DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 415–429, http://dx.doi.org/10.1515/JOS-2015-0025

Coverage Evaluation on Probabilistically Linked Data

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

The Capture-recapture method is a well-known solution for evaluating the unknown size of a population. Administrative data represent sources of independent counts of a population and can be jointly exploited for applying the capture-recapture method. Of course, administrative sources are affected by over- or undercoverage when considered separately. The standard Petersen approach is based on strong assumptions, including perfect record linkage between lists. In reality, record linkage results can be affected by errors. A simple method for achieving *linkage error-unbiased* population total estimates is proposed in Ding and Fienberg (1994). In this article, an extension of the Ding and Fienberg model by relaxing their conditions is proposed. The procedures are illustrated for estimating the total number of road casualties, on the basis of a probabilistic record linkage between two administrative data sources. Moreover, a simulation study is developed, providing evidence that the adjusted estimator always performs better than the Petersen estimator.

Key words: Linkage errors; capture-recapture method; Petersen estimator; administrative data.

1. Introduction

The problem of assessing the unknown size of a population is one that has long been grappled with, from the first experiments at measuring wild animal population size during the seventeenth century (Petersen 1896; Lincoln 1930) to applications for determining the number of people affected by specific diseases or using illegal drugs (Bartolucci and Forcina 2006), including the population census coverage (Wolter 1986). One well-known and widespread solution for this problem is the capture-recapture method. This method consists of comparing two (or more) independent counts ("capture" in the field of wild animal population estimation) of the same units, then evaluating, without error, the number of individuals in both the counts, and, as a result, counting the number of those caught only once.

In this framework, the standard Petersen estimator works well under some strong assumptions, such as the independence of the lists, the homogeneity of capture probabilities, and the lists' error-free linkage at record level.

Several extensions and adjustments of the Petersen estimator have been proposed over time in order to avoid bias due to failure of these assumptions, which causes the population to be under- or overestimated (e.g., Chao 2001, Chen and Kuo 2001).

Nowadays, the use of administrative data is emerging as a new opportunity in several statistical fields. Administrative data represent sources of several independent counts of a population. They can be exploited for the application of the capture-recapture method to estimate the unknown size of the population.

[1] Italian National Statistical Institute - Istat, Via Cesare Balbo, 16 00184 Rome, Italy. Email: diconsig@istat.it and tuoto@istat.it

Since records are collected for different purposes by different actors, the different administrative sources can be expected to be independent recaptures of the same (sub)population, in contrast to survey data, which are collected by the same organization. In fact, the independence assumption could be violated if the heterogeneity of capture probabilities of units is not properly encompassed in the statistical model.

Given their large size, data sets collected by administrative sources require a massive use of automatic tools, implementing record linkage techniques. Therefore, the error-free linkage assumption can be compromised, particularly in absence of unique identifiers for privacy issues.

In this article, we concentrate on failure of the perfect linkage hypothesis and we analyse different proposals that adjust the Petersen estimator by explicitly taking into account linkage errors.

In Ding and Fienberg (1994), a simple method to achieve linkage error-unbiased estimators of population total and undercoverage rate is proposed; moreover, different models for the two types of linking errors are described. The Ding and Fienberg (1994) adjustment considers the probability of missed true links and the probability of erroneous links, providing an alternative formula with respect to the Petersen estimator to assess the undercoverage and consequently the true population total.

We enhance the Ding and Fienberg (1994) model by defining the probabilities of being counted in both lists, handling the two lists in a symmetric way. These findings are subject to conditions of admissibility, which are discussed in the Appendix. The method is illustrated with an application to real data to estimate the number of casualties due to road accidents, integrating data from two registers: the "Causes of death" register and the "road accidents resulting in deaths (within 30 days) or injuries" register. Simulated data are used to show the benefit of the proposed new method over the existing ones in different linkage scenarios.

2. Capture-Recapture Background

The Petersen model (see Wolter 1986) is a standard well-known model for evaluating the population total. Let N be the unknown population total, and $N_1$and N_2 the population size reported in the first and second list, respectively. Let x_{11} be the number of units recorded in both lists, $x_{12} = N_1 - x_{11}$ the number of units reported only in List 1 and $x_{21} = N_2 - x_{11}$ the number of units reported only in List 2.

The counts can be organised in a 2 X 2 contingency table, with x_{22} the unknown number of units missed by both lists (Table 1).

Under the assumption of independent captures, the number of individuals in the contingency table follows the multinomial distribution.

Table 1. Contingency table of the counts in the two lists

		List 2	
		Present	*Absent*
List 1	*Present*	x_{11}	x_{12}
	Absent	x_{21}	x_{22}

Moreover, adding the following assumptions:

1. the population is closed, so the population being measured in both sources is the same
2. records from both sources can be linked without errors
3. units have the same capture probabilities within each source (homogeneity probability assumption)
4. overcount in both sources is negligible

an unbiased estimator of N, the well-known Petersen estimator, is given by

$$\widetilde{N}_P = N_1 \times N_2 / x_{11}. \tag{1}$$

The first list coverage is then given by

$$\widetilde{\tau}_{1,P} = x_{11}/N_2 \tag{2}$$

and similarly the second list coverage is

$$\widetilde{\tau}_{2,P} = x_{11}/N_1. \tag{3}$$

The previous assumptions' validity has been widely debated in a traditional survey context. Several extensions and adjustments have been proposed in order to avoid biases due to any failure of these assumptions that is under- or overestimation of the real population total amount.

As discussed above, on one hand, the independence of administrative sources could be guaranteed by different data collectors, while on the other hand, the heterogeneity of capture probabilities is a common issue in different settings due to inherent individual behaviour. When the individual capture propensity is not properly modelled, the dependence between lists can arise even in an administrative data context. Much literature focuses on including sources' dependencies and captures' heterogeneity by means of:

- extensions of the log-linear model (Fienberg 1972; Cormack 1989; Chao 2001, Agresti 1994; Coull and Agresti 1999)
- the conditional multinomial logit model (McFadden 1974; Bock 1975; Chen and Kuo 2001; Zwane and van der Heijden 2005)
- the latent class model (Bartolucci and Forcina 2006)
- the Bayesian capture-recapture model (Ghosh and Norris 2005).

More specifically, log-linear models explain the dependencies between data collections and the heterogeneity of capture probabilities by using categorical covariates, while the conditional multinomial logit model also allows continuous covariates to be included in the models.

The latent class model can be considered a conditional multinomial logit model extension and permits the modelling of both the observed heterogeneity using covariates and the unobserved heterogeneity by assuming that units belong to distinct latent classes. Finally, Bayesian capture-recapture models allow dependencies and heterogeneity to be formalised by means of suitable parameters for the distribution of individual capture probabilities.

When dealing with administrative data, compared to the survey context a change of perspective regarding the validity of previous assumptions is needed. In fact, overcoverage in the administrative lists may assume a relevant role. Recently, we have seen the failure of the last assumption 4), due to an observed significant level of list overcoverage affecting administrative data. Large et al. (2011) propose an adjustment to the Petersen estimator in order to correct bias due to overcount within the census context.

Another matter emerging when dealing with administrative sources concerns the unavailability of unique identifiers for maintaining privacy. In this framework, linkage errors could arise. This article considers extensions to deal with record linkage between lists affected by errors.

3. Including Linkage Errors in the Petersen Estimator

In this section, a short description of the most common probabilistic record-linkage framework is given, mainly in order to formalise linkage errors. Moreover, the Ding and Fienberg (1994) estimator to adjust the Petersen one for linkage errors is briefly reported; an extension is introduced to deal with more generic contexts, including those contexts typical for administrative data.

3.1. Linkage Model and Error Evaluation

A key step in applying the Petersen model is the integration of two (or more) sources at record level to identify the common units: this action is commonly referred to as record linkage.

A fundamental theory for record linkage is given in the seminal paper by Fellegi and Sunter (1969). Given two lists, say L1 and L2, of size N_1 and N_2, let $\Omega = \{(a, b), a \in L1$ and $b \in L2\}$ be the complete set of all possible pairs, of size $|\Omega| = N_1 \times N_2$. The linkage process between L1 and L2 can be viewed as a classification problem where the pairs in Ω have to be assigned to two independent and mutually exclusive subsets M and U, such that:

M is the link set ($a = b$)
U is the nonlink set ($a \neq b$).

In order to assign the pairs to the sets M or U, K common identifiers (the linking variables) are chosen and, for each pair, a comparison function is applied in order to obtain a comparison vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$. The ratio r of the conditional probability of $\boldsymbol{\gamma}$ given that the pair belongs to set M to the conditional probability of $\boldsymbol{\gamma}$ given that the pair belongs to set U

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

is used to classify the pairs. The probabilities m and u can be estimated by assuming the true link status is a latent variable, using, for instance, the EM algorithm (Jaro 1989). Hence, those pairs for which r is greater than the upper threshold value T_m are assigned to the set of linked pairs, M^*; those pairs for which r is smaller than the lower threshold value T_u are assigned to the set of unlinked pairs U^*; if r falls in the range (T_u, T_m), no decision is made and the pair is checked by clerical review.

The thresholds are chosen to minimise false link probability, β, and false nonlink probability, $1-\alpha$, defined as follows:

$$\beta = \sum_{\gamma\in\Gamma} u(\gamma)P(M^*|\gamma) = \sum_{\gamma\in\Gamma_{M^*}} u(\gamma) \quad \text{where} \quad \Gamma_{M^*} = \{\gamma : T_m \leq m(\gamma)/u(\gamma)\} \tag{4}$$

$$1-\alpha = \sum_{\gamma\in\Gamma} m(\gamma)P(U^*|\gamma) = \sum_{\gamma\in\Gamma_{U^*}} m(\gamma) \quad \text{where} \quad \Gamma_{U^*} = \{\gamma : T_u \geq m(\gamma)/u(\gamma)\}. \tag{5}$$

The linkage model also provides an evaluation of the probability of a link being a correct given that the link is assigned, the so-called true match rate:

$$\eta = 1 - \frac{\sum_{\gamma\in\Gamma_M} u(\gamma)P(M^*|\gamma)}{\sum_{\gamma\in\Gamma_M} m(\gamma)P(M^*|\gamma)} = 1 - \frac{\sum_{\gamma\in\Gamma_{M^*}} u(\gamma)}{\sum_{\gamma\in\Gamma_{M^*}} m(\gamma)}. \tag{6}$$

3.2. The Ding and Fienberg Estimator

In the context of probabilistically linked data, the coverage rates and population total estimates produced by the Petersen model may be biased and so they need to be "adjusted" in order to explicitly take into account the linkage errors.

A simple method for achieving "linkage error-unbiased" estimators of the population total and the coverage rates has been suggested by Ding and Fienberg (1994). They relax the perfect linkage assumption, propose models to describe linking errors and include those errors in the estimators derived by the Petersen model.

Under the following assumptions:

(a) true links between L1 and L2 are assigned with probability α
(b) false links between records belonging to M (see Subsection 3.1) are negligible
(c) false links can occur with a common probability β between records belonging to U (see Subsection 3.1)
(d) linkage direction from L1 to L2,

the adjustment proposed by Ding and Fienberg (1994) considers the false nonlink of linking cases probability (i.e., the probability of missing true link, 1-α) and the false link of nonlinking case probability (i.e., the probability of linking false pairs, β),

$$\widetilde{N}_{DF} = \frac{N_{1\cup 2}}{\hat{\tau}_{1,DF} + \hat{\tau}_{2,DF} - (\alpha-\beta)\hat{\tau}_{1,DF}\hat{\tau}_{2,DF} - \beta\hat{\tau}_{1,DF}} \tag{7}$$

where $\hat{\tau}_{1,DF}$ and $\hat{\tau}_{2,DF}$ are the estimates of probabilities of being recorded in lists 1 and 2, respectively. $N_{1\cup 2} = x_{11} + x_{12} + x_{21} = x_{11}^* + x_{12}^* + x_{21}^*$ is the number of records in list 1 or list 2, with x_{11} the number of *true* records in both lists, x_{12} the number of *true* records in list 1 and not in list 2 and, vice versa, x_{21} the number of *true* records in list 2 and not in list 1, while $x_{11}^*, x_{12}^*, x_{21}^*$ are the observed number of records in both lists, in list 1 and not in list 2, and in list 2 and not list 1, respectively, resulting from the linkage procedure.

The coverage of the first list is given by:

$$\hat{\tau}_{1,DF} = \frac{-x^*_{11} + \beta(x^*_{11} + x^*_{12})}{(\beta - \alpha)(x^*_{11} + x^*_{21})} \tag{8}$$

and similarly the coverage of the second list is

$$\hat{\tau}_{2,DF} = \frac{-x^*_{11} + \beta(x^*_{11} + x^*_{12})}{(\beta - \alpha)(x^*_{11} + x^*_{12})}. \tag{9}$$

The coverage rate estimates, $\hat{\tau}_{1,\mathrm{DF}}$ and $\hat{\tau}_{2,DF}$, are obtained by maximizing the conditional likelihood of $(x^*_{11}, x^*_{12}, x^*_{21})$ given $N_{1\cup 2}$,

$$L_1(p_{11}, p_{12}, p_{21}) = L_1(\tau_1, \tau_2) = \frac{N_{1\cup 2}!}{x^*_{11}!x^*_{12}!x^*_{21}!} \frac{p_{11}^{x^*_{11}} p_{12}^{x^*_{12}} p_{21}^{x^*_{21}}}{(p_{11} + p_{12} + p_{21})^{N_{1\cup 2}}}. \tag{10}$$

In this setting, a record is counted in both lists when it is actually in both lists and a link is made, and when the record is only in L1 but it is incorrectly linked with a record in L2. The former event has the probability $\alpha\tau_1\tau_2$, whereas the latter has $\beta\tau_1(1 - \tau_2)$, so the probability of observing a count in (1,1) is $p_{11} = \alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2)$. The probability of occurrence in cell (1,2) and (2,1) can be derived as $p_{12} = \tau_1 - p_{11}$and $p_{21} = \tau_2 - p_{11}$, respectively. See Ding and Fienberg (1994) for more details.

Note that the solutions are admissible under conditions on relationships of errors and counts.

The previous estimators are based on the assumptions: false links that occur when at least two errors are made (that is, records are incorrectly linked and the correct link is missed) have negligible probability of occurrence (assumption b). Moreover, a direction from L1 to L2 is assumed both in the linkage procedure (assumption d) and in the specification of the linkage errors. In the next subsection, generalised estimators for (7)–(9) achieved by relaxing assumption d are illustrated.

3.3. A Generalised Estimator

The Ding and Fienberg (1994) proposal was explicitly defined in the traditional census coverage evaluation context, where the linkage procedure between census data and the postenumeration survey results (Wolter 1986) works in one direction. When dealing with administrative data sources, this assumed one-way linkage direction is not guaranteed. Linkage errors, in particular false links, can occur in both directions, in contrast to what is assumed in d) of Subsection 3.2 according to Model B proposed by Ding and Fienberg (1994, 150). Note that in the context of administrative data, due to differences in unit and time reference, as well as variables' definitions, joint linkage errors (i.e., incorrect link and missed true links at the same time) may occur. Nevertheless, their probability can still be assumed negligible as at least three errors should be made, each one with small probability.

In the present proposal, assumption d) in Subsection 3.2 is relaxed, allowing for two-directional linkage. Hence, the probability of an occurrence in cell (1,1) is $p_{11} = \alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2) + \beta\tau_2(1 - \tau_1)$ where $\alpha\tau_1\tau_2$ is the probability that a unit is actually in both lists and a link is made, $\beta\tau_1(1 - \tau_2)$ is the probability that a unit actually registered only in L1 is incorrectly linked with a record in L2, and finally $\beta\tau_2(1 - \tau_1)$ is the

probability that a unit actually registered only in L2 is incorrectly linked with a record in L1. The probability of occurrence in cell (1,2) and (2,1) can be derived as $p_{12} = \tau_1 - p_{11}$ and $p_{21} = \tau_2 - p_{11}$, respectively.

Replacing p_{11}, p_{12}, p_{21} as defined above in the conditional likelihood (10) and maximizing with respect to τ_1 and τ_2, the Modified Ding and Fienberg (MDF) estimators are given by

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

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

Once $\hat{\tau}_{1,MDF}$ and $\hat{\tau}_{2,MDF}$ are obtained, the MDF ML estimator of N is given by:

$$\widetilde{N}_{MDF} = \frac{N_{1\cup 2}}{\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - (\alpha \hat{\tau}_{1,MDF}\hat{\tau}_{2,MDF} + \beta(\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - 2\hat{\tau}_{1,MDF}\hat{\tau}_{2,MDF}))} \tag{13}$$

Conditions for the admissibility of the estimates (11)–(12) also apply (see the Appendix).

The proposed estimators as well as the DF estimators are based on the assumption that linkage errors are constant. If this assumption holds at least in subgroups, the estimators can be applied within strata in which matching error probabilities (and capture probabilities) can be assumed to be more homogeneous than in the whole population.

4. Applications

4.1. Real Data Application

In this section, we present an application to data coming from two independent registers of deaths caused by road accidents. These data are exploited mainly because a complete analysis of the linkage status by clerical review is possible thanks to their small size.

In Italy, police authorities locally collect the road accidents resulting in deaths (within 30 days) or injuries and provide those data to the National Institute of Statistics. The Road Accident Register (denoted as RAR – or list 1, in the following) is an exhaustive, monthly-based register reporting the dynamics and circumstances of road accidents. Data collected by police are the main source for studying road traffic injuries. However, although the police usually collect very detailed information on crash dynamics and circumstances, relevant underreporting could occur due to the very complex situations related to the seriousness of the accidents. Therefore, the integration with health-care databases, such as mortality registers, can be very useful, complementing police data by capturing missing cases and also enriching them with detailed information on causes of death. For this purpose, a record linkage between the RAR and the data on causes of mortality, collected by the Italian National Vital Statistics Death Registry on causes of death (RCD – or list 2, in the following), was carried out.

The linkage procedure is not straightforward: a common personal identifying code is not available. Moreover, since RAR reference units are the road accidents, personal

identifying variables (i.e., names, surnames, ages) are sometimes missing or mistaken when more than one person is involved.

The reference year of the application is 2009. As far as the data from RAR are concerned, only records with at least one fatal casualty are considered, corresponding in that year to 4,237 records. Regarding RCD data, only road-accident deaths are considered, according to ICD-10 codes for traffic accidents involving motor vehicles on public roads. These correspond to a total of 4,642 records. The variables used for the linkage are: the road traffic victim/dead person name, surname and age, and the accident/death day, month, municipality and province.

The selected data sources' sizes do not require reduction procedures and the cross product of all records can be considered. The whole linking space is also exploited for the clerical review of links missed by the probabilistic procedure.

The linkage procedure identifies 3,129 linked records. The linkage errors estimated by the Fellegi-Sunter model (see (4) and (5)) are $\beta = 0.00$ and $1 - \alpha = 0.15$.

As is well known, in this approach the accuracy of linkage-error estimates is heavily dependent on the estimates' accuracy in the tails of the $m(\gamma)$ and $u(\gamma)$ distributions. Misspecifications in the model assumptions, errors or lack of information can cause a loss of accuracy in the latter. So, even though in most practical cases the linkage procedure is robust with respect to the links identification, the linkage error-estimates based on the linkage model are nevertheless generally too optimistic (Larsen and Rubin 2001).

As stated above, with these data, a clerical review of the linkage status is possible: this allows an evaluation of the proposed estimators knowing the true linkage-error values.

According to Table 2, the true $1 - \alpha$ is 0.1141 and β is 0.0009. On the basis of the true linkage status, the Petersen estimate of the total amount of road deaths is 5,572.

The results for the population size and the coverage list rates evaluation using the illustrated estimators are summarised in Table 3, where DF and MDF are defined in (7)–(9) and (11)–(13), respectively, and the naïve Petersen estimators are given by Equations (1)–(3), replacing the unobserved count x_{11} by the observed one x_{11}^{*}.

As expected, the DF and the MDF estimators give the same results when linkage errors are obtained from the linkage model due to the negligible value of β. All the compared estimators provide values close to the true one when linkage errors are known. Moreover, they are also less biased than the naïve Petersen estimates when linkage errors are estimated via the linkage model.

It is worth noting that even when a training set with known linkage status is available, the evaluation of β and $1 - \alpha$ is not straightforward. For instance, the well-known method

Table 2. Comparison between true linkage status and probabilistic linkage results

		True linkage status		
		Link	Nonlink	
Probabilistic linkage	Link	3,127	2	3,129
	Nonlink	403	2,218	2,621
		3,530	2,220	

Table 3. Amount of road deaths and coverage-rate estimates with estimated and true linkage errors

	True values	Petersen		DF	MDF
N	5,572	6,286	Estimated linkage errors	5,330	5,330
		–	True linkage errors	5,569	5,571
Coverage rate List 1	0.760	0.674	Estimated linkage errors	0.795	0.795
		–	True linkage errors	0.761	0.761
Coverage rate List 2	0.833	0.738	Estimated linkage errors	0.871	0.871
		–	True linkage errors	0.833	0.833

proposed by Belin and Rubin (1995) only provides estimates for β. In fact, detecting false links is more practicable than identifying missing links.

4.2. Simulation Study

The previous section showed an interesting real capture-recapture application that takes into account linkage errors. In that case, even with low linkage-error levels, the adjusted estimators perform better than the naïve Petersen estimator. However, the benefit of the proposed MDF over the DF is not sufficiently evident. In this section, a simulation is performed on fictitious data in order to compare the estimators in different linkage scenarios with variables of varying identifying power.

4.2.1. Description of the Simulated Setting

The simulation study was conducted on 100 replicated settings. Each one consists of a population of 1,000 units and two different lists that are generated mimicking the register undercoverage and the presence of errors in the common identifiers (the linking variables). The replicated pseudopopulations were independently randomly selected from the fictitious data on the UK population census. These data were created for the ESSnet DI (McLeod et al. 2011), which was a European project on data integration (Record Linkage, Statistical Matching, Microintegration Processing) run from 2009 to 2011. For each replicated pseudopopulation, the two lists were randomly generated according to the following coverage probabilities, $\tau_1 = 0.930$ and $\tau_2 = 0.924$, respectively.

Finally, on each replicated setting, the two lists were linked assuming three different scenarios to reflect different levels of informativeness in the linking variables. The Gold scenario uses linking variables with the highest identifying power, namely, *Name, Surname, Complete date of birth*. In this scenario, of course, the best results in terms of linked pairs and expected linkage errors are achieved.

The Silver scenario represents a situation where the strongest identifying variables – namely, *Name* and *Surname* – are not available, because, for instance, they are not released due to privacy issues. The linkage procedure can still be applied on variables with lower identification power than in the Gold Scenario, namely, the *Complete Date of Birth*. This causes linkage errors higher than in the previous scenario, in terms of both missing links and false links.

Table 4. Distribution of the linkage errors in the three scenarios

Scenario	Linkage results	Min	Q1	Median	Mean	Q3	Max
Gold	α	0.838	0.933	0.940	0.939	0.945	0.961
	β	0	0	0	0.001	0	0.057
Silver	α	0.807	0.842	0.853	0.851	0.861	0.884
	β	0.028	0.077	0.099	0.101	0.125	0.179
Bronze	α	0.808	0.822	0.833	0.833	0.843	0.874
	β	0.037	0.084	0.108	0.107	0.132	0.209

Finally, the Bronze scenario is the most unfavourable in terms of linkage errors; the set of variables used in the linkage procedure, namely *Surname, Day and Month of Birth*, has the lowest identifying power. In fact, in our data these variables are the ones most affected by typos and missing values. More precisely, in both lists, 16.7%, 2.6% and 4.3% of the records are affected by error in *Surname*, *Day of Birth* and *Month of Birth* respectively.

All the probabilistic record-linkage procedures were applied by means of the software RELAIS (see RELAIS 2011), according to the Fellegi and Sunter model summarised in Subsection 3.1.

Table 4 summarises the linkage results in terms of linkage errors, reporting the probability of nonmissing true matches (α) and the probability of false matches (β) as defined in Subsection 3.1. The true values of α and β can be evaluated in light of the true linkage status, which is known for each pair in each replication of the three scenarios.

4.2.2. Performance of the Alternative Estimators in the Simulation Study

From each linked set, we computed the counts x^*_{11}, x^*_{12} and x^*_{21} to apply the naïve Petersen estimator and the adjusted DF and MDF estimators described in Subsection 3.2 and 3.3, respectively. The DF and the MDF estimators are computed using the true values of the probability of nonmissing true matches (α) and the probability of false matches (β) obtained in each replication. The use of the true values of α and β allows the comparison of the estimators without the effect of linkage-error estimation.

To assess their performance, alternative estimates for each replicate in the three scenarios are reported in Figures 1–3.

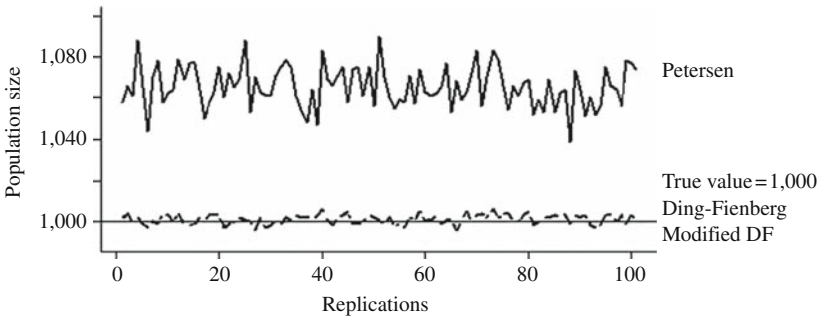

Fig. 1. Estimates in the Gold Scenario

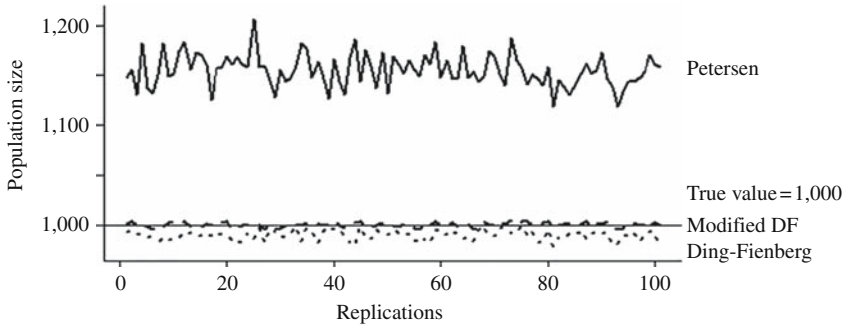

Fig. 2. Estimates in the Silver Scenario

In the Gold scenario, mimicking a situation where false linkage error is (nearly) absent, the adjusted estimators improve the naïve Petersen estimator in terms of bias as already shown with the real data application (Subsection 4.1). Again, as expected, the DF and the MDF are very close, as the extension in the MDF involves only the false linkage error β, as it results from a comparison of Equations (7) and (13) by simple algebra.

In the Silver scenario, where the false linkage error β is not negligible, the outperformance of the MDF with respect to the alternative estimators is clear. The comparison of Graphs 1–3 shows that the improvement by the MDF estimator is even more evident with higher levels of linkage error, as in the Bronze scenario.

The adjusted estimators' outperformance in terms of relative errors with respect to the naïve Petersen estimator is also shown in Table 5, where the minimum, the first quartile, the median, the mean, the third quartile and the maximum of the Percentage Relative Error over the 100 replications are reported for the three scenarios.

5. Concluding Remarks and Future Work

This work proposes a method for evaluating the unknown size of a population in the Petersen framework when the record linkage is not error free. This proposal overcomes the limitations of the Ding and Fienberg (1994) model tailored on the population census

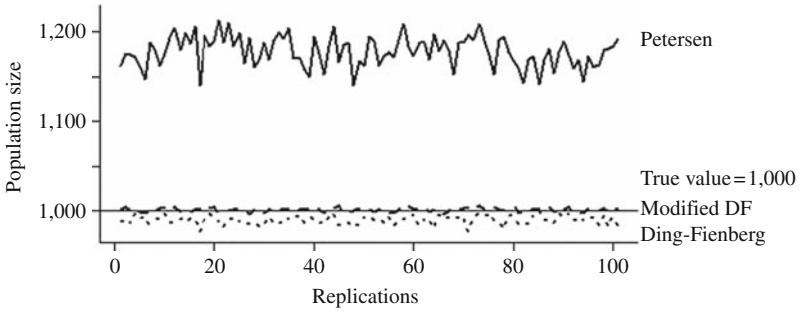

Fig. 3. Estimates in the Bronze Scenario

Table 5. Percentage Relative Error distribution in the three scenarios

Scenario	Estimator	Percentage Relative Error					
		Min	Q1	Median	Mean	Q3	Max
Gold	Petersen	3.9	5.9	6.4	6.5	7.3	9.0
	DF	−0.4	−0.1	0.1	0.1	0.3	0.6
	MDF	−0.4	−0.1	0.1	0.1	0.3	0.6
Silver	Petersen	11.8	14.5	15.6	15.5	16.4	20.6
	DF	−2.0	−1.3	−0.9	−0.9	−0.6	0
	MDF	−0.4	−0.1	0.1	0.1	0.3	0.6
Bronze	Petersen	14.0	16.7	17.9	17.8	19.0	21.3
	DF	−2.0	−1.4	−1.0	−1.0	−0.7	0
	MDF	−0.4	−0.1	0.1	0.1	0.3	0.6

coverage context. The application on real data showed an improvement of all the considered alternative methods in terms of bias with respect to the Petersen estimator. In this particular case, the model value of β was zero. When dealing with administrative data, this value is justified if personal identifying codes are available. In this case, the missed links are the most serious issue, since the omitting or erroneous reporting of identifying variables is not uncommon in administrative sources, in particular when they contain reference units and variables that differ from the statistical ones.

The simulation on fictitious data confirms the results of the real data application under more general frameworks, where different linkage-error levels are considered. Moreover, simulation results indicate that the MDF outperforms the other estimators when β is not negligible.

The adjusted methods depend on the correct evaluation of both kinds of linkage errors. This clearly appears in the real data application. In this application, the estimators' performances are assessed in both the following cases: linkage errors are estimated from the linkage model (Formulas 4 and 5); and the true linkage errors values are available. However, the adjusted estimators' improvement can also be observed with respect to the Petersen estimator in the first case. Further improvement in adjusting for linkage errors could be achieved by introducing individual values for the probability of correct links and missing links.

The evaluation of linkage errors is still an unresolved issue. The proposals that consider the linkage errors in analyses of linked data are often based on a training set to assess linkage quality. In any case, automatic probabilistic methods are necessary, particularly for detecting missing-link errors.

Moreover, a method for estimating the variance of the adjusted estimator is also needed. An interesting topic for future research would be the assessment of the trade-off between the gain in bias and the efficiency loss when linkage errors have to be estimated.

Finally, the effect of linkage-error adjustment should be studied for the extended models already proposed in the literature (see Section 2 for a short review) to overcome the other assumptions of the Petersen model.

Appendix

Conditions for admissibility of MDF

By straightforward algebra, estimates of the capture probabilities in (11) and (12) are positive under the following conditions for the linkage errors β and (1-α):

$$\text{a1)}\ x_{11}^*(1-2\beta) > \beta\left(x_{21}^* + x_{12}^*\right)\ \text{and}\ 2\beta-\alpha<0,\ \text{i.e.,}\ \beta < x_{11}^*/(N_1+N_2)\ \text{and}\ 2\beta-\alpha<0$$

or

$$\text{a2)}\ x_{11}^*(1-2\beta) < \beta\left(x_{21}^* + x_{12}^*\right)\ \text{and}\ 2\beta-\alpha>0,\ \text{i.e.,}\ \beta > x_{11}^*/(N_1+N_2)\ \text{and}\ 2\beta-\alpha>0.$$

In practical situations, the probability of linking false pairs, β, is close to zero, whereas probability of recognizing true links, α, is close to one, hence condition a1) will hold in common linkage contexts.

Furthermore, estimates of the capture probabilities in (11) and (12) are less than 1 under the following conditions for the linkage errors β and (1-α):

$$\text{b1)}\ x_{12}^* < x_{21}^*\ \text{and}\ < \frac{x_{11}^* - \alpha x_{11}^* - \alpha x_{12}^*}{x_{21}^* - x_{12}^*},\ \text{which in practice may hold only when}\ \alpha < \frac{x_{11}^*}{N_1}$$

or, on the contrary,

$$\text{b2)}\ x_{12}^* > x_{21}^*,\ \text{then}\ \beta > \frac{-x_{11}^* + \alpha x_{11}^* + \alpha x_{12}^*}{x_{12}^* - x_{21}^*},\ \text{which in practice may hold only when}\ \alpha > \frac{x_{11}^*}{N_1}$$

or

$$\text{b3)}\ x_{12}^* = x_{21}^*,\ \text{when}\ \alpha < \frac{x_{11}^*}{N_2} = \frac{x_{11}^*}{N_1},\ \text{i.e.,}\ \alpha < \hat{\tau}_1 = \hat{\tau}_2$$

6. References

Agresti, A. 1994. "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort". *Biometrics* 50: 494–500.

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.

Bock, R.D. 1975. *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

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.

Chen, Z. and L. Kuo 2001. "A Note on the Estimation of the Multinomial Logit Model with Random Effects". *The American Statistician* 55: 89–95. Doi: http://dx.doi.org/10.1198/000313001750358545.

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.

Ding, Y. and S.E. Fienberg 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error". *Survey Methodology* 20: 149–158.

Fellegi, I.P. and A.B. Sunter 1969. "A Theory for Record Linkage". *Journal of the American Statistical Association* 64: 1183–1210. 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: 591–603. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Ghosh, S.K. and J.L. Norris 2005. "Bayesian Capture-Recapture Analysis and Model Selection Allowing for Heterogeneity and Behavioral Effects". *NCSU Institute of Statistics, Mimeo Series* 2562: 1–27. Doi: http://dx.doi.org/10.1198/108571105X28651.

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.

Large, A., J. Brown, O. Abbott, and A. Taylor 2011. "Estimating and Correcting for Over-Count in the 2011 Census". *Survey Methodology Bulletin* 69: 35–48.

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.

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

McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior". In *Frontiers in Econometrics*, edited by P. Zarembka, 105–142, New York: Academic Press.

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 (accessed 20 July, 2015).

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.

RELAIS. 2011. *User's Guide Version 2.2* Available at: https://joinup.ec.europa.eu/software/relais/asset_release/relais-221 (accessed 20 July, 2015)

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. 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 February 2014
Revised December 2014
Accepted January 2015