Census like population size estimation based on administrative data ---

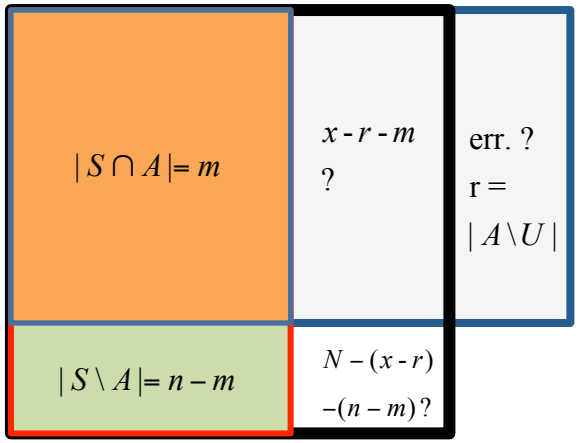
Some recent developments on capture recapture methodology in the presence of erroneous enumerations

Li-Chun Zhang
(L.Zhang@soton.ac.uk)
University of Southampton& Statistics Norway
John Dunne
Central Statistics Office, Ireland, John.Dunne@cso.ie

To start with, a traditional census costs too much, takes too long…

- **Alternative**: population register (**PR**) for some only, **administrative registers for all**
- **Q1**: What if one replaces the census with a register?
- Trimmed Dual System Estimation (TDSE)
  - Basic idea; additional complications; possibly combining with additional modelling
- Potential areas of application
  - Population size estimation: regular or irregular
  - Health screening , edit scoring, etc.
- The Irish question (**Q2**): Can one estimate population size **without census**, **nor PR**, **nor survey**?

Survey S
(size n)



List A
(size x)

Target population U
(size N, unknown)

Assume:

$$P(I_{iS}=0 \mid I_{iU}=1) = P\left(I_{iS}=0 \mid \begin{matrix} I_{iU}=1 \\ I_{iA}=1 \end{matrix}\right) = P\left(I_{iS}=0 \mid \begin{matrix} I_{iU}=1 \\ I_{iA}=0 \end{matrix}\right)$$

$$P(I_{iS}=1 \mid I_{iU}=1) = P\left(I_{iS}=1 \mid \begin{matrix} I_{iU}=1 \\ I_{iA}=1 \end{matrix}\right) = P\left(I_{iS}=1 \mid \begin{matrix} I_{iU}=1 \\ I_{iA}=0 \end{matrix}\right)$$

$$\frac{N-E(n)}{E(n)} = \frac{E(x-r-m)}{E(m)} = \frac{E[N-(x-r)-(n-m)]}{E(n-m)}$$

$$\frac{N}{E(n)} = \frac{E(x-r)}{E(m)}$$

$$N = E(n)\frac{E(x-r)}{E(m)}$$

$$\tilde{N} = n\frac{x}{m} \geq \tilde{N} = n\frac{x-r}{m}$$

Naive DSE Hypothetic DSE

Trimmed DSE (TDSE)

k_1 units in S and A + k_0 units in A\S

Scoring **k** units in A

TDSE_k

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

Ideal scoring
$(k, k_1) = (r, 0)$

$$\tilde{N} = n\frac{x-r}{m-0}$$

$$\tilde{N} \le \hat{N}_k \quad \text{iff}$$

$$\frac{x-r}{m} \le \frac{x-k}{m-k_1} \quad \Leftrightarrow \quad \frac{k-r}{x-r} \le \frac{k_1}{m}$$

OK as long as k < r
i.e. "safe" up to k = r

No scoring k=0
Naïve DSE

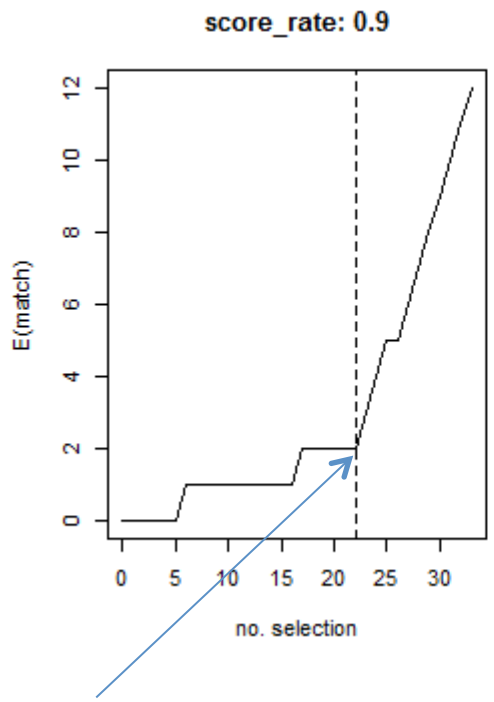
$$\hat{N}_0 = n\frac{x}{m}$$

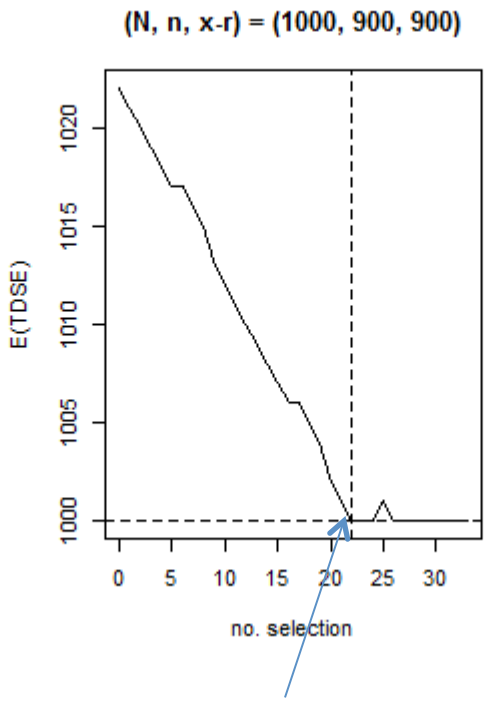
$$\hat{N}_k \le \hat{N}_0 \quad \text{iff}$$

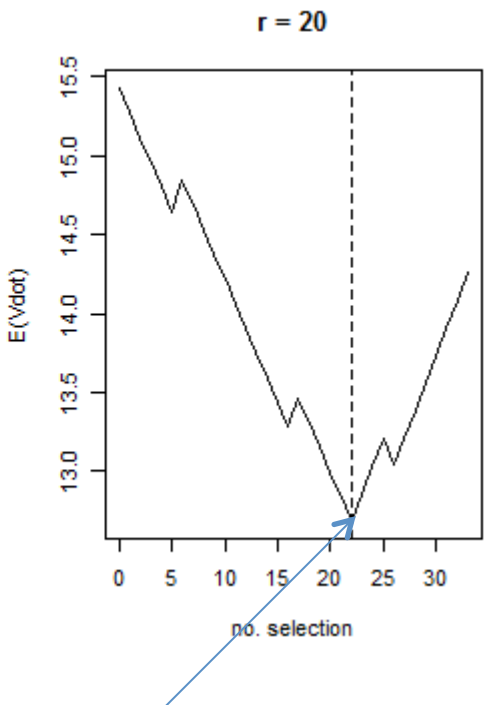
$$\frac{x-k}{m-k_1} \le \frac{x}{m} \quad \Leftrightarrow \quad \frac{k_1}{m} \le \frac{k}{x}$$

Needs to be more effective than random scoring

Stopping rule







$$k = k_1 + k_{0u} + k_{0r}$$

$$p_{rk} = E(k_{0r}) / k \quad \Rightarrow \quad k_r = r / p_{rk}$$

$$\frac{E(k_1)}{k} = \begin{cases} [(1 - p_{rk})\frac{m}{x-r}] & \text{if } k \le k_r \\ [(1 - \frac{r}{k})\frac{m}{x-r}] & \text{if } k > k_r \end{cases}$$

$$E(\hat{N}_k)$$

$$= E(n\frac{x-k}{m-k_1})$$

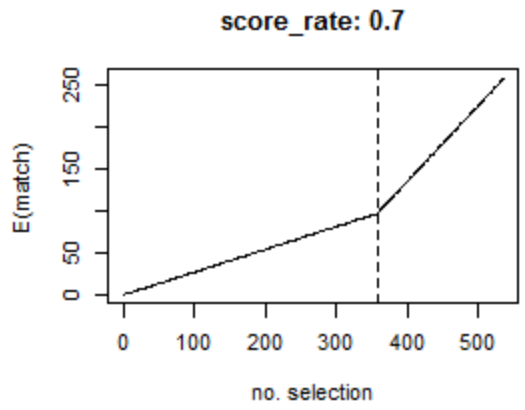
$$\approx n\frac{x-k}{m-\mu_{1k}}$$

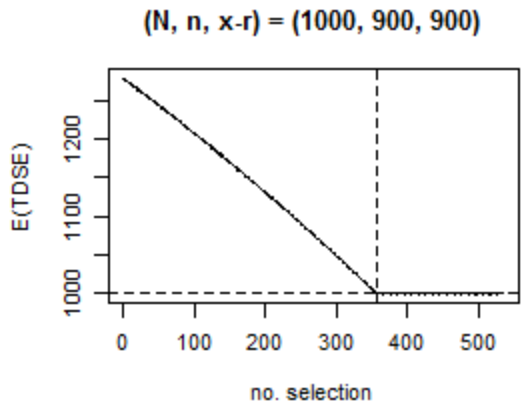
$$\mu_{1k} = E(k_1)$$

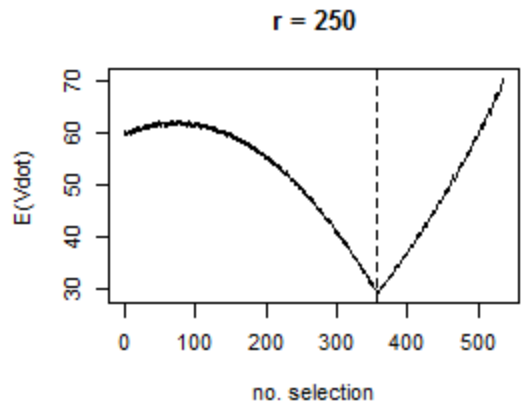
$$\dot{V}_k = (m-\mu_{1k})^{-3} n(n-m+\mu_{1k})$$
$$(x-k)(x-k-(m-\mu_{1k}))$$

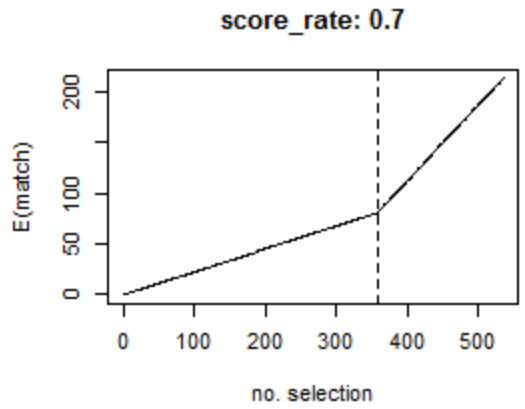
NB. had $\hat{N}_k$ been unbiased

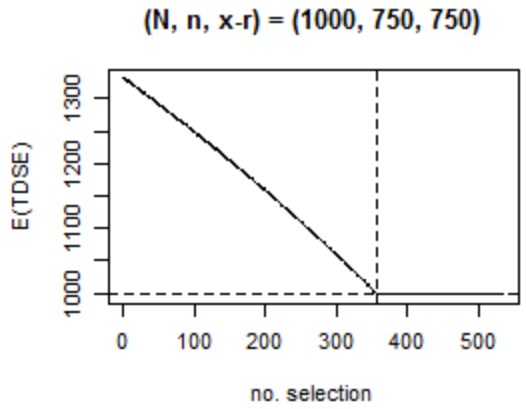
$$\hat{V}(\tilde{N}) = m^{-3} n(n-m)$$
$$(x-r)(x-r-m)$$

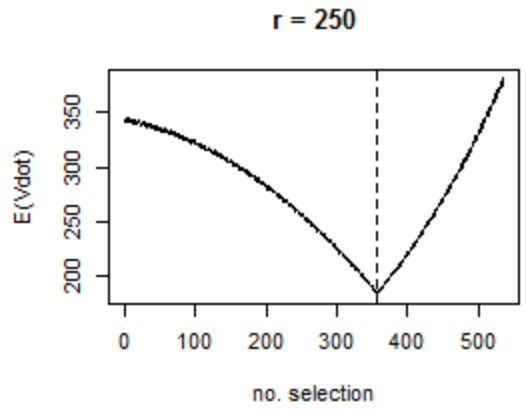

score_rate: 0.7
E(match)
0
50
150
250
0
100
200
300
400
500
no. selection

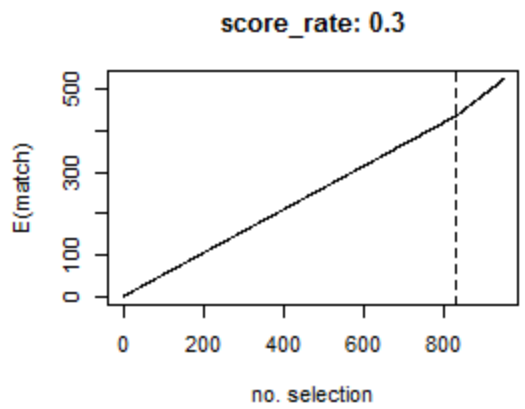

(N, n, x-r) = (1000, 900, 900)
E(TDSE)
1000
1100
1200
0
100
200
300
400
500
no. selection

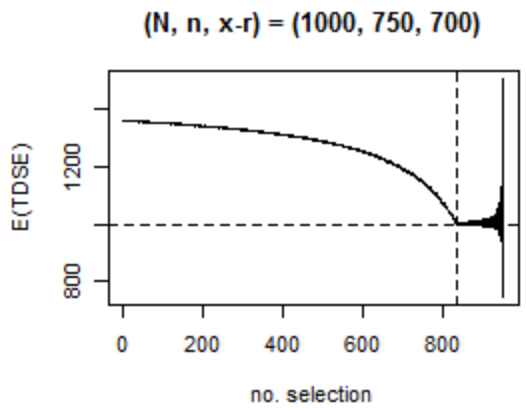

r = 250
E(Vdot)
30
40
50
60
70
0
100
200
300
400
500
no. selection

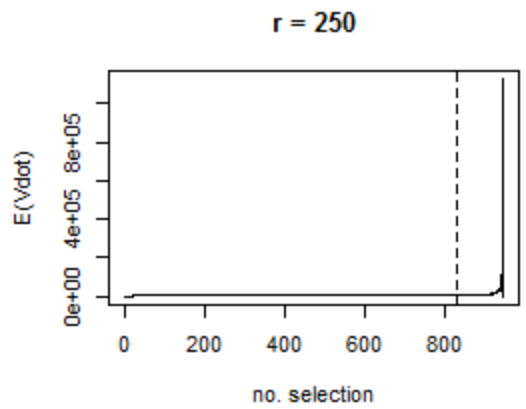

score_rate: 0.7
E(match)
0
50
100
200
0
100
200
300
400
500
no. selection


(N, n, x-r) = (1000, 750, 750)
E(TDSE)
1000
1100
1200
1300
0
100
200
300
400
500
no. selection


r = 250
E(Vdot)
200
250
300
350
0
100
200
300
400
500
no. selection


score_rate: 0.3
E(match)
0
100
300
500
0
200
400
600
800
no. selection


(N, n, x-r) = (1000, 750, 700)
E(TDSE)
800
1200
0
200
400
600
800
no. selection


r = 250
E(Vdot)
0e+00
4e+05
8e+05
0
200
400
600
800
no. selection

What if **both** lists have **erroneous enumerations**?

(1) Assumption of independence btw L_1 and L_2 necessary for trimmed DSE.

(2) Let the naive face-value DSE be $\hat{N}_0 = \frac{n_1 n_2}{n_{12}}$ in an obvious notation.

Let the ideal DSE be $\tilde{N} = \frac{(n_1 - r_1)(n_2 - r_2)}{n_{12} - r_{12}}$ after removing the errors.

We have $\tilde{N} < \hat{N}_0$ iff $\frac{r_{12}}{n_{12}} < \frac{r_1}{n_1} + \frac{r_2}{n_2} - \frac{r_1}{n_1}\frac{r_2}{n_2}$

i.e. provided relatively fewer errorneous records among the matched list, the face-value DSE remains biased upwards; whereas it is downwards biased

in the oppsite case; and unbiased if $\mathrm{E}(\frac{r_{12}}{n_{12}}) = E(\frac{r_1}{n_1}) + E(\frac{r_2}{n_2}) - E(\frac{r_1}{n_1})E(\frac{r_2}{n_2})$.

Notice the form P(A)+P(B)-P(A)P(B) on the right-hand side, which corresponds to the case where erroneous records occur randomly across L_1, L_2 and L_{12},

(3) Suppose scoring yields (k_1, k_2, k_{12}) units in L_1, L_2 and L_{12} respectively.

Put the trimmed DSE (TDSE) $\hat{N}_{k_1 k_2} = \dfrac{(n_1 - k_1)(n_2 - k_2)}{n_{12} - k_{12}}$

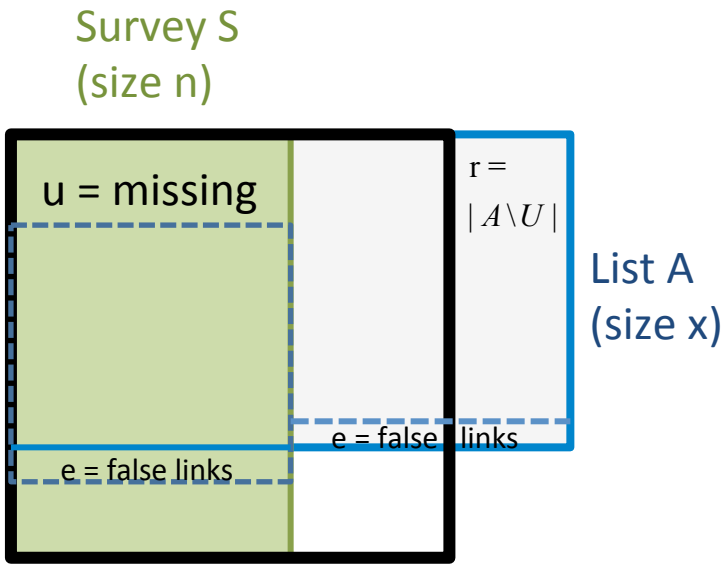
We have $\hat{N}_{k_1 k_2} < \hat{N}_0$ iff $\dfrac{k_{12}}{n_{12}} < \dfrac{k_1}{n_1} + \dfrac{k_2}{n_2} - \dfrac{k_1}{n_1}\dfrac{k_2}{n_2}$

i.e. TDSE reduces the bias of $\hat{N}_0$ provide scoring is more effective than random selection, and $\tilde{N} < \hat{N}_0$ to start with.

(4) Provided $\tilde{N} > \hat{N}_0$ to start with, which is unlikely in the census context, TDSE again reduces the bias of $\hat{N}_0$ provide scoring is more effective than random selection. But effective means $\dfrac{k_{12}}{n_{12}} > \dfrac{k_1}{n_1} + \dfrac{k_2}{n_2} - \dfrac{k_1}{n_1}\dfrac{k_2}{n_2}$ now.

(5) Trimming introduces bias if $\dfrac{r_{12}}{n_{12}} = \dfrac{r_1}{n_1} + \dfrac{r_2}{n_2} - \dfrac{r_1}{n_1}\dfrac{r_2}{n_2}$ to start with, unless scoring happens to yield $\dfrac{k_{12}}{n_{12}} = \dfrac{k_1}{n_1} + \dfrac{k_2}{n_2} - \dfrac{k_1}{n_1}\dfrac{k_2}{n_2}$.

What about TDSE in the presence of **linkage errors**?



Black: target population U

Blue: list A with under- and over-coverage

Green: survey S with only under-coverage

Dottet area: linked set S ∩ A with linkage error

no. missing matches = u

no. false links = e (invloing records in S and A)

Let the linked set be of size **M**, where M ≠ m,

and **m** is the size of the true matched set.

We have $E(M|m,S,A) = m - E(u|-) + E(e|-) = m - m \cdot f + E(M|-) \cdot q$

$\Rightarrow \quad \hat{m} = M \cdot \lambda \quad$ where $\lambda = (1-q)/(1-f)$

$f = m^{-1}E(u|m,S,A)$ is the missing rate and $q = E(e/M|m,S,A)$ the false link rate

$\Rightarrow \quad \tilde{N}_L = n(x-r)/(M\lambda)$ = approx. unbiased linkage DSE (LDSE)

Face-value DSE: $\hat{N}_0 = nx / \mathrm{M}$

Hypothetical ideal DSE: $\tilde{N} = n(x-r)/\mathrm{m}$ without linkage errors

Hypothetical linkage DSE (LDSE): $\tilde{N}_L = n(x-r)/(\mathrm{M}\lambda)$ with linkage errors

We have $\lambda = (1-q)/(1-\mathrm{f}) \approx (1-q)(1+\mathrm{f})$, so that

$$\hat{N}_0 > \tilde{N}_L \quad \Leftrightarrow \quad r > x(1-\lambda) \approx x[q(1+f)-f]$$

More simply, "conservative" linkage if q<f $\Leftrightarrow$ λ>1 $\Leftrightarrow$ E(M|m,S,A)<m

$\Rightarrow$ $\hat{N}_0 > \tilde{N}_L \approx \tilde{N}$

It follows that given the actual linkage result, with plug-in $\hat{\lambda}$,

the trimmed linkage DSE (TLDSE) can be given as

$$\hat{N}_k = n\frac{x-k}{(\mathrm{M}-k_1)\hat{\lambda}}$$

It can be expected to bahave similarly as the TDSE in the absence of linkage errors.

NB. A more subtle discussion requires one to know how λ changes under scoring.

What if trimming can not do all the job?

Need at least 2 lists, in addition to **coverage survey (S)**

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

NB. subscript = domain indicator (uab); $N_{000} = 0$ by definition

Zhang (2015) examine possible log-linear models for

- Target population U
- List-target universe U^* = union of U, A and B
- List universe U_L = union of A and B

Two well-defined, generalisable models

Model (10): put $\theta_{ab} = P(I_{iu} = 0 \mid I_{iab} = 1)$

unsaturated log-linear model for list-target universe

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

approximately the same as

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

$$\Leftrightarrow \quad P(i \notin U \mid i \in A \text{ and } i \in B) = P(i \notin U \mid i \in A \backslash B) \cdot P(i \notin U \mid i \in B \backslash A)$$

Model (11):

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

$$\Leftrightarrow \quad P(i \notin U \mid i \in A \text{ and } i \in B) = P(i \notin U \mid i \in A) \cdot P(i \notin U \mid i \in B)$$

Zhang (2015) refers this as pseudo conditional independence (PCI)

NB. for any random events Z, X and Y

CI: $P(X \cap Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$

PCI: $P(Z \mid X \cap Y) = P(Z \mid X) \cdot P(Z \mid X)$

Genera-lisation:

Generic log-linear models for error- or hit-rates of 3-list capture table

Model Restriction	Model Interpretation
-	Saturated model
$\alpha_{123} = 0$	Null 2nd-order PCI-interaction among (L_1, L_2, L_3)
$\alpha_{12} = 0$	PCI between L_1 and L_2
$\alpha_{123} = -\alpha_{12}$	Conditional PCI between L_1 and L_2 given L_3
$\alpha_{12} = \alpha_{123} = 0$	Both PCI and conditional PCI between L_1 and L_2
$\alpha_{12} = \alpha_{13} = \alpha_{123} = 0$	PCI between L_1 and (L_2, L_3)
$\alpha_{12} = \alpha_{13} = \alpha_{23} = \alpha_{123} = 0$	Mutual PCI between L_1, L_2 and L_3

Back to the 2-list situation…

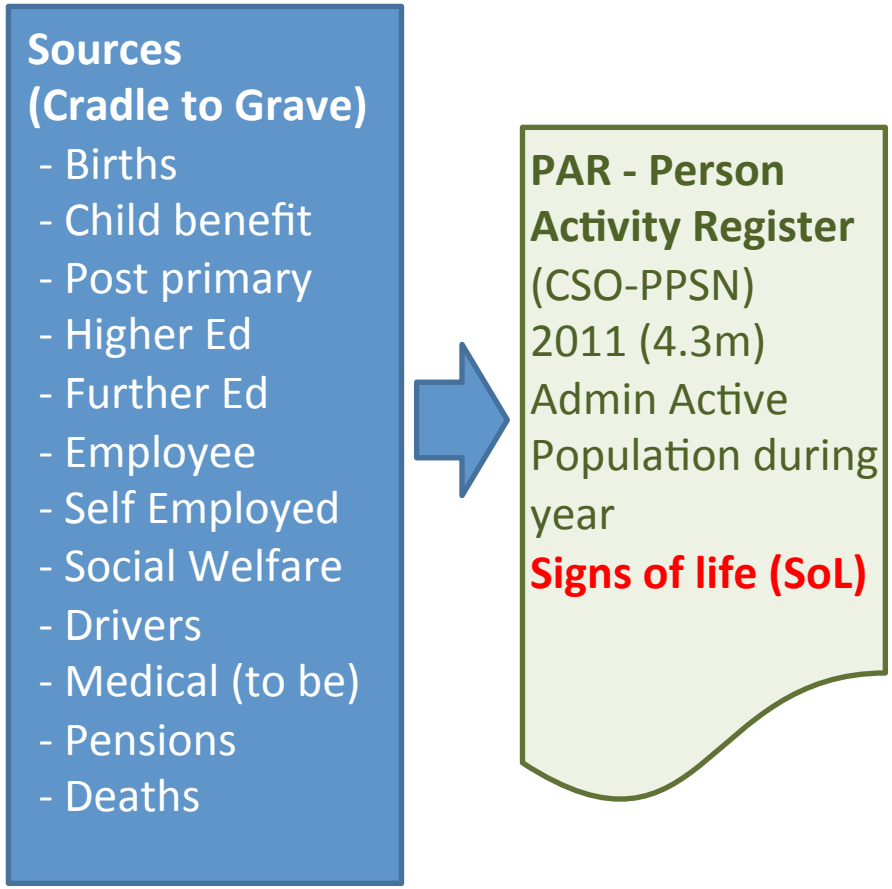
Table 3: Range of relative bias under model (10) and (11) for census enumeration error adjustment. Census enumeration = 1000, register enumeration = 1200, census-register enumeration = 900. Error rate of census errra(θ_{1+}), register enumeration (θ_{+1}), census-register enumeration (θ_{11}), where $0 < \theta_{11} < \theta_{1+}$. All numbers in %.

Model (10)	Register error rate			
Census error rate	1	5	10	20
0.2	(0.078, 0.078)	(-0.11, -0.11)	(-0.48, -0.48)	(-3.4, -3.4)
0.5	(-0.038, 0.43)	(-0.88, 0.32)	(-2.5, 0.095)	(-16, -1.6)
1	(-0.25, 1)	(-2.3, 1)	(-6.3, 1)	(-38, 1)
Model (11)	**Register error rate**			
Census error rate	1	5	10	20
0.2	(0.11, 0.11)	(0.11, 0.11)	(0.1, 0.1)	(0.089, 0.089)
0.5	(0.11, 0.45)	(0.091, 0.44)	(0.068, 0.44)	(0.014, 0.43)
1	(0.1, 1)	(0.065, 1)	(0.012, 1)	(-0.11, 1)

Combine Trimming e.g. with Model (11):

Trim the initial lists A and B in order to move them into the well-fitting range of model (11) ?

The Irish case: data sources



Person Activity Register (PAR) only includes those for whom there has been evidence of engaging in **key** administrative systems in a given year.

- **Over-coverage** can be "**defined away**", but is subjected to discussion.
- **Under-coverage** by and large **to be expected**

Population estimate = PAR enumeration?

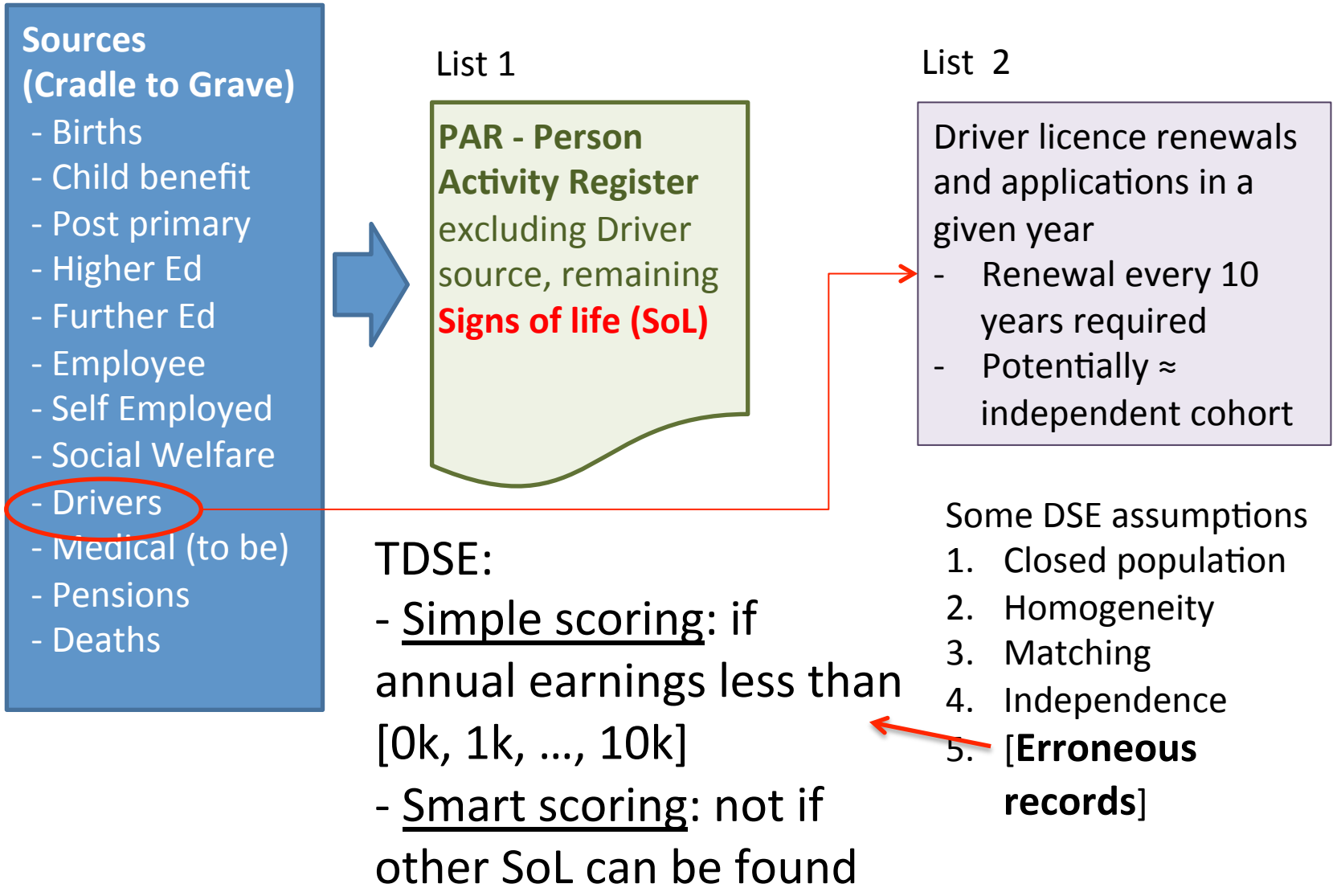
For the year **2011** in Ireland,
taking a "**signs of life**" approach
the Person Activity Register identified activity from

4.35m persons

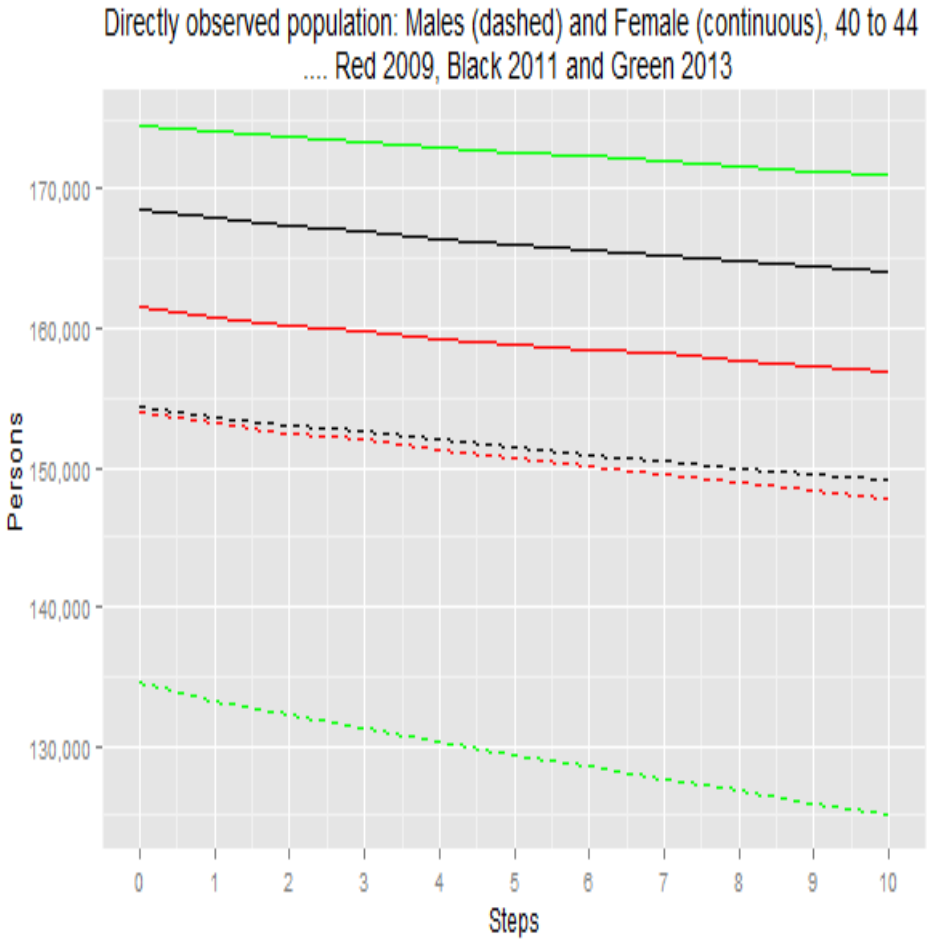
on Public Administration Systems.

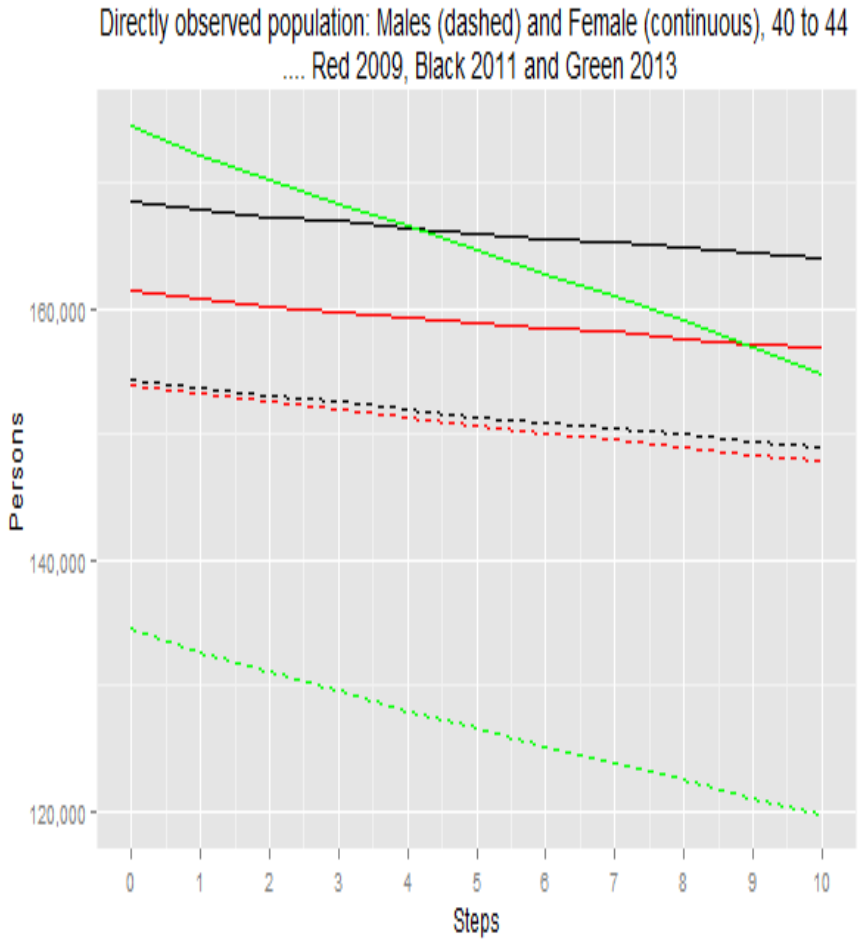
4.59m people were enumerated
on Census night in 2011

The Irish case: create a setting for DSE?



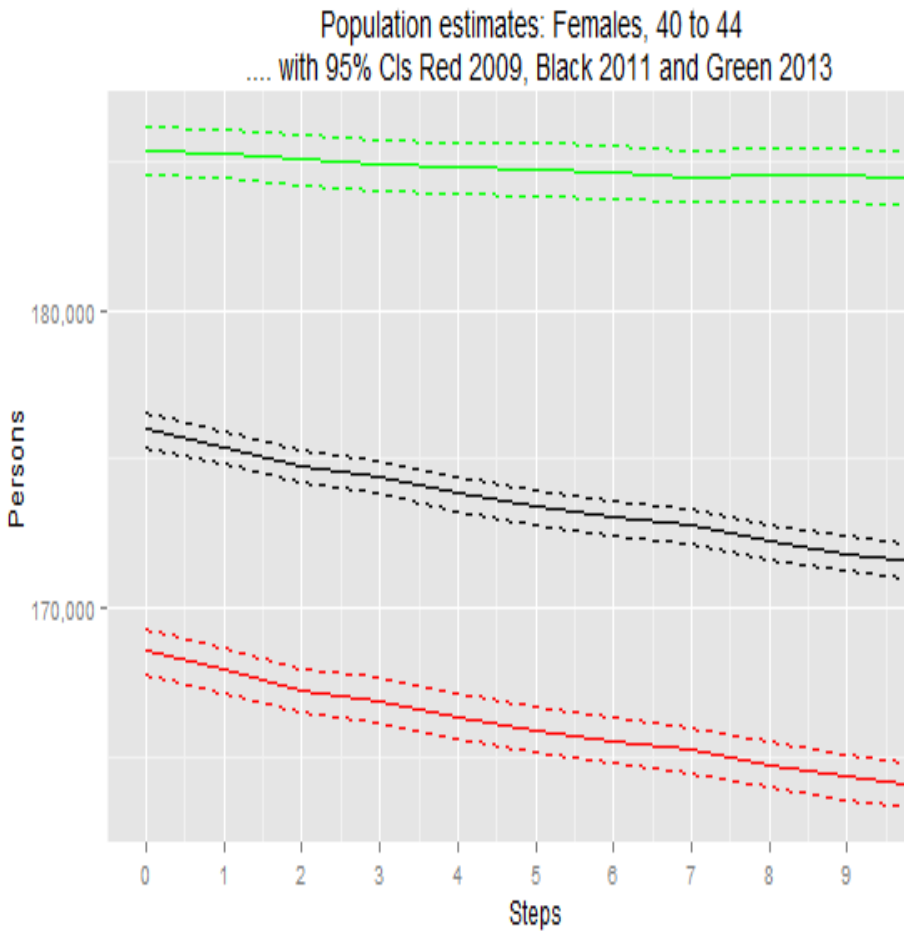
Age 40 - 44 PAR **scoring**: smart (left), simple (right)

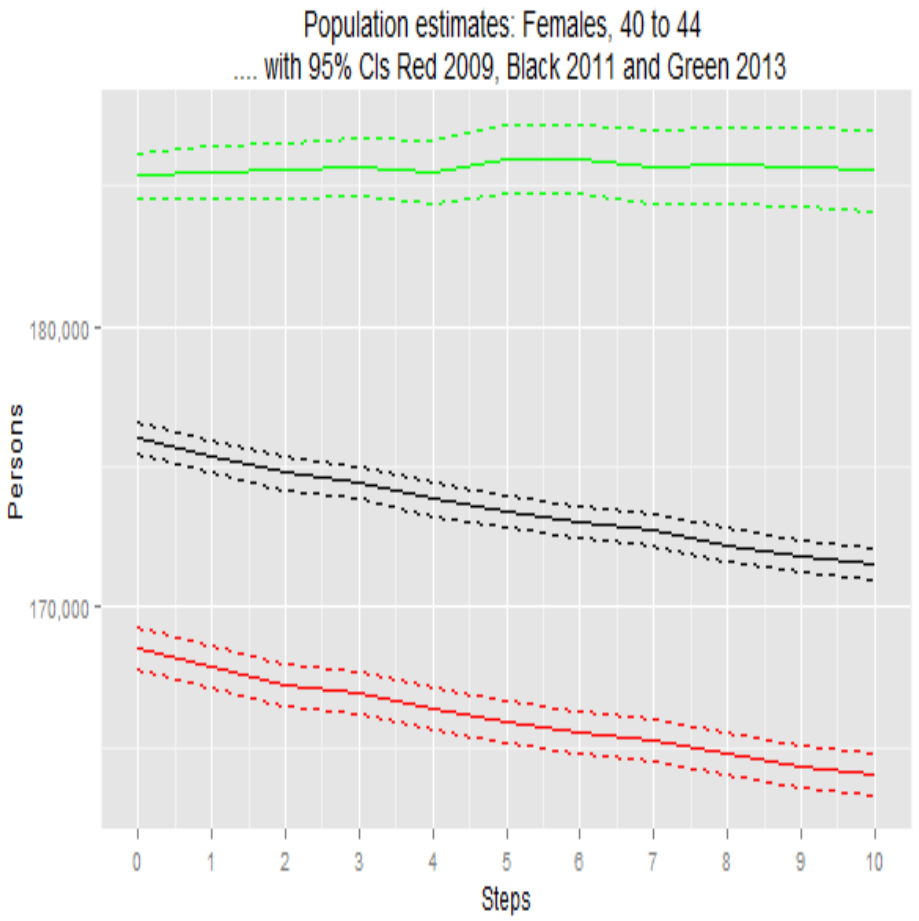




2013: PAR not finalised, lack of self-employed

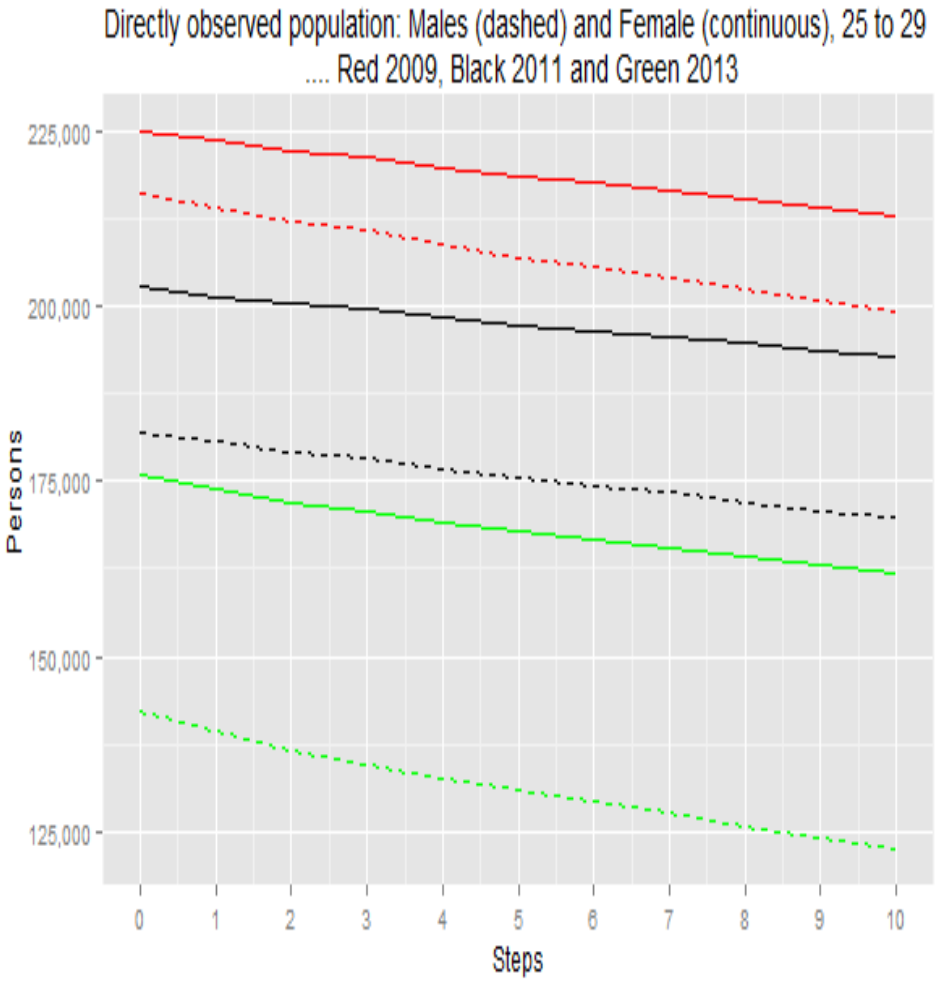
Female 40 - 44 PAR **TDSE**: smart (left), simple (right)

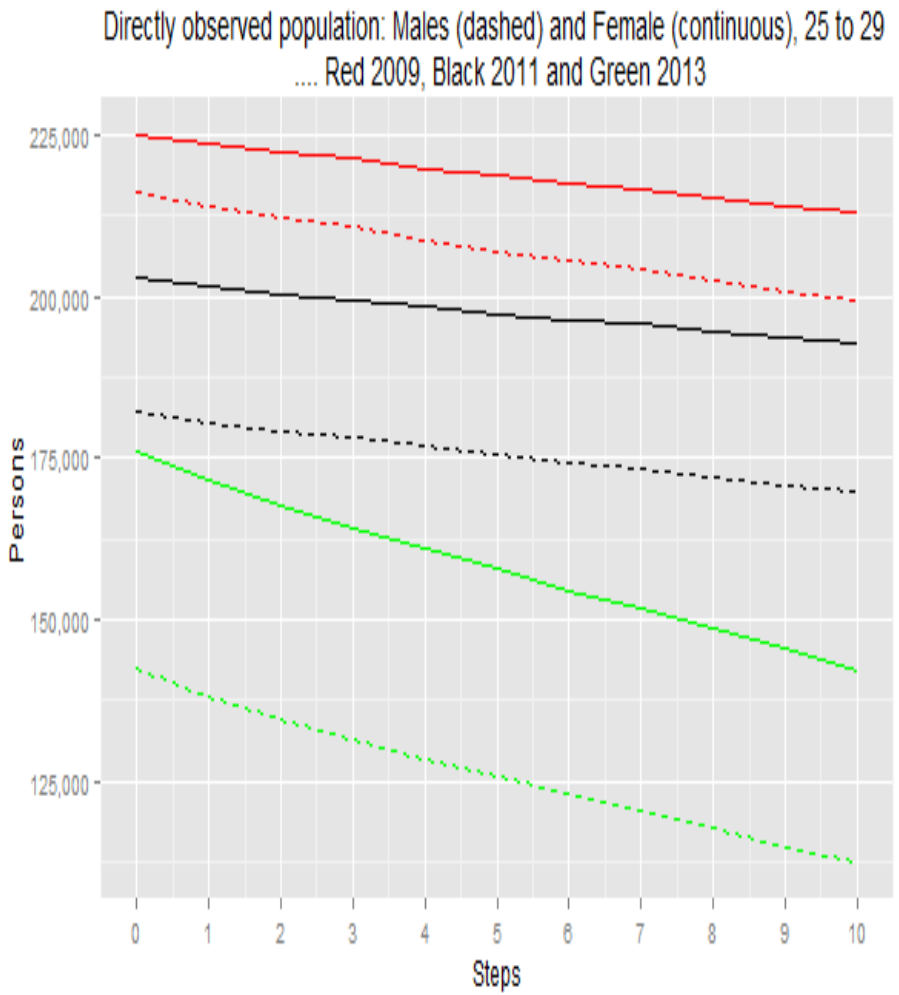




2013: PAR not finalised, lack of self-employed

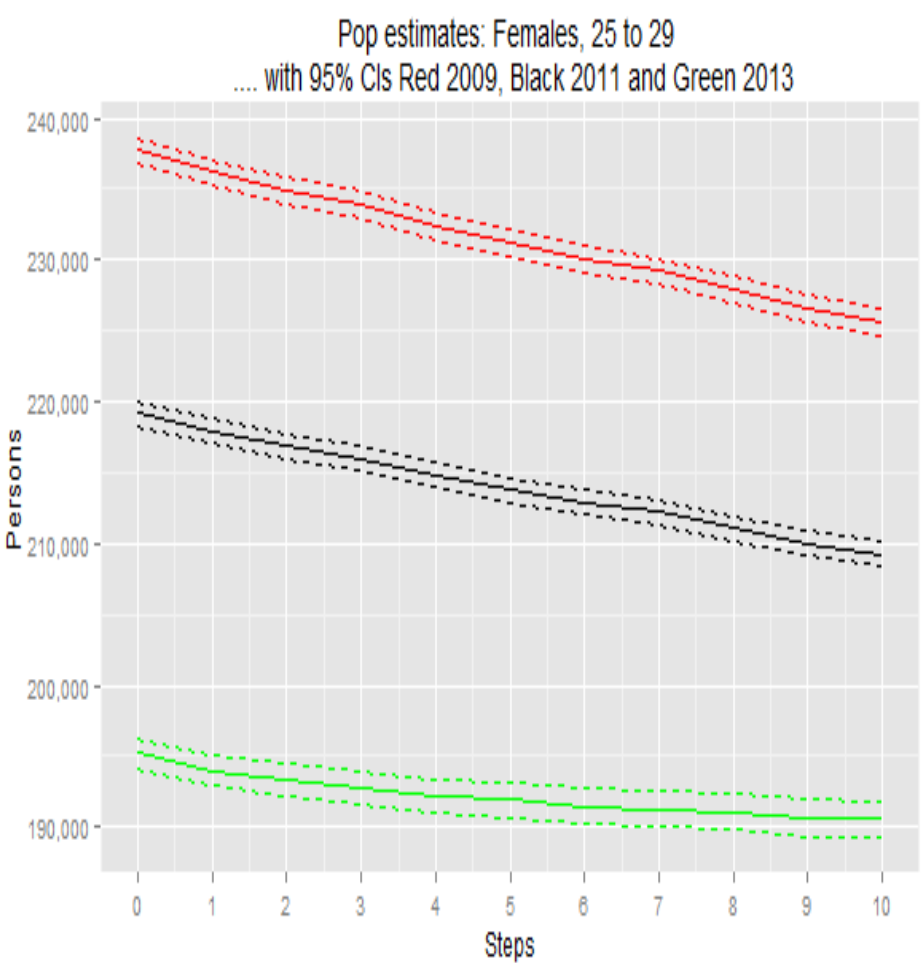
Age 25 - 29 PAR **scoring**: smart (left), simple (right)

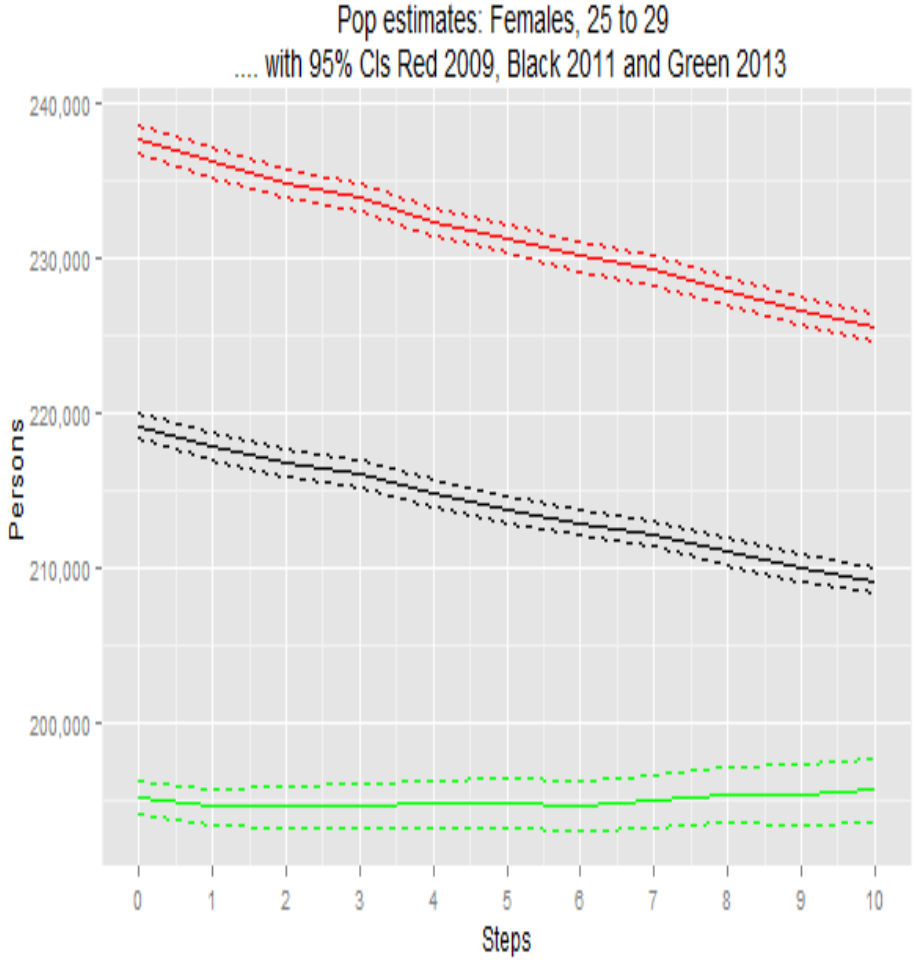




2013: PAR not finalised, lack of self-employed

Female 25 - 29 PAR **TDSE**: smart (left), simple (right)





2013: PAR not finalised, lack of self-employed

Discussion

- As by theory: larger k ≠ smaller N estimate
  - Bias: Scoring rate vs. prevalence of err. records
  - Variance: increases with extent of scoring
- Stability & composition of "Driver cohorts":
  - had everyone become a driver only at exactly the same point in their life, only a sub-population can be proved in any given year
- Implications of L_2 = Driver cohorts
  - Target population = one for which Driver cohorts have only under- but not over-coverage
  - Temporary workers or other special groups: random but not systematic under-enumeration
- Address/Immobility Register and address registration: Critical for census-like detailed population statistics