DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 889–908, http://dx.doi.org/10.2478/JOS-2018-0044

Population Size Estimation and Linkage Errors: the Multiple Lists Case

Loredana Di Consiglio[1] *and Tiziana Tuoto*[1]

Data integration is now common practice in official statistics and involves an increasing number of sources. When using multiple sources, an objective is to assess the unknown size of the population. To this aim, capture-recapture methods are applied. Standard capture-recapture methods are based on a number of strong assumptions, including the absence of errors in the integration procedures. However, in particular when the integrated sources were not originally collected for statistical purposes, this assumption is unlikely and linkage errors (false links and missing links) may occur. In this article, the problem of adjusting population estimates in the presence of linkage errors in multiple lists is tackled; under homogeneous linkage error probabilities assumption, a solution is proposed in a realistic and practical scenario of multiple lists linkage procedure.

Key words: Probabilistically linked data; capture-recapture model; multiple system estimation; log-linear model.

1. Introduction

The integration and combination of external sources with traditional statistical survey data is a pressing challenge for National Statistical Institutes. Micro-level integration of different sources is standard practice, generally performed by means of record linkage techniques. However, the linkage process is not completely error-free and statisticians must take linkage errors into account in subsequent analyses performed on integrated data (Chambers 2009). Linkage errors appear particularly relevant when the goal is to measure the size of a population (partially) enumerated in different lists, as shown in Di Consiglio and Tuoto (2015). A widespread method for population size estimation in the presence of two lists is the capture-recapture model (see Petersen 1896; Lincoln 1930; Pollock et al. 1990; Wolter 1986).

The capture-recapture method is subject to the following assumptions:

1. Perfect matching among lists,
2. Independence of lists,
3. Homogeneity of capture probabilities,
4. Closure of population, and
5. No out-of-scope units in the lists.

[1] Italian National Institute of Statistics (Istat), via Balbo 16, 00184 Roma, Italy. Emails: diconsig@istat.it and tuoto@istat.it

When more than two lists are considered, say k, the observations from multiple captures can be organized into a 2^k table, with the presence/absence on the ith list defining the category for the ith dimension. The cell count corresponding to no capture for all the k lists is unknown. Therefore, the goal of estimating the number of units in the population corresponds to the estimation of the unknown count of the missing cell in the 2^k incomplete contingency table.

Several procedures using log-linear models have been proposed (Fienberg 1972; Cormack 1989). When more than two lists are considered, the use of log-linear models enables the independence assumption to be weakened, even if higher order interactions are still subject to restrictions due to model identification. The original log-linear models proposed in Fienberg (1972) rely on the other assumptions: perfect linkage, homogeneity of capture probabilities, closed population, absence of over-coverage. Extensions to the basic log-linear models are provided. Just to mention a few examples, Cormack (1989) discusses the use of log-linear models for dependence and the detection of the presence of heterogeneity in capture probabilities; Darroch et al. (1993) and Agresti (1994) introduce models in the generalised class of Rasch models to explain the heterogeneity in capture probabilities; Coull and Agresti (1999) introduce generalised mixture models. Evans et al. (1994) suggest applying log-linear models when the heterogeneity effects can be explained by the observable covariates. IWGDMF (1995) reviews these approaches, see Chao (2001) for an overview. Zwane and van der Heijden (2005) propose conditional multinomial logit models allowing the inclusion of covariates in the models; Bartolucci and Forcina (2006) introduce latent class models that can be viewed as an extension of conditional multinomial logit models. These models permit accounting for both the observed heterogeneity using covariates and the unobserved heterogeneity, by assuming units to belong to distinct latent classes. Finally, a Bayesian approach can be found in Farcomeni and Tardella (2009).

When more than two lists are available, Di Cecco et al. (2017) discuss the use of a generalisation of the Latent Class models that can be expressed as log-linear models with a latent variable to deal with the problem of out-of-scope units.

Few contributions (Ding and Fienberg 1994, Lee et al. 2001; Di Consiglio and Tuoto 2015) have addressed the issue of matching errors in the population size estimation with two lists. This article explores adjustments for linkage errors in population size estimators, when $k > 2$ lists are considered. Extending the previous works of Di Consiglio and Tuoto (2015) and Fienberg and Ding (1996), this article takes into account both erroneous links and missing links in a realistic linkage error generation model.

The article is organised as follows: Section 2 briefly describes the linkage model and the errors when more than two lists are integrated, Section 3 presents the effects of linkage errors on the observed 2^k incomplete contingency table, as well as a formulation that relates the observed table with the true one, via the linkage errors. In Section 4, the procedure to estimate the population size using the log-linear model is defined. Section 5 discusses the definition of linkage errors used in this framework and reviews a few proposals for their estimation. In Section 6, the application of the proposed method is illustrated in the context of census and administrative data, whereas simulated data are used to analyse its statistical performance and to carry out a sensitivity analysis on the misspecification of linkage errors. Finally, Section 7 provides some concluding remarks and open issues to be tackled by future research.

2. Multiple Lists and Record Linkage

Record linkage is the activity of recognising the same real word entity, even if differently represented in the several data sources. When a common unique identifier is not available, the record linkage techniques exploit common attributes, potentially affected by errors and missing values, to identify the same unit. Therefore, at the end of a linkage procedure, records referring to the same real world entity may emerge unlinked (false negative). In a similar way, false matches may occur when the integration procedure links a pair of units that do not actually relate to the same real-world entity (false positive).

To exemplify, let us consider the two-list case, as in the seminal article of Fellegi and Sunter (1969), say L_1 and L_2, of size N_1 and N_2. Let $\Omega = \{(a, b), a \in L_1 \text{ and } b \in L_2\}$ be the Cartesian product of all possible pairs, of size $|\Omega| = N_1 \times N_2$. The record linkage between L_1 and L_2 is viewed as a classification problem, where the pairs in Ω have to be assigned to two subsets M and U, independent and mutually exclusive, such that:

M is the link set $(a = b)$

U is the non-link set $(a \neq b)$.

Common identifiers (linking variables) are chosen and, for each pair, a comparison vector, denoted by γ, is obtained. Let r be the ratio between the conditional probability of γ given that the pair belongs to the set M and the conditional probability of γ given that the pair belongs to the set U. The ratio r is the likelihood ratio test statistic for testing the null hypothesis H_0: $(a, b) \in M$ against the alternative hypothesis H_1:$(a, b) \in U$, that is

$$r = \frac{P(\gamma|(a,b) \in M)}{P(\gamma|(a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)}. \tag{1}$$

The pairs for which r is greater than an upper threshold value T_m are assigned to the set of linked pairs, M^*; the pairs for which r is smaller than a lower threshold value T_u are assigned to the set of unlinked pairs, U^*; if r falls in the range (T_u, T_m), a no-decision is made automatically and the pair is classified by a clerical review.

The previous thresholds are chosen to minimise the false link probability, denoted by β, and the false non-link probability, denoted by $1 - \alpha$, which are defined as follows:

$$\beta = \sum_{\gamma \in \Gamma} u(\gamma) P(M^*|\gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma) \quad \text{where} \quad \Gamma_{M^*} = \{\gamma : T_m \leq m(\gamma)/u(\gamma)\} \tag{2}$$

$$1 - \alpha = \sum_{\gamma \in \Gamma} m(\gamma) P(U^*|\gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma) \quad \text{where} \quad \Gamma_{U^*} = \{\gamma : T_u \geq m(\gamma)/u(\gamma)\}. \tag{3}$$

In applications, the probabilities m and u can be estimated by treating the true link status as a latent variable, and using the EM algorithm (Jaro 1989). Alternatively, Larsen (1996) applies a Bayesian latent class and Bayesian log-linear models to fit the mixture models (Larsen and Rubin 2001).

When more than two lists have to be linked, for instance, multiple administrative data sets, there are different ways to proceed. Indeed, the standard record-linkage methodologies in use at National Statistical Institutes deal mainly with pairs of lists.

Some proposals for simultaneously linking more than two lists are given by Sadinle et al. (2011); Sadinle and Fienberg (2013); Steorts et al. (2014); Ventura et al. (2014), and Fienberg and Manrique-Vallier (2009). However, currently these methods still need to be "industrialised", so they are not yet suitable for applications in the official statistics production systems due to their computational complexity (Fienberg 2015).

Alternatively, one can match all lists in pairs. A drawback of pairwise linkages is the risk of discrepancies in the linkage decisions. For instance, considering three lists, one can link the record of the individual a in list 1 and the record of an individual b in list 2 from a bipartite record linkage. Then, from a second bipartite record linkage, one links the record of b to the record of an individual c in list 3. Based on these two linkages, one might conclude that a, b, and c are the same individual. However, one also links the first and third lists, but the records a and c may emerge unmatched. If the records a, b, and c truly correspond to the same individual (entity), a nonmatch may occur due to measurement error or incomplete record information. On the other hand, if the records of a, b, and c do not refer to the same individual, we have four possibilities: a and b refer to the same individual but c refers to another one, a and c refer to the same individual but b refers to another one, b and c refer to the same individual but a refers to another one, or a, b, and c all refer to different individuals. By using bipartite record linkage for each pair of files, one cannot resolve the matching pattern. While there are various *ad hoc* approaches to resolve the results of multiple bipartite matchings, no formal methodology has appeared in the statistical literature (Herzog et al. 2007).

To solve multiple linkage, a widespread practice in the National Statistical Institutes is to consider a list as a master frame, and then to link each list sequentially into the master frame. In this case, the linkage procedure involving three lists consists of linking firstly list 1 and list 2, and then the resulting frame with list 3. This procedure has the advantages of needing only two linking operations, while the corresponding pairwise links involve three linkage operations; in addition it does not require solving potential discrepancies.

In the following, we consider the latter multiple-list linkage scenario. In the next session, we describe the linkage errors generated by these linkage operations and how they affect the capture-recapture model.

3. Capture-Recapture Model and Transition Matrix

3.1. Capture-Recapture Model

To focus on the effect of linkage errors in the multiple-capture framework, we consider the case of three captures (lists). In the absence of linkage errors, the capture-recapture data can be classified in the following incomplete 2^3 table (Fienberg 1972):

where n_{ijk} is the cell count of the presence/absence in the lists, with $i,j,k = 1,0$. Let π_{ijk} denote the corresponding cell probability. The table is incomplete, due to the fact that the count n_{000} is unobservable.

The linkage errors modify the counts in Table 1 in two ways: the number of observations may increase in some cells and decrease in others; and the total number of different individuals observed in the three lists may change, provided that the total number of observations in each list, $n_{1++}n_{+1+}$, n_{++1}, remains unchanged.

Table 1. True table for cell counts, without linkage errors.

	List 1			
	Present		*Absent*	
	List 3		List 3	
List 2	*Present*	*Absent*	*Present*	*Absent*
Present	n_{111}	n_{110}	n_{011}	n_{010}
Absent	n_{101}	n_{100}	n_{001}	n_{000}

Table 2 reports the observed counts, subject to linkage errors:
where $n^* = \{n^*_{ijk},\ i,j,k = 1,0\}$ denotes the observed counts after the linkage. Let $\pi^* = \{\pi^*_{ijk}, i,j,k = 1,0\}$ denote the corresponding probabilities. Finally, let $n^*_{\cup L}$ *be the sum of observed distinct units.*

3.2. Error Model with Missing and False Links

Fienberg and Ding (1996) propose a correction of the log-linear model that considers the possible transitions from the true configuration n to the observed one n^*, taking into account only the missing links. They assume that: (i) there are no erroneous matches in the linkage process; (ii) a transition can only go downwards by at most one level, and (iii) the probability of remaining at the original state (no missing error) equals α and the probability of a transition to any of the possible states is equal to $(1 - \alpha)/(m - 1)$, where m is the number of all possible states to which transitions are possible and allowed. For example, an individual truly recorded in all the three lists (111) can produce the following patterns {(110), (001)} or {(101), (010)} or {(110), (001)} with equal probability $(1 - \alpha)/3$.

In this article, we suppose that the transition probabilities are related to both the probability of missing a true match and the probability of a false link. Moreover, we apply a more realistic error model that mimics more closely a real three-list linkage process as described in Section 2, that is, we first assume a linkage step of list 1 and 2 and then a linkage to list 3, taking into account different linkage errors in the two linkage steps.

To this purpose, let $1 - \alpha_1$ be the probability of missing a match in the first linkage and $1 - \alpha_2$ be the probability of missing a match in the second linkage; moreover let β_1 be the probability of a false link in step 1 and let β_2 be the probability of a false link in step 2.

Table 2. Observed table for cell counts.

	List 1			
	Present		*Absent*	
	List 3		List 3	
List 2	*Present*	*Absent*	*Present*	*Absent*
Present	n^*_{111}	n^*_{110}	n^*_{011}	n^*_{010}
Absent	n^*_{101}	n^*_{100}	n^*_{001}	n^*_{000}

We study the effect of linkage errors, introducing first the probability of missing a true match. However, differently from Fienberg and Ding (1996), we aim at taking into account the realistic linkage process in two phases. Then, if one only considers the probability of missing a match, the possible alternative "decompositions" generated by a real unit observed in all the three lists (111), counted in n_{111}, result in the observed ones $(ijk)*$ counted in multiple cells, n^*_{ijk}, as follows:

a. $(111)^*$ with probability $\alpha_1\alpha_2$,
b. $(110)^*$ and $(001)^*$ with probability $\alpha_1(1-\alpha_2)$
c. $(101)^*$ and $(010)^*$ with probability $\frac{(1-\alpha_1)(\alpha_2)}{2}$
d. $(011)^*$ and $(100)^*$ with probability $\frac{(1-\alpha_1)(\alpha_2)}{2}$ and finally
e. $(100)^*$ and $(010)^*$ and $(001)^*$ with probability $(1-\alpha_1)(1-\alpha_2)$.

The five events above are complementary and mutually exclusive.

On the second line, for example, when we correctly link the first two lists but we miss the link with the third one, the event b results in the "decomposition" of (111) in $(110)*$ and $(001)*$ with probability $\alpha_1(1-\alpha_2)$. On the contrary, when an error occurs at the first linkage step, the individual is decomposed in two different units, then the third list is correctly linked to either the first or the second one with the same probability $\frac{(\alpha_2)}{2}$ (event c or d).

For convenience, following the terminology of Fienberg and Ding (1996), we call such a decomposition (or combination in case of false matches, discussed below) a "transition".

A similar reasoning for the decomposition of the other true individual patterns allows for a transition matrix M_1 to be obtained. Table 3 reports the matrix with the transition probabilities resulting from the different events that generate the observed patterns after linkage. For instance, the $(001)*$ is generated from (111) when either the event b or the event e of the above example occur. The probability of the transition from (111) to $(001)*$ is then $\alpha_1(1-\alpha_2)+(1-\alpha_1)(1-\alpha_2)=1-\alpha_2$, as in Table 3.

It is worth noting that the columns of the transition matrix M_1 do not necessarily add up to one. The probabilities of the alternative events (missingness/unmissingness of matches in one/two steps) obviously add up to one. However, when a linkage error occurs (e.g., a true match is missed) it affects more than one row of the matrix, generating decomposition/combination of the true unit of the population. This property is consistent with the observation that the sum of the distinct individuals enlisted in Table 1 differs from the sum of the observed distinct units in Table 2.

Table 3. Transition matrix M_1 from real to observed cells when only missing links occur.

	111	110	101	100	011	010	001
$(111)*$	$\alpha_1\alpha_2$	–	–	–	–	–	–
$(110)*$	$\alpha_1(1-\alpha_2)$	α_1	–	–	–	–	–
$(101)*$	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	–	α_2	–	–	–	–
$(100)*$	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$1-\alpha_1$	$1-\alpha_2$	1	–	–	–
$(011)*$	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	–	–	–	α_2	–	–
$(010)*$	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$1-\alpha_1$	–	–	$1-\alpha_2$	1	–
$(001)*$	$1-\alpha_2$	–	$1-\alpha_2$	–	$1-\alpha_2$	–	1

The transition matrix M_1 can be further extended to include the false linkage errors. As before, different linkage errors are assumed for the first and the second phase. In addition, we assume that whenever a true match is missed, the related records cannot be involved in false matches in the same phase, because this event happens when at least two errors occur: the records are incorrectly linked and the correct match is missed. Then, we assume it has a negligible probability of occurrence, as in Ding and Fienberg (1994) and in Di Consiglio and Tuoto (2015). Under the above assumptions, and, at the same time, treating the transitions caused by false and missing linkage errors, we obtain the transition matrix M_2 in Table 4. It is worth noting that the matrix M_2 can contain negative values due to algebra on the probabilities of composition/decomposition generated by the false links.

4. Estimation of Population Size

The true counts in Table 1 can be estimated by a linear combination of the observed counts via the inverse of the transition matrix:

$$n = M^{-1} n^* \tag{4}$$

The transition matrix M can be either M_1 or M_2 (see Tables 3 or 4) according to the adopted error model. Similarly, the cell probabilities can be estimated by $\pi = M^{-1}\pi^*$.

Once the true cell counts are obtained by (4), in order to estimate the population size N, one needs to estimate the unknown count of the missing cell in the 2^k incomplete contingency table, for example applying a suitable log-linear model. For instance, when dealing with three lists, one can use the log-linear saturated model

$$\log\left(E(n_{ijk})\right) = \lambda + \lambda_i^{L_1} + \lambda_j^{L_2} + \lambda_k^{L_3} + \lambda_{ij}^{L_1 L_2} + \lambda_{ik}^{L_1 L_3} + \lambda_{jk}^{L_2 L_3} \tag{5}$$

where the sum of any λ over any subscript is zero. The fitted count $\tilde{n}_{000}$ from the log-linear model is finally used to estimate the population size:

$$\tilde{N} = n + \tilde{n}_{000}. \tag{6}$$

Under the assumption of independence of each pair of lists, we have

$$\tilde{n}_{000} = \frac{n_{111} n_{001} n_{100} n_{010}}{n_{101} n_{011} n_{101}} \tag{7}$$

The assumption of independence of each pair of lists is equivalent to setting $\lambda^{L_u L_v} = 0$ for each u and v. The use of log-linear model, however, enables list pair dependency and its extensions to also take account of the heterogeneity of capture probabilities (see Section 1).

In general, to obtain an estimation of the population size N, we first compute the Maximum Likelihood (ML) estimates of the parameters from the conditional likelihood associated with observed cell count n^* given $n^*_{\cup L}$, as suggested in Fienberg and Ding (1996). Sanathanan (1972) shows that, under suitable regularity conditions, the conditional maximum likelihood estimates and the unconditional ones are both consistent and have the same asymptotic normal distribution. Once the conditional maximum likelihood estimates of the log-linear model are obtained, we use the log-linear model specified for the not-observed real values to compute the conditional maximum likelihood

Table 4. Transition matrix M_2 from real to observed cells with missing and false links.

	111	110	101	100	011	010	001
(111)*	$\alpha_1\alpha_2$	$\alpha_1\beta_2$	$\alpha_2\beta_1$	$\beta_1\beta_2$	$\alpha_2\beta_1$	–	–
(110)*	$\alpha_1(1-\alpha_2)$	$(\alpha_1)(1-\beta_2)$	$\beta_1(1-\alpha_2)$	$(\beta_1)(1-\beta_2)$	$\beta_1(1-\alpha_2)$	–	–
(101)*	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	$(1-\alpha_1)\beta_2/2$	$(1-\beta_1)(\alpha_2)$	$(1-\beta_1)(\beta_2)$	–	–	–
(100)*	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$(1-\alpha_1)\left(1-\frac{\beta_2}{2}\right)$	$(1-\beta_1)(1-\alpha_2)$	$(1-\beta_1)(1-\beta_2)$	$-\beta_1$	–	–
(011)*	$\frac{(1-\alpha_1)(\alpha_2)}{2}$	$(1-\alpha_1)\beta_2/2$	–	–	$(1-\beta_1)\alpha_2$	–	–
(010)*	$\frac{(1-\alpha_1)(2-\alpha_2)}{2}$	$(1-\alpha_1)\left(1-\frac{\beta_2}{2}\right)$	$-\beta_1$	$-\beta_1$	$(1-\beta_1)(1-\alpha_2)$	1	–
(001)*	$1-\alpha_2$	$-\beta_2$	$1-\alpha_2$	$-\beta_2$	$1-\alpha_2$	–	1

estimates of the expected cell counts $\tilde{n}_{ijk}$, including the one of the missing cell. Thus, $\tilde{N} = \sum_{ijk} \tilde{n}_{ijk}$.

5. Focus on Linkage Errors

The linkage errors defined in Formulas (2) and (3) are based on the Fellegi and Sunter (1969) theory for record linkage that is very effective for the link identification. Note that, conceptually, in (2) and (3) the probabilities β and α are defined for each element of the product space $\Omega = L_1 \times L_2$. However, as it is well known in practice, the Fellegi and Sunter (1969) linkage procedure is not reliable for estimating the linkage errors. Tuoto (2016) proposes a supervised learning method to predict both types of linkage errors, without relying on strong distribution assumptions, as in Belin and Rubin (1995). Alternatively, Chipperfield and Chambers (2015) apply a bootstrap method to the actual linkage procedure to evaluate the mismatch probabilities.

On the other hand, in the population size estimation context, it may be necessary to adopt alternative definitions of the linkage errors than (2) and (3). For instance, let us consider the multiple capture counts in Table 2 and the two linkage steps that produced it. At any of the linkage stages, if the true linkage status was known, the errors rates could be defined comparing the links made with the true ones. At the first stage, the results of this comparison could be reported as in Table 5.

Then to assess the quality of the linkage process, the following ratios could be defined:

$$\text{False nonmatched (missed match) rate: } 1 - \alpha = \frac{c}{a+c} = \frac{n_{11} \cap n_{11}^*}{n_{11}}; \tag{8}$$

$$\text{False match rate: } \beta = \frac{b}{b+d} = \frac{n_{11}^* - n_{11} \cap n_{11}^*}{(N_1 - n_{11}) + (N_2 - n_{11})}. \tag{9}$$

Clearly, the definition of false match error β in (9) is more pragmatic than in (2), because the set of all the unlinked pairs $U = (N_1 - n_{11}) \times (N_2 - n_{11})$ is a much larger set than $(N_1 - n_{11}) + (N_2 - n_{11})$, since the false matches in (9) are related to the unlinked cases of both the lists, rather than to the cross-product of the lists, as in (2), where the unlinked pairs set U is considered. Moreover, it is worth noting that one can expect the number of false links involving the actually linked records to be much lower than the number of false links between unlinked records, because the former implies two linkage errors simultaneously, that is, missing the true match and erroneously linking the matched record to a different record.

Finally, it should be pointed out that the false match rate defined by (9) is a different quantity to the false match rate used for adjusting regression analysis (e.g., in Chambers, 2009), where the latter is defined in relation to the number of actual links n_{11}^*. While both

Table 5. Comparison of true matches and assigned links.

	True matches	True non-matches
Links	a – true positives	b – false positives or false links
No links	c – false negatives or missing links	d – true negatives

quantities target the same number of false links among the links made, the two rates are not the same measure, because they have different denominators.

6. Applications

In this section, we present some applications of multiple capture estimation method in the presence of linkage errors. Firstly, in Subsection 6.1, the adjusted estimator derived applying transformation (4) with M_2 is applied in a real-life context, the census, post-enumeration survey and administrative data example already considered by Fienberg and Ding (1996). In Subsection 6.2, we propose a simulation study to analyse the empirical statistical properties of the suggested estimators; in Subsection 6.3, the simulation study provides a sensitivity analysis to show the robustness of the population size estimates with respect to the linkage error evaluation.

6.1. Example from Census, PES and Administrative Data

First, let us consider the data from the three lists previously used by Fienberg and Ding (1996): the 1990 U.S. Census, the corresponding post-census survey (PES), and the administrative list supplement (ALS). Data for sampling strata PES 11 at St. Louis are given in Table 6.

For the evaluation of the matching errors, Fienberg and Ding (1996) use the Matching Error study (see Mulry et al. 1989) to assess both the probability of missing a link in the linking procedure between the Census and the PES, and the probability of missing a link in the linkage involving the ALS, under the assumption of no errors in the rematch. The results of the Matching Error Study for 1990 U.S. Census in St. Louis stratum are reported in Table 7 (see Table 4, 562 in Fienberg and Ding, 1996).

Ignoring the unresolved cases, Fienberg and Ding (1996) estimate the probability of missing a true link as $(1 - \hat{\alpha}_1) = (1 - \hat{\alpha}_2) = 9/(2{,}667 + 9) = 0.3363\%$. Following the same reasoning, we evaluate the probability of a false link as $\hat{\beta}_1 = \hat{\beta}_2 = 7/(7 + 427) = 1.6129\%$. It is worth noting that the false linkage error is much greater than the missing linkage error, suggesting the need to correct also for false links.

For the estimation of the unknown size of the population, Fienberg and Ding (1996) examine various log-linear models with different dependency structures in order to better fit the data in Table 6. The model [CensusPes][PesALS] results to fit the data

Table 6. Three-sample data for stratum 11, St. Louis, 1990 U.S. Census.

	Census = List 1			
	Present		Absent	
	PES = List 2		PES = List 2	
ALS = List 3	Present	Absent	Present	Absent
Present	300	51	53	180
Absent	187	166	76	–

Table 7. St. Louis rematch study.

Original match classification	Rematch classification			
	Matched	Not matched	Unresolved	Total
Matched	2,667	7	8	2,682
Not matched	9	427	30	466
Unresolved	0	7	20	27
Total	2,676	441	58	3,175

much better. The corresponding naïve estimate is $\hat{N} = 1,599$. Applying their correction for missing links, Fienberg and Ding (1996) estimate $\tilde{N}_{DF} = 1,585$. Instead, including the false linkage errors as well, with the error matrix M_2 specified in Table 4, we get $\tilde{N}_{MDF} = 1,680$. This value is within the confidence interval of both of the previous estimates.

6.2. *Results on Simulated Data*

This section describes the results of a simulation on fictitious data. To simulate the linkage process in a realistic way, we use person identifiers from the fictitious population census data (McLeod et al. 2011) created for the ESSnet DI, which was a European project on data integration (Record Linkage, Statistical Matching, Micro integration Processing) running from 2009 to 2011.

The ESSnet DI provides three entirely fictitious data sources, which are supposed to have captured details of persons at the same reference time. The first data set consists of observations from the Patient Register Data of the National Health Service (PRD, in the following); the second data set contains observations from the Customer Information System (CIS), which combines administrative data from the tax and social security systems; the third data set reports observations from a decennial Census (CEN). In these data sets, which comprise over 26,000 records each, linking variables (names, dates of birth, addresses) for individual identification may be distorted by missing values and typos, to imitate real-life situations. These synthetic data reproduce the real data and the actual observed errors that make the linkage procedure difficult. For details on the generation of synthetic data and the perturbation of the key variables, see McLeod et al. (2011). The simulation setting lets us know the true match status to benchmark the linkage results. In the simulation, 500 populations of the size 1,000 were generated, sampling the data independently and randomly without replacement.

For each replicate, the three lists were randomly drawn by the PRD, CIS and CEN on the basis of the following capture probabilities: $\pi_{1++} = 0.65$, $\pi_{+1+} = 0.53$ and $\pi_{++1} = 0.57$, respectively.

At each replicate, the linkage was made as illustrated in Section 2: in the first phase, the PRD and CIS lists were linked; in the second phase, the linked and un-linked records of the first phase were linked to the third list (CEN).

Table 8. Distribution of the linkage error rates over the 500 replicates.

Linkage errors%	Min	Median	Mean	Max
First step				
$1\text{-}\alpha_1$	0.00	2.39	2.51	7.33
β_1	0.00	4.58	4.31	7.63
Second step				
$1 - \alpha_2$	0.90	2.86	2.91	5.93
β_2	0.20	4.53	4.10	8.56

In both steps, the linkage variables were Name, Surname, Day, Month and Year of Birth, and the probabilistic record linkage model (Fellegi and Sunter 1969, Jaro 1989) was implemented by the batch version of the software RELAIS (RELAIS, 2015).

Table 8 summarises the results of the linkage procedure in terms of realised linkage error rates, reporting the probability of missing a true match $1\text{-}\alpha$ and the probability of a false match β for both steps, as defined in Section 5 (see Formulas 8 and 9 for step 1). The realised $1\text{-}\alpha$ and β can be evaluated in light of the known true linkage status.

At each replicate, we compute the naïve log-linear estimator and the adjusted estimators, applying the transformation (4) with M_1 or M_2 as described in Section 4. Having generated the three lists independently, the log-linear model assumes the independency of the lists. The adjusted estimator was computed using the true values of the probability of nonmissing true matches α and the probability of false match β obtained in each replicate. The use of the true values of α and β allows us to compare the estimators without the effect of the linkage error estimation, hence focusing on the performance of the adjusted estimator.

Figure 1 shows the distributions over the 500 replicates of the several estimators: the naïve estimator, the adjusted estimator taking account of missing links only (DF as Ding and Fienberg) according to the matrix M_1 and the adjusted estimator taking account of the two types of linkage errors (MDF, modified DF) according to the matrix M_2 in Table 4. For comparison, the figure shows the estimates that can be obtained with the true counts unaffected by linkage errors.

The relative percentage errors of the estimators are summarised in Table 9. The table shows the minimum value, the first quartile, the median, the average, the third quartile

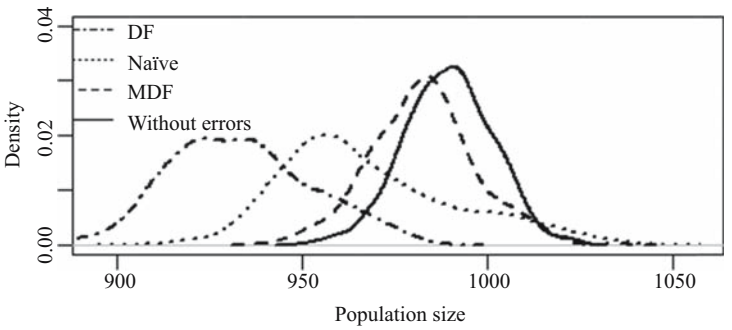

Fig. 1. Empirical density of the alternative estimates of the population size over the replicates (true N = 1,000).

Table 9. Distribution of percentage relative error.

	Percentage relative error					
Estimator	Min	Q1	Median	Mean	Q3	Max
Naïve	− 8.70	− 4.90	− 3.70	− 3.19	− 1.70	3.90
DF	− 11.70	− 8.11	− 6.70	− 6.67	− 5.40	− 1.60
MDF	− 5.90	− 2.70	− 1.75	− 1.74	− 0.90	3.50
True values	− 4.80	− 1.82	− 1.00	− 1.05	− 0.20	2.30

and the maximum value of the relative percentage of error calculated over the 500 replicates.

The results in Figure 1 and Table 9 show that the proposed adjustment reduces the bias of the naïve estimator without side effects on the variability of the estimator, even if the bias is not entirely removed due to the non-linear nature of the population size estimator. Likewise, the residual bias may be due to the misspecification of the linkage error model: it is observed in this simulation, as well as in other real applications (Tuoto et al. 2017) that the probability of double errors (i.e., missing a true link and false link of the records at the same time) may be not negligible, as assumed in the proposed transition matrix M_2.

6.3. A Sensitivity Analysis

The simulation setting is exploited for a sensitivity analysis of the proposed estimator with respect to the misspecification of the linkage errors. In the previous subsection, the MDF estimator was calculated under optimal conditions, that is, knowing the values of the linkage errors made. In this section, several values of α_1, α_2, β_1 and β_2 are tested to evaluate the statistical properties of the MDF estimator in different nonoptimal scenarios. First, we apply the MDF estimator with the four average linkage errors over the 500 replicates – we denote the estimator as MDF_{mmmm} in the following. Moreover, the variability of the linkage errors is accounted for in MDF estimates by evaluating the matrix M_2 with several combinations of the lower and upper bounds of the confidence intervals over the 500 replicates. We denote $\mathbf{MDF}_{\boldsymbol{\alpha_1\beta_1\alpha_2\beta_2}}$ where the subscripts take values in {o, m, l, u}, standing for "observed", "mean", "lower bound of the confidence interval", "upper bound of the confidence interval" respectively.

Figure 2 compares the true values, the naïve estimates and the adjusted estimators $\text{MDF}_{\alpha_1\beta_1\alpha_2\beta_2}$.

As expected, the MDF estimator with true observed linkage errors outperforms the MDF estimators with different values (m, l, u) of the linkage errors, both in terms of bias and variability. However, when we compare the naïve estimator and the MDF estimators with inaccurate values of the linkage errors, the results are diverse. Figure 2 shows that the MDF estimate still improves the naïve one, at a cost of a slight increase in variability, when using the linkage error averages. As expected, when using the lower bound of the confidence intervals of the errors, the MDF estimates tend to the naïve one. On the contrary, when applying the upper bound of the confidence intervals (i.e., on average

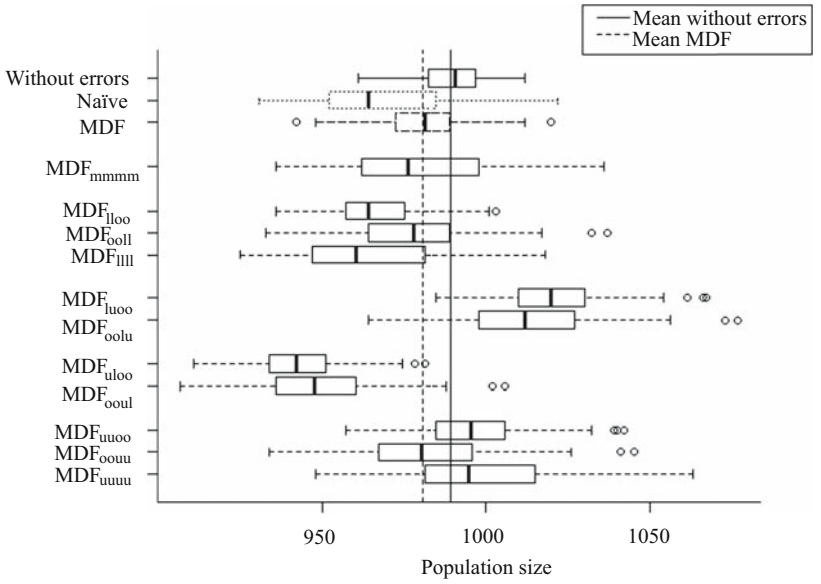

Fig. 2. Simulated alternative estimates of the population size (true $N = 1{,}000$) with different values of linkage errors.

applying an over-correction), there is a tendency to overestimate. Finally, the MDF correction is ineffective when the missing linkage errors are overestimated and the false linkage errors are underestimated, or viceversa.

This analysis also shows that the adjustment with an inaccurate evaluation of the second step linkage errors causes an increase in the variability but produces less bias in the estimates compared to the bias caused by an inaccurate evaluation of the first step errors, that is, once the linkage errors at the first step are misspecified we cannot adjust only with the second step error probabilities. On the other hand, this sensitivity analysis indicates that the independence assumption on linkage errors may not hold, as anticipated at the end of the previous section: in fact, the MDF_{uuoo} and MDF_{uuuu} are on average closer to the true values than MDF estimates.

7. Discussion and Concluding Remarks

This article proposes an extension of the Fienberg and Ding (1996) approach in order to take account of linkage errors in the evaluation of the population size when more than two lists are considered in a multiple system estimation framework.

However, the proposed estimator presents some open issues that need further investigation, partially inherited from the general context of multiple captures. Some reflections are briefly discussed in the next subsections.

7.1. A Note on Variance Estimation

In the estimation of the population size, it is assumed that the counts are distributed according to a log-linear model; using the delta method, Darroch (1958) derives an estimator of the variance of the population size estimator. For instance, when the three lists

are independent, the estimator proposed by Darroch (1958) is as follows

$$\widehat{Var}(\tilde{N}) = \tilde{N}\tilde{n}_{000}\left(\sum_{\{ijk\}\in S}\tilde{n}_{ijk}\right)^{-1} \tag{10}$$

where S contains all cells corresponding to individuals caught more than once.

However, in our context, a straightforward application of Formula (10) on the estimated counts would omit the additional source of variability introduced by the linkage errors process. In fact, when encountering linkage errors, the observations are subject to the multinomial process generating the true captures plus the additional probabilistic process of linking the lists. Then the variance evaluation needs to consider this additional probabilistic process generating the linkage errors. Simply replacing the counts in Formula (10) with their estimates obtained via the transformation (4) would not take into account the latter source of variation. Moreover, in practice, the linkage errors are not known and their estimation will introduce an additional source of error that should be considered.

As an alternative to analytical variance analysis, one can explore a bootstrap approach. The variance estimation of the adjusted estimator is an open issue for future research.

7.2. *Scalability*

This article explicitly evaluates a general adjustment for linkage errors when the population size is based on three sources. The method is readily applicable to the multiple list case; however, a generalisation to $k > 3$ lists requires the evaluation of the transition matrix M and the knowledge of the multiple step linkage mechanism. Considering only the missing link error α, the transition matrix for $k = 5$ is implemented in Link et al. (2010) – see below in Subsection 7.3 for more details. Obviously, when the false link errors are introduced into the analysis, the evaluation of the transition matrix is not straightforward.

It is worth noting that the trade-off between the risk of potential linkage errors and the advantages of increasing the number of lists for the population size estimation should be further investigated by means of case studies.

7.3. *A Bayesian Perspective on the Population Size Estimation*

Alternative approaches to record linkage are based on Bayesian methods. For instance, in Fortini et al. (2001) and Liseo and Tancredi (2011), the interest is focused on a matrix-valued parameter $\boldsymbol{C}$, which represents the true pattern of matches between the two lists. The sum of the elements of $\boldsymbol{C}$ is an estimate of the number of true matches between the two lists, given the following constraints on the parameter space of C that avoid multiple matches:

$$C_{ij} = \{0, 1\}, \qquad \sum_{L_1} C_{ij} \leq 1, \qquad \sum_{L_2} C_{ij} \leq 1.$$

The Bayesian approach enables the propagation of the uncertainty of the linkage process to subsequent analysis of the linkage data in a natural way. According to the knowledge of the authors, this method is only described in the two-list case, but similarly to the Fellegi-Sunter approach, it could be applied by incremental steps that consider an

augmented number of lists. A practical difficulty with the Bayesian approach is the lack of scalability to large data sets, which is the case of the population size estimation in official statistics.

Steorts et al. (2015) propose an alternative Bayesian approach that allows linking records from multiple lists simultaneously while de-duplicating the lists. Similarly to Parag and Domingos (2004), the linkage is considered as a process of recognising latent "entities" with a graphical representation, that is, each record in the lists can be linked to a latent unit from 1 to N_{max}, where N_{max} is the total number of units in all the lists, if no unit is present in more than one. A uniform prior is assumed on the linkage structure, that is, any observed unit is equally assigned to any of the latent individual. A hybrid MCMC algorithm is used to improve the computing performances. However, Steorts et al. (2015) do not utilise their model for the estimation of the population size in the presence of undercoverage. Further research is needed to apply their method in such setting.

Finally, we mention the linkage errors adjustment proposed by Link et al. (2010). They assume only missing matches and no erroneous links; they model the capture-recapture history with a vector where the components are indicators of:

(i) Presence in the given capture and correct identification of the individual,
(ii) Presence in the given capture but missing identification, and
(iii) Absence in the given capture.

However, this model is still subject to the specification of a matrix M. They define the recorded frequency vector n^* as a linear combination of true history n, which is considered as a latent variable. So, the application of the method still requires the actual specification of the M matrix that connects the observed values to the true one, similarly to what is described in this article.

7.4. Concluding Remarks

To summarise, this article first defines a realistic and widely used linkage setting for multiple sources, then the errors caused by both missing and erroneous links are included in the contingency table of the presence/absence of the units in the various sources. The originality of the proposal consists in adjusting for false matches in addition to missing matches, extending the previous works of Fienberg and Ding (1996) and Link et al. (2010). Indeed, the false matches are frequent, as well as missing matches; this fact is also observed in the Matching Error Study (Mulry et al. 1989) on the linkage between 1990 U.S. Census and PES, which is used to apply the proposed adjustment.

The suggested estimator allows reducing the bias of the naïve estimator without relevant effects on variability, even if the bias is not entirely cancelled out due to the nonlinear nature of the estimator. It is worth recalling the assumptions underlying the estimator (6): a. the linkage procedure acts in sequential steps, for instance, two steps in the description of the three-list case provided in Subsection 3.2; b. linkage errors are independent in different steps; c. at each step, the probability of missing a true match and erroneously linking the related records in false matches is negligible, as in Fienberg and Ding (1996); d. the linkage errors are either known or accurately estimated; and e. the linkage errors are homogeneous, at least in sub-groups.

The independence assumption b should be verified as, linkage errors are caused by errors in the matching variables, one can, given the occurrence of these errors, assume that linkage errors in different step are independent. However, the linkage mechanism can be such that if a link is missed (or a false link is introduced) in the first step, this may increase the probability of a linkage error in the second step. In our simulation setup, we tested the adjustment with known linkage errors, evaluating them by means of the known actual matches. The sensitivity analysis shows that the adjusted estimator outperforms the naïve one in several cases, even if the linkage errors are unknown. However, when the missing linkage errors are overestimated and the false linkage errors are underestimated, and viceversa, both at the first and the second step, the MDF correction is ineffective. The simulation and the sensitivity analyses are restricted to one population framework (i.e., Census and administrative data) and one linkage scenario. Other applications or simulation settings can provide further insights and prove the generalisability of the observed results. Moreover, the evaluation of linkage errors and the effect of these errors on the variability of the population size estimates are still open issues.

The proposed estimator is developed assuming constant linkage errors across the entire population. This may not always hold in practice; in those cases, the adjustment can still be applied considering strata in which homogeneous linkage errors occur. As linkage errors depend on errors in the key variables, then homogeneous groups can be built on the basis of them. The gain of the adjusted estimator in the presence of homogeneous strata compared to the use of average values of the errors over the entire population could be examined; this is an aspect to be tackled in future research. However, the sensitivity analysis already provides the insight that the adjustment can still be valuable compared to the naïve estimator, even with error values not corresponding to the true ones.

Finally, additional case studies should be carried out to analyse the statistical properties of the suggested adjustment when considering extensions to basic log-linear models.

8. References

Agresti, A. 1994. "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort." *Biometrics* 50: 494–500. Doi: http://dx.doi.org/10.2307/2533391.

Bartolucci, F. and A. Forcina. 2006. "A Class of Latent Marginal Models for Capture-Recapture Data with Continuous Covariates." *Journal of the American Statistical Association* 101: 786–794. Doi: http://dx.doi.org/10.1198/073500105000000243.

Belin, T.R. and D.B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90: 694–707. Doi: http://dx.doi.org/10.1080/01621459.1995.10476563.

Chambers, R. 2009. "Regression Analysis of Probability-Linked Data." *Official Statistics Research Series* 4. Available at http://www3.stats.govt.nz/statisphere/Official_Statistics_Research_Series/Regression_Analysis_of_Probability-Linked_Data.pdf (accessed November 2018).

Chao, A. 2001. "An overview of closed Capture-Recapture Models." *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158–175. Doi: http://dx.doi.org/10.1198/108571101750524670.

Chipperfield, J. and R. Chambers. 2015. "Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data." *Journal of Official Statistics* 31(3): 397–414. Doi: http://dx.doi.org/10.1515/jos-2015-0024.

Cormack, R.M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Coull, B.A. and A. Agresti. 1999. "The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies." *Biometrics* 55: 294–301. Doi: http://dx.doi.org/10.1111/j.0006-341X.1999.00294.x.

Darroch, J.N. 1958. "The Multiple-Recapture Census: I. Estimation of a closed population." *Biometrika* 45: 343–359. Doi: http://dx.doi.org/10.2307/2333183.

Darroch, J.N., S.E. Fienberg, G.F.V. Glonek, and B.W. Junker. 1993. "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability." *Journal of the American Statistical Association* 88: 1137–1148. Doi: http://dx.doi.org/10.2307/2290811.

Di Cecco, D., M. Di Zio, D. Filipponi, and I. Rocchetti. 2016. "Estimating Population Size from Multisource Data with Coverage and Unit Errors." In Proceeding of the ICES-V, Geneva, Switzerland, June 20–23, 2016. Available at http://ww2.amstat.org/meetings/ices/2016/proceedings/165_ices15Final00072.pdf (accessed November 2018).

Di Consiglio, L. and T. Tuoto. 2015. "Coverage Evaluation on Probabilistically Linked Data." *Journal of Official Statistics* 31(3): 415–429. Doi: http://dx.doi.org/10.1515/JOS-2015-0025.

Ding, Y. and S.E. Fienberg. 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error." *Survey Methodology* 20: 149–158. Available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994002/article/14422-eng.pdf?st=YtHflfaV (accessed November 2018).

Evans, M.A., D.G. Bonett, and L.L. McDonald. 1994. "A General Theory for Modeling Capture-Recapture Data from a Closed Population." *Biometrics* 50(2): 396–405. Doi: http://dx.doi.org/10.2307/2533383.

Farcomeni, A. and L. Tardella. 2009. "Reference Bayesian Methods for Recapture Models with Heterogeneity." *Test, May* 2010, 19(1): 187–208. Doi: http://dx.doi.org/10.1007/s11749-009-0147-9.

Fellegi, I. and A. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64: 1183–2010. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables." *Biometrika* 59: 409–439. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Fienberg, S.E. 2015. "Discussion." *Journal of Official Statistics* 31(3): 527–535. Doi: http://dx.doi.org/10.1515/JOS-2015-0032.

Fienberg, S.E. and Y. Ding. 1996. "Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Error." In Proceedings of 1994 Annual research conference and CASIC technologies Interchange, Bureau of Census, United States. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14385-eng.pdf?st=8LhKz2Tt (accessed November 2018).

Fienberg, S.E. and D. Manrique-Vallier. 2009. "Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation." *Advances in Statistical Analysis* 93: 49–60. Doi: http://dx.doi.org/10.1007/s10182-008-0084-z.

Fortini, M., B. Liseo, A. Nuccitelli, and M. Scanu. 2001. "On Bayesian Record Linkage." *Research in Official Statistics* 4(1): 185–198.

Herzog, T., F. Scheuren, and W. Winkler. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer-Verlag. Doi: http://dx.doi.org/10.1007/0-387-69505-2.

IWGDMF – International Working Group for Disease Monitoring and Forecasting. 1995. "Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Doi: http://dx.doi.org/10.1093/oxfordjournals.aje.a117558.

Jaro, M. 1989. "Advances in Record Linkage Methodology as Applied to Matching the 1985 Test Census of Tampa, Florida." *Journal of American Statistical Association* 84: 414–420. Doi: http://dx.doi.org/10.1080/01621459.1989.10478785.

Larsen, M.D. 1996. *Bayesian Approaches to Finite Mixture Models*, Ph.D. Thesis, Harvard University.

Larsen, M.D. and D.B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41. Doi: http://dx.doi.org/10.1198/016214501750332956.

Lee, A.J., G.A.F. Seber, J.K. Holden, and J.T. Huakau. 2001. "Capture-Recapture, Epidemiology, and List Mismatches: Several Lists." *Biometrics* 57: 707–713. Doi: http://dx.doi.org/10.1111/j.0006-341X.2001.00707.x.

Lincoln, F.C. 1930. *Calculating Waterfowl Abundance on the Basis of Banding Returns*. United States Department of Agriculture Circular, 118, 1–4.

Link, W.A., J. Yoshizaki, L.L. Bailey, and K.H. Pollok. 2010. "Uncovering a Latent Multinomial: Analysis of Mark-Recapture Data with Misidentification." *Biometrics* 66: 178–185. Doi: http://dx.doi.org/10.1111/j.1541-0420.2009.01244.x.

Liseo, B. and A. Tancredi. 2011. "Bayesian Estimation of Population Size Via Linkage of Multivariate Normal Data Sets." *Journal of Official Statistics* 27(3): 491–505. Available at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/bayesian-estimation-of-population-size-via-linkage-of-multivariate-normal-data-sets.pdf (accessed November 2018).

McLeod, P., D. Heasman, and I. Forbes. 2011. Simulated data for the on the job training. Essnet DI. Available at http://www.cros-portal.eu/content/job-training.

Mulry, M.H., A. Dajani, and P. Biemer. 1989. "The Matching Error Study for the 1988 Dress Rehearsal." In Proceedings of the Section on Survey Research Methods, ASA, 704–709. Available for instance at researchgate: https://www.researchgate.net/publication/267379153_THE_MATCHING_ERROR_STUDY_FOR_THE_1988_DRESS_REHEARSAL/download.

Parag and P. Domingos. 2004. "Multi-Relational Record Linkage." In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining. Available at: https://homes.cs.washington.edu/~pedrod/papers/mrdm04.pdf (accessed November 2018).

Petersen, C.G.J. 1896. *The Yearly Immigration of Young Plaice into the Limfiord from the German Sea*. Report of the Danish Biological Station 6: 5–84.

Pollock, K.H., J.D. Nichols, C. Brownie, and J.E. Hines. 1990. "Statistical Inference for Capture-Recapture Experiments." *Wildlife monographs* 107.

RELAIS. 2015. *User's Guide Version 3.0*. Available at http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/relais.

Sadinle, M., R. Hall, and S.E. Fienberg. 2011. "Approaches to Multiple Record Linkage." In Proceedings of the ISI World Statistical Congress, 21–26 August 2011, Dublin: 1064–1071. Available at: http://2011.isiproceedings.org/papers/450092.pdf (accessed November 2018).

Sadinle, M. and S.E. Fienberg. 2013. "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." *Journal of the American Statistical Association* 108: 385–397. Doi: http://dx.doi.org/10.1080/01621459.2012.757231.

Sanathanan, L. 1972. "Estimating the Size of a Multinomial Population." *Annals of Mathematical Statistics* 43: 142–152. Available at: https://projecteuclid.org/download/pdf_1/euclid.aoms/1177692709 (accessed November 2018).

Steorts, R., R. Hall, and S.E. Fienberg. 2014. "SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication." *Journal of Machine Learning Research* 33: 922–930. Available at: http://proceedings.mlr.press/v33/steorts14.pdf (accessed November 2018).

Steorts, R., R. Hall, and S.E. Fienberg. 2015. "A Bayesian Approach to Graphical Record Linkage and De-duplication." *Journal of the American Statistical Association*. Available at: URL http://arxiv.org/abs/1312.4645.

Tuoto, T. 2016. "New Proposal for Linkage Error Estimation." *Statistical Journal of the IAOS* 32(2): 413–420. Doi: http://dx.doi.org/10.3233/SJI-160995.

Tuoto, T., B.F.M. Bakker, L. Di Consiglio, D.J. van der Laan, P.-P. de Wolf, and D. Zult. 2017. "Two Improvements of the Method for Population Size Estimation." in Proceedings of the 61st World Statistics Congress 16–21 July 2017, Marrakech.

Ventura, S. and R. Nugent. 2014. "Hierarchical Clustering with Distributions of Distances for Large-Scale Record Linkage." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, 283–298. Berlin: Springer Link. Lecture Notes in Computer Science 8744.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zwane, E. and P.G.M. van der Heijden. 2005. "Population Estimation using the Multiple System Estimator in the Presence of Continuous Covariates." *Statistical Modelling* 5: 39–52. Doi: http://dx.doi.org/10.1191/1471082X05st086oa.

Received June 2017
Revised April 2018
Accepted August 2018