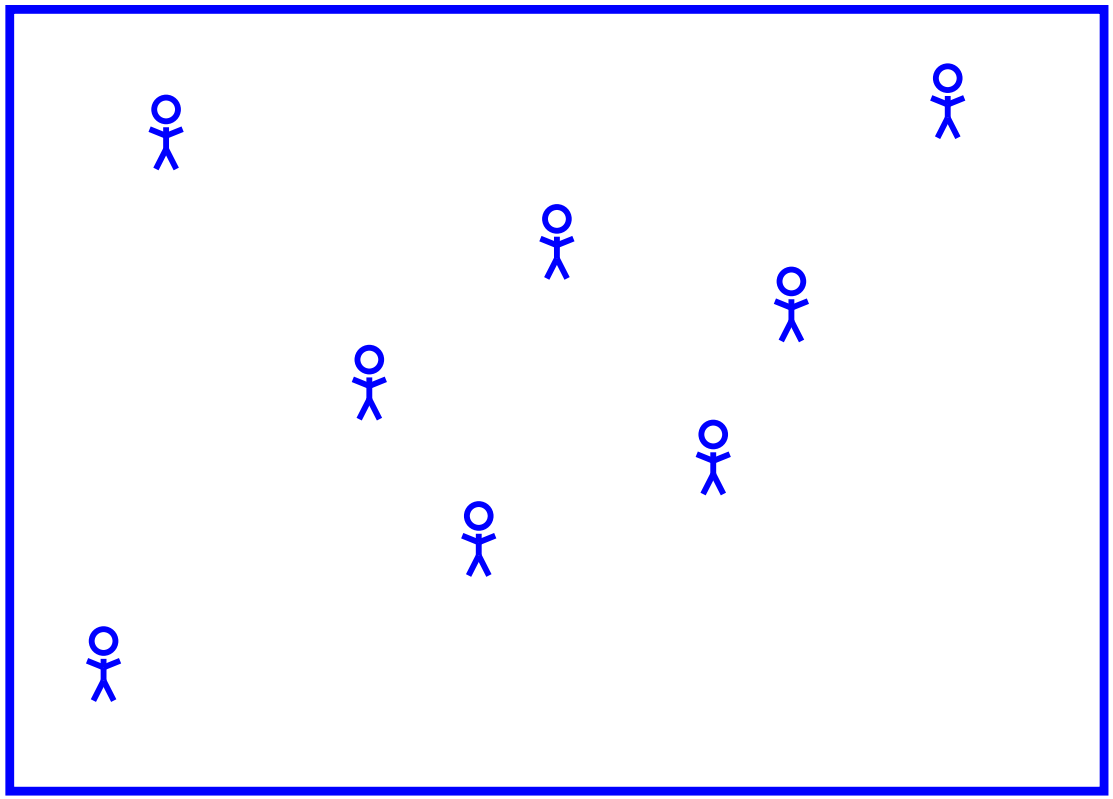
Keynote presentation
Recent advances in imputation methods
Imputation for Swiss cheese nonresponse

Yves Tillé and Audrey-Anne Vallée

University of Neuchatel

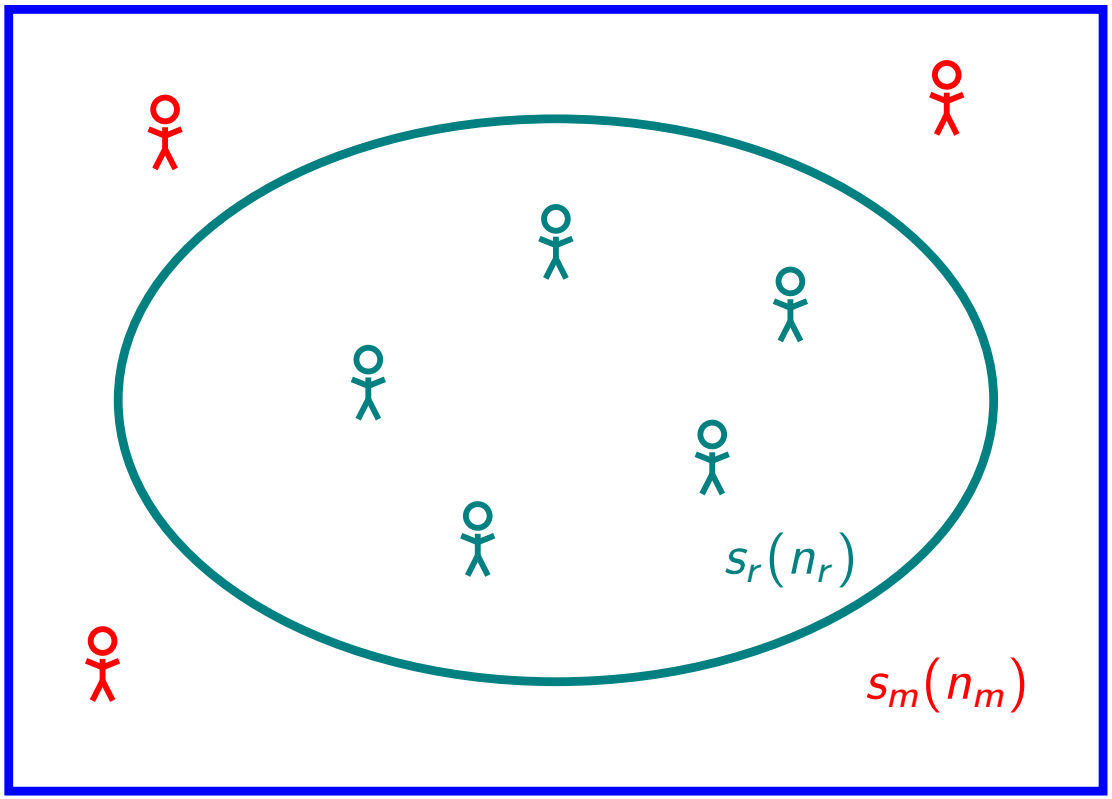
UNECE Statistical Data Editing Virtual Workshop 2020

Introduction to nonresponse



Unit on which we observe J variables $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})^\top$.

Introduction to nonresponse



- Respondant: $x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})^\top$ is completely observed.
- Nonrespondant: $x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})^\top$ is not completely observed.

Nonresponse can appear under different patterns:

- Questionaire nonresponse

Gender	Size	Weights
H	175	68
F	160	55
?	?	?
?	?	?

- Item nonresponse

Nonresponse can appear under different patterns:

- Questionaire nonresponse
- Item nonresponse
  - Only one variable is subject to nonresponse.

Gender	Size	Weights
H	175	68
F	160	55
H	180	?
F	165	?

  - All variables are subject to nonresponse.

Gender	Size	$P_{t=1}$	$P_{t=2}$	$P_{t=3}$
H	175	68	67	68
F	160	55	58	?
H	180	70	?	?
F	165	?	?	?

Monotone

Gender	Size	Weights
H	175	68
F	160	?
H	?	70
?	165	?

Non monotonic

Swiss cheese nonresponse

Swiss cheese nonresponse (non monotonic)

All the variables in a survey contain missing values without any particular pattern.

Traitements

- Imputation methods by donors (Andridge et Little, 2010; Judkins, 1997).
- Iterative imputation methods: a sequence of regression models between variables (Raghunathan et coll., 2001).

Swiss cheese nonresponse

Desired properties of an imputation method

- Impute by realistic values;
- Preserve the distributions of variables;
- Preserve relationships between variables

Swiss cheese nonresponse

Desired properties of an imputation method

- Impute by realistic values;
- Preserve the distributions of variables;
- Preserve relationships between variables

Imputation balanced by the K nearest neighbors

- Imputation in the univariate case (Hasler et Tillé, 2016);
- Donor method (random);
  - Continuous and categorical variables;
  - One donor per non-respondent;
- Imputation by close donors (neighbors);
- Balancing constraints.

Swiss cheese nonresponse

Desired properties of an imputation method

- Impute by realistic values;
- Preserve the distributions of variables;
- Preserve relationships between variables

Imputation balanced by the K nearest neighbors

- Imputation in the univariate case (Hasler et Tillé, 2016);
- Donor method (random);
  - Continuous and categorical variables;
  - One donor per non-respondent;
- Imputation by close donors (neighbors);
- Balancing constraints.

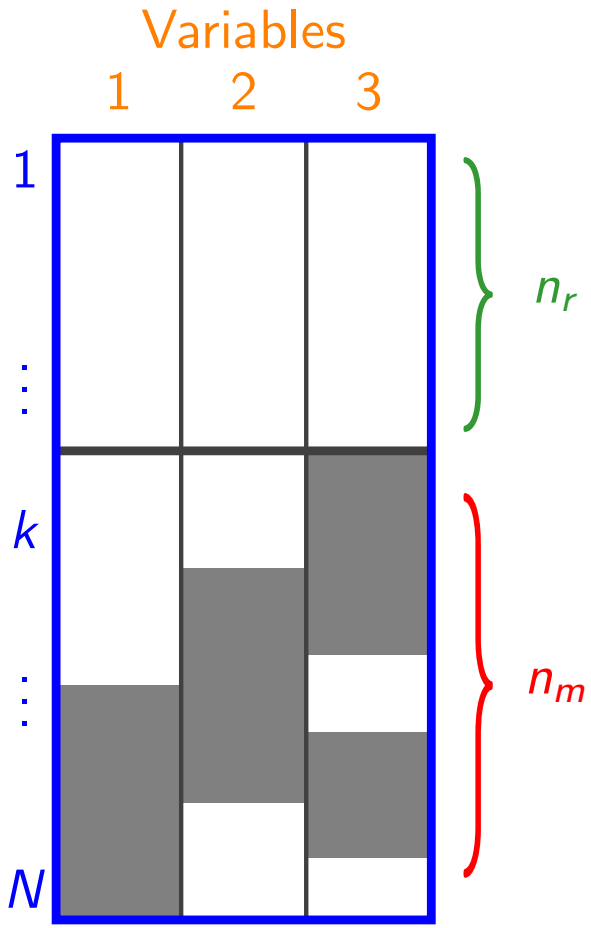
$\rightarrow$ Let's extend this method for the Swiss cheese nonresponse !

American vision of the Swiss cheese

True Swiss cheeses

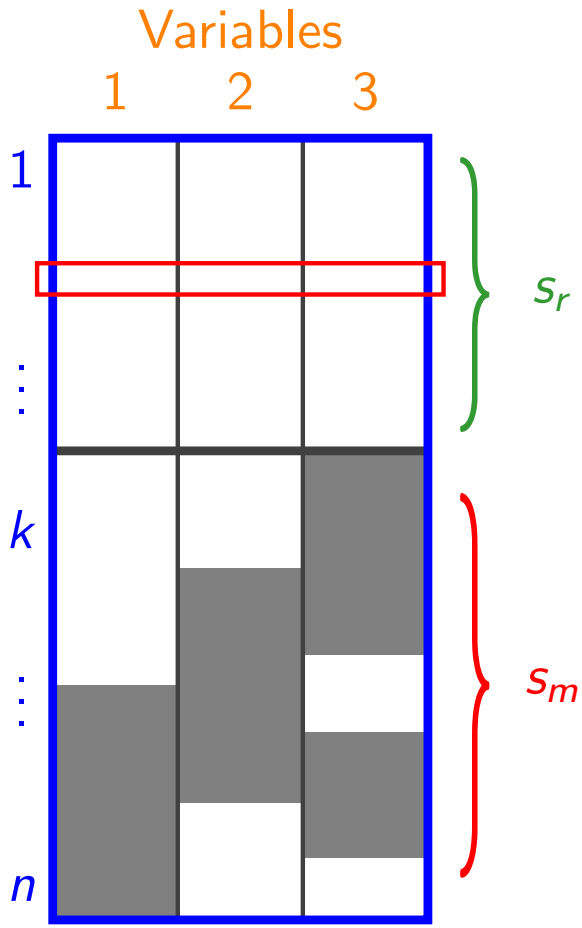
Swiss cheese nonresponse

- Population U of size N.
- J variables of interest,
  $x_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$.
- $s_r \subset U$,
  n_r completely observed units.
- $s_m = U - s_r$,
  $n_m = N - n_r$ units with missing values.
- Nonresponse non monotonic.



Requirements of the imputation method

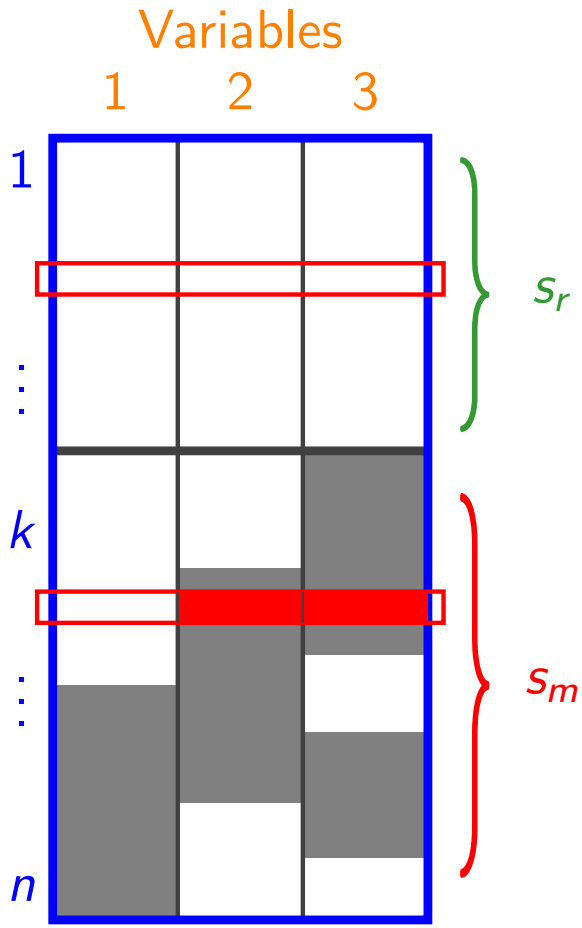
(i) Method by donors: choose donors among s_r.



Requirements of the imputation method

(i) Method by donors: choose donors among s_r.

(ii) Only one donor per unit.

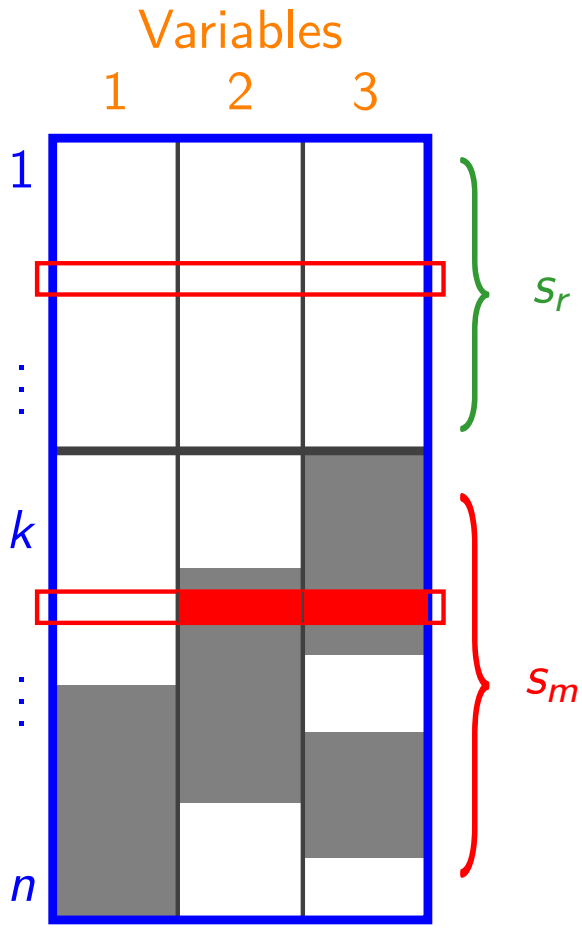


Requirements of the imputation method

(i) Method by donors: choose donors among s_r.

(ii) Only one donor per unit.

(iii) Donor selected from K closest neighbors of the unit with missing values.



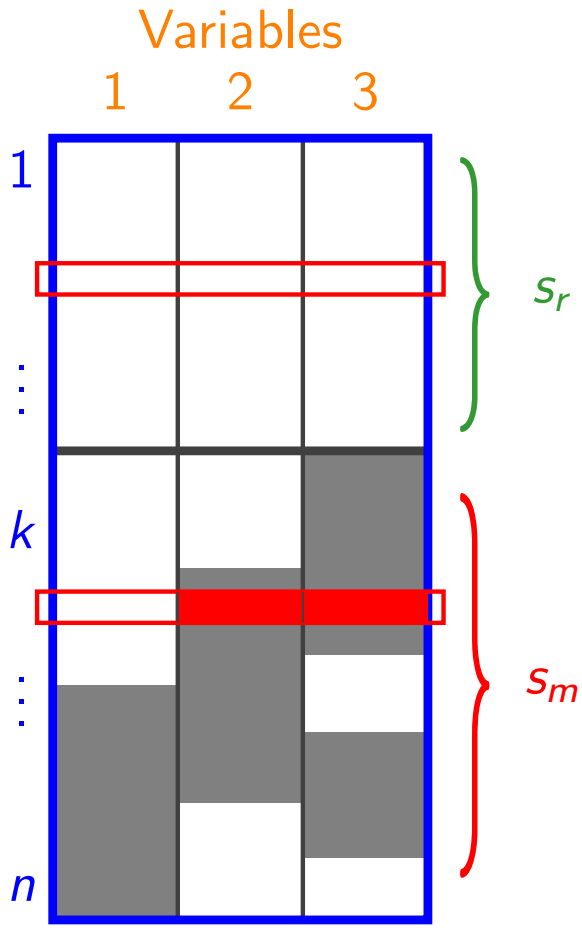
Requirements of the imputation method

(i) Method by donors: choose donors among s_r.

(ii) Only one donor per unit.

(iii) Donor selected from K closest neighbors of the unit with missing values.

(iv) Balancing constraints.



Matrix of imputation probabilities

(i) Method by donors: choose donors among s_r:

Matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.

- ψ_{ik}: probability that the respondent i gives his values to the non-respondent k;
- $\psi_{ik} \geq 0$.

$$\boldsymbol{\psi} = \begin{matrix} \\ \text{Respondants} \end{matrix} \overset{\text{Nonrespondants}}{\begin{pmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \\ \psi_{41} & \psi_{42} & \psi_{43} \end{pmatrix}} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0.3 & 0 & 0.4 \\ 0.2 & 0 & 0.1 \end{pmatrix}$$

Matrix of imputation probabilities

(i) Method by donors: choose donors among s_r:

Matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.

- ψ_{ik}: probability that the respondent i gives his values to the non-respondent k;
- $\psi_{ik} \geq 0$.

(ii) Only one donor per non-responding unit:

$$\sum_{i \in s_r} \psi_{ik} = 1.$$

Matrix of imputation probabilities

(i) Method by donors: choose donors among s_r:

Matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.

- ψ_{ik}: probability that the respondent i gives his values to the non-respondent k;
- $\psi_{ik} \geq 0$.

(ii) Only one donor per non-responding unit:

$$\sum_{i \in s_r} \psi_{ik} = 1.$$

(iii) Donor selected from K closest neighbors of the unit with missing values:

$$\psi_{ik} = 0 \text{ si } i \notin \text{Kpp}(k)$$

where $\text{Kpp}(\ell) = \{j \in s_r | \, \text{rank}(d(j, \ell)) \leq K\}$ et $d(.,.)$ is a function of distance.

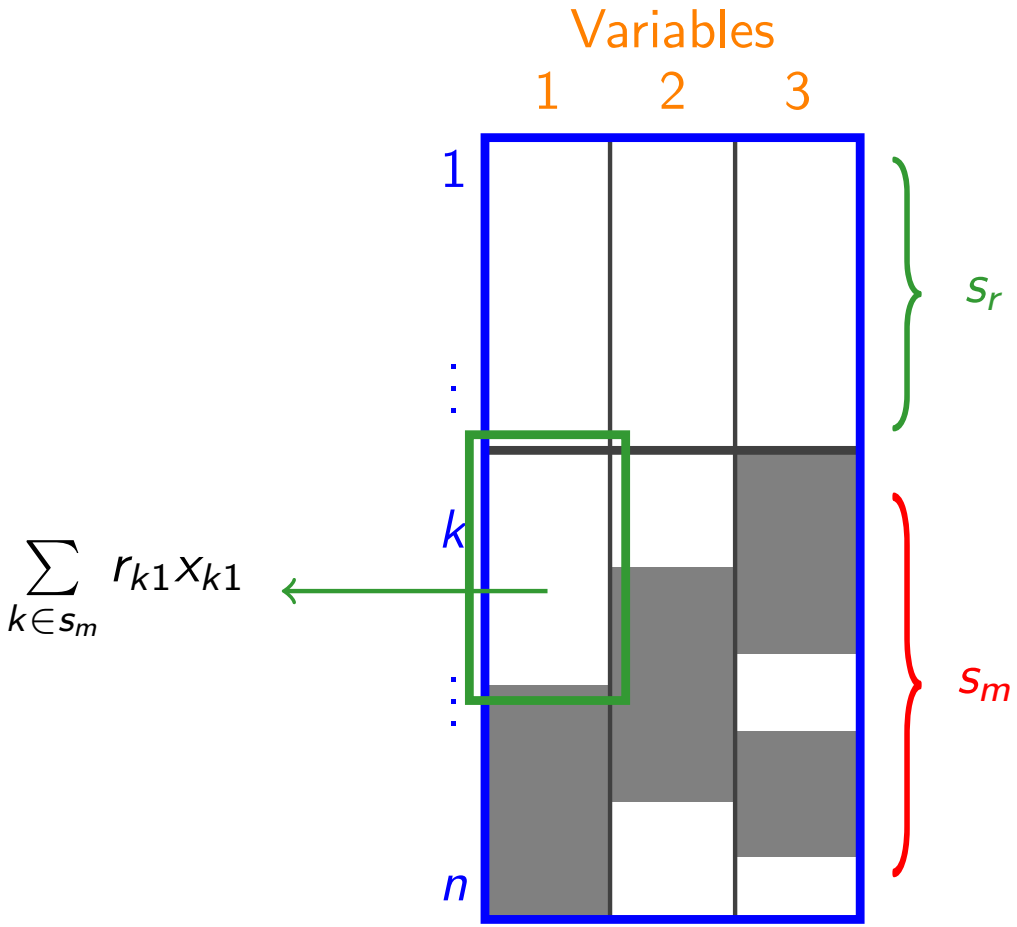
Matrix of imputation probabilities

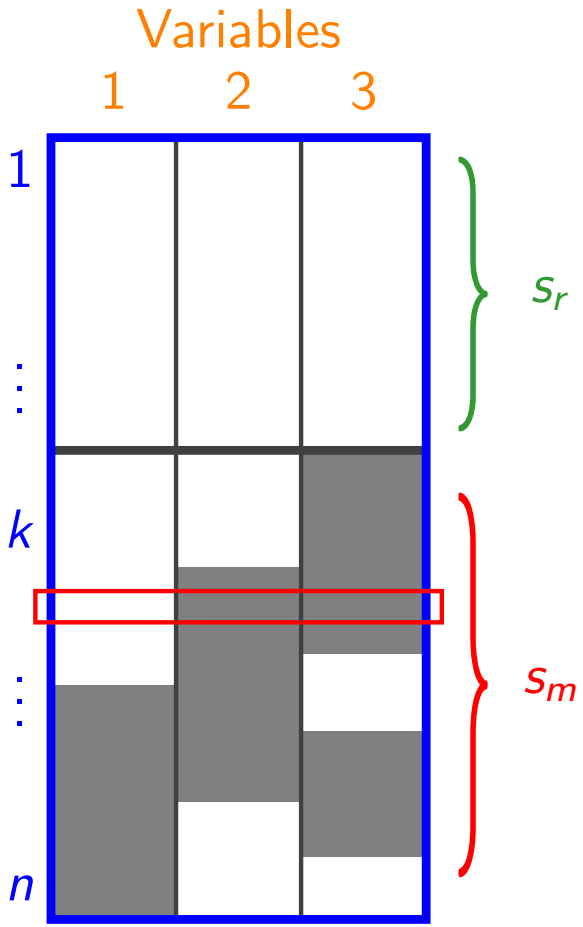
(iv) Balancing constraints:

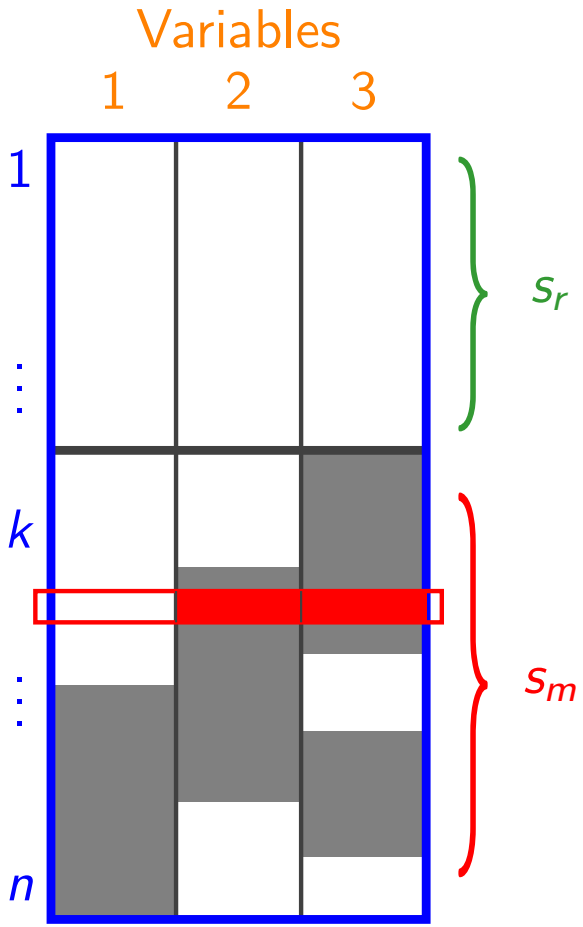
If the observed values of non-respondents were imputed, the estimator of the total of all observed values should be left unchanged.

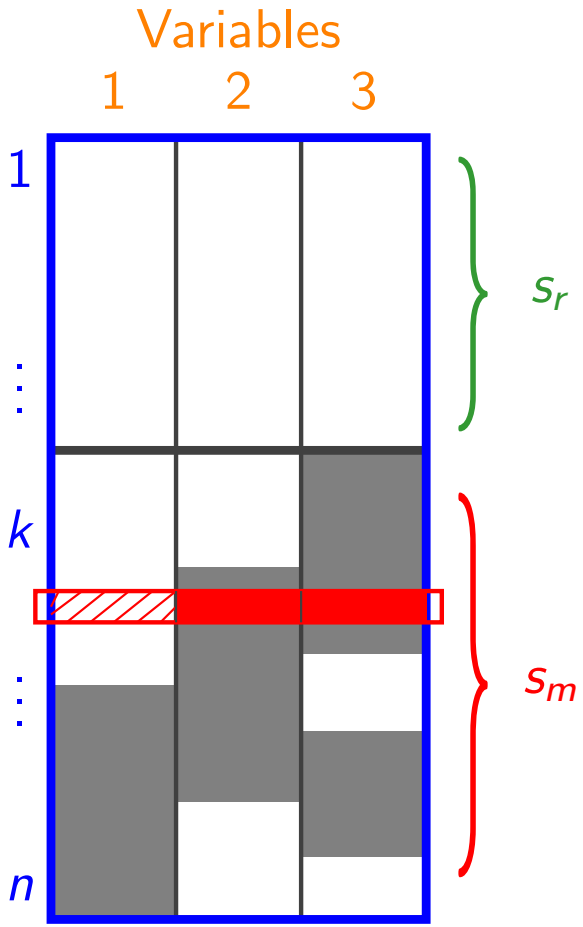
$$\sum_{k \in s_m} r_{kj} \underbrace{\sum_{i \in s_r} \psi_{ik} x_{ij}}_{x^*_{kj}} = \sum_{k \in s_m} r_{kj} x_{kj}, \quad \text{for} \quad j = 1, \ldots, J,$$

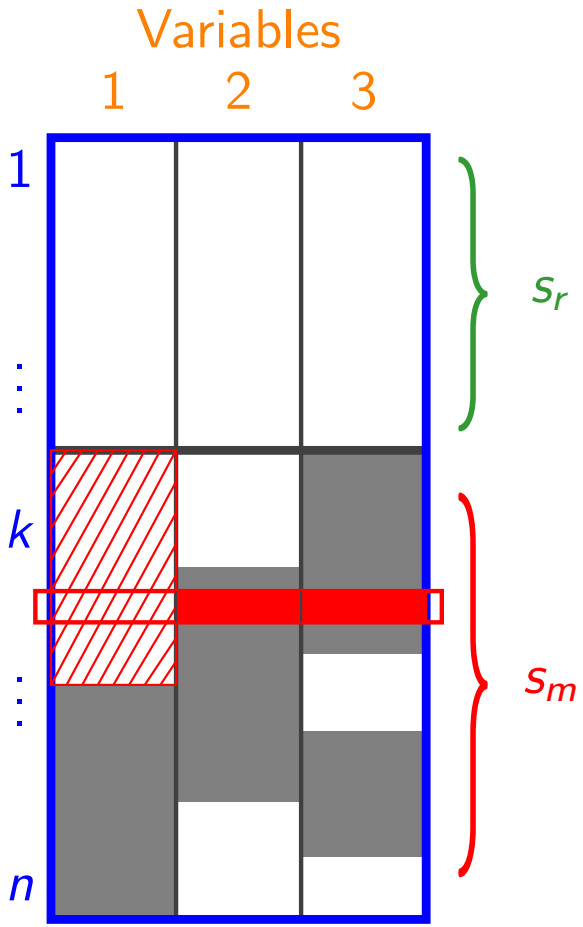
where r_{kj} equals 1 if unit k responded to variable j, 0 otherwise.

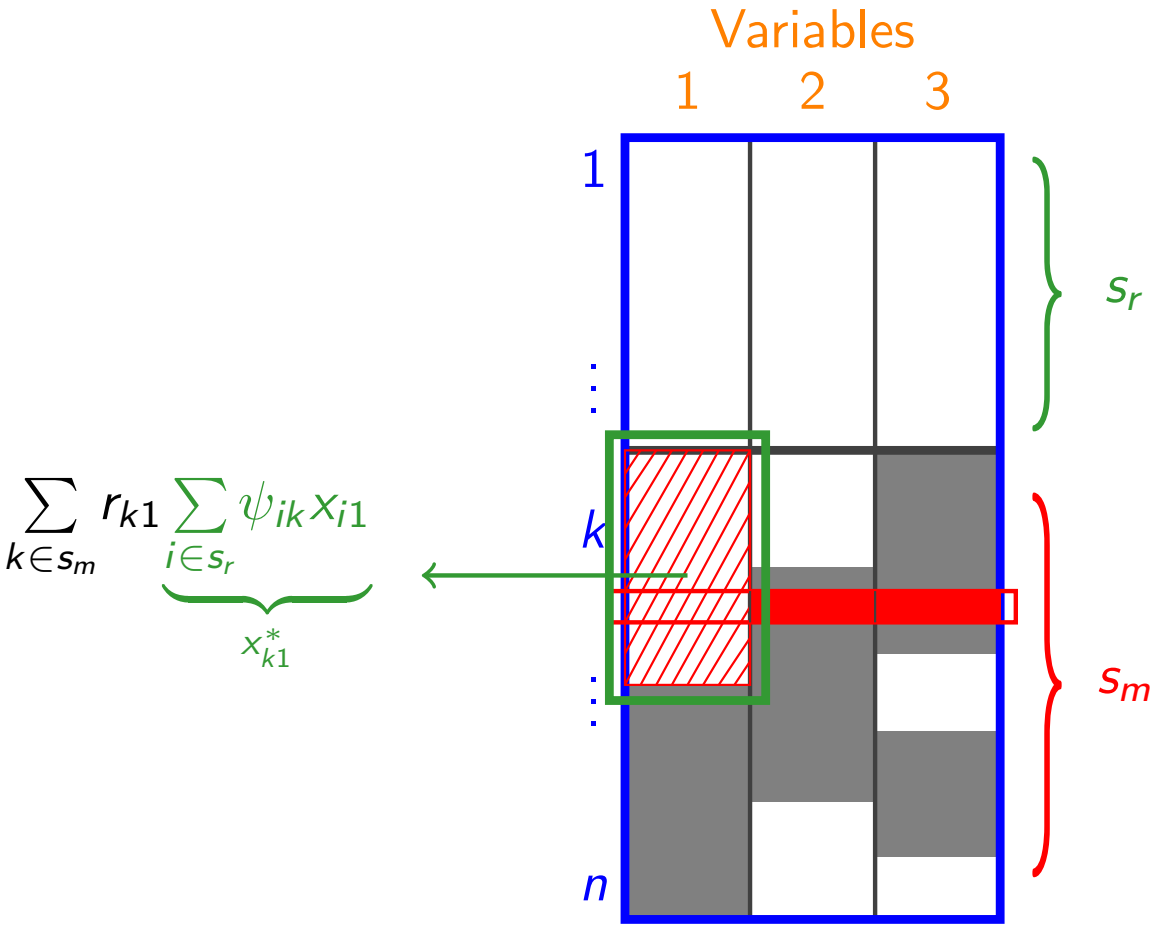

Variables
1
2
3
1
k
n
s_r
s_m
$\sum_{k \in s_m} r_{k1} x_{k1}$


Variables
1
2
3
1
k
n
s_r
s_m


Variables
1
2
3
1
k
n
s_r
s_m


Variables
1
2
3
1
k
n
s_r
s_m


Variables
1
2
3
1
k
n
s_r
s_m


Variables
1
2
3
1
k
n
s_r
s_m
$\sum_{k \in s_m} r_{k1} \underbrace{\sum_{i \in s_r} \psi_{ik} x_{i1}}_{x^*_{k1}}$

	Dataset incomplete		Dataset imputed		Dataset totally imputed	
	x_1	x_2	x_1	x_2	x_1	x_2
s_r	8	4	8	4	8	4
	6	2	6	2	6	2
s_m	?	3	*8*	3	*8*	*4*
	?	3	*8*	3	*8*	*4*
Total in s_m		6		6		8

	Dataset incomplete		Dataset imputed		Dataset totally imputed	
	x_1	x_2	x_1	x_2	x_1	x_2
s_r	8	4	8	4	8	4
	6	2	6	2	6	2
s_m	?	3	*8*	3	*8*	*4*
	?	3	*8*	3	*8*	*4*
Total in s_m		6		6		8
			8	4	8	4
			6	2	6	2
			6	3	*6*	*2*
			6	3	*6*	*2*
				6		4

	Dataset incomplete		Dataset imputed		Dataset totally imputed	
	x_1	x_2	x_1	x_2	x_1	x_2
s_r	8	4	8	4	8	4
	6	2	6	2	6	2
s_m	?	3	*8*	3	*8*	*4*
	?	3	*8*	3	*8*	*4*
Total in s_m		6		6		8
			8	4	8	4
			6	2	6	2
			6	3	*6*	*2*
			6	3	*6*	*2*
				6		4
			8	4	8	4
			6	2	6	2
			8	3	*8*	*4*
			6	3	*6*	*2*
				6		6

Matrix of imputation probabilities

(iv) For $j = 1, \ldots, J$,

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

Matrix of imputation probabilities

(iv) For $j = 1, \ldots, J$,

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

Matrix of imputation probabilities

(iv) For $j = 1, \ldots, J$,

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

$$\sum_{i \in s_r} \underbrace{\left(\sum_{k \in s_m} r_{kj} \psi_{ik} \right)}_{\text{Calibration weights}} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}.$$

Matrix of imputation probabilities

(iv) For $j = 1, \ldots, J$,

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

$$\sum_{k \in s_m} r_{kj} \sum_{i \in s_r} \psi_{ik} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

$$\sum_{i \in s_r} \underbrace{\left(\sum_{k \in s_m} r_{kj} \psi_{ik} \right)}_{\text{Calibration weights}} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}.$$

The imputation probabilities ψ_{ik} can be found using calibration methods.

Calibration within the framework of survey sampling

Auxiliary information is used to reweight a set of units. Consider:

- A population of units U;
- A random sample $s \subset U$;
- d_k, the sampling weight of unit $k \in s$;
- x_k, the value of variable x observed for $k \in s$;
- The known total of x.

The estimator $\widehat{X}_d = \sum_{k \in s} d_k x_k$ can be different from X. The units of s are reweighted in order to satisfy

$$\widehat{X}_w = \sum_{k \in s} w_k x_k \quad = \quad \sum_{k \in U} x_k = X,$$

where w_k is the calibration weights of unit k.

Calibration within the framework of survey sampling

$$\widehat{X}_w = \sum_{k \in s} w_k x_k \quad = \quad \sum_{k \in U} x_k = X.$$

$\widehat{X}_d = \sum_{k \in s} d_k x_k$ is an unbiased estimator of X.

- w_k should be close to d_k for minimize the bias due to calibration.

Deville et Särndal (1992) propose

$$w_k = d_k F_k(\lambda x_k),$$

where $F(.)$ depends on a pseudo-distance between w_k, d_k and λ is a Lagrange multiplier.

Matrix of imputation probabilities

Calibration weights

$w_k = d_k F_k(\lambda x_k)$, where $F(.)$ depends on a pseudo-distance between w_k and d_k.

(iv) For $j = 1, \ldots, J$,

$$\sum_{i \in s_r} \underbrace{\left(\sum_{k \in s_m} r_{kj} \psi_{ik} \right)}_{\text{Calibration weights}} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}.$$

The imputation probabilities ψ_{ik} are the researched calibration weights. We need initial weights ψ_{ik}^0 and a pseudo-distance.

Matrix of imputation probabilities

- Initial weights:

$$\psi_{ik}^0 = \begin{cases} \frac{1}{K} & \text{if } i \in \text{Kpp}(k), \\ 0 & \text{sinon.} \end{cases}$$

- Calibration weights:

$$\psi_{ik} = \psi_{ik}^0 \exp\left(\sum_{j=1}^{J} \lambda_j r_{kj} x_{ij}\right),$$

where $\lambda_1, \ldots, \lambda_J$ are Lagrange multipliers which can be found using an algorithm.

- Reminder: for $j = 1, \ldots, J$,

$$\sum_{i \in s_r} \sum_{k \in s_m} r_{kj} \psi_{ik} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}.$$

Algorithme 1 Imputation probabilities for the J variables simultaneously

Initialize:

- Transform ψ in a vector ψ_v of size $n_r n_m$ and use ψ_{ik}^0,
- Create a matrix $(n_r n_m \times J)$ with the calibration variables $(r_{kj} x_{ij})$,
- Create the vector of J known totals.

Iterate:

1: *Calibrate* for respecting the calibration constraints simultaneously for the J variables.
2: *Normalize* in such a way that each column sums up to 1.
3: *Repeat* steps 1 and 2 until ψ_v almost perfectly meets all requirements.

End:

Transform vector ψ_v into a matrix of imputation probabilities.

Matrix of imputation probabilities

For $j = 1, \ldots, J$,

$$\sum_{i \in s_r} \sum_{k \in s_m} r_{kj} \psi_{ik} r_{ij} x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}$$

$$\sum_{i \in s_r} \left(\sum_{k \in s_m} r_{kj} \psi_{ik} \right) x_{ij} = \sum_{k \in s_m} r_{kj} x_{kj}.$$

Algorithm 2 Probabilities for the J variables in a sequential way

Initialise

- Probability ψ_{ik} initialised by ψ_{ik}^0,
- For $j = 1, \ldots, J$, a vector of calibration variables: $(x_{1j}, \ldots, x_{n_r j})^\top$,
- For $j = 1, \ldots, J$, use the known total.

Iterate

1: *Calibrate* for each variable, one at a time. *Repeat* until ψ meets the calibration constraints.
2: *Normalize* for each column to sum to 1.
3: *Repeat* steps 1 and 2 until ψ statisticians almost perfectly all the requirements.

Terminer:

Matrix ψ contains imputation probabilities.

Imputation matrix

Matrix of imputation probabilities

$$\psi = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0.3 & 0 & 0.4 \\ 0.2 & 0 & 0.1 \end{pmatrix}$$

Imputation matrix

$$\phi = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- ϕ_{ik}: 1 if unit i is selected as the donor of unit k, 0 otherwise.
- Only one donor is selected by nonrespondant, $\sum_{i \in s_r} \phi_{ik} = 1$.
- Requirement (iv): the donors must be chosen is such a way that

$$\sum_{k \in s_m} \sum_{i \in s_r} \phi_{ik} \, r_{kj} x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} \psi_{ik} \, r_{kj} x_{ij} \left(= \sum_{k \in s_m} r_{kj} x_{kj} \right).$$

Imputation matrix

- Stratified sampling:

Matrix of imputation probabilities

$$\psi = \left(\begin{array}{c|c|c} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0.3 & 0 & 0.4 \\ 0.2 & 0 & 0.1 \end{array}\right)$$

Imputation matrix

$$\phi = \left(\begin{array}{c|c|c} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

A column (a nonrespondant) corresponds to a stratum.

Imputation matrix

- Stratified sampling:

Matrix of imputation probabilities

$$\psi = \left(\begin{array}{c|c|c} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0.3 & 0 & 0.4 \\ 0.2 & 0 & 0.1 \end{array}\right)$$

Imputation matrix

$$\phi = \left(\begin{array}{c|c|c} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

A column (a nonrespondant) corresponds to a stratum.

- Stratified balanced sampling:
  - select one donor per stratum ,
  - satisfy the requirement (iv)

$$\sum_{k \in s_m} \sum_{i \in s_r} \phi_{ik} \, r_{kj} x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} \psi_{ik} \, r_{kj} x_{ij}.$$

Balanced sampling

Auxiliary information is used for the selection of a sample. Consider

- A population of unit U;
- A sample $s \subset U$;
- π_k the probability of selecting k from the sample;
- x_k the value of the observed variable x for $k \in U$;
- Le total de x is known.

A sampling design is balanced on the variable x if

$$\widehat{X} = \sum_{k \in s} \frac{x_k}{\pi_k} \quad = \quad \sum_{k \in U} x_i = X.$$

Methods for selecting a balanced sample

- The cube method (Deville et Tillé, 2004).
  Fixed sample size if $\sum_{k \in U} \pi_k$ is an integer.
- Balanced sample for stratified population (Chauvet, 2009);
- Balanced sample for highly stratified population (Hasler et Tillé, 2014).
  Sample size fixe into the strata.

Imputation matrix

- n_m strata (nonrespondants) defined by;
-

$$\sum_{k \in s_m} \sum_{i \in s_r} \phi_{ik} \, r_{kj} x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} \psi_{ik} \, r_{kj} x_{ij}$$

$$\sum_{k \in s_m} \sum_{i \in s_r} \frac{\phi_{ik}}{\psi_{ik}} \psi_{ik} r_{kj} x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} \psi_{ik} \, r_{kj} x_{ij}.$$

- A donor is chosen per stratum $k \in s_m$.
- Inclusion probability used in stratified balanced sampling is ψ_{ik};
- Associated balancing variable is $\psi_{ik} \, r_{kj} x_{ij}$.

Imputation

Imputed value: $$x_{kj}^{*} = \sum_{i \in s_r} \phi_{ik} x_{ij}$$

Deterministic variant: $x_{kj}^{*} = \sum_{i \in s_r} \psi_{ik} x_{ij}$.

Description of the dataset

- Population `MU284` de Särndal et coll. (1992),
- Available in the `R sampling` package (Tillé et Matei, 2007),
- Population of $N = 284$ Swedish municipalities,
- Variables:
  - x_1: population in 1985, in thousands (`P85`);
  - x_2: population in 1975, in thousands (`P75`);
  - x_3: 1985 municipal tax bill income, in millions of SEK (`RMT85`);
  - x_4: number of conservative party seats on city council (`CS582`).

Simulations

- p_{ij}: responds to the variable j, generated according to a logistic model;
- r_{ij} equals 1 with probability p_{ij}, 0 otherwise;
- Realised imputation methods:
  - random hot deck,
  - K nearest neighbors (K-PP),
  - K balanced nearest neighbors (balanced K-PP),
  - deterministic version of the balanced K nearest neighbors (deterministic balanced K-PP);
- Estimation of the parameters:
  - for the 4 variables: total, 10^{th}, 50^{th} et 90^{th} percentiles,
  - variance-covariance matrix.

Simulations

- Simulations:
  - generate $M_R = 100$ times the matrix of response/nonresponse;
  - impute $M_I = 100$ times for each nonresponse.
- For each imputation method and for each parameter θ, the mean square error of the imputed estimator $\widehat{\theta}_I$ is

$$\mathrm{MSE}(\widehat{\theta}_I) = \frac{1}{M_R}\frac{1}{M_I}\sum_{r=1}^{M_R}\sum_{i=1}^{M_I}(\widehat{\theta}_I^{r,i} - \theta)^2,$$

  where $\widehat{\theta}_I^{r,i}$ is the estimator of θ at simulation r, i;
- Comparison of the 3 methods (K-PP, balanced K-PP, deterministic balanced K-PP) to the random hot deck:

$$\frac{\mathrm{MSE}(\widehat{\theta}_{I,\text{method}})}{\mathrm{MSE}(\widehat{\theta}_{I,\text{hot deck}})} \times 100.$$

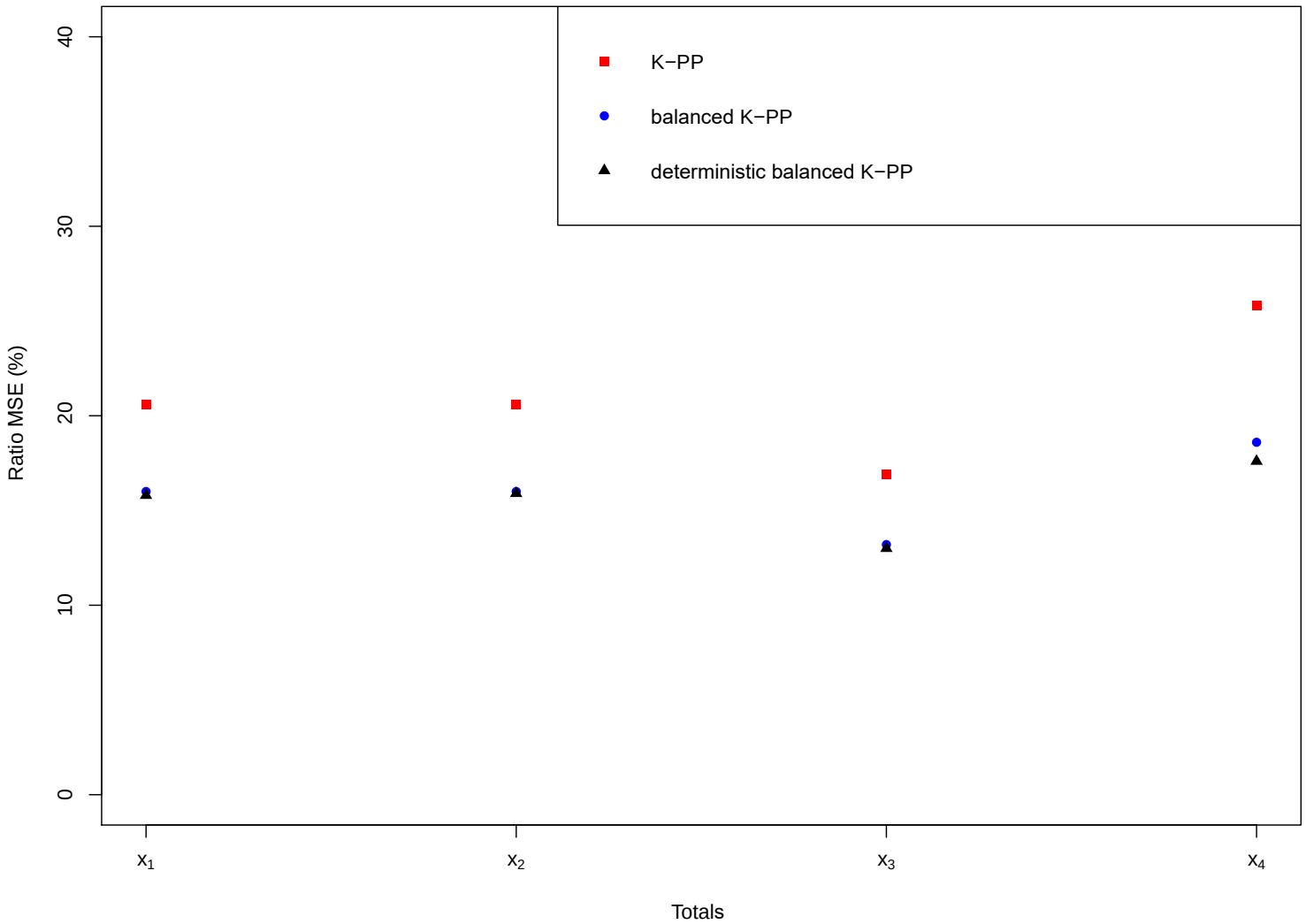


Figure: MSE of the three imputation methods divided by MSE of the hot deck for the total each variable.

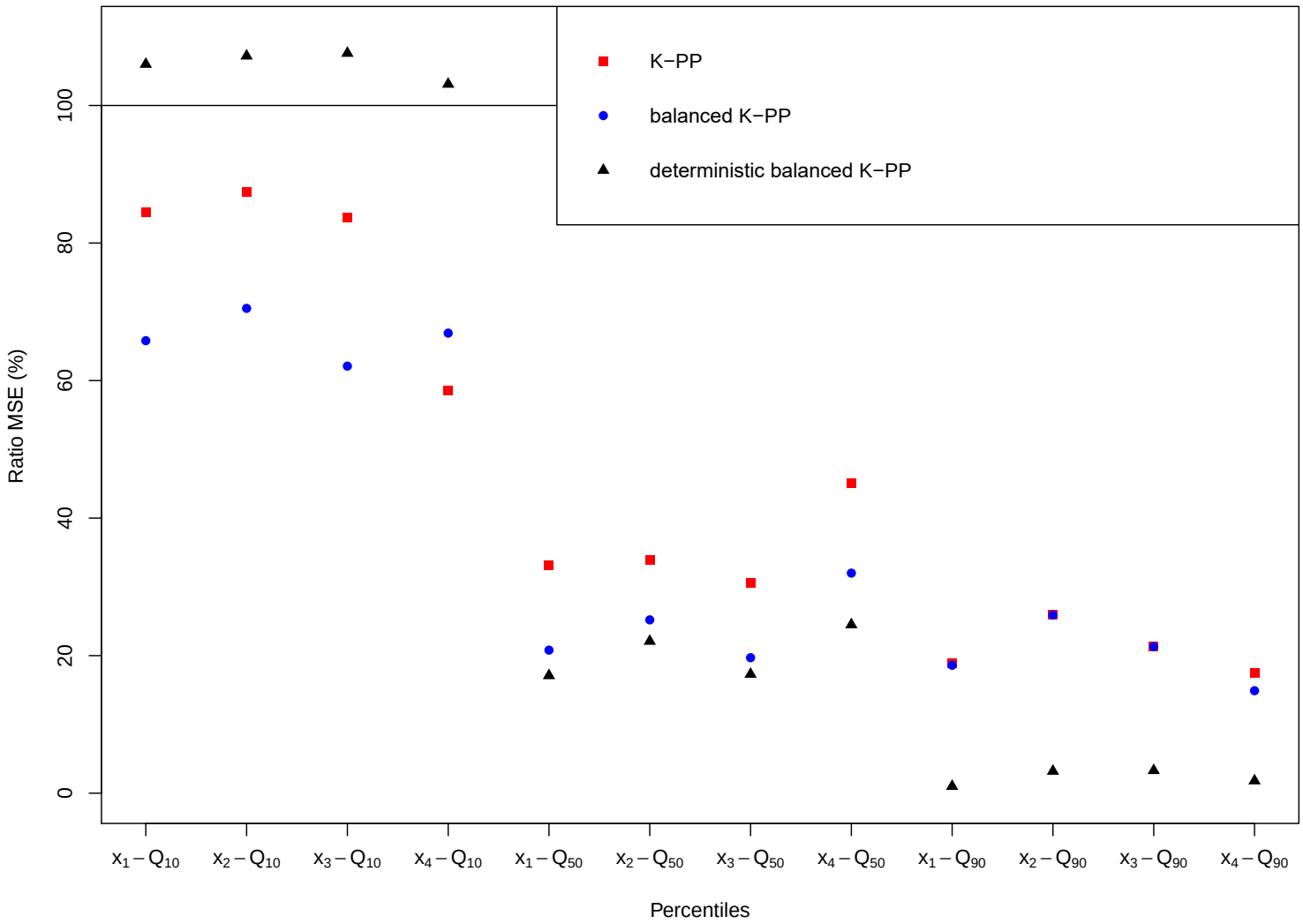


Figure: MSE of the three imputation methods divided by MSE of the hot deck for the 10^{ith}, the 50^{th} and the 90^{ith} percentile of each variable.

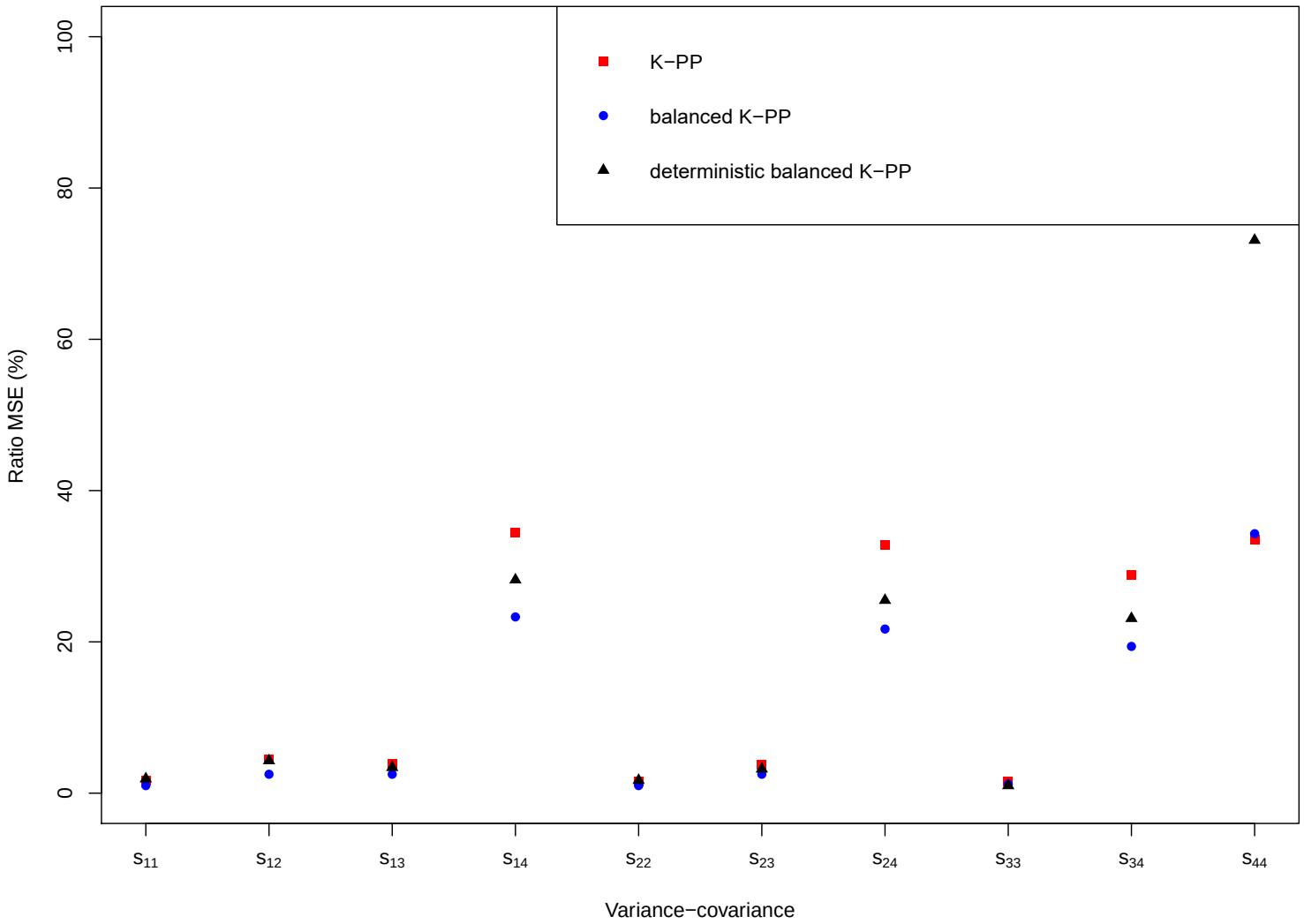


Figure: MSE of the three imputation methods divided by MSE of the hot deck for the vairance-covariance matrix.

Discussion

- Choice of K;
- Qualitative and quantitative Variables;
- Possibility to force $\psi_{ik} = 0$;
- Program `R`;
- Estimation of the variance;
- Additional research.

Andridge, R. R. et Little, R. J. A. (2010). A review of dot deck imputation for survey non-response. *International Statistical Review*, **78**, 40–64.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, **35**, 115–119.

Deville, J.-C. et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Deville, J.-C. et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, **91**, 893–912.

Hasler, C. et Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, **74**, 81–94.

Hasler, C. et Tillé, Y. (2016). Balanced k-nearest neighbor imputation. *Statistics*, **105**, 11–23.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium*, 97. Statistics Canada.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. et Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, **27**, 85–95.

Särndal, C.-E., Swensson, B. et Wretman, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Tillé, Y. et Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, http://cran. r-project. org/, Manual of the Contributed Packages.

The cube method (Deville et Tillé, 2004):

Sequence of random vectors updating the inclusion probability vector. $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_k, \cdots, \pi_N)^\top$ until the sample $s = (0, 1, 0, \ldots, 0, 1)^\top$ is selected.