Correction for Linkage Error in Population Size Estimation

B.F.M. Bakker, L. Di Consiglio, D.J. van der Laan, T. Tuoto,
P.-P. de Wolf, D. Zult

Capture – recapture method

- Estimate the number of fish in a pond

		first	
		yes	no
second	yes	$n_{1,1}$	$n_{0,1}$
	no	$n_{1,0}$	$n_{0,0}$

		first	
		yes	no
second	yes	14	106
	no	86	??

Capture – recapture method

- Estimate the number of fish in a pond

		first	
		yes	no
second	yes	$n_{1,1}$	$n_{0,1}$
	no	$n_{1,0}$	$n_{0,0}$

		first	
		yes	no
second	yes	14	106
	no	86	$n_{0,0}$

- Assume independency between capture and recapture

$$\mathrm{OR} = \frac{n_{1,1} / n_{0,1}}{n_{1,0} / n_{0,0}} = 1 \qquad \widehat{n_{0,0}} = \frac{n_{1,0} \times n_{0,1}}{n_{1,1}} = \frac{86 \times 106}{14} = 651$$

Assumptions

- Instead of samples, you can also use registers

- Assumptions if you use two sources:
  a. Independency
  b. Population is closed
  c. Positive inclusion probability
  d. Perfect linkage
  e. No erroneous captures

Assumptions

- Instead of samples, you can also use registers

- Assumptions if you use two sources:
  a. Independency
  b. Population is closed
  c. Positive inclusion probability
  d. Perfect linkage
  e. No erroneous captures

		first	
		yes	no
second	yes	$n_{1,1}$	$n_{0,1}$
	no	$n_{1,0}$	$n_{0,0}$

Meeting the assumptions

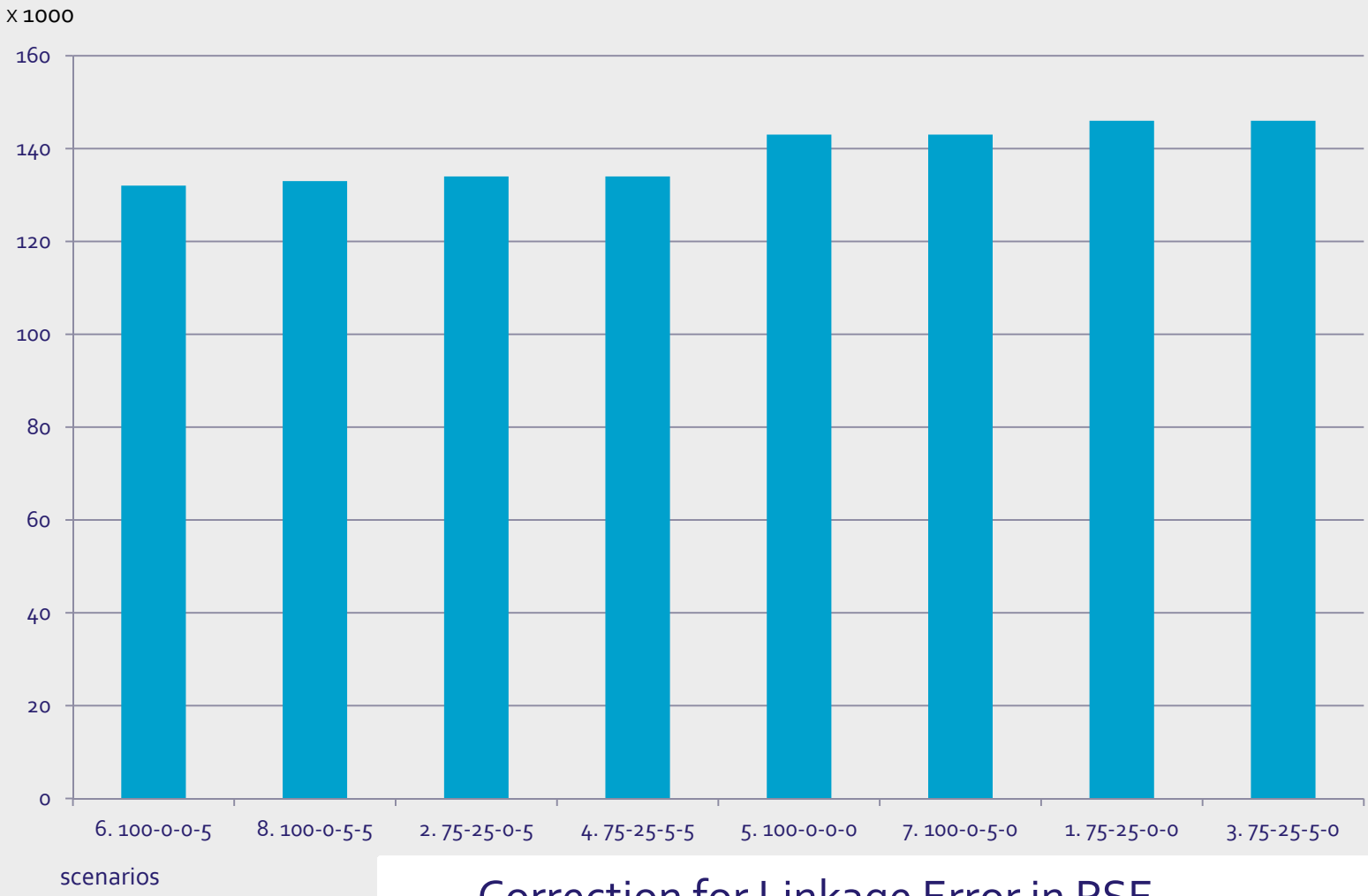
- Independency: three registers used (PR, ER and CSR) and covariates added to the model
- Closed population: ultimo September 2010 for PR and ER, second half of 2010 for CSR.
- Positive probability: only estimation of 15-65 years of age
- Perfect linkage: careful linkage
- Prevent erroneous captures: records which do not belong to the population removed

Results

- Estimation for each Nationality group (7 classes).
- With covariates Age (4 classes), Sex (2 classes), Residence duration (2 classes).
- Find the most parsimonious log-linear model that fits the data well with the Bayesian Information Criterion (BIC).

Usual residents not in PR: 470,000 (=5,2%).

Results with the traditional procedure



Probabilistic linkage

Probabilistic linkage:

- Generate all possible pairs
- Large datasets: blocking
- Compute (within blocks) weights for each pair
- Determine the threshold for appropriate linkage quality
- Select pairs with a weight above the threshold

Probabilistic linkage

m_i is the probability that variable i has the same value given that both records are from the same unit

u_i is the probability that variable i have the same value given that both records are not from the same unit

m_i and u_i are estimated by means of an EM-algorithm and are therefore approximations

Probabilistic linkage

Weights are determined by:

$$w_i = \begin{cases} \ln\left(\frac{m_i}{u_i}\right) & \text{if value i is similar} \\ \ln\left(\frac{1-m_i}{1-u_i}\right) & \text{if value i is different} \end{cases}$$

$$w = \sum_i w_i$$

Choose threshold in such a way that the number of false negatives and false positives are minimised

The MDF-correction

$$\hat{X}_{..MDF} = \frac{X_{...}}{\hat{\tau}_{1..MDF} - \hat{\tau}_{2..MDF} - (\alpha\,\hat{\tau}_{1..MDF}\hat{\tau}_{2..MDF} - \beta\,(\hat{\tau}_{1..MDF} - \hat{\tau}_{2..MDF} - 2\hat{\tau}_{1..MDF}\hat{\tau}_{2..MDF}))}$$, where

$$\hat{\tau}_{1..MDF} = \frac{2\beta x^{*}_{11.} + \beta x^{*}_{12.} + \beta x^{*}_{21.} - x^{*}_{11.}}{(2\beta - \alpha)(x^{*}_{11.} + x^{*}_{21.})}$$ and $$\hat{\tau}_{2..MDF} = \frac{2\beta x^{*}_{11.} + \beta x^{*}_{12.} + \beta x^{*}_{21.} - x^{*}_{11.}}{(2\beta - \alpha)(x^{*}_{11.} + x^{*}_{21.})}$$

		first	
		yes	no
second	yes	$n_{1,1}$	$n_{0,1}$
	no	$n_{1,0}$	$n_{0,0}$

α = probability that a true match is linked

β = probability that a false match is accidentally linked

The MDF-correction

Three problems to solve:

- MDF is developed for only two sources
- MDF is available for models without covariates
- We used several blockings for the probabilistic linkage
  The estimation of the α and β is not straightforward
- Ignoring this problem leads to implausible results: in all scenarios for removing erroneous captures the number of missed usual residents are negative

Conclusions

- CRC assumes among else perfect linked sources
- Two ways to correct: different scenarios and MDF-correction
- Theoretically the MDF-correction is superior
- In practice several problems have to be solved
- We will work on that the coming months