Quality of automatically edited data is still quite low: need to include more subject-matter knowledge

Automatic multisource editing

New pilot study (2024–2025):

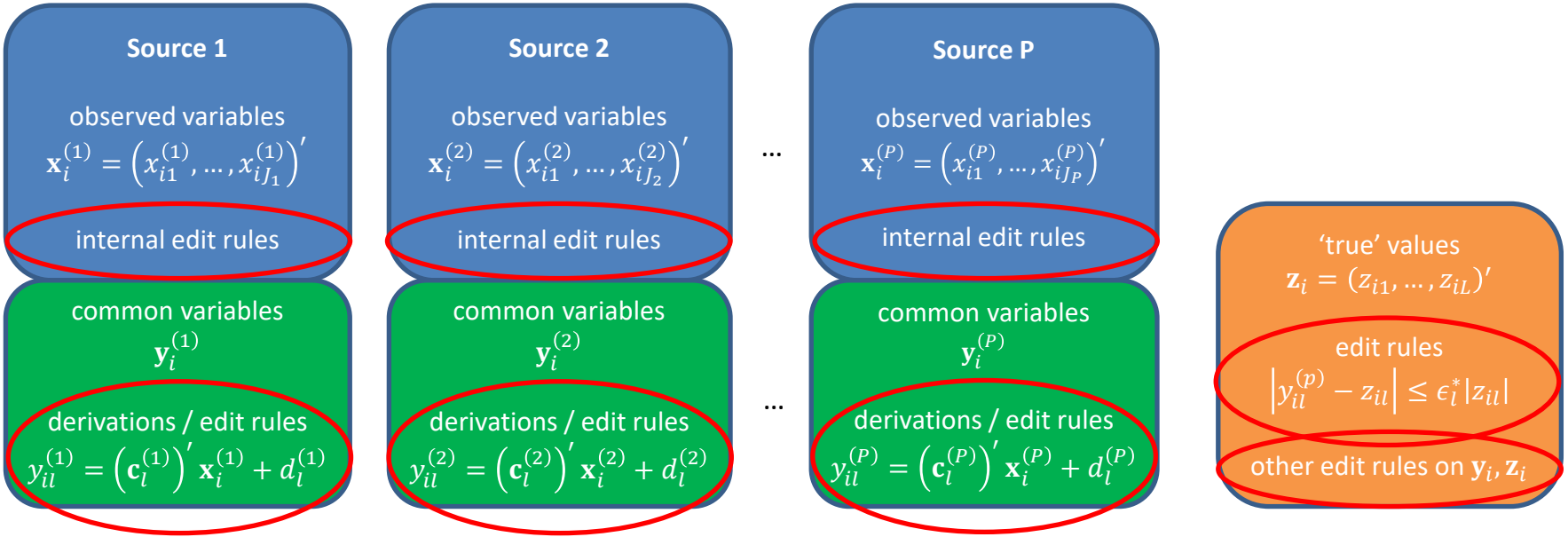
- 9 data sources (reference year 2022)
  - Structural Business Statistics (survey)
  - ProdCom (survey)
  - Statistics on Finances of Large Enterprise groups (survey)
  - Short-Term Statistics (admin. data)
  - Short-Term Statistics (survey)
  - Statistics on Employees and Salaries (admin. data)
  - Statistics on International Trade of Goods and Services (survey + admin. data)
  - Profit Declaration Tax Data (admin. data)
  - Investment Statistics (survey)
- 33 common variables

Incorporating subject-matter knowledge

Parameters that affects outcome of automatic editing

- Deductive correction rules
- Edit rules for error localization
- Reliability weights for error localization

Incorporating knowledge: edit rules



Incorporating knowledge: edit rules

Edit rules on ‘true’ values of common variables

- Statistical analysts currently have little experience in this area (important exception: Large Cases Unit)
- Proposal to derive additional edit rules:
  - Use a data-driven approach
  - Discuss findings with subject-matter experts

Incorporating knowledge: edit rules

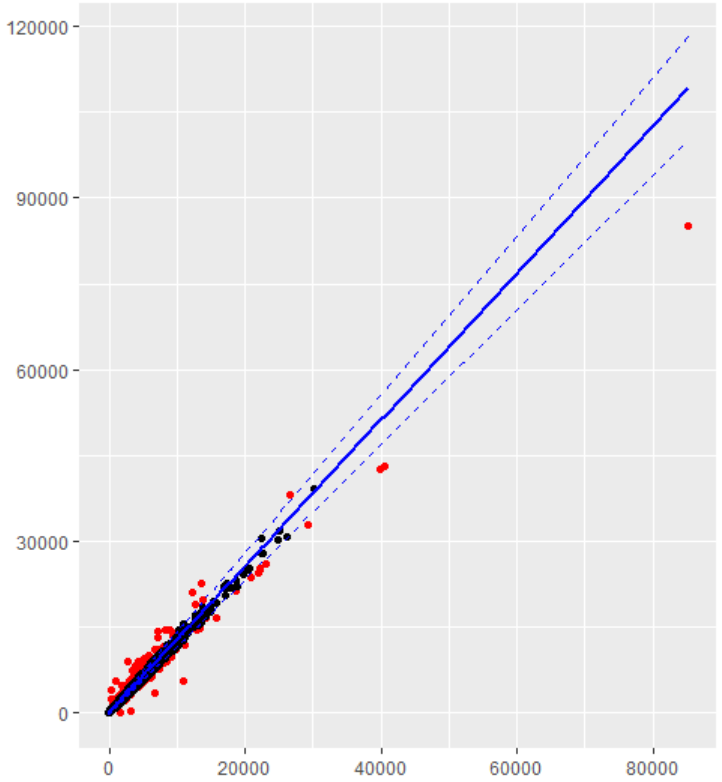
Example:
Edit rule based on prediction interval from a weighted linear regression

$$\hat{\alpha}^{(low)} + \hat{\beta}^{(low)} z_{i1} \leq z_{i2} \leq \hat{\alpha}^{(up)} + \hat{\beta}^{(up)} z_{i1}$$

Use $\left(y_{i1}^{(p)}, y_{i2}^{(p)}\right)$ as proxy for (z_{i1}, z_{i2})

Other approaches are possible:

- Decision trees
- Advanced models (machine learning)



Incorporating knowledge: reliability weights

Liepins (1980): Fellegi-Holt paradigm yields an approximate maximum-likelihood estimator of the true error pattern under a (very simple) error model:

- Random errors occurring independently across variables
- Reliability weight of value x_{ij} should be $-\log\frac{p_{ij}}{1-p_{ij}}$ with $p_{ij} = \mathrm{P}\left(x_{ij}\ \text{erroneous}\right)$

Currently a lack of knowledge on p_{ij}, in particular for common variables

Incorporating knowledge: reliability weights

Reliability weights for common variables

- Static weights: varying by stratum
  - Model the occurrence of large differences $y_{il}^{(p)} \gg y_{il}^{(q)}$ or $y_{il}^{(p)} \ll y_{il}^{(q)}$ as a function of background variables to define strata
  - Discuss findings with subject-matter experts to decide which source is more reliable in each stratum
- Dynamic weights: varying by unit
  - Nearest-neighbour approach (see paper)

Evaluating quality of multisource editing

Approaches in pilot study:

- Construct a sample of ‘gold standard’ edited data
  - Random sample of 350 units (7 x 50) selected from seven economic sectors
  - Manual editing: collaborative effort by statistical analysts across data sources
  - Allows estimation of evaluation measures such as recall, precision, accuracy, …
- Evaluate effects of automatic editing on statistical output
  - Compare stratum totals before and after editing
  - Large changes should be considered plausible by subject-matter experts

References

M. Bankier (2006), Imputing Numeric and Qualitative Variables Simultaneously. Memo, Statistics Canada, Social Survey Methods Division.

J. Daalmans & S. Scholtus (2018), A MIP Approach for a Generalised Data Editing Problem. Discussion Paper, Statistics Netherlands, The Hague.

I.P. Fellegi & D. Holt (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.

B. Ghosh-Dastidar & J.L. Schafer (2006), Outlier Detection and Editing Procedures for Continuous Multivariate Data. *Journal of Official Statistics* **22**, 487–506.

H.J. Kim, L.H. Cox, A.F. Karr, J.P. Reiter & Q. Wang (2015), Simultaneous Edit-Imputation for Continuous Microdata. *Journal of the American Statistical Association* **110**, 987–999.

G.E. Liepins (1980), A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis. Report ORNL/TM-7126, Oak Ridge National Laboratory.

References (continued)

R.J.A. Little & P.J. Smith (1987), Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association* **82**, 58–68.

S. Scholtus (2015), New Results on Automatic Editing using Hard and Soft Edit Rules. UNECE Work Session on Statistical Data Editing, Budapest.

S. Scholtus (2016), A Generalized Fellegi-Holt Paradigm for Automatic Error Localization. *Survey Methodology* **42**, 1–18.

S. Scholtus, W. de Jong, A. Vaasen-Otten & F. Aelen (2022), Towards a New Integrated Uniform Production System for Business Statistics at Statistics Netherlands: Automatic Data Editing with Multiple Data Sources. UNECE Expert Meeting on Statistical Data Editing (virtual).

A. Vaasen-Otten, F. Aelen, S. Scholtus & W. de Jong (2022), Towards a New Integrated Uniform Production System for Business Statistics at Statistics Netherlands: Quality Indicators to Guide Top-down Analysis. UNECE Expert Meeting on Statistical Data Editing (virtual).