Two improvements of the method for population size estimation

T. Tuoto[1], B.F.M. Bakker[2,3], L. Di Consiglio[4], D.J. van der Laan[2], P.-P. de Wolf[2], D. Zult[2]

[1] Istat, Rome, Italy;
[2] Statistics Netherlands, The Hague, Netherlands;
[3] VU University, Amsterdam, Netherlands;
[4]Eurostat, Luxembourg.

Abstract

Keywords: capture-recapture methodology, linkage error, census quality

1. Introduction

Population size estimation is important for census taking. It provides information on the number of residents at a particular moment in time and on how many residents a traditional door-to-door census or a register-based census missed. Capture-recapture methods are used to estimate the population size. In most cases, two or three sources are linked and a log-linear model is fitted in order to estimate the number of residents missed by all sources. To get accurate outcomes from these models, several assumptions have to be met. The main assumptions are independence of inclusion probabilities, homogeneous inclusion probabilities of at least one source or non-related heterogeneity inclusion probabilities, closed population, no erroneous captures and perfect record linkage. In practice, the assumption of perfect linkage is hardly ever met. False negative and false positive links lead to biased estimates.

Ding and Fienberg (1994) have developed a method to adjust the outcomes of capture-recapture analysis for linkage error. An important condition to apply this method is that you use probabilistic linkage to link your sources, because it makes use of the estimated probability of a linked pair being a correct link. The correction method was explicitly designed in the context of a traditional census, in which the census data are linked to a post-enumeration survey. It therefore assumes that there is a one direction linkage of the survey to the census data: false links between records from the survey that should be linked are negligible and each record of the this subset is actually linked to the census. A more general setting would be to use two different registers of approximately the same size and calculate a CRC estimate. Di Consiglio and Tuoto (2015) extended the method of Ding and Fienberg to allow that a record in one administrative source can be falsely linked to another administrative source, irrespective of which data source is the one or the other. They introduced their correction method as the Modified Ding-Fienberg (MDF) estimator.

However, the MDF correction method can still be generalized further. First, Di Consiglio and Tuoto have balanced the Ding-Fienberg solution by introducing two-directional linkage: it assumes that the probability of a false positive is equal in both directions. That seems reasonable in case the two administrative sources are of approximately equal size. Second, the MDF-estimator was developed for only two sources and without the use of covariates in the model. That has the disadvantage that the usual way to overcome violation of the independence assumption by linkage of three sources and the introduction of covariates was not possible. In this paper we introduce a general model to correct for linkage error in case of two sources, with DF and MDF as special cases. Moreover, we will extend the MDF estimator to the situation where three sources are linked.

2. General model to include linkage error in the CRC estimate with two registers

The record linkage process that we consider is the one as given in the famous Fellegi and Sunter (1969) paper. Their approach considers all possible pairs of records from the two sources and divides them into two sets: *M* containing all matched pairs (true matches) and *U* containing all unmatched pairs (true non-matches). Defining linkage errors as the probabilities that a true match will be missed or that a true non-match is matched nonetheless, they describe a process to link records from the two sources in such a way that it minimizes the linkage errors. However, in the CRC setting, perfect linkage is assumed, hence the still remaining linkage errors influence the estimate.

2.1 Underlying assumptions

To describe the assumptions more easily, we write the two sources as $R_1 = (M_1, U_1)$ and $R_2 = (M_2, U_2)$ where M_i is the set of records of source *i* that belongs to a true match and U_i the set of records in source *i* that does not belong to a true match. We then assume that

a) A matching pair between a record from M_1 and a record from M_2 remains a match with probability $0 < \alpha \leq 1$.
b) The probability that a record from M_1 is incorrectly linked to a record in M_2 is negligible.
c) A false link between a record from M_1 and U_2 occurs with negligible probability.
d) A false link between a record from M_2 and U_1 occurs with negligible probability.
e) A record from U_1 will be linked with a record from U_2 with probability $0 \leq \beta_1 < 1$.
f) A record from U_2 will be linked with a record from U_1 with probability $0 \leq \beta_2 < 1$.

2.2 General model and corresponding estimator

The general model can then be written as:

$$p_{11} = \alpha p_1 p_2 + \beta_1 p_1 (1 - p_2) + \beta_2 p_2 (1 - p_1)$$

where p_{11} is the probability that a pair of records will be considered to be a link and p_i the probability that a unit from the population is included in source *i*. Following the conditional ML approach as in Ding and Fienberg (1994), we may then derive a corrected CRC estimator for the population size as

$$\widehat{N} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \frac{n_{1+} n_{+1}}{n_{11}}$$

where n_{11} is the number of observed links, n_{1+} is the number of records in source 1 and n_{1+} the number of records in source 2. We will call this general estimator the M2DF estimator.
Note that, setting $\alpha = 1$ and $\beta_1 = \beta_2 = 0$ this results in the well-known Peterson estimator (see Peterson, 1896), setting $\beta_1 = \beta$ and $\beta_2 = 0$ this is the DF estimator with parameters α and β as defined in Ding and Fienberg (1994) and setting $\beta_1 = \beta_2 = \beta$ this is the MDF with parameters α and β as defined in Di Consiglio and Tuoto (2015).

2.3 Open issues

To evaluate the estimators, we would need to know the parameters α, β_1 and β_2. In practice, these parameters need to be estimated, using the observed counts and the information from the linkage process. In case we would know the 'true' parameters, we could compare the DF, MDF and M2DF without the effect of errors in estimating the parameters. However, in our view, there is not a unambiguous, straightforward way to define the 'true' parameters. Different definitions will result in different estimates: in our simulations we observed that the estimators are substantially affected by the way the 'true' parameters are defined.

It is still an open issue how the 'true' parameters should be defined, the main issue being how to incorporate (or not) the 'negligible' false and incorrect links. Moreover, choosing an appropriate definition would still raise the question how to estimate the parameters in practice.

3. Extension to three registers

The effect of linkage errors and relative adjustments in population size estimators has been extended in the general framework of multiple recapture methods based on log-linear models (Di Consiglio and Tuoto, 2017). They propose an extension of the previous work of Ding and Fienberg (1994) who propose a correction of the log-linear model considering the possible transitions from the real configuration $\mathbf{n}$ to the observed $\mathbf{n}^*$ taking into account missing links only. They assume that: (i) there are no erroneous matches in the linkage process; (ii) a transition can go only downwards by at most one level, (iii) the probability of staying at the original state (no missing error) equal to α and the probability to transit to any of a possible state is equal to $(1-\alpha)/(m-1)$, where m is the number of all possible states to which transitions are possible and allowed. Di Consiglio and Tuoto (2017) deal also with erroneous links, assuming that the transitions occur in function of the probability of missing a true match and probability of false match as well. Moreover, in Di Consiglio and Tuoto (2017) a more realistic linkage error model is defined, mimicking more closely the process when linking multiple lists in a real case. For instance, in the three lists case, they generalize the nature of the link process in the two phases, assuming first a linkage of list 1 and 2 and then a linkage with list 3, allowing for different linkage errors in the two linkage steps.

In this enriched framework, the transition matrix can be applied to the observed data in order to provide estimates of the cell probabilities of the real not-observable table, computing the Maximum Likelihood estimates of the parameters from the conditional likelihood associated with observed cell count $\mathbf{n}^*$. Then, the log-linear model is used to compute the conditional maximum likelihood estimates of the expected cell counts, including the one of the missing cell. In this setting, the use of log-linear models allows the introduction of covariates to overcome violation of the independence and homogeneous captures assumptions.

4. Conclusions

Elaborating on the literature on using the capture-recapture approach in case of population size estimation, we have proposed a general model to include linkage error in the CRC estimate in case one uses two sources. This resulted in an estimator that includes the estimator introduced in Ding and Fienberg (1994) and the one introduced in Di Consiglio and Tuoto (2015) as special cases. Moreover, we have extended the approach in Di Consiglio and Tuoto (2015) to the situation where three sources are used in the CRC estimate to weaken the independence assumption when using two sources.

To evaluate the general model in case of two sources, we have applied that model to some simulations based on data that were created for the ESSnet on Data Integration (McLeod, Heasman and Forbes, 2011). The results show that the actual estimates are affected by the exact way the 'true' parameters are calculated. This raises the question how those parameters should be defined and how they should be estimated in practice, even in case a clerical review on a subset is available. Moreover, we should pay close attention to the effects of the 'double errors' that are considered negligible (links counted while simultaneously being a missed match and a mismatch), but turn out to be substantially present in practice. Either this should be included in the definition of the 'true' parameters nonetheless, or we should extend the general model to include those 'double errors'.

In a multiple system estimation framework, the extension to the three lists case takes account of linkage errors in a realistic and widely used linkage setting for multiple sources. It explicitly introduces the errors caused by both missing and erroneous links generated by the linkage procedure into the contingency table counting the occurrences in the multiple sources. The generalization to the multiple lists case requires a not straightforward evaluation of the transition matrix, as well as the knowledge of the multiple steps in the linkage mechanism.

The adjustment allows to reduce the bias of the naive estimator without relevant effect on the variance, at least when the linkage errors are considered as known; however, the bias is not cancelled out completely due to the nonlinear nature of the CRC estimator. The evaluation of the linkage errors is still an open issue, also in the multiple lists case. One solution derives directly from the Fellegi-Sunter linkage model. Other proposals to assess linkage quality are based on a training set, providing more accurate evaluations (see Tuoto, 2016).

References

Di Consiglio, L. and Tuoto, T., 2015, "Coverage Evaluation on Probabilistically Linked Data", Journal of Official Statistics, 31: 415–429

Di Consiglio, L. and Tuoto, T., 2017, "Population Size Estimation and Linkage Errors: the Multiple Lists Case", proceeding of the NTTS 2107 Conference

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

Fellegi, I.P. and Sunter, A.B., 1969, "A Theory for Record Linkage", Journal of the American Statistical Association, 64: 1183–1210.

McLeod, P., Heasman, D. and Forbes, I., 2011, Simulated data for the on the job training. ESSnet DI, available at http://ec.europa.eu/eurostat/cros/content/job-training_en.

Peterson, C.G.J., 1896, "The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea", Report of the Danish Biological Station (1895), 6: 5–84.

Tuoto, T., 2016, "New proposal for linkage error estimation", Statistical Journal of the IAOS, Vol 32, no. 2