SAPIENZA
UNIVERSITÀ EDITRICE

ANNALI DEL DIPARTIMENTO DI METODI E MODELLI PER L'ECONOMIA IL TERRITORIO E LA FINANZA

2020

Direttore Responsabile – Director
Alessandra De Rose

Direttore Scientifico – Editor in Chief
Roberta Gemmiti

Curatore del numero – Managing Editor
Roberta Gemmiti

Comitato Scientifico – Editorial Board
Maria Giuseppina Bruno (Sapienza Università di Roma)
Adriana Conti Puorger (Sapienza Università di Roma)
Alessandra Faggian (The Ohio State University)
Francesca Gargiulo (Sapienza Università di Roma)
Roberta Gemmiti (Sapienza Università di Roma)
Cristina Giudici (Sapienza Università di Roma)
Ersilia Incelli (Sapienza Università di Roma)
Antonella Leoncini Bartoli (Sapienza Università di Roma)
Isabella Santini (Sapienza Università di Roma)
Marco Teodori (Sapienza Università di Roma)
Catherine Wihtol de Wenden (CERI-Sciences Po-CNRS Paris)

ISSN: 2385-0825 (print)
Registrazione presso il Tribunale di Roma n. 247/2016 del 30/12/2016

ISSN: 2611-6634 (online)
Registrazione presso il Tribunale di Roma n. 105/2019 del 01/08/2019

Copyright © 2020

Sapienza Università Editrice
Piazzale Aldo Moro 5 – 00185 Roma

www.editricesapienza.it
editrice.sapienza@uniroma1.it

Iscrizione Registro Operatori Comunicazione n. 11420

Pubblicato a dicembre 2020

Quest'opera è distribuita
con licenza Creative Commons 3.0 IT
diffusa in modalità *open access*.

Impaginazione/layout a cura del: Managing Editor

Ballerini Veronica*

Capture-recapture models for official statistics in presence of out-of-scope units: an overview

Abstract. During the last years, National Statistics Institutes have been exploring the possibility to produce statistics based on administrative data only. In particular, the interest in populations' size estimation is increasing. However, some issues naturally emerge since the aims of those who collect data and those who use them differ. On the one hand, it is very likely to have out-of-scope units in the datasets. On the other hand, some units that belong to the target population are not observed. In practice, this is the case of incomplete contingency tables that include some overcounts. The aim of this paper is to review the main and most recent literature dealing with the population size estimation problem, with an insight on the overcoverage issue.

Keywords: capture-recapture, population size estimation, log-linear models, overcoverage, Bayesian graphical models, erroneous enumeration

1. Introduction

In recent years, National Statistics Institutes (NSIs) have been more and more interested in producing statistics using administrative data only. The idea of replacing censuses with the integration of data coming from multiple sources is very attractive, due to the several shortcomings presented by surveys. First, there exists a well-known and unavoidable trade-off between timeliness and accuracy; second, the response rate is becoming lower and lower. Third, to conduct a survey implies huge costs. In these terms, a data integration approach would imply an enormous potential gain. Nevertheless, several methodological issues emerge, especially since the aims of those who collect data and those who use them differ. Consequently, more uncertainty and the risk of biasedness naturally arise. The Italian Statistical Institute (ISTAT) is among the NSIs that are paying more attention to the "paradigm shift" in official statistics; see Filipponi et al. (2017) and Chiariello and Tuoto (2018) among others.
In line with the NSIs interests, academia is producing a vast literature. One of the most recent examples is "Analysis of Integrated Data" edited by Zhang and Chambers (2019). The contributes therein deal with statistical uncertainty and inference issues that arise when a dataset is created via integration of multiple sources. Certainly, among the key issues there is that of *undercoverage*. A list is subject to undercoverage if it enumerates only a subset of the population of interest (or *target population*). In the case of multiple lists, we may face a situation in which the union of the lists is still a subset of the target population. Special attention is devoted to the *entity ambiguity* problem, which arises "whenever it is not possible to state with certainty that the integrated source corresponds to the target population of interest" (Zhang and Chambers (2019), Preface). One of the frameworks in which it is possible to encounter such entity ambiguity is when the units of a target population are partially or erroneously covered in each source. A particular case is that of *overcoverage*, which occurs when some out-of-scope units are erroneously included in one or more data sources.

On the one hand, the undercoverage issue has been largely treated in literature. The most widespread models are the so-called *capture-recapture* models. Although the idea of capture-recapture dates back to the eighteenth century (see Amorós (2014)), they have developed during the twentieth in the field of ecology. In such models, the capture (or miss) of specimens belonging to a target animal population is registered, and the registration is repeated for several capture occasions. A milestone is Fienberg (1972). In this work, Fienberg uses a contingency table to describe the capture histories of the registered units; the table is said to be *incomplete* since the cell referring to those units that have never been captured is unobserved. The use of log-linear models allows the estimation of the missing cell count

* Department Methods and Models for Territory, Economics and Finance - Sapienza University of Rome

and consequently of the total population size. Since 1972, a vast literature dealing with population size estimation in a capture-recapture framework has been developing, following both frequentist and Bayesian approaches; the most recent developments in capture-recapture models for social sciences can be found in Böhning et al. (2018).
On the other hand, the overcoverage issue has become relevant only recently, with the increase of the interest of the NSIs in the production of statistics through data integration. The approach which has spread the most is the latent class modelling. We will discuss this topic more in detail in section 5.

This work aims to review the main and most recent literature about the population size estimation problem. Section 2 introduces the basic notation we will use throughout the paper. Section 3 presents the most recent literature about capture-recapture models which follows Fienberg's line; from its milestone Fienberg (1972), to the Bayesian version of log-linear models. An example of interest will be presented in section 4. Section 5 extends the models previously introduced to the overcoverage issue and section 6 compares them with two examples simulating different scenarios. The conclusions follow.

2. Notation

In this section we mainly follow the notation presented in Zhang (2019). Consider K lists A_k, $k = 1, ..., K$, only partially enumerating a target population U of unknown cardinality $|U| = N$. Refer to $A = A_1, ..., A_K$ as the *lists' universe*. Now let $\delta_{A_k}(i)$ be a random variable assuming value 1 if $i \in A_k$, i.e. if the unit i, $i \in \{1, ..., N\}$ is "captured" in the k^{th} list, and 0 otherwise. This way, the vector

$$\delta_{\mathbf{A}}(i) = (\delta_{A_1}(i) \ ... \ \delta_{A_K}(i)) \, ,$$

represents the *capture history* (or capture vector) of unit i. Now define

$$\omega(i) = \{k; \delta_{A_k}(i) = 1, 1 \leq k \leq K\} \, ,$$

i.e. $\omega(i)$ is the index set of the lists in which the observed unit i is enumerated. For instance, when $K = 3$ the possible values for ω will be $\{1, 2, 3, 12, 13, 23, 123\}$ for any i. Define A_ω as the cross-classified *list domain* according to $\delta_{\mathbf{A}}$, with $\sum_{k=1}^{K} \delta_{A_k} \geq 1$. Similarly, $A_{\omega+}$ is the marginal list domain. Together the A_ω's form a partition of the lists' universe A, which is of size $|A| = \sum_{\omega} |A_\omega| = \sum_{\omega} x_\omega$, i.e., x_ω is the number of individuals observed in A_ω; e.g., x_{12} is the number of units observed in both (and only) A_1 and A_2, and x_1 is the number of those observed in A_1 only. The counts can be summarized in an incomplete contingency table, as shown in Table 1. In a similar fashion, $x_{\omega+}$ is the number of units in $A_{\omega+}$; for instance, x_{12+} is the number of units in both A_1 and A_2 but not exclusively, and x_{1+} is the number of units observed at least in A_1.

Now assume that some lists partly enumerate out-of-scope units, i.e. $\cup_{k=1}^{K} A_k \subset \{U \cup \tilde{U}\}$, $\tilde{U}$ being the set that includes units not belonging to the target population. Let

$$\delta_U(i) \begin{cases} 1 & \text{if } i \in U \\ 0 & \text{otherwise.} \end{cases}$$

Let $y_\omega = \sum_{i \in A_\omega} \delta_U(i)$ be the number of in-scope units in A_ω; e.g., y_{12} is the number of population units enumerated both in list A_1 and A_2. For all ω's including at least one of the lists not affected by overcoverage, the observed count x_ω is equal to y_ω. Yet, for the others $y_\omega = x_\omega - r_\omega$, with r_ω unobserved. Finally, indicate with y_0 the unknown number of population units enumerated in none of the lists. Hence, the total target population's size will be $N = \sum_{\omega} y_\omega + y_0$.

For convenience, we define $y = \sum_{\omega} y_\omega$ to indicate the total number of in-scope individuals captured in the lists.

Table 2 helps clarifying the notation; it shows a three-way incomplete contingency table where all sources are affected by overcoverage.

	$\delta_{A_1}(i)=1$		$\delta_{A_1}(i)=0$	
	$\delta_{A_2}(i)=1$	$\delta_{A_2}(i)=0$	$\delta_{A_2}(i)=1$	$\delta_{A_2}(i)=0$
$\delta_{A_3}(i)=1$	x_{123}	x_{13}	x_{23}	x_3
$\delta_{A_3}(i)=1$	x_{12}	x_1	x_2	?

Table 1. Three lists observed contingency table

		$\delta_{A_1}(i)=1$		$\delta_{A_1}(i)=0$	
		$\delta_{A_2}(i)=1$	$\delta_{A_2}(i)=0$	$\delta_{A_2}(i)=1$	$\delta_{A_2}(i)=0$
$\delta_{A_3}(i)=1$	$\delta_U(i)=1$	y_{123}	y_{13}	y_{23}	y_3
	$\delta_U(i)=0$	r_{123}	r_{13}	r_{23}	r_3
$\delta_{A_3}(i)=0$	$\delta_U(i)=1$	y_{12}	y_1	y_2	y_0
	$\delta_U(i)=0$	r_{12}	r_1	r_2	?

Table 2. Latent structure of three lists contingency table where all sources are affected by overcoverage

3. Capture-recapture

Capture-recapture models were born in ecology at the end of 19^{th} century, when the need of accurate tools able to estimate the number of specimens belonging to a target animal population became stronger. Indeed, C. G. J. Petersen[1] and F. C. Lincoln[2] are the ones who can be considered the "fathers" of capture-recapture methods; a marine biologist and an ornithologist respectively. Starting from the popular Lincoln-Petersen estimator, capture-recapture models have evolved and found one of their main applications in social sciences. In the following subsection we briefly review the most recent models suitable for official statistics' needs.

3.1 Log-linear models' setup

Log-linear models have been the most popular representation for count data so far. The way these models work is self-explanatory; considering the observed counts as random variables' realizations, we may express the natural logarithm of their expected values as a linear function of a set of unknown parameters.
Assume to be able to classify y units belonging to a particular population in a contingency table according to some characterizing factors, e.g. A_1, A_2, A_3. Also assume that each factor have different levels, $l_1 = 1, ..., c_1, l_2 = 1, ..., c_2$ and $l_3 = 1, ..., c_3$ respectively.
Let $Y_{l_1 l_2 l_3}$ be the random variable indicating the number of counts for the cell corresponding to level $l_1 l_2 l_3$. In case of independence among factors, i.e. $P(i \in A_{123}) = P(i \in A_{1++})P(i \in A_{+2+})P(i \in A_{++3})$, its expected value is

$$\lambda_{l_1 l_2 l_3} = \frac{y_{l_{1++}}}{y} \frac{y_{l_{+2+}}}{y} \frac{y_{l_{++3}}}{y} y$$

where $\lambda_{l_1 l_2 l_3} = \mathbb{E}(Y_{l_1 l_2 l_3})$ and $y = \sum_{l_1} \sum_{l_2} \sum_{l_3} y_{l_1 l_2 l_3}$ is the total number of counts.
Analogously to the analysis of the variance, Fienberg (1970) expresses the natural logarithm of such expected value as

$$\log(\lambda_{l_1 l_2 l_3}) = \phi + \theta_{l_1} + \theta_{l_2} + \theta_{l_3}$$

where the θ's represent deviations from the grand mean, i.e. ϕ. Equivalently,

$$\log(\lambda_{l_1 l_2 l_3}) = [1] + [A_1]_{l_1} + [A_2]_{l_2} + [A_3]_{l_3} \ .$$

Yet, in case the factors' independence assumption does not hold, we need to introduce additional parameters, i.e. the interaction terms. In the case of three factors, the so-called *saturated model* will include three two-factor interaction terms and one three-factor:

$$\log(\lambda_{l_1 l_2 l_3}) = \phi + \theta_{l_1} + \theta_{l_2} + \theta_{l_3} + \theta_{l_1 l_2} + \theta_{l_1 l_3} + \theta_{l_2 l_3} + \theta_{l_1 l_2 l_3}$$

or

$$\log(\lambda_{l_1 l_2 l_3}) = [1] + [A_1]_{l_1} + [A_2]_{l_2} + [A_3]_{l_3} + [A_1 A_2]_{l_1 l_2} + [A_1 A_3]_{l_1 l_3} + [A_2 A_3]_{l_2 l_3} + [A_1 A_2 A_3]_{l_1 l_2 l_3} \ .$$

Any model which does not include all the interaction terms is said to be *unsaturated*. Generalizing, we may classify units according to K factors A_k, $k = 1, ..., K$, each of which assumes l_k levels, $l_k = 1, ..., c_k$. Therefore, the most general log-linear model for the expected value of the number of counts corresponding to level $l_1 ... l_K$ will be

$$\log(\lambda_{l_1 ... l_K}) = \phi + \sum_k \theta_{l_k} + \sum_k \sum_{j>k} \theta_{l_k l_j} + ... + \theta_{l_1 ... l_K} \ .$$

The over parameterization of the model emerges clearly, asking for a constraint which will allow for the model's identification. One possibility is the *sum-to-zero* constraint:

$$\sum_{l_k=1}^{c_k} \theta_{l_k} = \sum_{l_1=1}^{c_1} \theta_{l_1 l_k} = ... = 0 \qquad \forall k \ .$$

[1] see Petersen (1985)
[2] see Lincoln (1930)

Model specification	Highest order interactions
$\phi+\theta_1+\theta_2+\theta_3+\theta_{12}+\theta_{13}+\theta_{23}+\theta_{123}$	$[A_1A_2A_3]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{12}+\theta_{13}+\theta_{23}$	$[A_1A_2][A_1A_3][A_2A_3]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{12}+\theta_{13}$	$[A_1A_2][A_1A_3]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{12}+\theta_{23}$	$[A_1A_2][A_2A_3]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{13}+\theta_{23}$	$[A_1A_3][A_2A_3]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{12}$	$[A_1A_2]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{13}$	$[A_1A_3]$
$\phi+\theta_1+\theta_2+\theta_3+\theta_{23}$	$[A_2A_3]$
$\phi+\theta_1+\theta_2+\theta_3$	$[A_1][A_2][A_3]$

Table 3. Different model specifications for $\log(\lambda_{123})$

Another option is the *corner point* constraint, which will be used throughout this paper:

$$\theta_{(l_k=1)}=0,...,\theta_{(l_k=1)...l_j}=0 \qquad \forall k\neq j \ .$$

Log-linear models have been applied to capture-recapture data in Fienberg (1972) for the first time. Here, the factors A_k are the capture occasions, or lists. For each list only two levels l_k are possible: captured ($l_k=1$) or missed ($l_k=0$). As a result, the observed units can be classified in an incomplete 2^K contingency table. By incomplete we mean that it will presents a missing cell, the one corresponding to $\{l_1=0\ ...\ l_K=0\}$, by construction. In this framework, three assumptions play a key role. First, the population is assumed to be closed (Fienberg (1972)). Second, the probability of being captured in one or more lists is the same for any individual i belonging to the target population; in other words, there is *capture homogeneity*. Another crucial assumption is the units' unique labelling: the entire multiple recapture history of any observed individual can be inferred from its label anytime. In the conclusion, we will see how literature has addressed the cases of deviation from the last two assumptions.
To my knowledge, the entire capture-recapture literature on log-linear models has only focused on the hierarchical ones, i.e. those models where the higher-order relatives of a zero term are constrained to be zero as well (see Fienberg (1972)). For instance, assume $K=3$; the saturated model results being

$$\log(\lambda_{l_1=1\ l_2=1\ l_3=1})=\phi+\theta_{l_1=1}+\theta_{l_2=1}+\theta_{l_3=1}+\theta_{l_1=l_2=1}+\theta_{l_1=l_3=1}+\theta_{l_2=l_3=1}+\theta_{l_1=l_2=l_3=1}.$$

Since the only admissible levels for each list are 0 and 1, with a little abuse of notation we replace the subscript l_k with its index k when $l_k=1$ and we omit it when $l_k=0$. The result is in line with notation described in section 2. Therefore, we can rewrite the equation above as

$$\log(\lambda_{123})=\phi+\theta_1+\theta_2+\theta_3+\theta_{12}+\theta_{13}+\theta_{23}+\theta_{123}.$$

Generalizing,

$$\log(\lambda_\omega)=\phi+\sum_{\nu\in\Omega(\omega)}\theta_\nu$$

where $\Omega(\omega)$ is the set of all non-empty subsets of ω; equivalently,

$$\log(\lambda_\omega)=\phi+\mathbf{d}_\omega{}^T\theta$$

where $\phi\in\mathbb{R}$ is the grand mean and $\mathbf{d}_\omega$ is the design vector that indicates which elements of the regression parameters vector θ apply to the cell indexed by ω. θ is the vector of the factors' effects and interaction terms:

$$\theta=(\theta_1,...,\theta_K,...,\theta_{k_1k_2},...,\theta_{k_1k_2k_3},...,\theta_{1...K})^T$$

Table 3 shows all possible model specifications for the number of counts y_{123} when $K=3$.

Fienberg's approach consists in estimating the most parsimonious log-linear model, restricted to the incomplete table, and use it to predict the count of the missing cell. To give an insight of the effectiveness of such estimation procedure,

we briefly describe the main steps; see Fienberg (1972) for the details.
Let $\{y_\omega\}$ be Multinomial with parameters N, p_ω, where p_ω is function of some parameters $\boldsymbol{\zeta}$, i.e. $p_\omega = p_\omega(\boldsymbol{\zeta})$, and let $L(N; \boldsymbol{\zeta})$ be the relative likelihood function:

$$L(N; \boldsymbol{\zeta}) = \frac{N!}{y_0!}(1 - \sum_\omega p_\omega(\boldsymbol{\zeta}))^{y_0} \prod_\omega \frac{p_\omega(\boldsymbol{\zeta})^{y_\omega}}{y_\omega!}$$

As shown in Sanathanan (1972), the likelihood can be factorised and expressed as

$$L(N; \boldsymbol{\zeta}) = L_1(N; \sum_\omega p_\omega(\boldsymbol{\zeta}))L_2(\boldsymbol{\zeta}) \ ,$$

where

$$L_1(N; \sum_\omega p_\omega(\boldsymbol{\zeta})) = \frac{N!}{(\sum_\omega y_\omega)! y_0!}(1 - \sum_\omega p_\omega(\boldsymbol{\zeta}))^{y_0}(\sum_\omega p_\omega(\boldsymbol{\zeta}))^{\sum_\omega y_\omega}$$

and

$$L_2(\boldsymbol{\zeta}) = (\sum_\omega y_\omega)! \prod_\nu \frac{p_\nu(\boldsymbol{\zeta})^{y_\nu}}{y_\nu!(\sum_\omega p_\omega(\boldsymbol{\zeta}))^{y_\nu}}$$

Maximizing $L(N; \zeta)$ we obtain the unrestricted estimates of N, $\hat{N}_U$, and ζ, $\hat{\zeta}$. Yet, maximizing $L_1(N, \sum_\omega p_\omega(\hat{\zeta}_C))$ where $\hat{\zeta}_C$ is the MLE of L_2, we obtain a conditional estimate of N, $\hat{N}_C$. Sanathanan (1972) proves that $(\hat{N}_U, \hat{\zeta}_U)$ and $(\hat{N}_C, \hat{\zeta}_C)$ are both consistent estimators; hence, Fienberg (1972) suggests to use $(\hat{N}_C, \hat{\zeta}_C)$ to assess the appropriateness of a given model.
The maximum likelihood estimator for N_C is

$$\hat{N}_C = \left[\frac{\sum_\omega y_\omega}{\sum_\omega p_\omega} \right] \ ,$$

where [.] stands for the the closest integer. After some algebraic manipulation, we get

$$\hat{N}_C = \sum_\omega y_\omega + \hat{\lambda}_0$$

where, for any K,

$$\hat{\lambda}_0 = \frac{\hat{\Lambda}_{odd}}{\hat{\Lambda}_{even}} \ ,$$

$\hat{\Lambda}_{odd}$ ($\hat{\Lambda}_{even}$) being the product of all $\hat{\lambda}_\omega$ whose ω has an odd (even) number of elements. $\hat{\lambda}_\omega$'s are the MLE's obtained from the incomplete contingency table, given by setting the expected values of the marginal totals corresponding to the highest order interaction terms in the model equal to their observed value (see Fienberg (1972)). A confidence interval can be computed relying either on the asymptotic normality of the estimator $\hat{N}$ Bishop et al. (1975), or the profile likelihood of $\hat{N}$ (see Cormack (1992)).
The appropriateness of a given model can be assessed using either the Pearson's Chi-squared or the likelihood ratio statistics. Certainly, it is possible to use any information criterion, such as the Akaike or the Bayesian, as well. However, notice that for small K and large number of model's parameters the available degrees of freedom may be very few (0 in case of $K = 2$ and independence model). Moreover, as proved by Regal and Hook (1991), it may happen that more than one specification can fit the data perfectly, even with same number of parameters, giving very different confidence intervals for the population size. Zhang (2019) proposes a model selection criterion based on a so-called *latent likelihood ratio* which may help to select a model in cases of zero degrees of freedom.
Another limitation of Fienberg's approach consists of the difficulty of including any information on the population's size that may be available a priori.

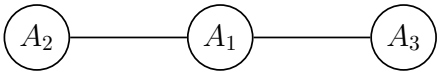

Figure 1

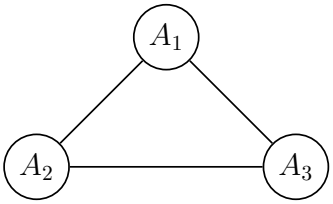

Figure 2

3.2 Decomposable graphical models

With the aim of overcoming the aforementioned limitations of Fienberg (1972), Madigan and York (1997) presented the Bayesian approach to population size estimation problem which has deeply influenced the following literature. This approach is hierarchical log-linear models based, but it focuses only on a subset of such models, namely the so-called *decomposable graphical models*. A statistical model is said to be *graphical* if it embodies a set of conditional independence relationships which can be summarized by a graph. For the sake of clarity, we briefly introduce the basic terminology of graph thoretic, mainly relying on Madigan and York (1995).

Define a graph as a pair $G = (V, E)$, with V being a finite set of vertices and E being the set of edges, i.e. a subset of $V \times V$ of ordered pairs of distinct vertices. In practise, the vertices represent the model's variables and the edges the dependence relations among them. A graphical model consists of a statistical model describing the conditional independence relationships among variables via a graph. Two variables may be just correlated, or there might be a casual relation between them. In the former case, both $(V_i, V_j), (V_j, V_i) \in E, \forall\, i, j$ and such edge is represented as a plain line; V_i and V_j are said to be *neighbours* and the resulting is a graph so-called *undirected*. Yet, if $(V_i, V_j) \in E$ but $(V_i, V_j) \notin E$, the edge is a directed arrow, and the graph is said to be *directed*. When all pairs of vertices are joined by an edge, the graph is *complete*. A complete subset of the vertex set that is not contained in any other complete subset is a maximal *clique* C. In an undirected graph, $G = \cup_{i=1}^{g} C_g$, $(C_1, ..., C_g)$ is a *perfect ordering* of the cliques when the vertices of each clique C_i also contained in previous cliques are all members of one previous clique only; the sets $S_i = C_i \cap (\cup_j^{i-1} C_j)$ are called *clique separators*. Figure 1 shows an undirected graph composed by two cliques, $C_1 = \{A_2, A_1\}$ and $C_2 = \{A_1, A_3\}$, whose ordering is perfect and where A_1 is a separator. Now focus on undirected graphs and define a *path* as a sequence $V_0, ..., V_n$ of distinct vertices such that (V_i, V_{i-1}) are neighbours for all $i = 1, ..., n$. If V_0 and V_n coincide, that path is said to be a *cycle*. An undirected graph is chordal when it contains no cycles of four or more vertices without a chord, i.e. two non-consecutive vertices that are neighbours. Only if a graph is chordal, it admits a perfect ordering of its cliques. A decomposable model is represented by an undirected chordal graph.

Let us go back to log-linear models. With a little abuse of notation, let A_k be the variable indicating a unit's presence or absence in each of the K lists; allowing for all the pairwise interactions $[A_1 A_2]...[A_{K-1} A_K]$ the resulting graph would be a cycle, and there is no way to exclude the K^{th}-order interaction $[A_1 ... A_K]$. Figure 2 shows this concept in the case $K = 3$. It results that decomposable graphical models are only a subset of the log-linear models. Clearly, limiting ourselves to the decomposable models is restrictive. However, the analysis of solely log-linear models that can be represented as decomposable graphs is much more tractable from a computational point of view. Indeed, Dawid and Lauritzen (1993) show that if a model is decomposable, the joint distribution can be factorized into a product of

conditional distributions; following the notation in Di Cecco (2019):

$$p_G = \prod_{i=1}^{g} p_{C_i} \left(\prod_{j=2}^{g} p_{S_j} \right)^{-1} = p_{C_1} \prod_{i=2}^{g} \frac{p_{C_i}}{p_{S_i}} = p_{C_1} \prod_{i=2}^{g} p_{C_i|S_i}$$

where p_{C_i} (p_{S_j}) is the marginal distribution of over the variables included in the i^{th} clique (j^{th} separator), and $p_{C_i|S_i}$ is the conditional distribution of the i^{th} clique given the relative separator. This way, both the maximization and the integration tasks can be solved in closed form; see Dawid and Lauritzen (1993). Focusing on the Bayesian approach, it is possible to set a prior distribution on cells probabilities which are conjugate with multinomial sampling, as proved by Dawid and Lauritzen (1993). Particularly, a Dirichlet marginal distribution on the probability of each clique C_i, ρ_{C_i} must be placed following the perfect ordering of the cliques. Such prior distribution is the so-called *hyper-Dirichlet*. To account for model uncertainty, Madigan and York (1997) suggests to average all posterior distributions of N conditional on different models m weighted by their posterior model probabilities in order to obtain an unconditional posterior distribution. On model averaging for graphical models see Madigan and Raftery (1994). An illustration of the model in Madigan and York (1997) follows.

Indicate with M the class of possible models for the cell probabilities of the contingency table, indexed by $\mathcal{M} = \{1, 2, ..., s\}$. Define $\mathbf{p}(m)$ as being the vector of cell probabilities for each model $m \in \mathcal{M}$.
Let $y|N, \theta, \mathcal{M} = m$ be Multinomial with parameters $(N, \mathbf{p}(m))$, where

- the prior on N might be
  - $\pi(N) \propto \frac{1}{N}$, i.e. Jeffreys prior;
  - $\pi(N) \propto 2^{-\log^*(N)}$, with $\log^*(N)$ is the sum of the positive terms in $\{\log_2(N), \log_2\{\log_2(N)\}, ...\}$, i.e. the Rissanen's prior;
  - $N \sim$ Poisson(λ) and $\lambda \sim$ Gamma(a, b) if needed;
- the prior on M is Uniform, and
- $\mathbf{p}(m) \sim$ hyper-Dirichlet, as stated before, with $p_C(m) \sim$ Dirichlet(ρ) for each clique. As non-informative choice, ρ is set equal to 1 or $\frac{1}{2}$.

The model has been implemented in the **R** package **dga** (Johndrow et al. (2015)). Considering decomposable models only, all posterior distributions of N conditional on different model M could be easily calculated when K is low. However, as K increases, the number of parameters grows exponentially; thus the calculation becomes cumbersome and the model averaging impracticable. Madigan and York (1995) suggests to use Markov Chain Monte Carlo Model Composition to approximate the average of the posterior distributions under each of the models. On the other hand, Madigan and Raftery (1994) proposes to average over a small set of models, facilitating the communication of model uncertainty (Madigan, Raftery, et al. (1994)). A completely different approach is suggested by Green (1995), that introduces the *reversible jump MCMC*. The RJMCMC is a sampler able to move ("jump") across different dimensions' parameters spaces, thus allowing to explore different dimensions' models in a single chain. Dellaportas and Forster (1999) and King and Brooks (2001) apply the RJMCMC to the log-linear models, going beyond the class of decomposable graphical ones.
In the next subsection, we will review Bayesian hierarchical log-linear models for capture-recapture and give an insight of the RJMCMC.

3.3 Bayesian log-linear models and the Reversible Jump sampler

Dealing with the decomposable class of graphical log-linear models from a Bayesian perspective requires a prior specification for the cell probabilities. Yet, specifying a prior for the model parameters implies the possibility to go beyond the decomposable model and take into account the broader class of log-linear ones.
Dellaportas and Forster (1999) and King and Brooks (2001) are the first main references for a detailed specification of a fully Bayesian log-linear model. Nowadays, Overstall and King (2014a) is a widespread approach; mainly based

on Forster (2010), it is the main theoretical support of the popular **R** package **conting**[3] (see also Overstall and King (2014b)). Notice that these works mainly deal with general log-linear models; they can be easily adapted to capture-recapture problems though. Below we outline a model common to the most recent literature on Bayesian log-linear models, but plugging it in the particular framework of incomplete contingency tables. We mainly refer to Overstall and King (2014b), dwelling on the differences when not negligible.
Following the notation of the previous sections, let y_ω be the number of population units count in cell ω. Moreover,

$$y_\omega|\phi, \boldsymbol{\theta}, m \sim \text{Poisson}(\lambda_\omega)$$

$$\log(\lambda_\omega) = \phi + \mathbf{d}_\omega^T \boldsymbol{\theta} \, .$$

Then, introducing the model indicator m as in subsection 3.2 we obtain

$$\log(\lambda_\omega) = \phi + \mathbf{d}_{m,\omega}^T \boldsymbol{\theta}_m$$

that may be written in compact form as

$$\log(\boldsymbol{\lambda}) = (\mathbb{1}_{2^\kappa}, \mathbf{D}_m)\boldsymbol{\gamma}_m$$

where $\mathbf{D}_m$ is a the matrix whose rows are given by $\mathbf{d}_{m,\omega}$, and $\boldsymbol{\gamma}_m = (\phi, \boldsymbol{\theta}_m)^T$. As an alternative model for the data counts we can also consider

$$\mathbf{y}|N, \boldsymbol{\theta}, m \sim \text{Multinomial}(N, \mathbf{p})$$

with $\mathbf{p}$ being the vector of p_ω's, and each $p_\omega = \frac{\lambda_\omega}{\sum_\omega \lambda_\omega}$. Whatever the model specification, we need to specify the joint prior

$$\pi(\phi, \boldsymbol{\theta}_m, m) = \pi(\phi, \boldsymbol{\theta}_m|m)\pi(m)$$

For the first factor, Overstall and King (2014b) follows Sabanés Bové and Held (2011) using the hyper-g prior, i.e. it decomposes the joint prior on ϕ and $\boldsymbol{\theta}_m$ as

$$\pi(\phi, \boldsymbol{\theta}_m|m) = \pi(\phi)\pi(\boldsymbol{\theta}_m|m)$$

with $\pi(\phi) \propto 1$, and

$$\boldsymbol{\theta}_m|\sigma^2, m \sim N(\mathbf{0}, \Sigma)$$

$$\Sigma = \sigma^2 n(\mathbf{d}_m^T \mathbf{d}_m)^{-1}$$

$$\sigma^2 \sim \text{Inverse Gamma}\left(\frac{a}{2}, \frac{b}{2}\right)$$

Yet, it sets a Uniform prior over the model space, i.e.

$$\pi(m) = \frac{1}{|\mathcal{M}|} = \frac{1}{s}$$

Overstall, King, et al. (2014) proves that under the prior specifications depicted above and choosing the Jeffreys prior for N, the joint posterior for $\boldsymbol{\theta}_m$, m and y_0 are identical under Poisson or Multinomial models.
We report the presence of sensible differences with respect to the other references exclusively concerning the specification of the variance-covariance matrix Σ.
We refer the reader to Dellaportas and Forster (1999) and King and Brooks (2001) for more details.
The joint posterior distribution results being

$$\pi(y_0, \phi, \boldsymbol{\theta}_m, m|\mathbf{y}) \propto \pi(\mathbf{y}, y_0|\phi, \boldsymbol{\theta}_m, m)\pi(\boldsymbol{\theta}_m|\sigma^2, m)\pi(\sigma^2)\pi(m)$$

Updating y_0 and σ^2 from their full conditional distributions is straghtforward; yet, to simulate from the full conditional distribution of the other parameters, Overstall and King (2014b) implements a RJMCMC algorithm.

Let m be the current model at iteration t; denote $\boldsymbol{\gamma}_{\boldsymbol{m}}^{(\boldsymbol{t})}$ the current parameter vector.

[3] available from the CRAN at https://cran.r-project.org/web/packages/conting/index.html

		$\delta_{A_1}=1$		$\delta_{A_1}=0$	
		$\delta_{A_2}=1$	$\delta_{A_2}=0$	$\delta_{A_2}=1$	$\delta_{A_2}=0$
$\delta_{A_3}=1$	$\delta_{A_4}=1$	27	32	42	123
	$\delta_{A_4}=0$	18	31	106	306
$\delta_{A_3}=1$	$\delta_{A_4}=1$	181	217	228	936
	$\delta_{A_4}=0$	177	845	1131	?

Table 4. Number of casualties during the conflict in Kosovo, March-June 1999. Ball et al. 2002

1. Propose a move from model m to model $m' \in \mathcal{M}$ with probability $\pi_{m,m'}$. Typical jumps are those to models with parameters' dimension close to that of model m; in the case of log-linear models this is equivalent to limit the moves to models with one more or one less interaction term. A move to model m itself is also allowed. If $m' = m$, the algorithm turns up being a Metropolis-Hastings, otherwise step 2 follows;
2. generate a vector of *innovation variables* $\mathbf{u}_{m,m'}$ from a proposal distribution $q_{m,m'}(u)$;
3. apply a mapping function T to $(\boldsymbol{\gamma_m}, \mathbf{u}_{m,m'})$ to obtain $\boldsymbol{\gamma}_{m'}$;
4. set $\boldsymbol{\gamma}^{(t+1)} = \begin{cases} \boldsymbol{\gamma}_{m'} & \text{with probability } \alpha \\ \boldsymbol{\gamma}_m & \text{with probability } 1-\alpha \end{cases}$
   where
$$\alpha = \min\left\{1; \frac{\pi(\mathbf{y}, y_0|\boldsymbol{\gamma}_{m'}, m')\pi_{m',m}q_{m',m}(u_{m',m})}{\pi(\mathbf{y}, y_0|\boldsymbol{\gamma}_m^{(t)}, m)\pi_{m,m'}q_{m,m'}(u_{m,m'})}\left|\frac{\partial T(\boldsymbol{\gamma}_m^{(t)}, u_{m,m'})}{\partial(\boldsymbol{\gamma}_m^{(t)}, u_{m,m'})}\right|\right\}$$

See Green (1995) or Robert and Casella (2004) for more details. Clearly, the RJMCMC can be adapted to the restricted class of decomposable models; see King and Brooks (2001).

Once simulated from the posterior distribution, model adequacy can be assessed via information criteria or via the computation of the Bayesian p-value (see Gelman et al. (2004)), as in Overstall, King, et al. (2014).

4. A comparing example: killings in Kosovo

We compare the methods outlined in the previous section using the dataset reported in Ball et al. (2002) about killings in Kosovo during the period March-June 1999. A total of 4400 deaths have been documented by four different sources, i.e. the interviews conducted by the American Bar Association/Central and East European Law Initiative, the exhumation reports produced on behalf of the International Criminal Tribunal for Former Yugoslavia, the Human Rights Watch, and the Organization for Security and Cooperation in Europe. We refer to the sources as A_1, A_2, A_3 and A_4 respectively. The number of casualties recorded by the four sources is summarized in Table 4.
A decade after the conflict, the Humanitarian Law Center (HLC) has published a near-exhaustive list of victims (Humanitarian Law Center (2014)) for the whole period 1998-2000. Manrique-Vallier (2016) uses these data to compute the total number of casualties for the period considered by Ball et al. (2002), giving us a point of reference for the "true" N, which results $N_{HLC} = 10401$.
Ball et al. (2002) estimates $\hat{N}$ for all possible hierarchical log-linear models, computing the confidence interval according to the profile likelihood method of Cormack (1992). According to the adjusted Pearson Chi-square statistic, the best model has one three-factor and two two-factor interaction terms, as shown in Table 5. The 95% confidence interval contains N_{HLC}. To obtain comparative estimates of the total number of casualties, we use the **R** packages **dga** and **conting** implementing Madigan and York (1997) and Overstall and King (2014a) methods respectively. We

	Model	$\hat{N}$	95% CI
Cormack 1992	$[A_1\ A_2\ A_4][A_2\ A_3][A_3\ A_4]$	10356	(9002 12122)
Madigan and York 1997	–	11257	(9352 14318)
Overstall and King 2014	–	12672	(9888 15728)

Table 5. Killings in Kosovo: results

remind that, according to the former, $\hat{N}$ represents the posterior mean of the unconditional posterior distribution of N; for the latter, $\hat{N}$ is the posterior mean of a distribution sampled via a reversible jump algorithm. The results are shown in Table 5. We decided to use the default priors when implementing these models, which are noninformative priors. The credible intervals contain N_{HLC} as well, but they are much wider than the confidence interval obtained with Cormack (1992) method due to the fact that they incorporate the uncertainty linked to the model.

5. Dealing with out-of-scope units

The overcoverage issue has become relevant only recently, with the increase of the interest of the NSIs in the production of statistics through data integration. Such problem naturally arises due to the fact that the aims of those who collect the data and those who use them differ. Out-of-scope units cannot be ignored; the estimate would result strongly biased otherwise. In the next subsections we show how the models discussed in the previous section have since been extended to deal with the presence of latent out-of-scope units in the lists.

5.1 Log-linear models

In the log-linear models framework, Zhang (2015) addresses the overcoverage issue directly modelling the probability of being erroneously classified, i.e. the *error rate*. The author introduces two main alternatives to deal with out-of scope units when $K = 2$, the first of which relies on the conditional independence assumption (CIA) at the base of the standard log-linear models; we briefly discuss the details.
Consider the latent structure of a contingency table of two lists, both affected by overcoverage. Assume the cell counts to be Multinomial with parameters N^*, defined as the sum of N and the captured out-of-target units, and $\mathbf{p}_{\delta_U\omega}$. The error rates $\{\xi_\omega\}$, where $\xi_\omega = P(i \notin U | i \in A_\omega)$, can be defined as functions of $\mathbf{p}_{\delta_U\omega}$, i.e.

$$\xi_{12} = \frac{p_{0\{12\}}}{p_{+\{12\}}}$$

$$\xi_1 = \frac{p_{0\{1\}}}{p_{+\{1\}}}$$

$$\xi_2 = \frac{p_{0\{2\}}}{p_{+\{2\}}}$$

Since we only observe x_1, x_2, x_{12}, the vector $\mathbf{p}_{\delta_U\omega}$ can not be estimated without further assumptions. Hence, assume that a coverage survey S exclusively affected by undercoverage is available and that its inclusion probability is equal to τ_S; also, let y_S be the number of units listed in S and $y_{S\omega}$ be the number of units captured by both S and the set of lists indexed by ω. Then, Zhang (2015) introduces a system of moment equations to model the observations as function of the error rates:

$$\begin{cases} \mathbb{E}(y_{S12}|\mathbf{x}) = x_{12}(1-\xi_{12})\tau_S \\ \mathbb{E}(y_{S1}|\mathbf{x}) = x_1(1-\xi_1)\tau_S \\ \mathbb{E}(y_{S2}|\mathbf{x}) = x_2(1-\xi_2)\tau_S \\ \mathbb{E}(y_{S0}|\mathbf{x}) = (\mathbb{E}(N|\mathbf{x}) - x_{12}(1-\xi_{12}) - x_1(1-\xi_1) - x_2(1-\xi_2))\tau_S \end{cases} \tag{1}$$

The system is underidentified, due to the presence of four parameters in the first three equations; the additional unknown in the fourth equation, $\mathbb{E}(N|\mathbf{x})$, can be derived given the estimates of the others. The idea is to impose a

constraint on the ξ_ω's in order to make the system identifiable. Let us define a log-linear model for the $p_{\delta_U \omega}$'s; the largest nonsaturated model will be

$$\log(p_{\delta_U \omega}) = \beta + \beta_{\delta_U} + \beta_{A_1} + \beta_{A_2} + \beta_{\delta_U A_1} + \beta_{\delta_U A_2} + \beta_{A_1 A_2}$$

Now consider the logit of ξ_{12}; after some algebra, it results

$$\text{logit}(\xi_{12}) = \text{logit}(\xi_1) + \text{logit}(\xi_2) + p_{1\{\mathbf{0}\}} \ ;$$

the model above is said to be *incidental* since it introduces a constraint between the error rate and the population size, thus it can not be taken into account. To overcome this issue, Zhang (2015) set a log-linear model for a transformation of $p_{\delta_U \omega}$, namely $q_{\delta_U \omega} = \frac{\mathbf{p}_{\delta_U \omega}}{1 - p_{1\{\mathbf{0}\}}}$. Again, the error rate can be expressed as a function of the $q_{\delta_U \omega}$'s, i.e. $\xi_\omega = \frac{q_{0\{\omega\}}}{p_{+\{\omega\}}}$ and $\text{logit}(\xi_{12})$ becomes

$$\text{logit}(\xi_{12}) = \text{logit}(\xi_1) + \text{logit}(\xi_2)$$

which is not an incidental model and makes the system (1) identifiable. For small ξ_ω, $\text{logit}(\xi_\omega) \approx \log(\xi_\omega)$; hence,

$$\xi_{12} = \xi_1 \xi_2 \ . \tag{2}$$

However, in case of good data quality and low error rates, it is reasonable to assume that the domain $\{12\}$ is much larger than both $\{1\}$ and $\{2\}$, and that the error rate among the units in $\{12\}$ is much lower than that in $\{1+\}$ and $\{+2\}$. Therefore, as an alternative to the previous model, Zhang (2015) suggests to express ξ_{12} as the product of the marginal error rates rather than the cross-classified ones; in other words, it can be assumed that

$$P(i \notin U | i \in A_{12}) = P(i \notin U | i \in A_{1+}) P(i \notin U | i \in A_{+2}) \ ,$$

or

$$\xi_{12} = \xi_{1+} \xi_{+2} \ . \tag{3}$$

Contrarily to identity (2), the condition above can not be derived from a standard log-linear model, thus it does not rely on the concept of conditional independence; Zhang (2015) calls that in (3) *pseudo conditional independence* (PCI) assumption, whose concept is extended to the case of $K \geq 2$ in Zhang (2019). The idea is to define a general log-linear model for the marginal $\xi_{\omega+}$

$$\log(\xi_{\omega+}) = \sum_{\nu \in \Omega(\omega)} \log \psi_{\nu+}$$

such that each unsaturated model corresponds to a different specification of the PCI assumption; e.g. the model including none of the interaction terms corresponds to the mutual PCI between the marginal list domains. ξ_ω is estimated via ML; we refer the reader to Zhang (2019) for the details.

5.2 Decomposable graphical models

Di Cecco (2019) extends the decomposable model described in subsection 3.2 introducing a latent class (LC) approach developed both from a frequentist and a Bayesian perspective - the latter proposed also in Di Cecco et al. (2020).

There exists a vast literature on the use of LC models in the capture-recapture framework, particularly addressing the heterogeneity problem; among others, see Agresti (1994), Bartolucci and Forcina (2001) and Bartolucci and Pennoni (2007). The first dealing with erroneous enumeration using LC models were Biemer, Brown, Judson, and Wiesen (2001), followed by Biemer, Woltmann, et al. (2001) and Biemer, Brown, and Judson (2004). Such strand of literature has led to frame the identifiability problems arising in this context, highlighting the fact that at least four lists are needed to estimate any LC model that include interactions among lists; see Brown et al. (2004) and Biemer (2011) for further details.

Here we describe the approach by Di Cecco (2019) mainly because of its computational advantages.

Recalling the notation in section 2, define $x_{\delta_\mathbf{A}}$ as the number of observed units with capture history $\delta_\mathbf{A}$. Introduce the latent class δ_U; let $x_{\delta_U \delta_\mathbf{A}}$ be the number of observed units with capture history $\delta_\mathbf{A}$ belonging ($\delta_U = 1$) or not ($\delta_U = 0$) to the target population. Coherently with the notation introduced, we may also rename $x_{1\delta_\mathbf{A}} = y_{\delta_\mathbf{A}}$, and

$x_{0\delta_{\mathbf{A}}} = r_{\delta_{\mathbf{A}}}$. The objective is to estimate $N = \sum_{\delta_{\mathbf{A}}} y_{\delta_{\mathbf{A}}} + y_{\mathbf{0}}$.
The class of model considered can be expressed as

$$p_{\delta_{\mathbf{A}}} = \sum_{\delta_U} p_{\delta_U} p_{\delta_{\mathbf{A}}|\delta_U}$$

Restrict the interest to decomposable model only. Since the latent variable δ_U interacts with all other variables, the likelihood in subsection 3.2 can be written as

$$p_G = p_{\delta_U} p_{C_1|\delta_U} \prod_{i=2}^{g} p_{C_i|S_i} \ .$$

Such likelihood function may be maximized via EM algorithm (see Di Cecco (2019) for a detailed description) or used to compute population size's posterior distribution in a fully Bayesian analysis. In the latter case, the prior specification is similar to that described in section 3.2; it is sufficient to add a Beta prior on p_{δ_U}. MCMC methods can be used to sample from the posterior distribution; in particular, a Gibbs sampler is appropriate in case of Jeffreys prior on N, a Metropolis-within-Gibbs is needed otherwise.

5.3 Bayesian log-linear models

The last approach to overcoverage we analyse is that in Overstall, King, et al. (2014), which is proposed also in Overstall and King (2014b) as an extension of the basic model. Assume that $J < K$ lists enumerate also units that are not part of the target population. For those cells ω affected by overcoverage, y_ω can be seen as the true value of a *left censored* cell count, since only its upper bound is observed, i.e. x_ω. Let $\mathbf{y}$ indicate the vector of counts such that $x_\omega = y_\omega$; let $\mathbf{x}^c$ and $\mathbf{y}^c$ be the vectors of observed counts and number of population units respectively such that $y_\omega < x_\omega$ (c stands for *censored*). Recall the joint posterior introduced in subsection 3.3; now it becomes

$$\pi(y_0, \mathbf{y}^c, \phi, \theta_m, \sigma^2, m|\mathbf{y}, \mathbf{x}^c) \propto \pi(\mathbf{y}, y_0, \mathbf{y}^c|\phi, \theta_m, m)\pi(\mathbf{x}^c|\mathbf{y}^c)\pi(\theta_m|\sigma^2, m)\pi(\sigma^2)\pi(m)$$

The additional step to include in the algorithm relative to the model described in subsection 3.3 is the sampling of the latent true count for those ω's affected by overcoverage;

$$y^c_\omega|\phi, \theta_m, x^c_\omega, m \sim \text{Truncated Poisson}(\lambda_\omega, x^c_\omega)$$

6. Comparing examples with simulated data

As for the general capture-recapture setting, we may want to compare the methods described in the previous section. Nevertheless, while the three models introduced in the general framework are applicable to the same context and aim at the same objective, the models proposed for facing the overcoverage issue differ in their motivations. E.g. the idea behind the strand of LC models for capture-recapture is to identify the different behaviors of different (sub)populations captured in the same occasions in order to reliably estimate the size of the population of interest. In Overstall, King, et al. (2014) the cross-classified overcount can be seen as a noise, a measurement error deriving from no specific underlying behavior; yet Zhang (2019) models the erroneous enumerations relying on the goodness of data. Moreover, Zhang (2019) models a situation in which all the sources - at least two - are affected by overcoverage but the enumeration survey, whereas in Overstall, King, et al. (2014) a maximum number of cell counts with erroneous enumeration is defined depending on the total number of observations, to preserve the identifiability of the model. For the same reason, LC models with less than four lists can be estimated only if local independence is assumed.

Having this premise in mind, in the next paragraphs we present two different simulated scenarios and for each of them we compare the models described in section 5 to see how different assumptions impact the results. The first scenario simulates the situation in which the sources are homogeneously affected by the presence of erroneous enumeration; the second scenario is the (typical) case in which a post-enumeration survey is conducted in order to assess the goodness of administrative lists in covering the target population and the data quality of such lists is pretty good.
We follow the frequentist approach in the estimation of the decomposable graphical models (see Di Cecco (2019)). The estimates for Overstall, King, et al. (2014) model are obtained using the **R** package **conting** (see Overstall and King (2014b)).

			$\delta_{A_1}=1$		$\delta_{A_1}=0$	
			$\delta_{A_2}=1$	$\delta_{A_2}=0$	$\delta_{A_2}=1$	$\delta_{A_2}=0$
$\delta_{A_3}=1$	$\delta_{A_4}=1$	$\delta_U=1$	25	40	22	446
		$\delta_U=0$	74	30	25	55
	$\delta_{A_4}=0$	$\delta_U=1$	148	245	134	2697
		$\delta_U=0$	200	81	67	148
$\delta_{A_3}=0$	$\delta_{A_4}=1$	$\delta_U=1$	164	270	148	2981
		$\delta_{A_4}=0$	148	60	49	110
	$\delta_{A_4}=0$	$\delta_U=1$	992	1636	898	18034
		$\delta_U=0$	403	164	134	

Table 6. Data from scenario 1

Model	$\hat{N}$	$\sum_\omega \hat{y}_\omega$
Di Cecco 2019 - $[XA_1A_2][XA_3][XA_4]$	28943	10903
Zhang 2019; error free list - A_4	19238	9599
Overstall, King et al. 2014 - $[A_1A_2][A_1A_3][A_1A_4][A_2A_3][A_2A_4][A_3A_4]$	28428	27925
True values	28880	10846

Table 7. Results from scenario 1

6.1 Scenario 1: capturing two groups

First we generate the contingency table in Table 6 from a decomposable graphical LC model $[XA_1A_2][XA_3][XA_4]$. The coverage rates of lists A_1, A_2, A_3, A_4 are set to be between 9% and 15%, yet their marginal error rates are equal to 0.25, 0.3, 0.15 and 0.12 respectively; the target population size N amounts to 28880 and the number of units captured by the four sources is equal to 30927.
We compare the estimates obtained via the EM algorithm for capture-recapture LC models by Di Cecco (2019) using the true data model with those obtained using Zhang (2019) algorithm and the **R** package **conting**; Table 7 shows the best results in terms of AIC or Bayesian p-value. For Zhang (2019) we indicate which of the sources is preferred as being the error free source, whereas for Overstall, King, et al. (2014) the maximal model we fixed.

In this scenario, the PCI based model from Zhang (2019) performs poorly in terms of y_0 estimation, whatever the choice of the error free list. However, it is good in detecting the amount of out of scope units in the sample; according to the information criterion, such model best performs assuming list A_4 to be the overcoverage free source, which is in fact the list with the lowest error rate. On the other hand, for the Bayesian log-linear model by Overstall, King, et al. (2014) the population size estimate improves as the number of interaction terms included in the maximal model increases, regardless of which cells are censored. Note that to include at least all the two-ways interactions, the number of censored cells can not be more than three; the best performing model allows for the censoring of cells (x_1, x_2).

			$\delta_{A_1}=1$		$\delta_{A_1}=0$	
			$\delta_{A_2}=1$	$\delta_{A_2}=0$	$\delta_{A_2}=1$	$\delta_{A_2}=0$
$\delta_{A_3}=1$	$\delta_{A_4}=1$	$\delta_U=1$	768	5474	164	1145
		$\delta_U=0$	0	0	6	380
	$\delta_{A_4}=0$	$\delta_U=1$	3660	3703	711	721
		$\delta_U=0$	0	0	172	2259
$\delta_{A_3}=0$	$\delta_{A_4}=1$	$\delta_U=1$	4563	4951	843	1006
		$\delta_U=0$	0	0	403	1696
	$\delta_{A_4}=0$	$\delta_U=1$	4112	1659	834	343
		$\delta_U=0$	0	0	1372	

Table 8. Data from scenario 2

6.2 Scenario 2: post-enumeration survey and accurate administrative data

Table 8 summarizes the capture histories of 40945 units during four capture occasions. A_1 is an error free source with a high population coverage rate, equal to 0.83, whereas the others are affected by the presence of erroneous enumerations, with marginal error rates set to be between 10% and 15%, with a total of 6288 out-of-target units. Since a post-enumeration survey should capture target units uniformly, we assume A_1 to be independent from the other sources, but lists A_2, A_3 and A_4 target captures are dependent on each other, i.e. the target counts in Table 8 are generated from the log-linear model $[A_1][A_2A_3A_4]$. Counts of out-of-scope units are added such that the error rates domains $\{23+\}$, $\{24+\}$ and $\{34+\}$ are smaller than $\{2+\}$, $\{3+\}$ and $\{4+\}$ and larger than $\{234+\}$; moreover, cross-sectional error rates are much bigger than their respective marginal ones. We estimate the population size implementing the three models described in section 5; results are shown in Table 9.

Zhang (2019) algorithm correctly detects list A_1 as the error free source, although it overestimates the number of out-of-target units. Nevertheless, it performs well in the estimation of y_0. On the other hand, Overstall, King, et al. (2014) overestimates both the number of out-of-target units and y_0, despite it recognizes the true data model; in this case it allows for the censoring of cells $(x_2, x_3, x_4, x_{23}, x_{24}, x_{34})$. Concerning the LC model, we experimented the functioning of the EM algorithm in Di Cecco (2019) allowing for the interaction between the latent variable and list A_1; results are far from the true values. Hence, we split the estimation procedure in two steps. Firstly the EM for LC in capture-recapture is implemented considering the lists affected by erroneous enumeration only; given the identifiability problems previously discussed, we are constrained to the local independence model, i.e. A_2, A_3 and A_4 are set to be independent given the latent variable. Then, we use the vector $\hat{\boldsymbol{y}}$ to get an estimate of y_0 fitting a log-linear model. Here, the main issue consists of the impossibility to compare (or average) different models and to allow for the manifest variables' interaction; this leads to the underestimation of the number of out-of-scope units.

7. Conclusions

In this paper we reviewed the main and most recent literature dealing with the population size estimation problem, with an insight on the overcoverage issue. Nevertheless, we do not claim to be exhaustive. Many methodological

Model	$\hat{N}$	$\sum_{\omega} \hat{y}_{\omega}$
Di Cecco 2019 - local independence	37291	36876
Zhang 2019 - error free list: A_1	33345	33089
Overstall, King et al. 2014; $[A_1][A_2A_3A_4]$:	35258	33807
True values	34657	34314

Table 9. Results from scenario 1

	$\hat{N}$	95% CI
HLC count	10401	–
Manrique-Vallier 2016	10442	(9020 13637)
Ball et al. 2002 $[A_1\ A_2\ A_4][A_2\ A_3][A_3\ A_4]$	10356	(9002 12122)

Table 10. Killings in Kosovo: Manrique-Vallier 2016 results

issues that have not been covered in this paper deserves attention. Among them, how to include available covariates in the model is one of the most interesting. See Part III of Böhning et al. (2018) for some recent advances. Furthermore, in real applications the assumptions at the basis of log-linear models introduced in section 3.1 may not hold. On the one hand, there might be heterogeneity in the population, i.e. the capture probabilities vary among individuals. On the other hand, it may occur that the captured units are not uniquely labelled across the multiple lists.

The latter case, i.e. the lack of unique labelling, implies a non-exact match of the units observed in multiple lists, and hence a potential overestimation of the population of interest; in this framework, linkage uncertainty must be taken into account. Di Consiglio, Tuoto, and Zhang (2019) reviews some of the approaches taking into account such uncertainty in the population size estimation procedure, from natural extensions of Fienberg (1972) model (see Di Consiglio and Tuoto (2018)), to fully Bayesian models as that in Tancredi and Liseo (2011). A unique labelling may miss within the sources as well, implying the presence of duplicates in some lists; see Tancredi, Steorts, et al. (2019) for an approach to this kind of issue.

The former case includes scenarios in which heterogeneity is either due to some measurable attributes, e.g. sex, age, or given by some unmeasurable characteristics (Johnson et al. (1986)). If the source of heterogeneity is known and the attributes are recorded in the available data, a convenient strategy is to stratify the population to obtain different homogeneous groups. Otherwise, literature offers a variety of methods to address the issue. Among them, we cite Fienberg et al. (1999), which encompasses the log-linear models and the Rasch model (Rasch (1960)) in a fully Bayesian hierarchical framework. Moreover, the latent variable approach described in subsection 5.2 can be easily adapted to the heterogeneity problem; actually, in Di Cecco (2019) and Di Cecco et al. (2020) the mixture model is allowed to have more than two components.

A nonparametric Bayesian approach dealing with population heterogeneity in capture-recapture experiments can be found in Manrique-Vallier (2016). The underlying idea is that, in case of heterogeneous population and in the absence of covariates allowing to appropriately stratify the sample, it is convenient to assume that the population may be partitioned into an unknown number of homogeneous strata within which the independence model holds. In this case, the generating mechanism of the capture vectors consists of a general capture-recapture Multinomial model in which the probability mass function of each capture vector is a Dirichlet-process mixture of product-Bernoulli distributions. Indeed, letting the weights of the mixture be generated from a stick-breaking process allows to avoid the specification of the number of mixture components in advance; the identification of the number of homogeneous groups within the heterogeneous population occurs in an unsupervised way. Manrique-Vallier (2016) applies the nonparametric latent class model to the data killings in Kosovo shown in section 4, and it results to perform very well. The point estimate of N is incredibly close to the count N_{HLC}. See Table 6 to compare Manrique-Vallier (2016) results to the real count of casualties and to the best model in Ball et al. (2002). We refer the reader to Manrique-Vallier (2016) for model's details. From the same author, and specifically for casualties estimation in capture-recapture experiments, see Manrique-Vallier et al. (2019). Yet, for a review of most of the literature on heterogeneous animal populations in capture-recapture models, see Gimenez et al. (2018).

References

AGRESTI A. (1994), Simple capture-recapture models permitting unequal catchability and variable sampling effort, *Biometrics*, 494–500.

AMORÓS J. (2014), Recapturing Laplace, *Significance*, 11(3), 38–39.

BALL, P. et al. (2002), Killings and Refugee Flow in Kosovo – March-June 1999, *Report to ICTY*.

BARTOLUCCI F., FORCINA A. (2001), Analysis of capture-recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality, *Biometrics*, 57(3), 714–719.

BARTOLUCCI F., PENNONI F. (2007), A class of latent Markov models for capture–recapture data allowing for time, heterogeneity, and behavior effects, *Biometrics*, 63(2), 568–578.

BIEMER P. P. (2011), *Latent class analysis of survey error*, Vol. 571, John Wiley Sons.

BIEMER P. P., BROWN G. G., JUDSON D. H. (2004), Latent Class Models for Evaluating the Accuracy of Census Counts, *2004 Proceedings of the American Statistical Association, Section on Survey Research Methods*, Alexandria.

BIEMER P. P., BROWN G. G., JUDSON D. H., WIESEN C. (2001), Triple System Estimation with Erroneous Enumerations in the Administrative Records List, *2001 Proceedings of the American Statistical Association, Survey Research Methods Section* [CD-ROM], Alexandria, VA: American Statistical Association.

BIEMER P. P., WOLTMANN H., et al. (2001), Enumeration accuracy in a population census: An evaluation using latent class analysis, *Journal of Official Statistics*, 17(1), 129.

BISHOP Y. M., FIENBERG S. E., HOLLAND P. W. (1975), *Discrete Multivariate Analysis - Theory and Applications*, Interdisciplinary Statistics Series, Cambridge, Massachusetts: The MIT Press.

BÖHNING D., VAN DER HEIJDEN P. G. M., BUNGE J., eds. (2018), *Capture-Recapture Methods for the Social and Medical Sciences*, Interdisciplinary Statistics Series. Chapman and Hall/CRC.

BROWN G. G., BIEMER P. P., JUDSON D. H. (2004), Estimating Erroneous Enumeration in the US Decennial Census using Four Lists, paper presented at the Joint Statistical Meetings, Survey Research Methods Section.

CHIARIELLO V. TUOTO T. (2018), Estimation of Criminal Populations Using Administrative Registers in the Presence of Linkage, URL: https://www.istat.it/it/files/2018/11/Chiariello_originalpaper.pdf.

CORMACK R. M. (1992), Interval Estimation for Mark-Recapture Studies of Closed Population, *Biometrics*, 48(2), 567–576.

DAWID A. P., LAURITZEN S. (1993), Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models, *Annals of Statistics*, 21, 1272–1317.

DELLAPORTAS P., FORSTER J. J. (1999), Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-Linear Models, *Biometrika*, 86(3), 615–633.

DI CECCO D. (2019), Estimating Population Size in Multiple Record System with Uncertainty of State Identification, in ZHANG L-C., CHAMBERS R. L., *Analysis of Integrated Data*, Chapman and Hall/CRC, Chap. 8.

DI CECCO D., DI ZIO M., LISEO B. (2020), Bayesian latent class models for capture–recapture in the presence of missing data, *Biometrical Journal*.

DI CONSIGLIO L. TUOTO T. (2018), Population Size Estimation and Linkage Errors: the Multiple Lists Case, *Journal of Official Statistics*, 34(4), 889–908.

DI CONSIGLIO L., TUOTO T., ZHANG L-C. (2019), Capture-Recapture Methods in the Presence of Linkage Errors, in ZHANG L-C., CHAMBERS R. L., *Analysis of Integrated Data*, Chapman and Hall/CRC., Chap. 2.

FIENBERG S. E. (1970), The Analysis of Multidimensional Contingency Tables, *Ecology*, 51(3), 419–433.

— (1972), The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables, *Biometrika*, 59(3), 591–603.

FIENBERG S. E., JOHNSON M. S., JUNKER B.W. (1999), Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(3), 383–405.

FILIPPONI D., GUARNERA U., VARRIALE R. (2017), Integration of Administrative Sources and Survey Data Through Hidden Markov Models for the Production of Labour Statistics, URL: https://www.istat.it/it/ files/2018/11/ Filipponi_original-paper.pdf.

FORSTER J. J. (2010), Bayesian Inference for Poisson and Multinomial Log-linear Models, Working Paper M09/11

University of Southampton, Southampton Statistical Sciences Research Institute.

GELMAN A. et al., eds. (2004), *Bayesian Data Analysis*, 2nd ed. Chapman Hall: Boca Raton.

GIMENEZ O., CAM E., GAILLARD J.-M. (2018), Individual heterogeneity and capture–recapture models: what, why and how? *Oikos*, 127(5), 664–686.

GREEN P. J. (1995), Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, 82(4), 711–732.

HUMANITARIAN LAW CENTER (2014), The Kosovo Memory Book Project 1998-2000: List of Killed, Missing and Disappeared 1998-2000, URL: http://www.kosovskaknjigapamcenja.%20org/db/kkp$_en/index.html$.

JOHNDROW, J., LUM K., BALL P. (2015), dga: Capture-Recapture Estimation Using Bayesian Model Averaging, URL: https://CRAN.R-project.org/package=dga.

JOHNSON D. H., BURNHAM K. P., NICHOLS J. D. (1986), The role of heterogeneity in animal population dynamics.

KING R., BROOKS S. P. (2001), On the Bayesian Analysis of Population Size, *Biometrika* 88(2), 317–336.

LINCOLN F. C. (1930), Calculating Waterfowl Abundance on the Basis of Banding Returns, Circular of the United States Dept. of Agriculture – no. 118.

MADIGAN D., RAFTERY A. E. (1994), Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window, *Journal of the American Statistical Association*, 89, 1535–1546.

MADIGAN D., RAFTERY A. E., et al. (1994), Strategies for Graphical Model Selection, in CHEESMAN P., OLDFORD R. W. *Selecting Models from Data: AI and Statistics*, New York: Springer-Verlag, 91–100.

MADIGAN D., YORK J. C. (1995), Bayesian Graphical Models for Discrete Data, *International Statistical Review*, 63(2), 215–232.

— (1997), Bayesian Methods for Estimation of the Size of a Closed Population, *Biometrika*, 84(1), 19–31.

MANRIQUE-VALLIER D. (2016), Bayesian Population Size Estimation Using Dirichlet Process Mixtures, *Biometrics*, 72(4), 1246–1254.

MANRIQUE-VALLIER D., BALL P., SADINLE M. (2019), Capture-Recapture for Casualty Estimation and Beyond: Recent Advances and Research Directions.

OVERSTALL A. M., KING R. (2014a), A Default Prior Distribution for Contingency Tables with Dependent Factor Levels, *Statistical Methodology*, 16, 90–99.

— (2014b), conting: An R Package for Bayesian Analysis of Complete and Incomplete Contingency Tables, *Journal of Statistical Software*, 58(7), pp. 1–27.

OVERSTALL A. M., KING R., et al. (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.

PETERSEN C. G. J. (1985), The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea, *Report of the Danish Biological Station*, 6, 5–84.

RASCH G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Interdisciplinary Statistics Series.

REGAL R. R., HOOK E. B. (1991), The Effects of Model Selection on Confidence Intervals for the Size of a Closed Population. *Statistics in Medicine*, 10, 717–721.

ROBERT C. P., CASELLA G., eds. (2004), *Monte Carlo Statistical Methods*, Springer Text in Statistics, Springer.

SABANÉS BOVÉ D., HELD L. (2011), Hyper-g Priors for Generalized Linear Models, *Bayesian Analysis* 6(3), 387–410.

SANATHANAN L. (1972), Estimating the Size of a Multinomial Population, *The Annals of Mathematical Statistics*, 43(1), 142–152.

TANCREDI A., LISEO B. (2011), A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems, *The Annals of Applied Statistics* 5(2B), 1553–1585.

TANCREDI A., STEORTS R., LISEO B. (2019), A Unified Framework for De-duplication and Population Size Estimation, *Bayesian Analysis*, 1–26.

ZHANG L-C. (2015), On Modelling Register Coverage Errors, *Journal of Official Statistics*, 31(3), 381–396.

— (2019), Log-linear Models of Erroneous List Data, in ZHANG L-C., CHAMBERS R. L., *Analysis of Integrated Data*, Chapman and Hall/CRC, Chap. 9.

ZHANG L-C., CHAMBERS R. L., eds. (2019), *Analysis of Integrated Data*, Statistics in the Social and Behavioral Sciences Series, Chapman and Hall/CRC.