Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Measuring the undercoverage of two data sources with a nearly perfect coverage through capture and recapture in the presence of linkage errors

by A. Dasylva, A. Goussanou and C.O. Nambeu

Release date: October 22, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Measuring the undercoverage of two data sources with a nearly perfect coverage through capture and recapture in the presence of linkage errors

A. Dasylva, A. Goussanou, and C.O. Nambeu[1]

Abstract

In the context of its "admin-first" paradigm, Statistics Canada is prioritizing the use of non-survey sources to produce official statistics. This paradigm critically relies on non-survey sources that may have a nearly perfect coverage of some target populations, including administrative files or big data sources. Yet, this coverage must be measured, e.g., by applying the capture-recapture method, where they are compared to other sources with good coverage of the same populations, including a census. However, this is a challenging exercise in the presence of linkage errors, which arise inevitably when the linkage is based on quasi-identifiers, as is typically the case. To address the issue, a new methodology is described where the capture-recapture method is enhanced with a new error model that is based on the number of links adjacent to a given record. It is applied in an experiment with public census data.

Key Words: dual system estimation, data matching, record linkage, quality, data integration, big data.

Disclaimer: The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that may not reflect those implemented by the Agency.

1. Introduction

There is an increasing reliance on non-survey data including administrative data and big data in official statistics. However measuring the coverage of these sources is essential, e.g. when providing quality indicators for a given source or when estimating population sizes from administrative data . To this end, one may use the method of capture and recapture that rests on many assumptions including that of an error-free linkage. In practice, this assumption must be relaxed, despite the challenge when trying to account for linkage errors, including false negatives and false positives, where a false negative is failing to link records from the same population unit, and a false positive is linking records from different units.

In previous work, Ding and Fienberg (1994) have addressed the problem in the context of dual system estimation based on a census and a coverage survey, when assuming that the false positives do not involve units (i.e. their records) that are selected in both sources, with a known probability that unit has at least one false positive, given that it is only included in the coverage survey. Thus a false positive only occurs between units that are selected either by the census or the coverage survey and not by both. Di Consiglio and Tuoto (2015) have extended this solution for dual system estimation based on two administrative lists, under the same assumption about the units included in both lists, while accounting for the known probability that unit has at least one false positive, given that it is included in a single list. In practice, it may be hard to satisfy the assumption of no false positives for the units selected in both lists. Besides it is difficult to estimate the probability that a unit has at least one false positive, given that it is included in a single list. This challenge remains even when the recall and precision are given. Indeed, deriving the probability of interest from these accuracy measures is not straightforward, because they apply to record pairs, while this probability involves all the pairs that are formed between a record and all the records in the other list. Račinskij et al. (2019) have described a solution when the linkage variables satisfy assumptions of conditional independence and the linkage errors are estimated through a log-linear mixture. A related solution relies on two linkage decisions that are assumed independent

[1]Abel Dasylva, Statistics Canada, 100 Tunney's pasture driveway, Ottawa ON, K1A0T6, abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistics Canada, 100 Tunney's pasture driveway, Ottawa ON, K1A0T6; Christian-Olivier Nambeu, Statistics Canada, 100 Tunney's pasture driveway, Ottawa ON, K1A0T6

for records that are from the same units (see Brown et al., 2020). However, the resulting estimators may be biased, where the linkage decisions depart from this assumption. This is connected to the problem of correlation bias previously mentioned by Newcombe (1983, chap E.6, p. 149), Belin and Rubin (1995) and Blakely and Salmond (2002).

This communication describes a new methodology to the problem of capture and recapture with linkage errors, while relaxing the assumption of no false positives for the units captured in both lists and the assumption that the linkage variables are conditionally independent. The new methodology yields a lower-bound on the coverage of each list. It also yields an estimator of the actual coverage if it can be assumed that there are no false negatives or if the two lists are linked with multiple linkage decisions, where the recall of each decision is given by a shared parametric model. The methodology is an adaptation of the error model by Dasylva and Goussanou (2020) and an example of "linkage free dual system estimation" (Račinskij et al., 2019), because no file of linked pairs is produced.

The remaining sections are organized as follows. The next section provides some background. It is followed by sections describing the proposed methodology and the experiment with public census data, in sequence. The last section contains the conclusion and future work.

2. Background

The basic capture-recapture setup includes a finite population U with an unknown population size N, two independent lists s_A and s_B from this population, where the probability of inclusion in s_A is uniform. The goal is how to estimate the population size N and the coverage of each list. With the methodology by Petersen (1896) and Lincoln (1930), the population size and the coverage of the two lists are estimated as follows.

$$\begin{aligned} \widehat{N} &= |s_A||s_B|/|s_A \cap s_B|, \\ \hat{P}(i \in s_A) &= |s_A \cap s_B|/|s_B|, \\ \hat{P}(i \in s_B) &= |s_A \cap s_B|/|s_A|. \end{aligned}$$

However, these estimators rest on the following key assumptions that must be relaxed in practice.

1. *Closed population*: i.e. no births, no deaths and no migration.
2. *Independent lists*: inclusion in the first list is independent of inclusion in the second list.
3. *Independent units*: each unit is included in the first or the second list independently of the other units.
4. *Homogeneous capture*: a uniform inclusion probability for at least one list.
5. *No overcoverage*: no duplicates and no out-of-scope units in each list.
6. *Perfect linkage*: no errors when identifying the units in both lists.

Relaxing the perfect linkage is required when the linkage variables are nonunique and have typos, so that linkage errors do occur including false positives and false negatives. A false positive leads to overestimating the intersection, while a false negative leads to underestimating this intersection. These errors are measured by the recall and the precision, where the recall is the proportion of links among the record pairs that comprise records from the same unit, while the precision is the proportion of pairs with records from the same unit among the links.

Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015) have described solutions for capture-recapture with linkage errors, when there are no false positives for the units that are selected in both lists, with a known probability that unit has at least one false positive, given that it is included in a single list. However these requirements are hard to meet in practice. Račinskij et al. (2019) have described a different solution, which assumes the conditional independence of the linkage variables and estimates the population size and the coverage without producing a file of linked pairs. However the conditional independence assumption may not apply, resulting in some bias.

2. Methodology

The proposed methodology aims to relax the assumption of no false positives for the units in both lists and that of conditionally independent linkage variables. It is another example of "linkage free dual system estimation" that adapts the error model by Dasylva and Goussanou (2020). This methodology produces a lower-bound on the coverage of each list. It also yields consistent estimators of the coverage of the lists if it can be assumed that there are no false negatives or if the files are linked with multiple linkage decisions, where the recall of each decision is given by a shared parametric model.

Obtaining a lower-bound on the coverage: To motivate the coverage lower-bound, first observe that the coverage of s_A is no less than the conditional joint probability that unit i is in s_A and a true positive (TP) given that the unit is in s_B, where a true positive is linking two records from the same unit and the recall corresponds to the conditional probability of a TP given that the unit is in both lists, i.e. $P(a\,TP\,|i \in s_B \cap s_A)$.

$$P(i \in s_A) \quad = \quad P(i \in s_A\,|i \in s_B) \quad \geq \quad P(i \in s_A\; and\; a\; TP\,|i \in s_B).$$

From this lower-bound, it is straightforward to derive an upper-bound on the population size and a lower-bound on the coverage of s_B. The lower-bound on the coverage of s_A may be estimated by reusing the finite mixture model described by Dasylva and Goussanou (2020), while giving a new meaning to the model parameters. To this end, the first step is noting the important connection between the number of links n_i from record $i \in s_B$ and the linkage errors involving this record, as shown in Table 3-1. This table differs from the similar table described by Dasylva and Goussanou (2020) because each list has some undercoverage. When $n_i = 0$, there is no or one false negative because the unit may be outside s_A or it may be on s_A but with a false negative, with no way of knowing which case applies. However it is known that there are no false positives. When n_i is positive, there is still no or one false negative but there are $n_i - 1$ or n_i false positives, which does provide a lot of useful information. Clearly a statistical model is required where some uncertainty remains, i.e. where n_i is positive.

Table 3-1
Neighbours and errors.

n_i	False negatives	False positives
0	0 or 1	0
$1 \leq n_i \leq \lvert s_B \rvert - 1$	0 or 1	$n_i - 1$ or n_i

The model arises as a limit in distribution under regularity conditions that slightly differ from those previously given by Dasylva and Goussanou (2020). In order to detail these conditions, let the two lists be samples drawn from notional registers A and B, and let unit i be associated with record v_i in B and record $v'_{\pi(i)}$ in A, for some uniformly random permutation $\pi(.)$ of $\{1, \ldots, N\}$. The recording and list inclusion processes are assumed to be such that $\left[\left(I(i \in s_A), I(i \in s_B), v_i, v'_{\pi(i)}\right)\right]_{1 \leq i \leq N}$ are independent and identically distributed and independent of the random permutation $\pi(.)$. The records are assumed to take their values from some finite but possibly large set $\mathcal{V}$. The decision to link two records is such that it is characterized by a set-valued function $\mathcal{B}(.)$ from $\mathcal{V}$ into the power set $2^{\mathcal{V}}$, i.e. the

set of all subsets of $\mathcal{V}$, such that v_i is linked to v'_j if and only if $v'_j \in \mathcal{B}(v_i)$. For $v \in \mathcal{V}$, call $\mathcal{B}(v)$ the neighbourhood of v and define the conditional probabilities

$$\begin{aligned} p_N(v) &= P\big(i \in s_A \text{ and } v'_{\pi(i)} \in \mathcal{B}(v_i) | i \in s_B \text{ and } v_i = v\big), \\ \lambda_N(v) &= P\Big(i' \in s_A \text{ and } v'_{\pi(i')} \in \mathcal{B}(v_i) | i \in s_B \text{ and } v_i = v\Big), \quad i' \neq i, \end{aligned}$$

where the functions $p_N(.)$ and $\lambda_N(.)$ incorporate the inclusion mecanisms in the two lists with a partial overlap, unlike the related functions in Dasylva and Goussanou (2020), where both sources are complete registers, or Dasylva and Goussanou (2021), where one source is a complete register and the other is a file. For $i \in s_B$, let

$$n_i = \sum_{j \in s_A} I\Big(v'_j \in \mathcal{B}(v_i)\Big),$$

denote the number of neighbours. With the above notation, a TP corresponds to the event $\{i \in s_A \text{ and } v'_{\pi(i)} \in \mathcal{B}(v_i)\} \cap \{i \in s_B\}$, a false positive corresponds to the event $\{i' \in s_A \text{ and } v'_{\pi(i')} \in \mathcal{B}(v_i)\} \cap \{i \in s_B\}$ with $i' \neq i$, and

$$P(i \in s_A \text{ and a } TP \mid i \in s_B) = P\big(i \in s_A \text{ and } v'_{\pi(i)} \in \mathcal{B}(v_i) | i \in s_B\big).$$

With the above refined definitions of $p_N(.)$, $\lambda_N(.)$ and n_i, the same regularity conditions can be assumed as by Dasylva and Goussanou (2020), i.e. a piecewise-constant function $\big(p_N(.), (N-1)\lambda_N(.)\big)$, with a finite number of levels, a bounded expected number of false positives (i.e. $sup_{v \in \mathcal{V}}(N-1)\lambda_N(v) \leq \Lambda$ for some positive Λ) and a joint distribution of $p_N(v_i)$ and $(N-1)\lambda_N(v_i)$ that is invariant with respect to N. Then, with arguments similar to those used by Dasylva and Goussanou (2020), it can be shown that

$$n_i | i \in s_B \xrightarrow{d} \sum_{g=1}^{G} \alpha_g Bernoulli(p_g) * Poisson(\lambda_g),$$

where * denotes the convolution operation. Hence $P(i \in s_A \text{ and a } TP \mid i \in s_B) \to \bar{p} = \sum_{g=1}^{G} \alpha_g p_g$, where the parameters $\big[(\alpha_g, p_g, \lambda_g)\big]_{1 \leq g \leq G}$ may be estimated by maximizing the composite likelihood of the n_i's, with G determined through the minimization of Akaike information criterion.

The coverage lower-bound estimator is appealing because it applies regardless of the correlation structure of the linkage variables and requires no clerical reviews. Thus it is quite inexpensive. It is of interest in applications that require one or both sources to have a minimum coverage, or if checking that the union of multiple independent lists provide a near complete coverage of the target population. For example, when building a frame from L independent lists where list l has a coverage of at least ε_l, then the union of the lists has a coverage of at least $1 - \prod_{l=1}^{L} \varepsilon_l$ (e.g. 99.22% when $\varepsilon_l = 1/2$ and $L = 7$), which provides a way of checking that enough lists are included for the intended purpose. Of course, duplicate records (i.e. records from the same unit regardless of whether they are identical) must be identified and accounted for when producing estimates, e.g. with the fractional counting technique by Zhang (2019). Yet it is still of interest to estimate the actual coverage of each list.

Estimating the actual coverage: The coverage is obviously an estimator of the actual coverage if it can be assumed that there are no false negatives, i.e. assuming $P(a\,TP\,|i \in s_B \cap s_A) = 1$ or $p_N(v) = 1$ on $\mathcal{V}$. The actual coverage can also be estimated from multiple linkage decisions and a shared parametric model for the recalls of these different decisions. To be specific consider Γ decisions and let $\mathcal{B}^{(\gamma)}(.)$ denote the set-valued function associated with the decision $\gamma = 1, \ldots, \Gamma$, such that v_i is linked to v_j' by this decision, if $v_j' \in \mathcal{B}^{(\gamma)}(v_i)$. Let $n_i^{(\gamma)}$ denote the corresponding number of neighbours. Each decision is assumed to satisfy the regularity conditions given above so that

$$n_i^{(\gamma)}|i \in s_B \xrightarrow{d} \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} Bernoulli\left(p_g^{(\gamma)}\right) * Poisson\left(\lambda_g^{(\gamma)}\right)$$

and the coverage lower-bound may be estimated as described above. This lower-bound is denoted by $\bar{p}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} p_g^{(\gamma)}$ and it is assumed to be of the form $\bar{p}^{(\gamma)} = P(i \in s_A\,|i \in s_B) r^{(\gamma)}(\boldsymbol{\beta})$, where $r^{(\gamma)}(\boldsymbol{\beta})$ is the recall; $r^{(\gamma)}(.)$ being a known function and $\boldsymbol{\beta}$ being a d-dimensional parameter that is shared among all the decisions, where $d \leq \Gamma - 1$. Then the coverage may be estimated by the following method of moments. Let $\hat{\bar{p}}^{(\gamma)}$ denote the estimated lower-bound for the decision γ. Then $\boldsymbol{\beta}$ may be estimated by the solution of the following system of equations.

$$\left(\sum_{\gamma=1}^{\Gamma} \hat{\bar{p}}^{(\gamma)}\right)^{-1} \begin{bmatrix} \hat{\bar{p}}^{(1)} \\ \vdots \\ \hat{\bar{p}}^{(\Gamma)} \end{bmatrix} = \left(\sum_{\gamma=1}^{\Gamma} r^{(\gamma)}(\hat{\boldsymbol{\beta}})\right)^{-1} \begin{bmatrix} r^{(1)}(\hat{\boldsymbol{\beta}}) \\ \vdots \\ r^{(\Gamma)}(\hat{\boldsymbol{\beta}}) \end{bmatrix}.$$

Consequently, the coverage may be estimated by

$$\hat{P}(i \in s_A\,|i \in s_B) = \frac{\sum_{\gamma=1}^{\Gamma} \hat{\bar{p}}^{(\gamma)}}{\sum_{\gamma=1}^{\Gamma} r^{(\gamma)}(\hat{\boldsymbol{\beta}})}.$$

In general the Γ decisions may be built from K elementary decisions based on $\mathcal{B}_1(.)$,…, $\mathcal{B}_K(.)$, with $\mathcal{B}^{(\gamma)}(.)$ characterized by a nonempty subset $S^{(\gamma)}$ of $\{1, \ldots, K\}$, such that

$$\mathcal{B}^{(\gamma)}(v) = \left(\bigcap_{k \in S^{(\gamma)}} \mathcal{B}_k(v)\right) \cap \left(\bigcap_{k \in \{1,\ldots,K\} - S^{(\gamma)}} \mathcal{B}_k(v)^c\right),$$

where $\Gamma \leq 2^K - 1$. Then $\boldsymbol{\beta}$ is related to the parameters of the loglinear model for $I\left(i' \in s_A\ and\ v'_{\pi(i)} \in \mathcal{B}_1(v_i)\right)$,…, $I\left(i' \in s_A\ and\ v'_{\pi(i)} \in \mathcal{B}_K(v_i)\right)$, including interactions if the elementary linkage decisions are correlated for two records that are from the same unit. For example, one may consider $\Gamma = 3$ decisions, which are built from two decisions based on $\mathcal{B}_1(.)$ and $\mathcal{B}_2(.)$, which are independent for records from the same unit. Let these decisions correspond to $\mathcal{B}^{(1)}(v) = \mathcal{B}_1(v)$, $\mathcal{B}^{(2)}(v) = \mathcal{B}_2(v)$ and $\mathcal{B}^{(3)}(v) = \mathcal{B}_1(v) \cap \mathcal{B}_2(v)$. Let $\bar{p}_1$ and $\bar{p}_2$ denote the coverage lower-bounds associated with $\mathcal{B}_1(.)$ and $\mathcal{B}_2(.)$ and $\boldsymbol{\beta} = [\bar{p}_1 \quad \bar{p}_2]^T$. Then, it is easily shown that the method of moment estimator is $\hat{P}(i \in s_A\,|i \in s_B) = \hat{\bar{p}}^{(1)}\hat{\bar{p}}^{(2)}/\hat{\bar{p}}^{(3)}$. In practice, it may be challenging to select the interactions because the n_i's are correlated such that the standard likelihood ratio test does not apply. A solution is to base inferences on a subset of $o\left(\sqrt{N}\right)$ n_i's that are then approximately independent, as suggested by Dasylva et al. (2019).

Relaxing the homogeneous capture assumption: The above discussion also applies when the capture in s_A is homogeneous within post-strata that are defined based on the auxiliary variables, which are available on list s_B. In this case, a coverage lower-bound and point estimator is obtained within each post-stratum. Then the overall coverage is obtained by aggregating this information across all the post-strata.

3. Data experiment

The proposed methodology is evaluated with simulations in two scenarios, with 100 Monte-Carlo repetitions in each scenario. In each repetition, a synthetic population is generated including one million individuals with the last name and birthdate based on public data from the 2010 US census. The lists are created by drawing independent Bernoulli samples with the inclusion probability of 0.95, and injecting typos into the variables according to Copas and Hilton (1990). For the last name, the errors are generated by first drawing a number t of typos according to a Poisson distribution with intensity ε, where $\varepsilon = 0.01$ in the first scenario and $\varepsilon = 0.0025$ in the second scenario. The last name is then transformed by applying t steps of a finite state machine, where the initial state is the original surname and the state at step $i \leq t$ is the surname modified by the first i typos. At a given step, a random typo is generated independently of the typos in the previous steps, such that it is equally likely to be a deletion or an insertion, at a uniformly random position in the string. When the typo is an insertion, the inserted character is drawn uniformly from the alphabet. For the birth date, each component is modified by adding an independent random error, which is equal to 0, -1 or 1 with probability $1 - \varepsilon$, $\varepsilon/2$ and $\varepsilon/2$ respectively. For simplicity, the modified date is recorded even if it is no longer legitimate, e.g. due to a null day or month component. The typos are generated independently for the last name on one hand and the birth day and birth month on the other hand. Then the files are linked using the decisions described in Table 4-1, where it can be noted that the decisions 1 and 2 are not independent (for records from the same unit) because they both involve the blocking criterion. It would be the case if there were no blocking false negatives, i.e. a recall of 1.0 for decision 0. The coverage lower-bound is estimated by fitting the proposed model with the constraint $p_1 = \cdots = p_G$, after replacing each selected n_i by $\min(\tau, n_i)$ to protect against outliers, with $\tau = 10$. The parameters estimates are computed by nonlinear optimization with the R constrOptim() procedure under the constraints $p_1 \in [0,1[$, $\lambda_g \geq 0$, $\alpha_g \geq 0$ and $\sum_{g=1}^{G-1} \alpha_g \leq 1$. The coverage may be estimated by using the lower-bounds of decisions 1, 2 and 3, as described above, and under the assumption that decisions 1 and 2 are independent.

The performance of the estimated lower-bound (relative to the actual lower-bound) appears on Tables 4-2 and 4-3, while that of the estimated coverage (relative to the actual coverage) appears on Table 4-4. They show that the coverage lower bound is estimated with a small relative bias and a small variance. They also show that the coverage is estimated with a smaller bias in the second scenario, where there is less correlation between decisions 1 and 2 because there are fewer blocking false negatives, i.e. a higher recall for decision 0.

Table 4-1
Linkage decisions.

Decision	Description
0	Blocking criterion based on the same birth year, same last name SOUNDEX and direct or cross agreement on the day and month of birth
1	Blocking criterion and perfect agreement on the last name
2	Blocking criterion and perfect agreement on the birth day and birth month
3	Perfect agreement on all the variables (implies the satisfaction of the blocking criterion)

Table 4-2
Estimated lower-bound for the first scenario.

Decision	Recall	Precision	Actual lower-bound	Estimated lower-bound	
				Relative bias (%)	Variance ($\times 10^{-7}$)
0	0.971	0.207	0.922	-0.540	2.58
1	0.964	0.369	0.915	-0.478	1.58
2	0.933	0.940	0.886	-0.029	1.30
3	0.926	0.972	0.880	-0.030	1.73

Table 4-3
Estimated lower-bound for the second scenario.

Decision	Recall	Precision	Actual lower-bound	Estimated lower-bound	
				Relative bias (%)	Variance ($\times 10^{-7}$)
0	0.993	0.210	0.943	-0.135	1.98
1	0.991	0.372	0.941	-0.119	0.886
2	0.983	0.942	0.934	-0.008	0.693
3	0.981	0.973	0.932	-0.007	0.634

Table 4-4
Estimated coverage for both scenarios.

Scenario	Relative bias (%)	Variance ($\times 10^{-7}$)
1	-3.34	8.73
2	-0.845	1.02

5. Conclusion

A new methodology is proposed for capture-recapture estimation with linkage errors, without clerical-reviews, while relaxing the assumption that there are no false positives for units included in both lists and the assumption that the linkage variables are conditionally independent. It yields a lower-bound on the coverage of each list. It also yields an estimator of the actual coverage if it can be assumed that there are no false negatives or if the two lists are linked with multiple linkage decisions, with a shared parametric model for the recall of these decisions. It is an example of "linkage-free" solution because it estimates the coverage by exploiting the connection between this parameter and the measures of linkage accuracy including the recall and the precision, without producing a file of linked pairs. This work also suggests that a high recall should be prioritized over a high precision when linking for capture-recapture estimation, unlike linkages that have an analytical purpose.

References

Belin, T., and Rubin, D (1995), "A method for calibrating false-match rates in record linkage", *Journal of the American Statistical Association*, 90, 694-707.

Blakely, T., and Salmond, C. (2002). "Probabilistic record linkage and a method to calculate the positive predicted value", *Journal of Epidemiology*, 31, 1246-1252.

Brown, J., Bycroft, C., Di Cecco, D., Elleouet, J., Powell, G., Račinskij, V., Smith, P., Tam, S.-M., Tuoto, T., and Zhang, L.-C. (2020), "Exploring developments in population size estimation", *Survey Statistician*, 82, pp. 27-39.

Copas, J., and Hilton, F. (1990), “Record linkage: Statistical models for matching computer records”, *Journal of the Royal Statistical Society A*, 153, 287-320.

Dasylva, A., Goussanou, A., Ajavon, A. and Abousaleh, H. (2019), “Revisiting the probabilistic method of record linkage”, arXiv:1911.01874.

Dasylva, A. and Goussanou, A. (2020), “Estimating linkage errors under regularity conditions”, *Proceedings of the Survey Methods Section, American Statistical Association*, pp. 687-692.

Dasylva, A. and Goussanou, A. (2021), “Estimating the false negatives due to blocking in record linkage”, *Survey Methodology*, 47.

Ding, Y. and Fienberg, S.E. (1994), “Dual System Estimation of Census Undercount in the Presence of Matching Error”, *Survey Methodology*, 20, pp. 149-158.

Di Consiglio, L., and Tuoto, T. (2015), “Coverage Evaluation on Probabilistically Linked Data”, *Journal of Official Statistics*, 31, pp. 415-429.

Lincoln, F.C. (1930). “Calculating Waterflow Abundance on the Basis of Banding Returns”, *United States Department of Agriculture Circular*.118:1-4.

Newcombe, H. (1988), *Handbook of Record Linkage*, Oxford University Press.

Petersen, C. G. J. (1896). “The Yearly Immigration of Young Plaice Into the Limfjord From The German Sea”, *Report of the Danish Biological Station (1895)* **6**, 5-84.

Račinskij, V., Smith, P. A., and van der Heijden, P. (2019), “Linkage Free Dual System Estimation”, arXiv:1903.10894.

Zhang, L. C. (2019), On provision of UK neighbourhood population statistics beyond 2021. Report for the ONS.