Estimating population census tables and their accuracy using Multiple Imputation of Latent Classes (MILC) with multi-source data

Laura Boeschoten (Tilburg University / Statistics Netherlands)
Sander Scholtus (Statistics Netherlands)
Jacco Daalmans (Statistics Netherlands)
Jeroen Vermunt (Tilburg University)
Ton de Waal (Statistics Netherlands / Tilburg University)

Outline

- Dutch virtual Census
- MILC method
- Simulation study
- Conclusions

Dutch virtual Census

- Decennial Population and Housing Census
- Based on available data in the Netherlands:
  - Central Population Register
  - Other administrative datasets
  - Labour Force Survey (educational attainment, occupation)
- Current approach (2011 Census):
  - Micro-integration
  - Repeated weighting

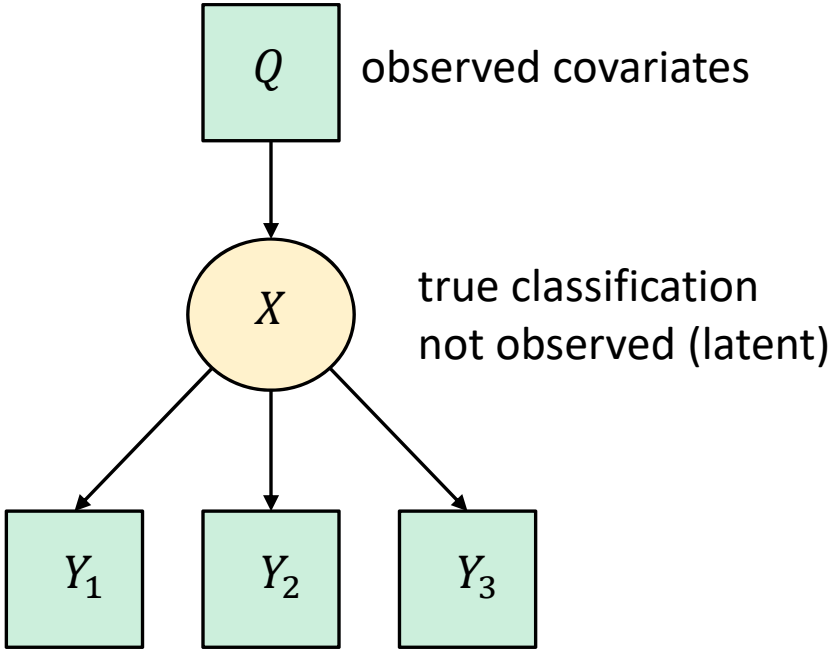
Dutch virtual Census

- Drawbacks of current approach:
  - Conflicting data often resolved by prioritising data sources → implicit assumption that some observed variables are error-free
  - Data processing and estimation steps are order-dependent
  - Uncertainty due to measurement errors not taken into account in accuracy of estimates

MILC

- Multiple Imputation of Latent Classes
  (Boeschoten et al., 2017)
  - Correct for measurement error in observed data using Latent Class (LC) analysis
  - Evaluate variance of resulting estimated target parameters using Multiple Imputation (MI)
  - Requires measures of the same variable(s) originating from different data sources that can be linked on unit level

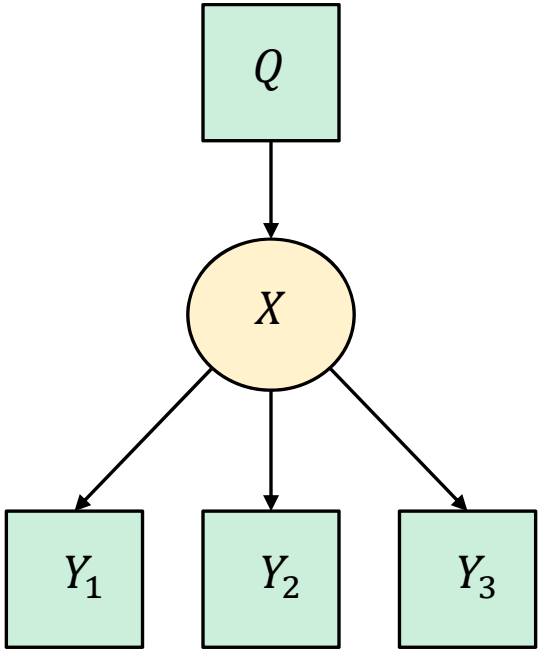
Latent Class analysis



Basic LC assumptions:

- Latent variable X represents true classification (error-free)
- Observed variables $Y_1, \dots, Y_L$ are "locally independent"
- Covariates Q are error-free
- Covariates Q do not affect measurement errors in Y_l

Latent Class analysis



Observable probabilities:

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{Q} = \boldsymbol{q}) = \sum_{x=1}^{K} P(\boldsymbol{Y} = \boldsymbol{y}, X = x|\boldsymbol{Q} = \boldsymbol{q})$$

Under assumptions of LC model:

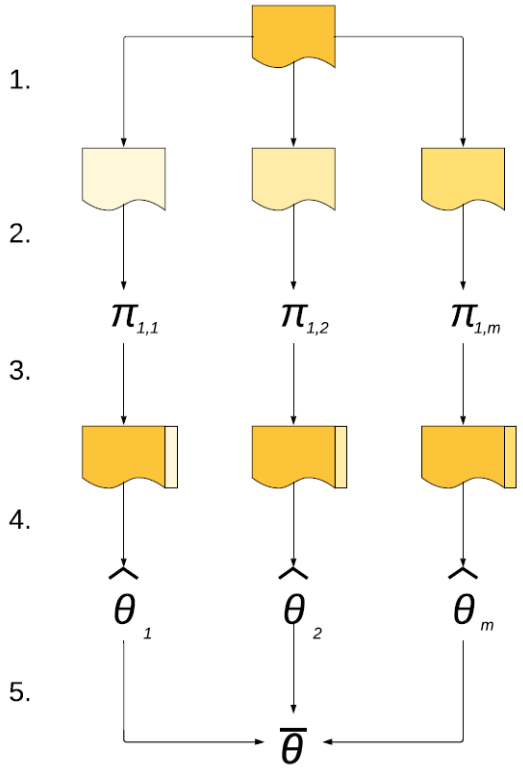
$$P(\boldsymbol{Y} = \boldsymbol{y}, X = x|\boldsymbol{Q} = \boldsymbol{q})$$
$$= P(X = x|\boldsymbol{Q} = \boldsymbol{q}) \times \prod_{l=1}^{L} P(Y_l = y_l|X = x)$$

Model parameters:

- Class membership probabilities $P(X = x|\boldsymbol{Q} = \boldsymbol{q})$
- Error probabilities $P(Y_l = y_l|X = x)$

MILC

1. Create m bootstrap samples from the original dataset
2. For each bootstrap sample, estimate the LC model
3. From each estimated LC model, draw one column of imputed true values in the original dataset
4. From each imputed dataset, estimate the target parameters
5. Pool the estimated parameters and estimate the associated variance



MILC

Step 3: Imputing predicted true values

Posterior membership probabilities:

$$\tau_x \equiv P(X = x|\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Q} = \boldsymbol{q}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, X = x|\boldsymbol{Q} = \boldsymbol{q})}{\sum_{x=1}^{K} P(\boldsymbol{Y} = \boldsymbol{y}, X = x|\boldsymbol{Q} = \boldsymbol{q})}$$

For each unit i with observed values $(\boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{Q} = \boldsymbol{q}_i)$, impute a predicted true value W by drawing $x \in \{1, \ldots, K\}$ from a multinomial distribution with

$$P(W = x|\boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{Q} = \boldsymbol{q}_i) = P(X = x|\boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{Q} = \boldsymbol{q}_i) \equiv \tau_{xi}$$

MILC

Step 4: Data with imputations $W_1, \ldots, W_m$ yield estimates $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_m$

Step 5: Apply Rubin's rules for multiple imputation

Pooled estimate: $\widehat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\boldsymbol{\theta}}_j$

Associated variance estimate: $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{U}} + \left(1 + \frac{1}{m}\right)\widehat{\boldsymbol{B}}$

$\widehat{\boldsymbol{U}} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\boldsymbol{U}}_j$: average estimated variance based on completed data

$\widehat{\boldsymbol{B}} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}\right)'$: between-imputation variance

$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\theta}})$ reflects uncertainty due to missing values *and* measurement errors

MILC

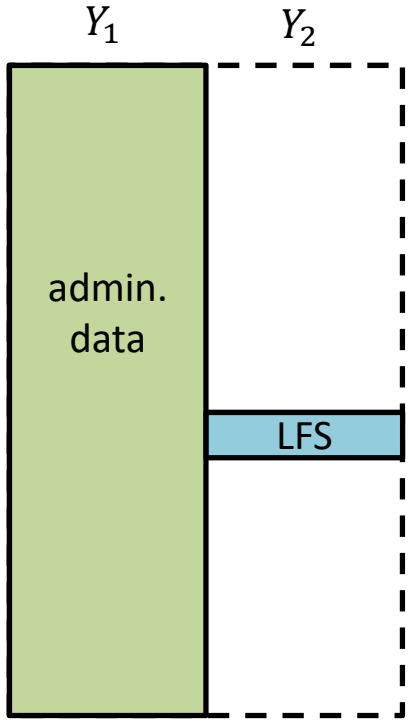
- Other aspects of MILC:
  - Edit rules can be taken into account as restrictions on parameters of LC model
  - Relative entropy of posterior membership probabilities:

$$R^2 = 1 - \frac{-\sum_{i=1}^{N}\sum_{x=1}^{K}\tau_{xi}\log\tau_{xi}}{N\log K}$$

    ($0 \leq R^2 \leq 1$, where $R^2 = 1$ means perfect prediction)
  - Simulation study in Boeschoten et al. (2017): performance of MILC method strongly related to R^2 of LC model
    - Approx. unbiased estimation of a frequency table for $R^2 > 0.9$
    - Approx. unbiased estimation of a logistic regression model for $R^2 > 0.6$

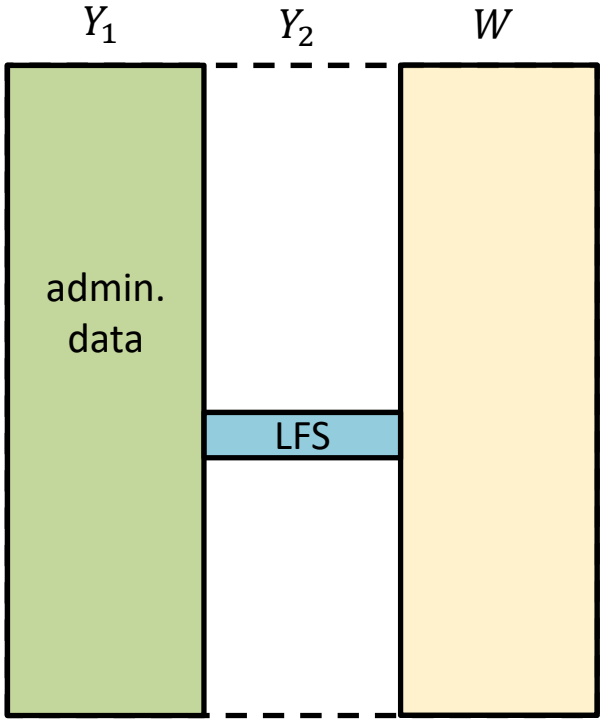
Applying MILC to Census

- Census:
  - Target parameters: (often high-dimensional) frequency tables for finite population
  - Finite population ‘completely’ covered by register data
- Implication for MILC method: evaluate all variances w.r.t. finite population



Applying MILC to Census

- Here: simple approach
  - LC model estimated only on units with overlapping data
  - MI applied to all units in population
- Consequences for MILC:
  - In Step 1: bootstrap applied to units with overlapping data
  - In Step 5: $\widehat{\boldsymbol{U}}_1 = \cdots = \widehat{\boldsymbol{U}}_m = 0$ and therefore $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\theta}}) = \left(1 + \frac{1}{m}\right)\widehat{\boldsymbol{B}}$



Simulation study: setup

- Starting point:
  - Six-dimensional table from 2011 Census (42,000 cells)
  - 2,691,477 persons living in region ‘Noord-Holland’
  - Variables:

covariates
- Age (21 five-year classes)
- Marital status (eight classes)
- Place of birth (NL; Within EU; Outside EU; Other; Not stated)

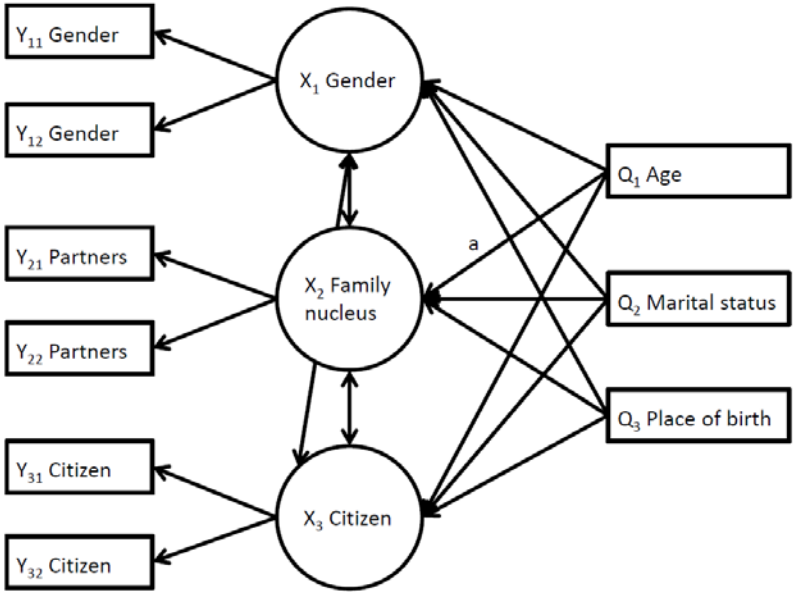
target variables
- Gender (Male; Female)
- Family nucleus (Partners; Lone parents; Sons/daughters; Not stated; N.A.)
- Country of citizenship (NL; Within EU; Outside EU; Stateless; Not stated)

Simulation study: setup

- For target variables (Gender, Family nucleus, Country of citizenship) create two indicators:
  - Indicator 1: 5% random misclassification, no missing values
  - Indicator 2: 5% random misclassification, 90% missing values
- First indicator represents administrative data
- Second indicator represents sample survey data
  - Two conditions:
    - MCAR (all units have same probability of being observed in survey)
    - MAR (probability of being observed in survey increases with age)

Simulation study: setup

LC model:



Edit restriction (a):

Age < 15 years $\Longrightarrow$ Family nucleus cannot be ‘Partners’ or ‘Lone parents’

Simulation study: setup

- Number of multiple imputations: $m \in \{5,10,20\}$
- Simulations repeated 500 times
- Results evaluated in terms of:
  - Bias and RMSE of estimated counts
  - Correctness of estimated standard errors

Simulation study: results

- Marginal frequencies for Family nucleus:

Type of family nucleus			MCAR			MAR		
	Frequency	$Y_{2,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
			Bias					
Lone parents	97,360	2,670	185	182	176	224	226	220
N.A.	604,032	8,985	-957	-975	-989	-1,601	-1,612	-1,611
Partners	1,272,339	-19,686	401	411	427	932	935	932
Sons/daughters	717,746	8,030	371	381	386	446	451	459
			RMSE					
Lone parents	97,360	2,672	425	408	395	426	421	414
N.A.	604,032	8,989	1,337	1,318	1,312	1,837	1,833	1,818
Partners	1,272,339	19,688	954	914	904	1,256	1,235	1,218
Sons/daughters	717,746	8,034	630	624	617	715	692	688

Simulation study: results

- Marginal frequencies for Country of citizenship:

Citizen								
				MCAR			MAR	
	Frequency	$Y_{3,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias				
EU	79,212	51,365	-5	-7	-12	-199	-211	-216
NL	2,511,214	-116,899	-555	-546	-545	117	124	107
not EU	89,592	58,085	512	502	507	62	69	89
Not stated	11,459	7,448	49	51	49	21	18	20
				RMSE				
EU	79,212	51,365	410	398	388	488	486	475
NL	2,511,214	116,899	925	894	883	767	756	720
not EU	89,592	58,086	800	770	767	618	611	590
Not stated	11,459	7,449	201	197	190	204	205	198

Simulation study: results

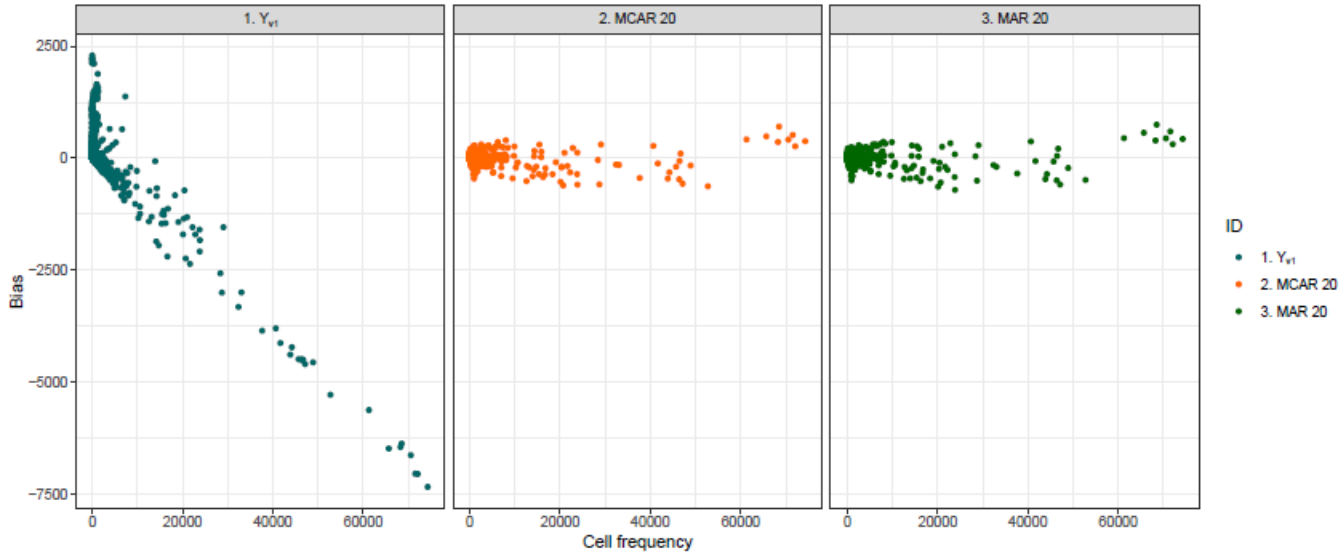
- Marginal frequencies for Gender:

Gender			MCAR			MAR		
	Frequency	$Y_{1,1}$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
				Bias				
F.	1,367,167	-2,126	3,386	3,308	3,325	3,231	3,153	3,109
M.	1,324,310	2,126	-3,386	-3,308	-3,325	-3,231	-3,153	-3,109
				RMSE				
F.	1,367,167	2,154	6,008	5,888	5,760	5,914	5,637	5,512
M.	1,324,310	2,154	6,008	5,888	5,760	5,914	5,637	5,512

- Relative entropy of LC model (MCAR case):
  - $R^2(\text{Gender}) = 0.74$
  - $R^2(\text{Country of citizenship}) = 0.86$
  - $R^2(\text{Family nucleus}) = 0.92$

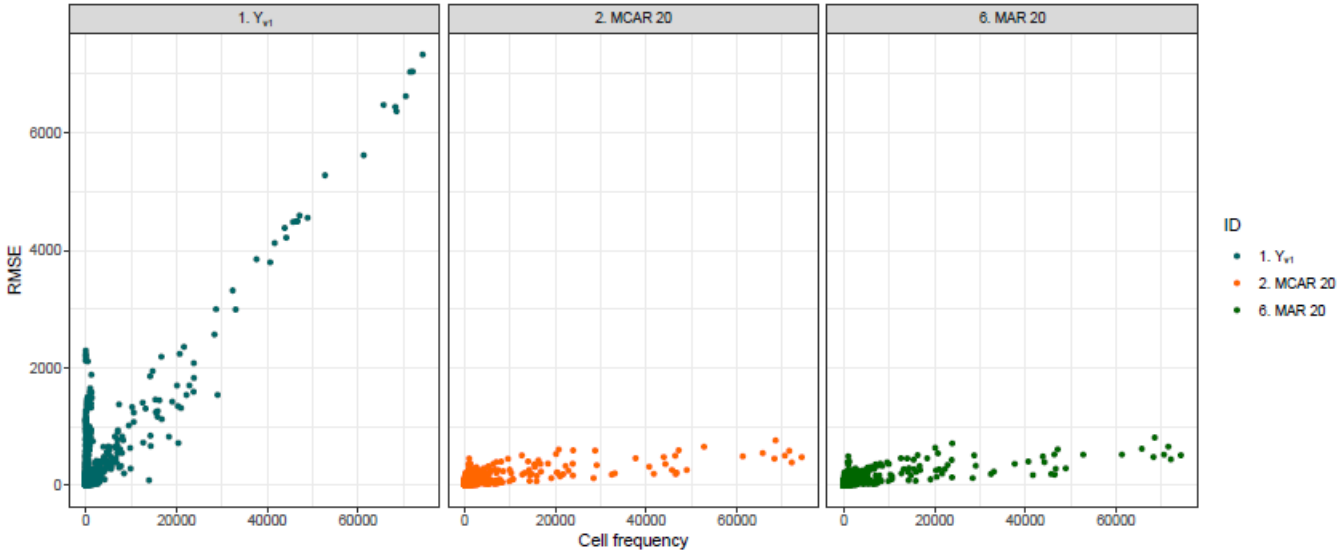
Simulation study: results

- Bias for all counts in six-dimensional table:



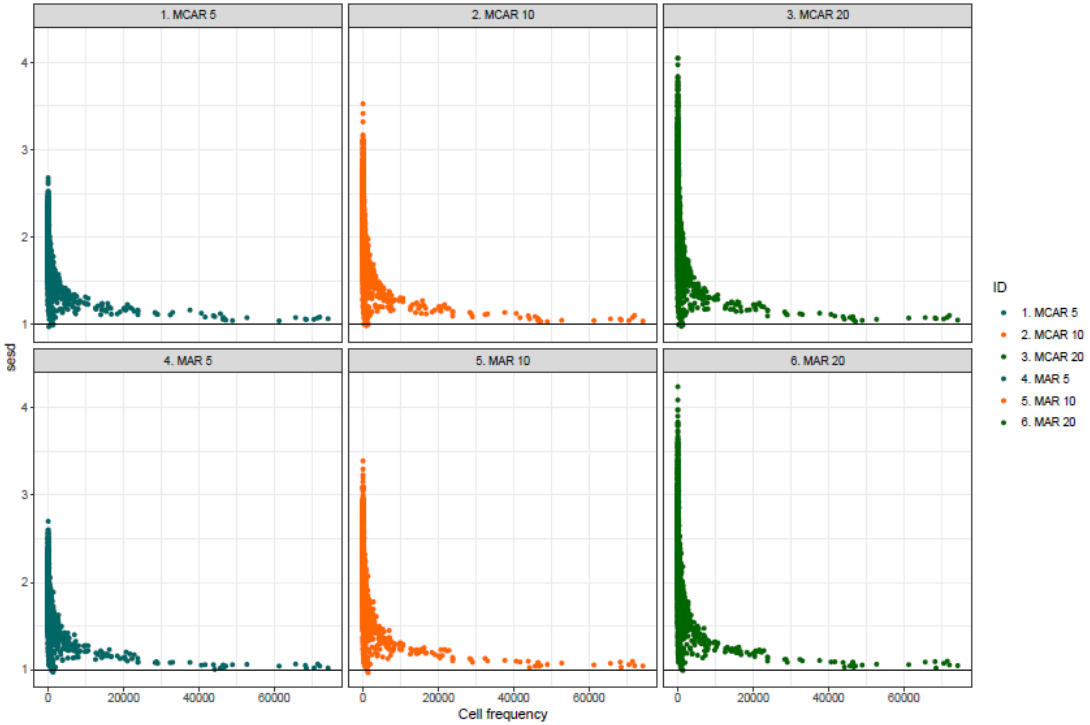
Simulation study: results

- RMSE for all counts in six-dimensional table:



Simulation study: results

- Ratio of estimated standard error to true standard deviation:



Conclusions

- MILC method can be used to correct for measurement error when estimating Census tables
  - Requires linked data from multiple sources on same variables
  - Bias corrected in comparison to original (single-source) data
  - Can account for edit rules (cells restricted to zero)
  - Important assumption: measurement errors independent between different data sources

Conclusions

- MILC method also provides variance estimates
  - Includes uncertainty due to measurement errors
  - Estimates for finite population: only between-imputation variance
  - Variance tends to be over-estimated for cells with small counts
  - Results might be improved by using a finite-population bootstrap
- Only minor differences between MCAR and MAR data and between $m = 5$, $m = 10$ and $m = 20$ imputations

References

L. Boeschoten, D. Oberski & T. de Waal (2017), Estimating classification errors under edit restrictions in composite survey-register data using Multiple Imputation Latent Class modelling (MILC). *Journal of Official Statistics* **33**, 921–962.

L. Boeschoten, S. Scholtus, J. Daalmans, J. Vermunt & T. de Waal (2019), Using Multiple Imputation of Latent Classes (MILC) to construct population census tables with data from multiple sources. *Submitted for publication*.