62nd ISI World Statistics Congress
ISI 2019 Kuala Lumpur
18 - 23 August 2019

PROCEEDING

SPECIAL TOPIC SESSION

VOLUME 4

62nd ISI WORLD STATISTICS CONGRESS 2019

18 - 23 August 2019, Kuala Lumpur

Come | Connect | Create

www.isi2019.org

Published by:
Department of Statistics Malaysia
Block C6, Complex C
Federal Government Administrative Centre
62514 Putrajaya
MALAYSIA

Central Bank of Malaysia
Jalan Dato' Onn
P.O. Box 10922
50929 Kuala Lumpur
MALAYSIA

Malaysia Institute of Statistics
Department of Mathematical Sciences
Faculty of Sciences and Technology
43600 UKM Bangi, Selangor
MALAYSIA

Portal : https://www.isi2019.org
Email : lpc@isi2019.org

Published in February 2020

Copyright of individual papers resides with the authors.

Suggested citation:
Department of Statistics Malaysia (DOSM). 2019. Proceeding of the 62nd ISI World Statistics Congress 2019: Special Topic Session: Volume 4, 2019. 419 pages

Disclaimer:
The views expressed in this proceeding are those of the author(s) of the respective paper and do not necessarily represent the opinions of other participants of the congress, nor the views or policy of the Department of Statistics Malaysia.

ISIWSC2019

Organised by :

DEPARTMENT OF STATISTICS MALAYSIA
MINISTRY OF ECONOMIC AFFAIRS

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

MALAYSIA INSTITUTE OF STATISTICS

Supported by:

MINISTRY OF TOURISM, ARTS & CULTURE MALAYSIA

Malaysia Convention & Exhibition Bureau

Visit Truly Asia Malaysia 2020

ISBN 978-967-2000-66-2

#ISIWSC2019

Population size estimation from incomplete multisource lists: A Bayesian perspective on latent class modelling

Davide Di Cecco[1], Marco Di Zio[1], Brunero Liseo[2]
[1] ISTAT, via Cesare Balbo, 16, 00184 Rome
[2] MEMOTEF, Sapienza Rome University, viale del castro laurenziano 9, 00161 Rome

Abstract
We propose a capture–recapture model for estimating the size of a population of interest based on a set of administrative sources and/or surveys in the presence of out-of-scope units (false captures). Our Bayesian approach makes use of a certain class of log - linear models with a latent structure. We also address the presence of sources providing partial information implementing a Gibbs Sampler algorithm which generates from the posterior distribution of the population size in presence of missing data. The proposed method is applied to simulated data sets.

Keywords
Bayesian Analysis, Capture–Recapture, Latent Class

1. Introduction

The use of administrative data for the production of official statistics is providing many new opportunities and methodological challenges. In estimating the size of the usual resident population by municipality, in almost all national statistics institutes the use of traditional censuses is gradually being replaced with the use of administrative sources, which provide "signs of life" for the population of interest. While undercoverage was the main issue in the former approach, overcoverage is the main concern with administrative data. By overcoverage we mean the erroneous inclusion in the lists of units which do not belong to our population, i.e., out-of-scope units. Of course, overcoverage can be encountered in surveys and census too, but almost always it consists of duplicated records generated by linkage errors, which are now commonly addressed even in capture–recapture contexts. In administrative data, on the other hand, linkage errors constitute just one of the factors, in a number of possible reasons for erroneous captures. In general, administrative data are gathered by other organizations for non-statistical purposes. Hence, units and variable definitions may not align perfectly. For example, the available information pertaining the registered events, their temporal description, their legal definition may vary in each source, and their harmonization can be difficult. As a consequence, each list may contain different subpopulations of out-of-scope units, and the assignment of the units to our target population may not be error free. Obviously, any piece of

available information should be included in the process of identification of the erroneous cases in the lists. Ideally, recognizing and deleting spurious cases should constitute a first phase of our analysis, after which some capture-recapture technique might be used on the "cleaned" data. However, in many cases, the available information does not suffice to single out every false capture, and there will remain a certain portion of uncertainty for which we have no capability of discerning the cause of error. In practice, the main approach in official statistics is the following: all available administrative sources are integrated into a unique population statistical registers. The register is coupled with an ad-hoc coverage survey (in the same way as censuses were coupled with an additional post enumeration survey) to exploit a Dual Systems Estimator (DSE). Then, the overcoverage rate is estimated on the basis of the comparison between the (supposedly) error-free survey and the administrative data via some supervised model, and then used to "correct" the DSE in some way. An original approach, called Trimmed DSE, and proposed in Zhang et al (2017), consists in an iterative procedure which removes units and estimate a DSE until a stopping criterion is satisfied. The authors prove that, if the survey has no overcoverage, the procedure has some optimal properties of convergence. The Dual System approach, including the aforementioned, has the remarkable property of being partic¬ularly robust (see, e.g., Chao et al 2001), and it does not rely on any complex model specification. Our approach, on the converse, relies on a Multiple Record System, where one considers the various adminis¬trative sources separately, in order to exploit the information redundancy. There exist various proposals in literature which use complex model to deal with false captures in multiple lists, particularly in animal abun¬dance problems, see, e.g., da Silva (2009), Wright et al (2009), and Link et al (2010). However, in all those works, the false captures are essentially duplicate linkage errors. To our knowledge, the only contributions dealing with false captures with no restrictive hypothesis on the source of error in multiple record systems are Overstall et al (2014) and Fegatelli et al (2017). The former proposes a Bayesian log-linear model, the latter extends that work in order to include latent variables. However, in both cases, only a single source list is assumed to suffer from false captures. When considering administrative sources separately, a series of methodological issues arises:

- It is necessary to take into account possible dependencies among the various sources.
- While DSE is known to be robust with respect to violation of basic hypotheses (e.g., the homogene¬ity of capture probabilities), this is not true in general in Multiple Record Systems.
- In our framework, administrative sources often target specific categories of citizens (e.g., people in a certain age range), leaving subset of the population with null probability of being captured.

Our proposal relies on the following assumption: all possible erroneous captures are defined as random classification errors under a binary classification model. That is, we hypothesize two subpopulations: one comprising the out-of-scope units, and the other the in-scope units. Then a two-component latent class model would adequately describe our data. To model possible dependencies among captures of a same individuals in different sources, we relax the classic conditional independence assumption of latent class models and assume a general log-linear model for the joint distribution. To address the problem of subpopulations that are uncatchable for some sources, we treat the uncatchable units as missing information and develop an inferential approach to deal with missing data. This model has been proposed in Di Cecco et al (2018). Here we present a Bayesian approach to estimate the size of the population, addressing the challenges listed above.

2. Methodology

Assume k lists or capture occasions are available, and let Yi be the random variable indicating whether a unit is included in the i–th list, $i = 1, ..., k$ (i.e., has been captured in the i–th occasion):

$$Y_i = \begin{cases} 1 \; if \; a \; unit \; is \; captured \; in \; the \; i-th \; list; \\ 0 \qquad\qquad\qquad\qquad otherwise. \end{cases}$$

Let $Y = (Y_1,\dots, Y_k)$ denote the capture profile of a unit, and let $\{P(Y = y) = p_y\}_{y\in\{0,1\}^k}$ be the associated probability distribution. Let $U(i)$ be the set of units that are catchable by list i, and let U be $U_i U(i)$. Let U_1 be our target population, with $U1 \subset U$. The cardinality of U is N, the one of U_1 N_1. Let X be the latent variable identifying the units belonging to our target population:

$$X = \begin{cases} 1 \qquad\qquad if \; a \; unit \; belongs \; to \; U1 \; ; \\ 0 \qquad\qquad otherwise. \end{cases}$$

Let n_y be the number of units having capture profile $\mathbf{y}$, of which $n_{x,y}$ belong to the latent class x so that $n_{0,y} + n_{1,y} = n_y$. The total number of observed unit is n_{obs}, while the units having capture history $\mathbf{y} = \mathbf{0} = (0, \dots, 0)$ are unobserved, so that $\sum_{y\neq 0} n_y = n_{obs}$, and $N = n_{obs} + n_0$. Note that $n_{1,0}$ is the number of units in U_1 that are not captured, while $n_{0,0}$ is the number of uncaptured units which are in U but not in U_1. We are interested in estimating $N_1 = \sum_y n_{1,y}$. The latent class model under the conditional independence assumption (CIA) can be equivalently expressed as the mixture model

$$p_{\mathbf{y}} = \sum_{x=0,1} p_x \prod_{i=1}^{k} p_{y_i|x}, \tag{1}$$

where $p_{y_i|x}$ indicates the conditional probability $P(Yi = y_i|X = x)$, or as in the log-linear model notation

(2) $$[XY_1][XY_2]\cdots[XY_k],$$

which reports only the higher order interactions (generators) of the model. Any additional interaction term in (2) represents a relaxation of the CIA.

2.1 Prior distributions: The usual priors for log–linear models are based on Multivariate Gaussian distributions. Here we propose a different prior based on Dirichlet distributions. We find this approach easier in terms of elicitation of prior knowledge, and also from a computational point of view, since it allows us to develop a Gibbs sampler for obtaining a sample from the posterior distribution of $N1$, so avoiding the use of a Metropolis–Hastings algorithm. To illustrate our proposal we start with decomposable models. In this case the prior distribution is simply the product of Dirichlet densities. In Dawid et al (1993) it has been demonstrated that, if G is the dependence graph of the decomposable model, $\{\mathcal{L}_1,\ldots,\mathcal{L}_g\}$ are the maximal cliques of G, and $\{\mathcal{L}_1,\ldots,\mathcal{L}_g\}$ are defined as

$$\mathcal{S}_i = \mathcal{C}_i \cap \bigcup_{j=1}^{i-1} \mathcal{C}_j \qquad i = 2,\ldots,g,$$

the joint distribution can be written as the product of conditional distributions:

(3) $$p_G = \prod_{i=1}^{g} p_{\mathcal{C}_i} \left(\prod_{j=2}^{g} p_{\mathcal{S}_j} \right)^{-1} = p_{\mathcal{C}_1} \prod_{i=2}^{g} p_{\mathcal{C}_i|\mathcal{S}_i},$$

where p over a (sub)graph is the (marginal) distribution over the variables included in the (sub)graph. Let θ be the vector of parameters $\theta = (P_{\mathcal{C}_1}, P_{\mathcal{C}_2|\mathcal{S}_2}, \ldots, P_{\mathcal{C}_g|\mathcal{S}_g})$. We define a prior distribution on θ as follows: for each $P_{\mathcal{C}_i|\mathcal{S}_i}$ and for each value of $\mathcal{S}_i$ we set a Dirichlet distribution defined for each possible combination of values $\mathbf{y}_{\mathcal{C}} \in \{0,1\}^{|\mathcal{C}|}$ of the variables in $\mathcal{C}$. The Dirichlet densities are independent by construction, and this class of priors is conjugate to (3). In the case of a general log–linear model, we made use of the "Bayesian iterative proportional fitting" described in Schafer (1997) in order to sample from a "Constrained Dirichlet". That is, we generate samples from a Dirichlet distribution which satisfies the constraints given by the log–linear model. This prior has been rarely utilized in literature, and, as far as we know, has never been utilized in capture–recapture analysis. Regarding N, in

accordance with the literature on Bayesian capture–recapture, sensible options include:

i) Jeffreys' prior, i.e. $\pi(N) \propto 1/N$;

ii) a hierarchical Poisson prior: $N \sim Poi(\lambda)$, $\lambda \sim Gamma(a, \beta)$;

iii) Rissanen's prior (Rissanen 1983), $\pi(N) \propto 2^{-\log^*(N)}$, where $\log^*(N)$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), ...\}$.

We further assume that N and θ are a priori independent.

2.2 Missing data: We propose a strategy useful to properly include sources which do not operate over certain subpopulations ("incomplete lists"). In fact, if we treat the uncatchable units as sampling zeros, the final population size estimate would be biased. The idea is to treat the incomplete lists as Missing at Random (MAR) information, i.e. assuming that, if they could operate on the whole population, they would retain the same joint distribution as in the observed subpopulations. In addition, we assume that we can distinguish whether a unit has not been captured in a list by chance or because it is out of the scope of that list, i.e., we can divide the population in strata where different set of lists operates. Then, certain profiles of the captured units are considered as partially observed, and we develop a data augmentation algorithm that imputes the complete capture histories using the rest of the data given the model. We distinguish completely observed capture profiles, $\mathbf{y}$, from the partially observed capture profiles $\mathbf{y}_{mis}$. In addition, for each stratum, we have a structural zero $\mathbf{z}$ consisting in a different combination of zeros and missing values. For example, in a 4-lists scenario with 2 strata, one where all lists operate and one where the first list does not operate, we have the structural zero $n_{0,0,0,0}$ in the first strata, and $n_{*,0,0,0}$ in the second, where the asterisk denotes the missing information. Then, our Gibbs algorithm at iteration $t + 1$ has the following steps:

(1) we sample the components of $\theta^{(t+1)}$ from their posterior conditional Dirichlet distributions (constrained or not);

(2) for each observed $\mathbf{y}$ and $\mathbf{y}_{mis}$, we randomly divide all the observed values n_y and $n_{y\,mis}$ into the corresponding consistent complete sequences $n_{x,y}$ according to their conditional probabilities;

(3) if we adopt $\pi(N) \propto 1/N$, it has been demonstrated in Manrique-Vallier et al (2014) that we can sample all structural zero cells counts n_z from a Negative Multinomial distribution. Otherwise, if we choose an informative prior for N, we can use a Metropolis-Hasting step to generate a value for $N(t+1)$ and then conditionally sample the structural zero cells such that $\sum_z n_z = N - n_{obs}$;

(4) for each generated n_z, we sample all complete sequences $n_{x,y}$ consistent with $\mathbf{z}$.

3. Simulations

We report the results of a simulation for empirically assessing the proposed algorithm. We considered 5 lists, *A*, *B*, *C*, *D*, *E*, and defined a scenario with three strata: one with all sources, one where 4 out of five sources operate, and another one where just three sources operate. We set N = 10000 and a proportion of out–of–scope units (both captured and non-captured) equal to 40%, so that the desired total $N1$ is 6000 in expectation. The model parameters have been set in such a way that the proportion of unobserved units (both in–scope and out–of–scope) is about 30%.

Model selection is a critical issue for capture–recapture modeling as population size estimate can be sensitive to changes in the parameterization. To have a hint on the robustness of the procedure under mis-specification of the model, we generated a sample from model *[XABC][XD][XE]*, and estimated $N1$ under two different models: the CIA model *[XA][XB][XC][XD][XE]*, and the (non – decomposable) model including all 15 second order interactions but no higher order parameters. Results regarding the second model can be viewed in Figure 1, where one sees that the true value of $N1$ is comprised in the 95% credibility interval, despite 5 parameters are missing (those relative to *[ABC], [XAB], [XAC], [XBC]* and *[XABC]*).

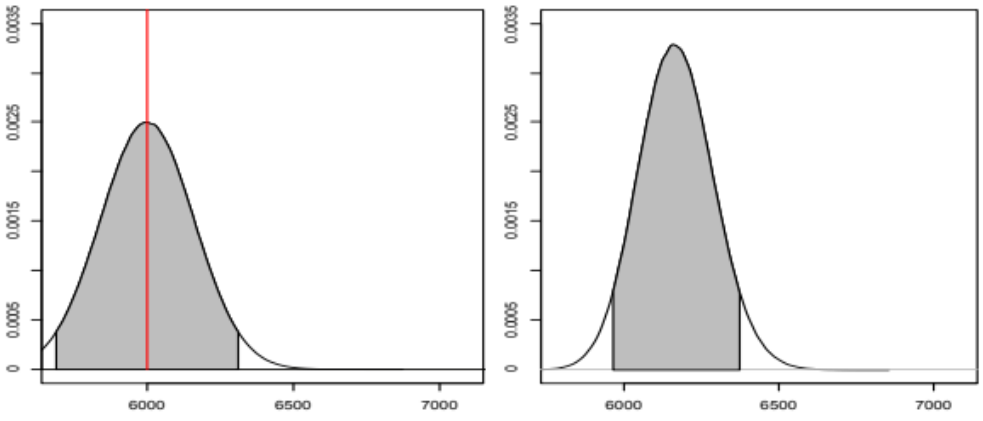

FIGURE 1. Posterior distributions of *N1* under the generating model *[XABC][XD][XE]* (left) and under the all–second–order–interactions model (right). The orange line indicates the true value of *N1,* the gray area the 95% HPD.

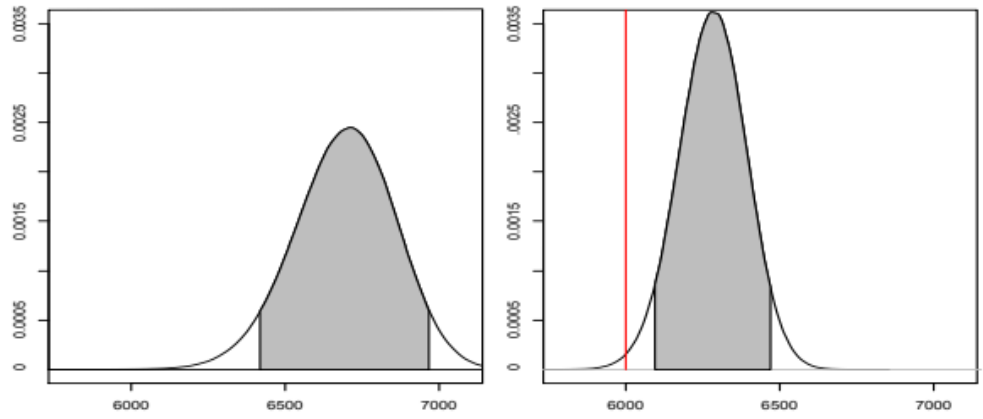

FIGURE 2. Posterior distributions of *N1* under the CIA model, flat priors(left) and informative priors (right). The orange line indicates the true value of *N1,* the gray area the 95% HPD.

On the converse, the left panel of Figure 2 shows that the estimated posterior distribution of *N1* under the CIA model is far from the real value. To evaluate the influence of the prior distributions to compensate for the model misspecification, we set an informative prior in the following way: we mimicked an informative context coming from an audit sample by taking a 5% sample of the generated complete population *[XABCDE],* and fixed the parameters of the Dirichlet prior equal to the observed counts in that sample. As one can see in the right panel of Figure 2, even though informative priors influence the posterior in the right direction, their contribution seems insufficient to even include the true value of N *1* in the credibility interval.

References

1. Chao, A., Tsay, P., Lin, S., Shau, W., and Chao, D. (2001). The applications of capture-recapture models to epidemiological data. Statistics in medicine, 20(20):3123–3157.
2. da Silva, C. Q. (2009). Bayesian analysis to correct false-negative errors in capture-recapture photo-ID abundance estimates. Brazilian Journal of Probability and Statistics, 23(1):36–48.
3. Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. The Annals of Statistics, 21(3):1272–1317.
4. Di Cecco, D., Di Zio, M., Filipponi, D., and Rocchetti, I. (2018). Population size estimation using multiple incomplete lists with overcoverage. J. Off. Stat., 34(2):557–572.
5. Fegatelli, D. A., Farcomeni, A., and Tardella, L. (2017). Bayesian population size estimation with censored counts. In Capture-Recapture Methods for the Social and Medical Sciences, pages 371–385. Chapman and Hall/CRC.

6. Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. Biometrics,66(1):178–185.
7. Manrique-Vallier, D. and Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. Journal of Computational and Graphical Statistics, 23(4):1061–1079.
8. Overstall, A. M., King, R., Bird, S. M., Hutchinson, S. J., and Hay, G.(2014). Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Statistics in medicine, 33(9):1564–1579.
9. Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. The Annals of statistics, 11(2):416–431.
10. Schafer, J. L. (1997).Analysis of incomplete multivariate data. CRC press.
11. Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., and Gleeson, D. M. (2009). Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using DNA samples. Biometrics, 65(3):833–840.
12. Zhang, L.C., and Dunne, J. (2017) Trimmed dual system estimation. In Capture-Recapture Methods for the Social and Medical Sciences, pages 237–257. Chapman and Hall/CRC