The integration of educational attainment data from administrative and survey data

(part 1)

Sander Scholtus

22 August 2023

Overview

- Educational Attainment File
- Different estimators
- Variance estimation

Educational Attainment File

- Educational Attainment File (EAF) is intended to be used for all statistical analyses on education at CBS
- Main target variable: **highest attained education level**
- EAF is produced annually, reference date 1 October
- Combination of administrative and survey data

Educational Attainment File

Main potential data sources on educational attainment:

- Labour Force Survey (LFS)
  - sample survey
  - (used to have) retrospective questions on education
  - data used from 2004 onwards
- Central registers on education:
  - Higher education: registration started in 1983
  - Secondary education: registration started in the late 1990s
  - Primary education: registration started around 2010
  - Only publicly financed education in the Netherlands

Educational Attainment File

Available input data (fictional example):

Person-ID	Date of birth	Type of education	Start date	End date	Certificate	Source
001	02-02-1986	Lower secondary vocational	----------	25-06-2002	Yes	Register A
001	02-02-1986	?	01-08-2002	31-07-2003	?	Register B
001	02-02-1986	?	01-08-2003	31-07-2004	?	Register B
001	02-02-1986	Non-academic tertiary bachelor	01-09-2007	31-08-2008	No	Register C
001	02-02-1986	Non-academic tertiary bachelor	01-09-2008	31-08-2009	No	Register C
001	02-02-1986	Non-academic tertiary bachelor	01-09-2009	31-08-2010	Yes	Register C
002	15-12-1989	Primary education	01-09-1994	30-06-2002	Yes	LFS 2004
002	15-12-1989	Secondary general	01-09-2002	----------	No	LFS 2004

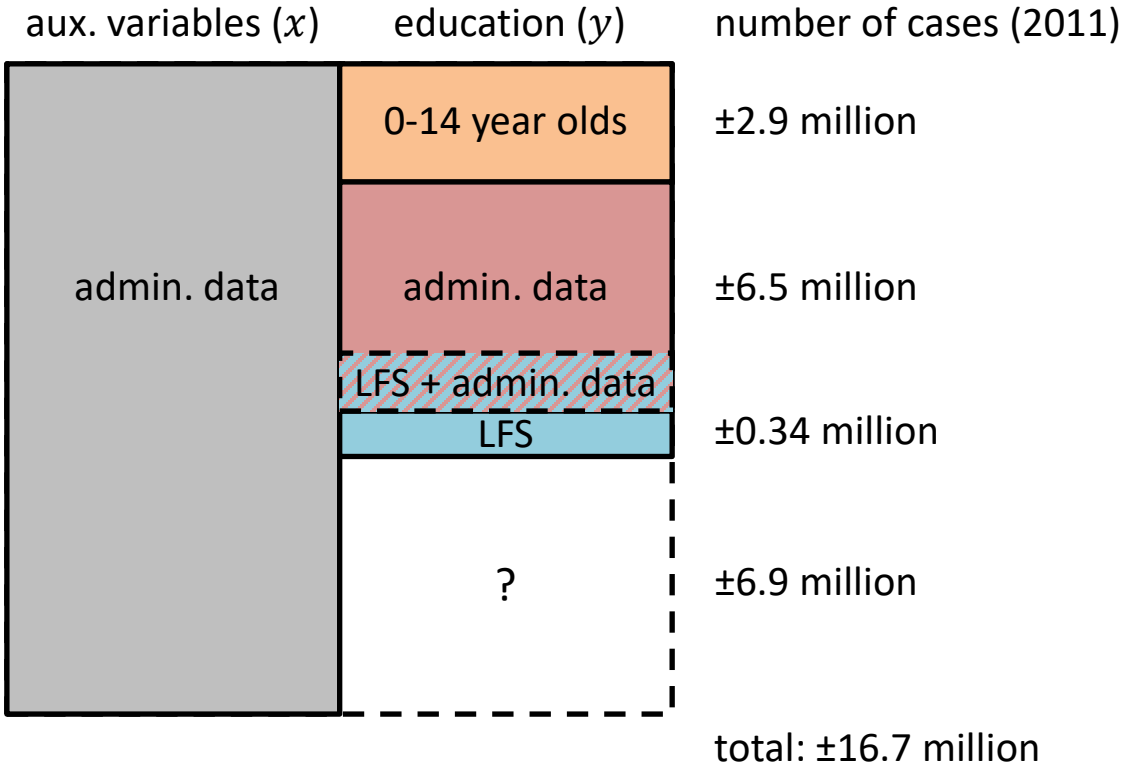
(based on: Linder, van Roon & Bakker, 2011)

Educational Attainment File

Micro integration to create EAF at reference time T

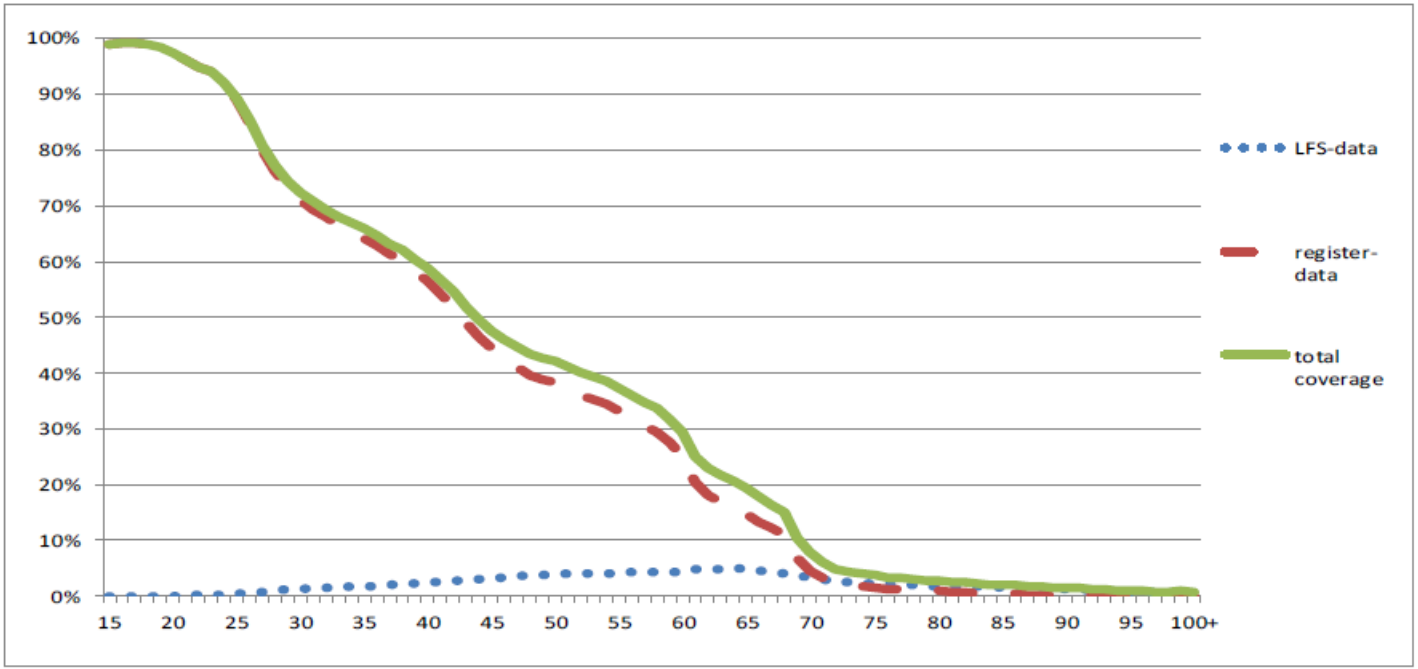
- Which observed values are (still) valid at time T?
  - Deterministic rules:
    - Example: if a person has reached the highest possible education level at time T_0, then this remains valid for all $t \geq T_0$
    - Example: “downgrading”
  - Probabilistic rules:
    - Non-parametric survival analysis on historical LFS data
    - Rule: Probability that education level recorded at time T_0 is still correct after a period $T - T_0$ has passed should be at least 0.95
- Choose highest valid education level for each person

Educational Attainment File



Educational Attainment File

Population coverage by age for the EAF of 2011 (for age ≥ 15 years)

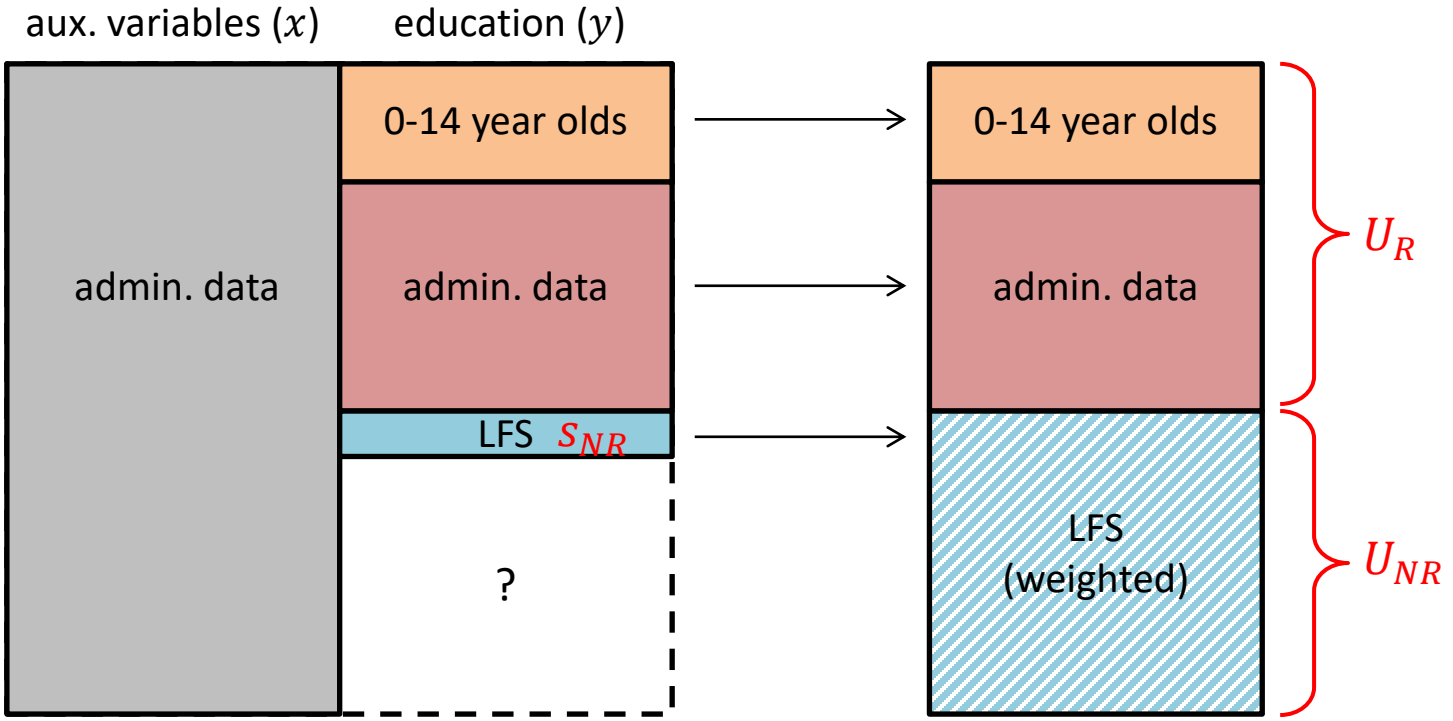


(source: De Waal, Daalmans & Linder, 2020)

Different estimators

- Weighting (current production method)
- Mass imputation
- Combined model-based estimator

Weighting


aux. variables (x)
education (y)
admin. data
0-14 year olds
admin. data
LFS s_{NR}
?
0-14 year olds
admin. data
LFS
(weighted)
U_R
U_{NR}

Weighting

Let $y_{ci} = 1$ if person i has education level c and $y_{ci} = 0$ otherwise.

Target parameter:

$$\theta_c = \sum_{i \in U} y_{ci} = \sum_{i \in U_R} y_{ci} + \sum_{i \in U_{NR}} y_{ci}$$

Weighted estimator:

$$\hat{\theta}_{cW} = \sum_{i \in U_R} y_{ci} + \sum_{i \in s_{NR}} w_i y_{ci}$$

Weighting

$$\hat{\theta}_{cW} = \sum_{i \in U_R} y_{ci} + \sum_{i \in s_{NR}} w_i y_{ci}$$

Weights w_i determined by **regression estimator**
(special case of **calibration**; Deville & Särndal, 1992):

- Starting weights v_i taken from LFS (design weights corrected for non-response in original survey)
- Final weights $w_i = v_i g_i$
- Minimize $\sum_{i \in s_{NR}} (w_i - v_i)^2 / v_i = \sum_{i \in s_{NR}} v_i (g_i - 1)^2$ under the restriction that $\sum_{i \in s_{NR}} w_i \mathbf{x}_i = \sum_{i \in U_{NR}} \mathbf{x}_i$ for auxiliary variables

Weighting

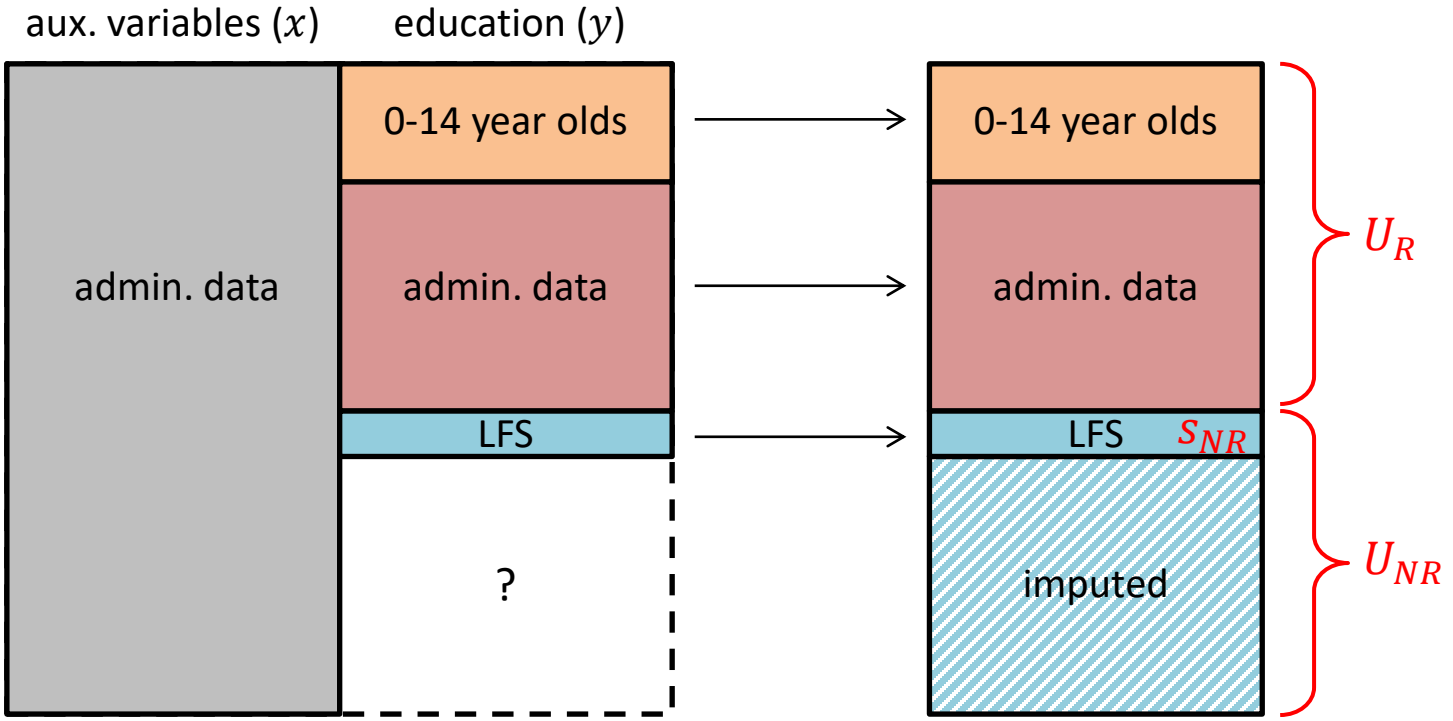
$$\hat{\theta}_{cW} = \sum_{i \in U_R} y_{ci} + \sum_{i \in s_{NR}} w_i y_{ci}$$

Auxiliary variables used for calibration include:

- demographic variables (gender, age, marital status, region, …)
- socio-economic variables (income level, type of income, …)

More details: see Linder, van Roon & Bakker (2011)

Mass imputation



Mass imputation

Target parameter:

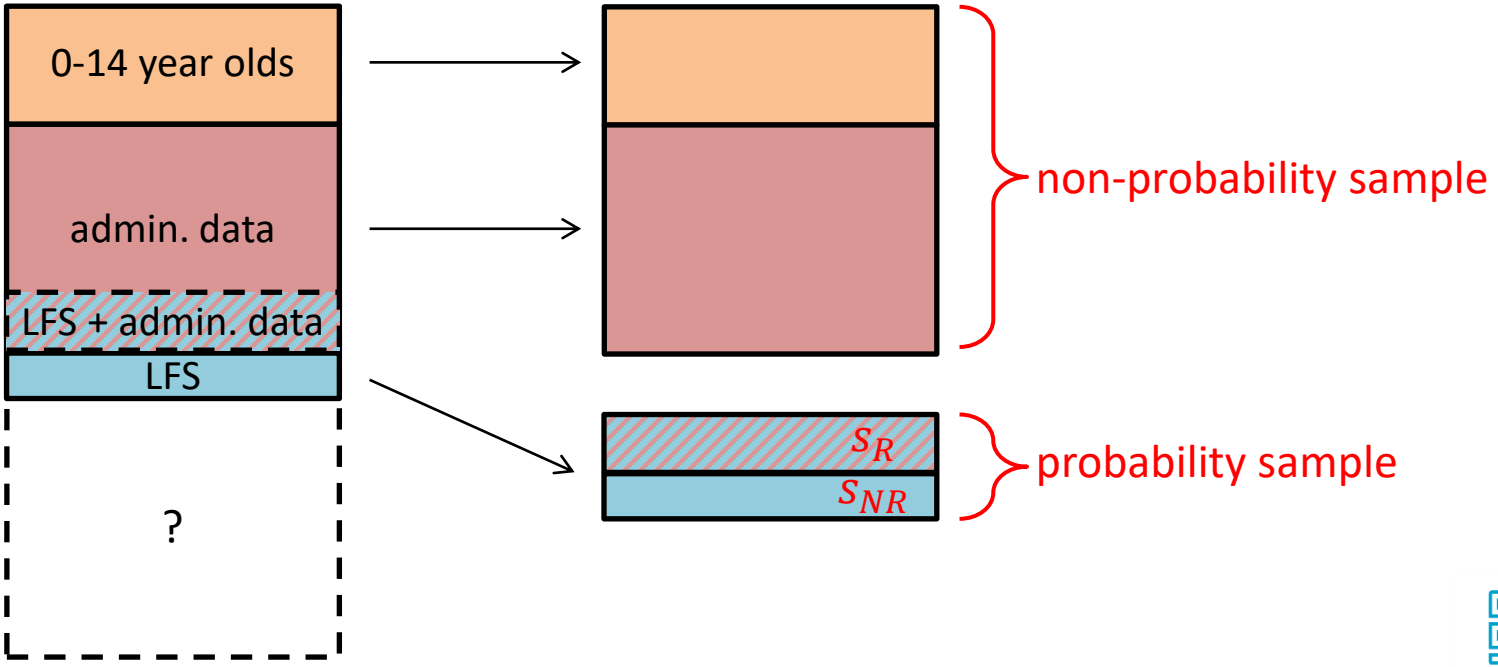
$$\theta_c = \sum_{i \in U} y_{ci} = \sum_{i \in U_R} y_{ci} + \sum_{i \in U_{NR}} y_{ci}$$

Mass-imputed estimator:

$$\hat{\theta}_{cI} = \sum_{i \in U_R} y_{ci} + \sum_{i \in s_{NR}} y_{ci} + \sum_{i \in U_{NR} \setminus s_{NR}} \tilde{y}_{ci}$$

$\tilde{y}_{ci}$ based on an imputation model estimated on s_{NR}

Combined model-based estimator



Combined model-based estimator

Individual estimators:

$$\hat{\theta}_c^{(NP)} = \sum_{i \in U_R} y_{ci}$$

$$\hat{\theta}_c^{(P)} = \sum_{i \in s_R \cup s_{NR}} v_i y_{ci}$$

Combined estimator (Villalobos-Aliste, 2022; see morning lecture):

$$\hat{\theta}_c = W_c \hat{\theta}_c^{(P)} + (1 - W_c)\hat{\theta}_c^{(NP)}$$

$$W_c = \frac{EMSE\left(\hat{\theta}_c^{(NP)}\right)}{EMSE\left(\hat{\theta}_c^{(NP)}\right) + EMSE\left(\hat{\theta}_c^{(P)}\right)}$$

Combined model-based estimator

Experimental application to EAF 2019 (Villalobos-Aliste, 2022):

Municipality	n(P)	n(NP)	regular EAF estimates (%)			only non-prob. sample (%)			combined estimator (%)		
			L	M	H	L	M	H	L	M	H
Amsterdam	11 051	473 324	24	29	46	23	30	46	24	29	46
Amstelveen	1 365	37 815	23	33	44	20	34	45	24	33	43
Krimpenerwaard	1 128	23 925	34	43	23	29	44	27	34	42	23
Medemblik	845	19 961	35	45	20	34	43	23	35	42	23
Teylingen	684	16 990	26	42	33	25	39	37	26	43	31
Boxtel	683	14 174	34	40	26	31	40	30	31	44	24
Wassenaar	336	10 177	26	33	41	25	33	42	26	33	41
Sluis	448	8 822	35	46	19	32	45	23	34	48	17
West Maas en Waal	329	8 900	33	46	21	30	45	26	30	44	26
Ouder-Amstel	264	6 552	23	34	43	21	33	45	21	34	46
Terschelling	60	2 474	27	51	22	25	49	25	26	49	25

Variance estimation

- Accuracy of weighted / mass-imputed estimator
  - Analytical variance approximation
  - Bootstrap method

Variance estimation: analytical approximation

Weighted estimator for a subpopulation:

$$\hat{\theta}_{cW}(g) = \sum_{i \in U_R(g)} y_{ci} + \sum_{i \in s_{NR}(g)} w_i y_{ci} \equiv \hat{\theta}_{cW,R}(g) + \hat{\theta}_{cW,NR}(g)$$

For small subpopulations, the following simple variance estimator turns out to work quite well in practice:

$$\widehat{\text{var}}\left(\hat{\theta}_{cW}(g)\right) \approx \frac{\hat{\theta}^2_{cW,NR}(g)}{n_{NR}(g)}\left\{1 + CV^2_{W,NR}(g)\right\}$$

Mass-imputed estimator: see Scholtus & Daalmans (2021)

Variance estimation: bootstrap

- Classical bootstrap does not account for
  - Finite-population sampling
  - Complex survey design
- Different extensions of the bootstrap available (Mashreghi et al., 2016)
- Here: extension based on pseudo-populations

Variance estimation: bootstrap

Simplifying assumption: starting weight $d_i = 1/\pi_i$ is integer-valued

Bootstrap algorithm:

1. Create a pseudo-population $\widehat{U}^*$ by taking d_i copies of each unit $i \in s = s_R \cup s_{NR}$.

2. For each $b = 1, \ldots, B$ do the following:
   - Draw sample s_b^* from $\widehat{U}^*$ analogous to design used to draw s from U.
   - Analogously to $\hat{\theta} = t(s, U_R)$, construct replicate $\hat{\theta}_b^* = t(s_b^*, U_R)$.

3. Compute the variance estimate for $\hat{\theta}$ based on pseudo-population $\widehat{U}^*$ as:
   $\widehat{\mathrm{var}}_{\mathrm{boot}}(\hat{\theta} - \theta) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2$, with $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$.

Extension to non-integer weights:
see Kuijvenhoven & Scholtus (2011) or Scholtus & Daalmans (2021)

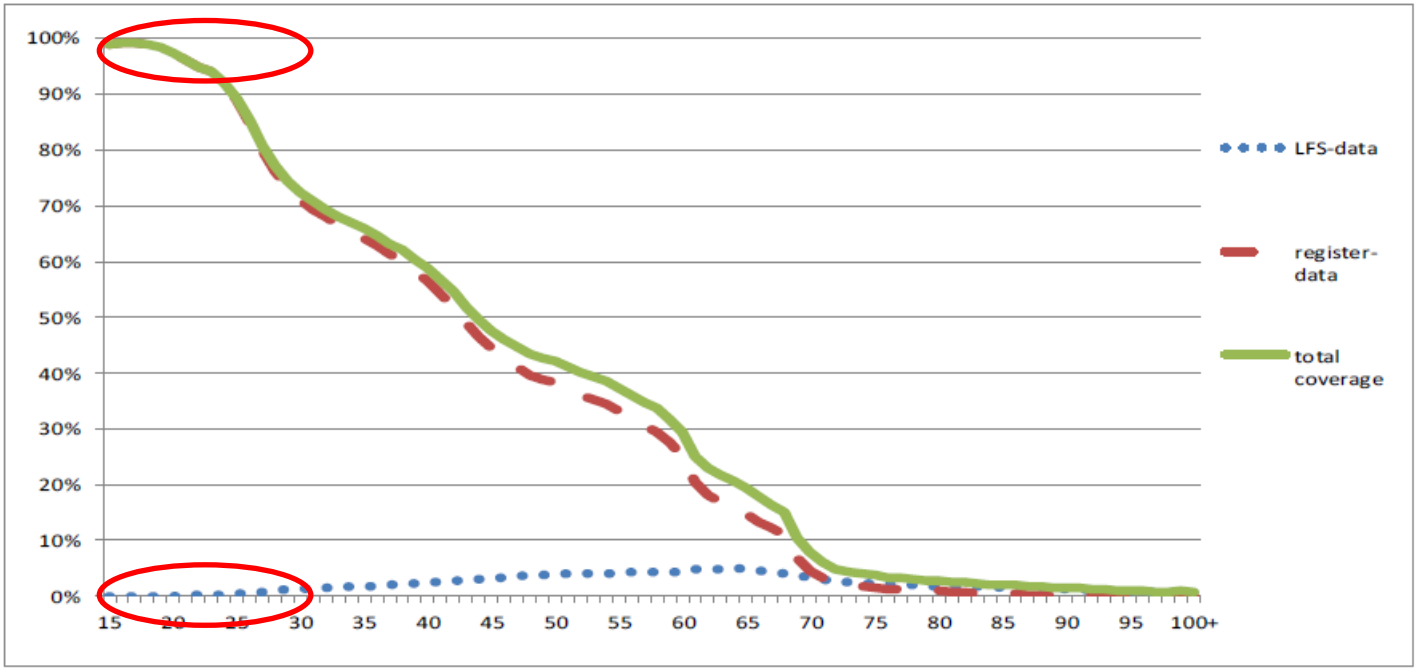
Variance estimation: bootstrap

Key step: Analogously to $\hat{\theta} = t(s, U_R)$, construct replicate $\hat{\theta}_b^* = t(s_b^*, U_R)$

- Example: weighted estimator
- Original estimator:
  $\hat{\theta}_{cW} = \sum_{i \in U_R} y_{ci} + \sum_{i \in s_{NR}} w_i y_{ci}$
- Construction of bootstrap replicate $\hat{\theta}_{cW,b}^*$:
  - $\hat{U}_{NR}^*$ is the subpopulation of $\hat{U}^*$ consisting of copies of units from s_{NR}
  - $s_{NR,b}^* = s_b^* \cap \hat{U}_{NR}^*$; note: size of overlap is random
  - Compute final weights $w_{k,b}^*$ for units in $s_{NR,b}^*$ by calibration, using known totals of auxiliary variables in $\hat{U}_{NR}^*$
  - Compute: $\hat{\theta}_{cW,b}^* = \sum_{k \in U_R} y_{ck} + \sum_{k \in s_{NR,b}^*} w_{k,b}^* y_{ck}$

Educational Attainment File

Ongoing work…



(source: De Waal, Daalmans & Linder, 2020)

References

J.-C. Deville & C.-E. Särndal (1992), Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* **87**, 376–382.

L. Kuijvenhoven & S. Scholtus (2011), Bootstrapping Combined Estimators based on Register and Sample Survey Data. Discussion Paper, Statistics Netherlands, The Hague.

F. Linder, D. van Roon & B. Bakker (2011), Combining Data from Administrative Sources and Sample Surveys; the Single-Variable Case. Case Study: Educational Attainment. Report for Work Package 4.2 of the ESSnet project Data Integration. URL: https://ec.europa.eu/eurostat/cros/content/wp4-case-studies_en

Z. Mashreghi, D. Haziza & C. Léger (2016), A Survey of Bootstrap Methods in Finite Population Sampling. *Statistics Surveys* **10**, 1–52.

S. Scholtus & J. Daalmans (2021), Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data. *Journal of Official Statistics* **37**, 433–459.

S. Villalobos-Aliste (2022), Combining Probability and Non-Probability Samples for Estimation. Master thesis, Utrecht University.

T. de Waal, J. Daalmans & F. Linder (2020), Mass imputation for Census estimation: Methodology. Report, CBS, The Hague.