Dealing with erroneous enumeration and misplacement in registers

Li-Chun Zhang[1,2]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***
[2] ***Statistisk sentralbyrå, Norway***
[3] ***Universitetet i Oslo***

Erroneous enumeration & misplacement in admin sources

Main types of over-counting:

- Duplicates (negligible provided CPR)
- Misplacement: inter-locality over-/under-counting at once
- Erroneous (overall): out-of-scope or non-existent individuals

Erroneous enumeration with or without CPR, e.g.

- Estonia: about 2.3% under-count in 2011 Census, 3% over-count in CPR (Tiit & Maasing, 2016)
- UK: Patient Register about 4% over adjusted 2011 Census count (ONS, 2013)

Misplacement can be a major issue in CPR, e.g.

- Israel: Integrated Census 2008 (Nirel and Glickman, 2009)
- Norway: register-based household statistics

Problem in the absence of under-coverage overall

Population	Statist. Population Dataset					
Locality	1	$\cdots$	j	$\cdots$	m	Total
Erroneous	N_{01}	$\cdots$	N_{0j}	$\cdots$	N_{0m}	N_{0+}
1	N_{11}	$\cdots$	N_{1j}	$\cdots$	N_{1m}	N_{1+}
$\vdots$	$\vdots$	$\ddots$	$\cdots$	$\cdots$	$\vdots$	$\vdots$
j	N_{j1}	$\cdots$	N_{jj}	$\cdots$	N_{jm}	N_{j+}
$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\vdots$	$\vdots$
m	N_{m1}	$\cdots$	N_{mj}	$\cdots$	N_{mm}	N_{m+}
Total	N_{+1}	$\cdots$	N_{+j}	$\cdots$	N_{+m}	N_{++}

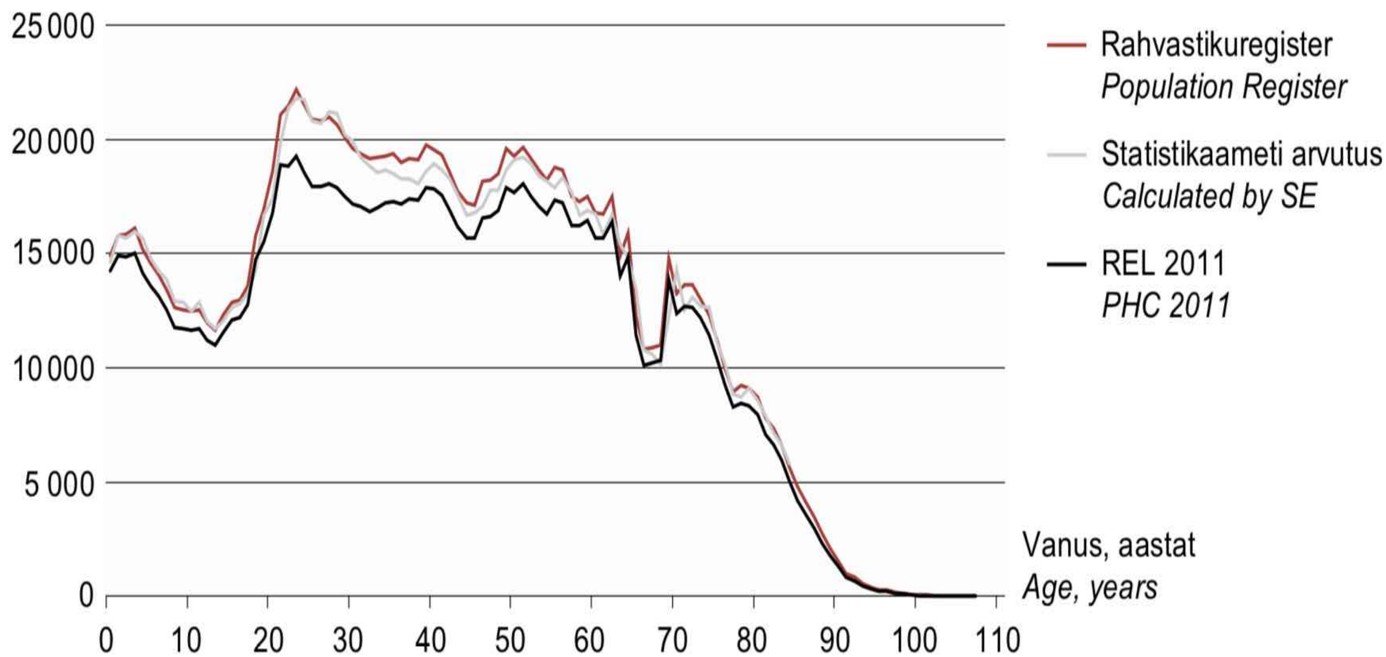
NB. SPD may or may not be the CPR

Known SPD-totals $N_{+1}, ..., N_{+m}$ and $N_{++} = \sum_{j=1}^{m} N_{+j}$

Unknown population total $N = \sum_{i=1}^{m} N_{i+} = N_{++} - N_{0+}$

Residency Index in Estonia (Tiit & Maasing, 2016)

Figure 1. Age distribution of population according to data of Population Register, Statistics Estonia's population calculations based on PHC 2000, and data of PHC 2011, 1 January 2012



Census under-count adjustment using 12 admin registers incl. CPR

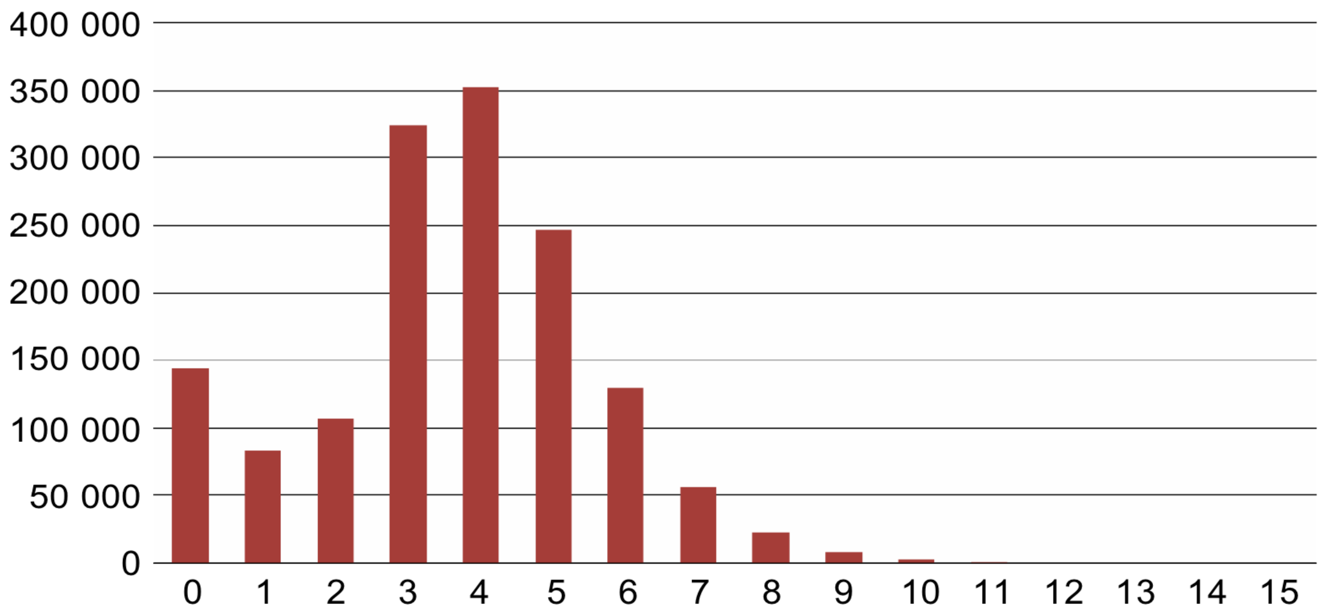
- Use PIN for identification: duplicates/linkage errors negligible
- Find people in various registers who are missing in census
- Regression modelling to obtain probability of missing in census

Residency Index in Estonia (Tiit & Maasing, 2016)

Sign-of-Life (SoL) register sources

- based on *events* in given time duration (e.g. a calendar year)
  e.g. Dunne (2015), Zhang and Dunne (2017) for approach in Ireland
- 27 SoL-registers: special care, parental leave, dental care... digital prescription... prison visit, change of vehicle, ..., residence permit

Figure 3. Distribution of simple sum of signs of life, 2015



Residency Index in Estonia (Tiit & Maasing, 2016)

Construct SPD as *extended* population. For person k in year t, let

$$R(k,t) = d \cdot R(k,t-1) + g \cdot X(k,t-1)$$

$d =$ stability rate: for classifier by threshold-c, choose $d^2 < c < d$

$g =$ SoL rate: for minimum impact, $g(1 + d + \cdots + d^h) > c$ given h

$$X(k,t) = \sum_{\ell=1}^{q} a_\ell \delta_\ell(k,t)$$

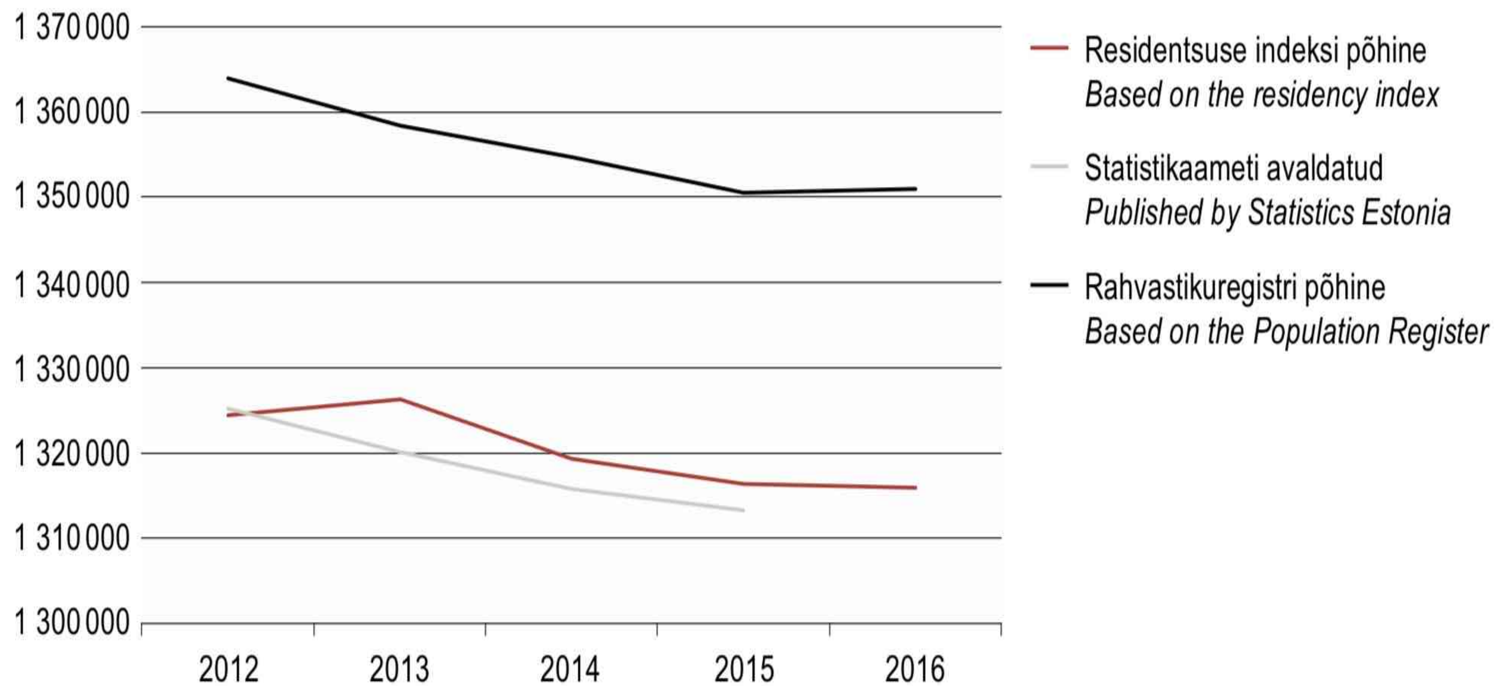
$\delta_\ell(k,t) = 1$ if there is sign of life in source ℓ, and 0 otherwise

$a_\ell =$ weight of source ℓ, e.g. special care more powerful than pension

NB. Choice of a_ℓ based on $b_\ell = \sum_{k \in A(t)} \delta_\ell(k,t) / \sum_{k \in B(t)} \delta_\ell(k,t)$, with "almost surely" residents $A(t)$ and non-residents $B(t)$, respectively.

Residency Index in Estonia (Tiit & Maasing, 2016)

Figure 6. Population figure based on calculations of Statistics Estonia, Population Register and residency index, 1 January 2012–2016



Residency-Index-based population statistics since 2016:

2.5% down from CPR count, 0.3% up from trad. method

Fractional counting: A basic theory

- For each $k \in \text{SPD}$, let

$\mathbf{a}_k = q$-vector containing all the available CPR- and SoL-addresses

$\mathbf{z}_k =$ vector of all the relevant auxiliary data, including known family relationships, previous addresses, emigration status, etc.

- Let an address *classifier* be

$$\mathbf{y}_k = g(\mathbf{a}_k, \mathbf{z}_k) \in \{0,1\}^q \qquad \text{where} \quad \mathbf{y}_k^\top \mathbf{1} = 1$$

NB. In case more than one component of $\mathbf{y}_k$ refer to the same address, by convention only one of them is set to 1 if this address is chosen.

- Let an address *predictor* be

$$\boldsymbol{\mu}_k = h(\mathbf{a}_k, \mathbf{z}_k) \in [0,1]^q \qquad \text{where} \quad \boldsymbol{\mu}_k^\top \mathbf{1} = 1$$

NB. The idea is for each component of $\boldsymbol{\mu}_k$ to be probability that the corresponding address is the true *usual resident address*.

Statistical register-based population counts

- Based on the classifier:

$$\widehat{N}^C_{ij} = \sum_{k \in U_j} \mathbf{y}_k^\top \boldsymbol{\delta}_k \quad \text{and} \quad \boldsymbol{\delta}_k = \boldsymbol{\delta}(\mathbf{a}_k \in A_i)$$

U_j = SPD-population in locality j

A_i = the set of admissible addresses in locality i

$\boldsymbol{\delta}(\mathbf{a}_k \in A_i)$ = is the q-vector of 0/1 indicators

- Based on the predictor, or ***fractional counting***:

$$\widehat{N}^P_{ij} = \sum_{k \in U_j} \boldsymbol{\mu}_k^\top \boldsymbol{\delta}_k \quad \text{and} \quad \boldsymbol{\delta}_k = \boldsymbol{\delta}(\mathbf{a}_k \in A_i)$$

Fractional counting: A basic theory

Let adr_k be true usual resident address, for $k \in \text{SPD}$.

Fractional counting is unbiased for N_{i+}, provided

$$\begin{cases} \Pr(\text{adr}_k \in \mathbf{a}_k) = 1 \\ \boldsymbol{\delta}(\text{adr}_k = \mathbf{a}_k) \perp \delta(\text{adr}_k \in A_i) | \mathbf{a}_k, \mathbf{z}_k \end{cases}$$

- The 1st condition is necessary because it is impossible get $\boldsymbol{\mu}_k$ right, where $\boldsymbol{\mu}_k^\top \mathbf{1} = 1$, as long as there are people whose usual resident address is outside the set of available addresses.

- The 2nd condition then implies that the probability of $\boldsymbol{\delta}(\text{adr}_k = \mathbf{a}_k)$ does not depend on $\delta(\text{adr}_k \in A_i)$, i.e. whether k is in locality i.

NB. The matter depends on how good the available addresses are, e.g. how powerful the SoL sources are, and how well $\boldsymbol{\mu}_k$ is estimated.

Fractional counting: A basic theory

Provided the $\boldsymbol{\mu}_k$'s, the prediction variance of fractional counting is

$$V(\widehat{N}_i - N_i) = \sum_{k \in U} \boldsymbol{\mu}_k^\top \boldsymbol{\delta}_k \big(1 - \boldsymbol{\mu}_k^\top \boldsymbol{\delta}_k\big) \tag{1}$$

where it is assumed that $\delta(\text{adr}_k \in A_i)$ is independent across the different persons, conditional on the corresponding $(\mathbf{a}_k, \mathbf{z}_k)$'s.

NB. possible to allow for clustering effects (e.g. family neuclus)

NB. possible to incorporate estimation uncertainty of $\boldsymbol{\mu}_k$ in addition

Some methods of supervised learning:

- Decision rules
- Regression modelling
- Machine Learning Methods

Data for continuous learning

a) The CPR and SoL-registers, basically every time there is an update of either $\mathbf{a}_k$ or $\mathbf{z}_k$ in these sources

b) On-going surveys: introduce a question on adr_k & related protocol

It will be helpful to enhance the *collection*, *organisation* and *usage* of adr_k, for $k \in U$, across the NSO.

In addition, purposely designed Coverage Survey can be used to validate the method of register-based population counts and possibly to provide adjusted counts.

Problem in the presence of under-coverage overall

Some recent developments:

- Residency Index combine over-/under-coverage adjustment, e.g.

$$\sum_{g=1}^{G}\sum_{h=1}^{H} x_{gh}\hat{\beta}_g \frac{n_h}{m_h} = \sum_{k\in A} R_k \quad \text{where } R_k = \hat{\beta}_g \frac{n_h}{m_h} \text{ for } i \in A_{gh}$$

- TDSE: Trimmed Dual System Estimation (Zhang & Dunne, 2017)

$$\hat{N}_k = n\frac{x-k}{m-k_1}$$

- Models: K-lists with both over- and under-coverage, for $K \geq 2$, and S with only under-coverage (Zhang, 2015; Zhang, 2018)

- Models: K-lists with over-/under-coverage, for $K \geq 4$ (Di Cecco et al., 2018; Di Cecco, 2018)

Application of DSE & TDSE to admin data in Ireland

The Irish case (Dunne, 2015; Zhang & Dunne, 2017)

- Traditional census every 5 years; the latest one in 2016

  No census coverage survey/adjustment; No CPR

- SPD = PAR (Person Activity Register), entirely SoL sources

  linkage based on PIN with negligible errors

  excl. Driving License Dataset (DLD), renewal every 10 years

First known application of entirely register-based DSE

- DSE set-up: fixed A = PAR, random B = DLD
- TDSE: exploring potential erroneous enumeration

Application of DSE & TDSE to admin data in Ireland

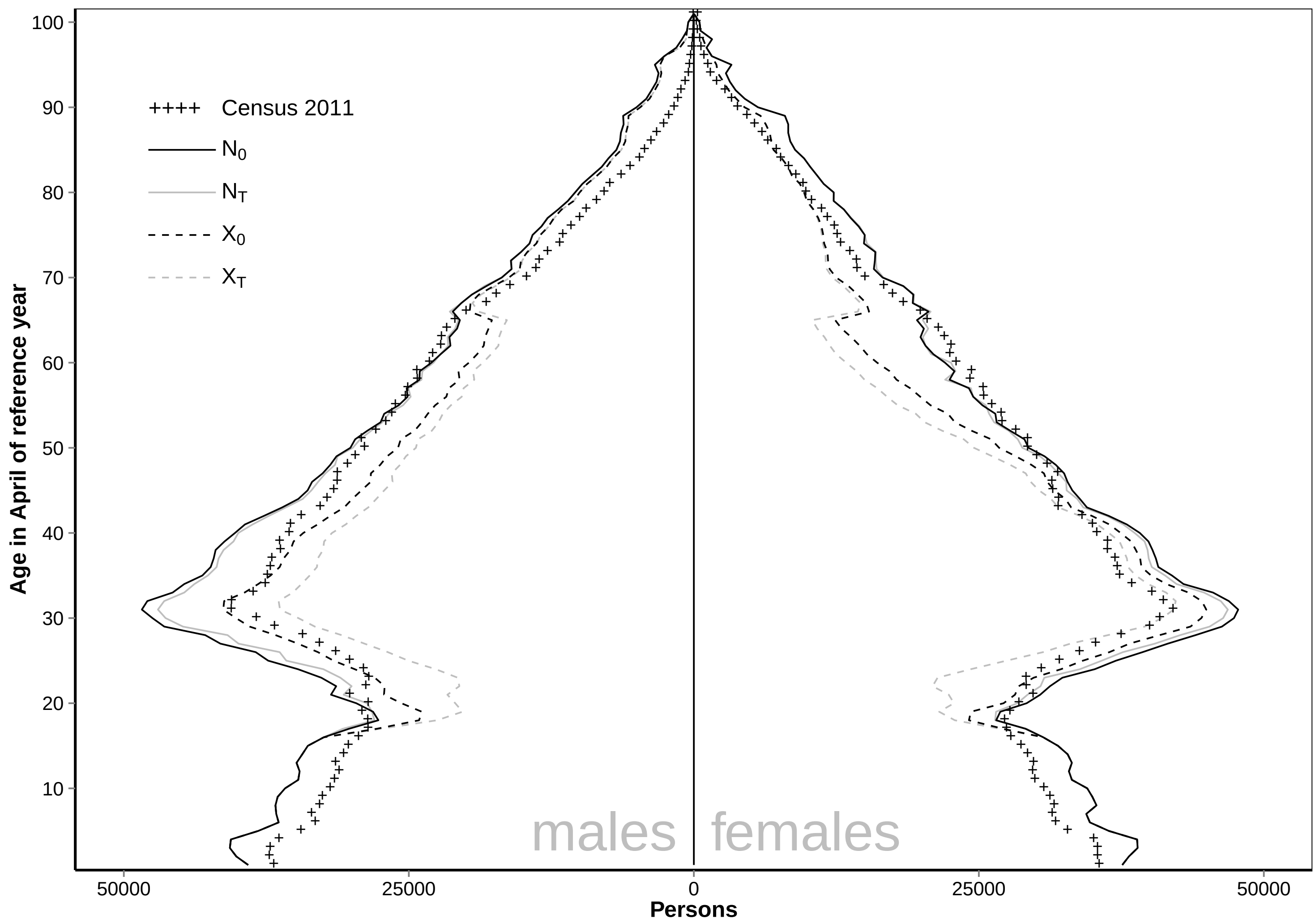
Four relevant population concepts:

- Census night population (U_I): *de facto* definition
- Usually resident population (U_{II}): difference across countries, e.g. reference date using CPR, reference year using SoL sources
- Hypothetical PAR population (U_A): any person who have had or *in principle* could have had interactions with public administration

  Underenumeration: could-haves, delays of registration, etc.

  Potential erroneous enumeration: e.g. leavers post SoL-activity
- Hypothetical DL population (U_B): any person who holds or *in principle* could have held an Irish driving licence

Application of DSE & TDSE to admin data in Ireland



NB. TDSE: trimming by Employment payment; can trim by sources

Application of DSE & TDSE to admin data in Ireland

TDSE: Scoring k records in A, of which k_1 in AB

$$\hat{N}_k = n\frac{x-k}{m-k_1}$$

NB. naïve DSE $\boxed{\hat{N} = \hat{N}_0 = n\frac{x}{m} > \tilde{N} = n\frac{x-r}{m}}$ ideal DSE

1. If $\frac{k_1}{m} < \frac{k}{x}$, then $\hat{N}_k < \hat{N}_0$. If $\frac{k_1}{m} = \frac{k}{x}$, then $\hat{N}_k = \hat{N}_0$.

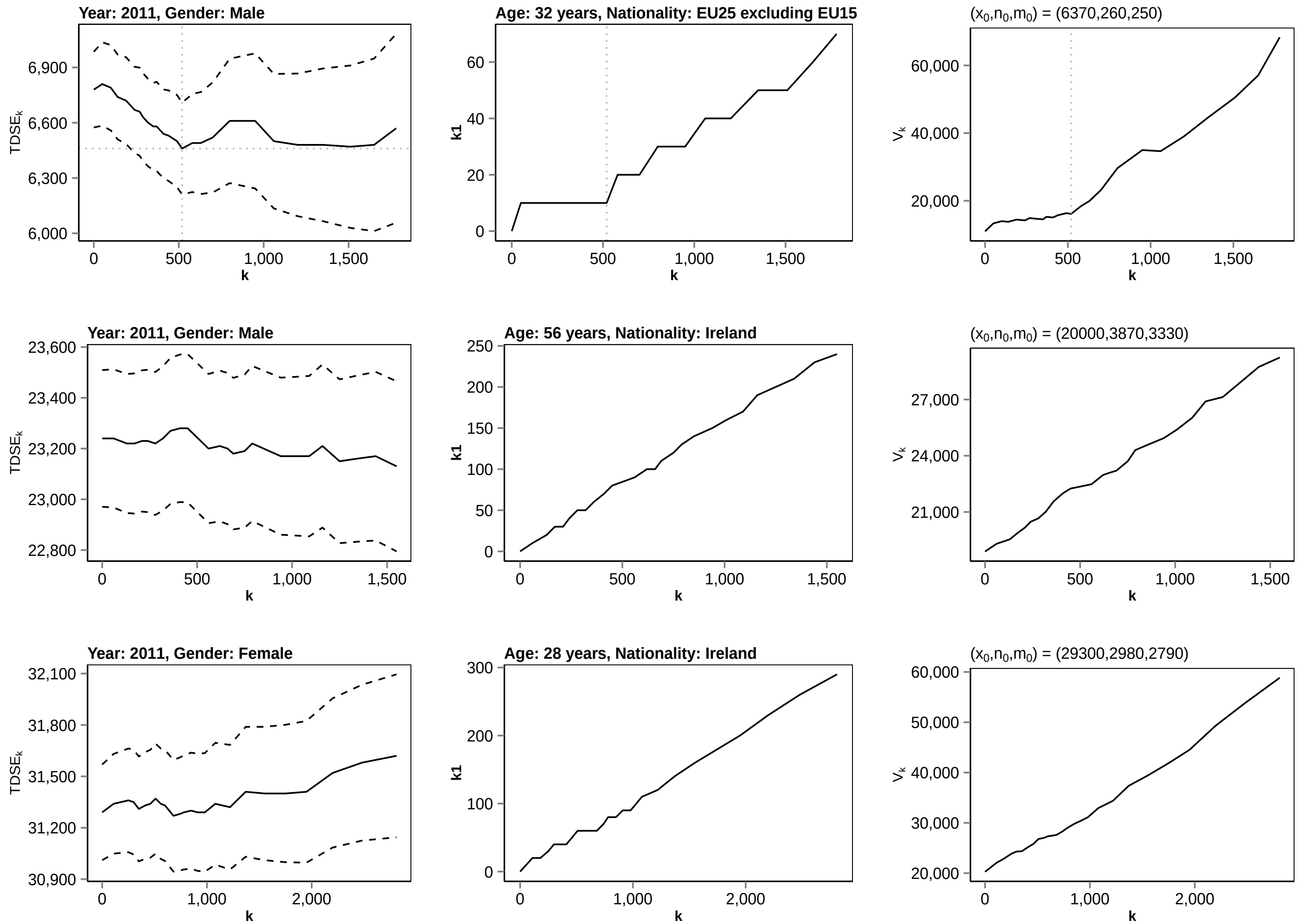
NB. trimming helps if scoring more effective than random sampling

2. If $k < r$, then $\tilde{N} < \hat{N}_k$.

NB. no 'over-adjustment' if no 'over-trimming'

3. If all the r erroneous records are among the k scored ones, then $\lim_{n\to\infty} E(\hat{N}_k) = \lim_{n\to\infty} E(\tilde{N})$.

Application of DSE & TDSE to admin data in Ireland



Modelling register coverage errors in $K+1$ lists

Target-list universe $U^* = U \cup A \cup B$ with $K = 2$:

In U

		List B: in	List B: out	
List A	in	p_{111}	p_{110}	p_{11+}
	out	p_{101}	p_{100}	p_{10+}
		p_{1+1}	p_{1+0}	p_{1++}

Out of U

		List B: in	List B: out	
List A	in	p_{011}	p_{010}	p_{01+}
	out	p_{001}	—	p_{001}
		p_{0+1}	p_{010}	

Modelling register coverage errors in $K+1$ lists

For $K = 2$ lists containing erroneous enumerations, let

$$\theta_{1+} = \Pr(i \notin U | i \in U^*_{+1+}) \qquad \text{[error rate in A]}$$

$$\theta_{+1} = \Pr(i \notin U | i \in U^*_{++1}) \qquad \text{[error rate in A]}$$

$$\theta_{11} = \Pr(i \notin U | i \in U^*_{+11}) \qquad \text{[error rate in AB]}$$

For instance, A = Tax Register, B = Patient Register

Q: As $\theta_{1+} \to 0$ *and* $\theta_{+1} \to 0$, *how fast does* $\theta_{11} \to 0$*?*

Investigation of *all* possible log-linear models (Zhang, 2015):

- set of units/model space = U
- set of units/model space = U^*
- set of units/model space = $A \cup B$

Modelling register coverage errors in $K+1$ lists

- Largest non-saturated model of target universe U implies

$$\frac{(1-\theta_{11})}{(1-\theta_{1+})(1-\theta_{+1})} = \frac{E(x_{1+})E(x_{+1})}{E(x_{11})E(N)}$$

i.e. *incidental* constraints between errors rates and N

- Largest non-sat. model of target-list universe U^* implies

$$\operatorname{logit}\theta_{11} = \operatorname{logit}\theta_{10} + \operatorname{logit}\theta_{01} + \big(\log E(N_{100}) - \log(N_{+++})\big)$$

i.e. again leading to incidental constraints

- Largest non-sat. model of list universe $A \cup B$ implies

$$\operatorname{logit}\theta_{11} = \operatorname{logit}\theta_{10} + \operatorname{logit}\theta_{01}$$

i.e. standard $\lambda_{uab}^{UAB} = 0$ assumption of three-way table, non-incidental and generalisable to $K > 2$

Modelling register coverage errors in $K + 1$ lists

For small error rates, $\text{logit}\theta_{ab} \approx \log \theta_{ab}$; assumption

$$\log \theta_{11} = \log \theta_{10} + \log \theta_{01} \quad \Leftrightarrow \quad \theta_{11} = \theta_{10}\theta_{01}$$

$$P(i \notin U | i \in A \cap B) = P(i \notin U | i \in A \setminus B) P(i \notin U | i \in B \setminus A)$$

However, as $\theta_{1+} \to 0$ and $\theta_{+1} \to 0$ in two 'good' lists, it may be likely that $\theta_{10} \to 1$ and $\theta_{01} \to 1$, whereas $\theta_{11} \to 0$, i.e. contrary to above!

A model that accommodates such situations is given by

$$\log \theta_{11} = \log \theta_{1+} + \log \theta_{+1} \qquad \Leftrightarrow \qquad \theta_{11} = \theta_{1+}\theta_{+1}$$

$$P(i \notin U | i \in A \cap B) = P(i \notin U | i \in A) P(i \notin U | i \in B)$$

A *Pseudo conditional independence (PCI)* assumption, unlike cond. ind., e.g. $\Pr(X \cap Y | Z) = \Pr(X|Z)\Pr(Y|Z)$

Modelling register coverage errors in $K+1$ lists

For generalisation to $K > 2$ (Zhang, 2018), let

$$\log \mu_{\omega\delta_U} = \lambda + \sum_{\nu \in \Omega(\omega)} \lambda_{\mathbf{1}_\nu}^{A_\nu} + \lambda_1^U + \sum_{\nu \in \Omega(\omega)} \lambda_{\mathbf{1}_\nu 1}^{A_\nu U}$$

for the contingency table arising from cross-classifying the target-list universe $\cup_{k=1}^K A_k \cup U$, and $\mu_{\omega\delta_U} = \mu_{\delta_1 \cdots \delta_K \delta_U}$ is the expected cell count, where $\omega = \{\delta_1, ..., \delta_K\}$, and $\Omega(\omega)$ consists of all the non-empty subsets of ω, and as the parameter constraints, set $\lambda_{\omega\delta_U}$ to 0 if there is at least one 0 among $\delta_1 \cdots \delta_K \delta_U$.

NB. See Zhang (2018) for model interpretation, maximum likelihood estimation and an application to Dutch homelessness data ($K = 3$).

References

[1] Di Cecco, D. (2018). Estimating population size in multiple record systems with uncertainty of state identification. In *Analysis of Integrated Data, eds. L.-C. Zhang and R-L Chambers*. Chapman & Hall/CRC. *To appear.*

[2] Di Cecco, D., Di Zio, M., Filipponi, D., and Rocchetti, I. (2018). Population size estimation using multiple incomplete lists with overcoverage. *Journal of Official Statistics*, **34**, 557-572.

[3] Dunne, J. (2015). The Irish Statistical System and the emerging Census opportunity. *Statistical Journal of the IAOS*, **31**, 391-400. `DOI:10.3233/SJI-150915`

[4] Nirel, R. and Glickman, H. (2009). Sample surveys and censuses. In *Sample Surveys: Design, Methods and Applications, Vol 29A (eds. D. Pfeffermann and C.R. Rao)*, Chapter 21, pp. 539-565.

[5] ONS - Office for National Statistics (2013). *Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice*. ONS Internal Report, available at: `http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html`

[6] Tiit, E.-M. and Maasing, E. (2016). Residency index and its applications in censuses and population statistics. Eesti statistika kvartalikri. (Quarterly Bulletin of Statistics Estonia). 3/16:41-60. `http://www.stat.ee/publication-2016_quarterly-bulletin-of-statistics-estonia-3-16`

[7] Zhang, L.-C. (2018). Log-linear models of erroneous list data. In *Analysis of Integrated Data, eds. L.-C. Zhang and R-L Chambers*. Chapman & Hall/CRC. *To appear.*

[8] Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, **31**, 381-396.

[9] Zhang, L.-C. and Dunne, J. (2017). Trimmed Dual System Estimation. In *Capture-Recapture Methods for the Social and Medical Sciences, eds. D. Böhning, J. Bunge and P. v. d. Heijden*, Chapter 17, pp. 239-259. Chapman & Hall/CRC.