DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 557–572, http://dx.doi.org/10.2478/JOS-2018-0026

Population Size Estimation Using Multiple Incomplete Lists with Overcoverage

Davide Di Cecco[1], *Marco Di Zio*[1], *Danila Filipponi*[1], *and Irene Rocchetti*[1]

The quantity and quality of administrative information available to National Statistical Institutes have been constantly increasing over the past several years. However, different sources of administrative data are not expected to each have the same population coverage, so that estimating the true population size from the collective set of data poses several methodological challenges that set the problem apart from a classical capture-recapture setting. In this article, we consider two specific aspects of this problem: (1) misclassification of the units, leading to lists with both overcoverage and undercoverage; and (2) lists focusing on a specific subpopulation, leaving a proportion of the population with null probability of being captured. We propose an approach to this problem that employs a class of capture-recapture methods based on Latent Class models. We assess the proposed approach via a simulation study, then apply the method to five sources of empirical data to estimate the number of active local units of Italian enterprises in 2011.

Key words: Multisource integration; capture-recapture models; latent class models; missing data.

1. Introduction

Traditionally, official statistics use primary data obtained from sample and complete enumeration surveys, whereas secondary data (namely administrative data) are auxiliary sources of information. Nowadays, National Statistical Institutes (NSIs) are investigating the possibility of producing official statistics solely from administrative data, such as register-based statistics (Wallgren and Wallgren 2007). However, since administrative data are gathered by other organizations for their specific aims, units and variable definitions may not align perfectly with those of the official statistics program (Zhang 2012; Zhang 2015).

In this article, we focus on population size estimation that uses multiple data sources, where all sources are incomplete (they do not list all units, and some unobserved units are not registered in any list), and overlapping (a unit can be registered in several sources). Our methodological framework is capture-recapture modeling with multiple lists, where the event of being captured corresponds to the event of being registered in one or more lists.

This scenario is frequently encountered in practice, as the number of available administrative data sources for NSIs has been constantly increasing. On one hand, the

[1] Italian National Statistical Institute, via Cesare Balbo 16, Rome 00184, Italy. Emails: dicecco@istat.it, dizio@istat.it, dafilipp@istat.it, and irocchetti@istat.it.

increased number of administrative data sources provides the opportunity to use and refine existing statistical methodologies that exploit the information redundancy. On the other hand, the different sources of administrative data are rarely of uniform quality. In these cases, we have a trade-off between the number of sources included in the analysis, and the overall quality of our data. That is, if we order the sources by quality (according to our past experience or according to a set of quality indicators), as the number of lists included in the construction of population estimates increases, the likelihood of violating the classical assumptions of a capture-recapture setting grows correspondingly. In particular, the main coverage issues that we encounter in practice are:

1. partial information, where the list only contains information on a specific subpopulation (subset) of our target population ("incomplete sources/lists"). This occurs often, as many of the new available sources are obtained from organizations collecting data for their own purposes, typically targeting a specific set of units (e.g., specific categories of workers, enterprises having certain legal form);
2. misclassification, which may be due to differences in the definition of the units or to delays in the registration/cancellation from a list.

The first problem leads to subsets of units with null probability of being included in some lists; the second leads to "false captures", that is, units that do not belong to our target population, but are erroneously included. As a consequence, each separate list may be subject to under- and overcoverage. In the vast literature of capture-recapture, these two problems have rarely been addressed and, to the best of our knowledge, have never been addressed simultaneously.

The problem of incomplete lists is studied in Sutherland (2003), Sutherland and Schwarz (2004) and, in different terms, in Zwane et al. (2004). These studies show that ignoring the incompleteness of the lists, that is, treating the uncatchable units as sampling zeros for the incomplete lists results in biased estimates of the population size. They suggest treating the unobservable captures of the units not covered by the incomplete lists as missing information under a Missing at Random (MAR) assumption. Then, each such unit is considered as partially classified, that is, as if the capture history is partially missing, and an Expectation-Maximization (EM) algorithm is presented to estimate the missing part. Note that this approach allows us to use all records even in the presence of incomplete lists. This is particularly important for us, since the model we are proposing requires a certain minimum number of sources to be identifiable (details will be given in a later section). Hence, in many cases, we cannot avoid using such incomplete lists. If we wanted to limit the estimation of the model on the subset of the population where all sources operate, we would have a great loss of information. In fact, we frequently encounter situations where the missing patterns in the lists are complementary and there is little or no subset where all sources operate.

As for the second point, in practice, overcoverage is often preliminarily treated by clerical review of spurious events and duplicated records, or by ad hoc studies that identify and remove units not belonging to the target population according to a set of deterministic rules. Afterwards, capture-recapture methodologies are applied to the treated data, which have been classified as entirely comprising target units. Sometimes, ad hoc surveys are deployed to estimate the overcoverage. This approach is common in

censuses, where O-sample coverage surveys are an integral part of the process (Zhang 2015). For example, in the 2008 Israeli Population Census, a 20% sample was selected to independently validate the correctness of the individual legal address in the Population Register. The sample is used to estimate individual weights representing under- and overcoverage parameters. Finally, coverage estimates are based on an extension of the classic Dual System Estimation, where "false captures" in the Population Register are accounted for by means of the weights (Kamen 2005).

Whenever a training set of validated data (such as a sample survey) is not available, we cannot fit a supervised model, and the possibility of an unsupervised approach can be considered. We treat overcoverage as being entirely induced by misclassification, without further analysis of the source of error, and propose the use of an unsupervised approach based on Latent Class Models (LCM), as LCM are particularly suitable for handling misclassification errors in an unsupervised fashion (e.g., Biemer 2011). The use of finite mixture models in capture-recapture applications to account for unobserved heterogeneity in capture probabilities is well known. The logic behind this is to improve the goodness of fit of the model by partitioning units into two or more homogeneous groups, according to a discrete latent variable (see e.g., Pledger 2000). In particular, the use of LCM in capture-recapture dates back to Agresti (1994). Since then, several extensions to the LCM models have been proposed to include covariates to model observed heterogeneity, and to relax the local independence assumption of the LCM, that is, the hypothesis of independence of captures of the same unit in different sources conditionally on the latent variable (e.g., Bartolucci and Forcina 2001). Relaxing the local independence hypothesis is of particular interest in our applications, as it is, in most cases, hardly tenable. For example, consider the case where it is a legal obligation to be registered on a source in order to be registered in a second source (e.g., any enterprise registered in the Chambers of Commerce should have a VAT and be registered in the Tax Revenue Agency). For these reasons, we opted for a generalization of the LCM that includes dependencies between captures in different lists. These models are sometimes called Local Dependence Models (Hagenaars 1988) or modified Lisrel models (Hagenaars 1993) and can be expressed as loglinear models with a latent variable. For the use of these models in capture-recapture, see Biggeri et al. (1999), and Stangehellini and Van der Heijden (2004).

The use of a latent variable to directly model overcoverage has been largely ignored in the literature. A contribution is given in Biemer (2011, Ch. 6.3), who proposes the use of these models for a Triple System Estimate, in which the classical Dual System Estimator based on Census data and a coverage survey is extended to a situation where a third administrative source is added.

In this article, we present some results on the use of these models to jointly estimate under- and overcoverage in the presence of incomplete lists. The article is structured as follows. In Section 2, we describe the model used for the estimation of population size with incomplete lists and false capture, and then we illustrate the algorithm used to estimate unknown parameters. Section 3 presents a simulation study designed to assess the performance of the proposed estimation method. An application on empirical data from an Istat business survey is presented in Section 4. In Section 5, we provide general conclusions and discuss areas of future research.

2. Model and Estimation

First of all, let us formalize a simple capture-recapture problem based on loglinear models. Suppose we have f lists, and let $Y_1, \ldots, Y_f$ be the binary random variables (r.v.s) associated with each list such that $Y_i = 1$ when a given unit is observed in the ith list, and 0 otherwise. The set composed of the union of the observations included in the lists has size n_{obs} and can be arranged in a 2^f contingency table $T = \left[n^{Y_1,\ldots,Y_f}_{y_1\ldots y_f} \right]_{y_1\ldots y_f \in \{0,1\}^f}$ where each cell represents the number of observed units presenting a certain pattern $(y_1, \ldots, y_f)$ of inclusion in the lists (also known as capture history). Hereafter, the superscript of n is omitted when the reference to the r.v.s is clear from the context. Cell $n_{0\ldots0}$, corresponding to $(Y_1 = 0, \ldots, Y_f = 0)$, is a structural zero cell, since no units can be observed for this combination. Our goal is to estimate the population size N, where $N = n_{obs} + n_{0\ldots0}$.

The use of loglinear models is typical in situations where each list has a different capture probability and captures of the same unit in different lists are not independent. Actually, in a loglinear model we explicitly model the dependencies between the r.v.s $Y_1, \ldots, Y_f$ by means of interaction parameters. The estimate of the unobserved count $n_{0\ldots0}$ (and, in general, of any structural zero cell) is obtained by the maximum likelihood (ML) estimates of the loglinear model conditional on the observed data (Fienberg 1972).

For ease of understanding, we describe the problem in the simplest case of two lists. The structure of the data is defined as in Table 1, where the generic cell represents $n^{Y_1Y_2}_{y_1y_2}$ for $y_1, y_2 \in \{0,1\}^2$. For instance, n_{10} represents the number of units captured only by list Y_1. We denote the row sum by n_{i+} and the column sum by n_{+j}, where the subscript "+" denotes the sum over the index it replaces. The shadowed cells in Table 1 are the observed counts (their sum is n_{obs}), while the structural zero cell is n_{00}.

In our problem, units observed in the lists do not all belong to the target population. In order to model this, we add a dichotomous latent variable X, identifying the in-scope and the out-of-scope units, to the loglinear model, letting

$$X = \begin{cases} 1 & \text{when a unit is in the target population;} \\ 0 & \text{otherwise.} \end{cases}$$

In this scenario, we are interested in estimating the target population size, that is, the number N_1 of units for which $X = 1$, $(N_1 + N_0 = N)$.

In every loglinear model considered, the latent variable X interacts with all observed variables $Y_1, \ldots, Y_f$. Hence, the simplest model that we look at (corresponding to a simple LCM) is:

$$[XY_1][XY_2]\ldots[XY_f], \tag{1}$$

Table 1. Example of the contingency table defined by two lists Y_1 and Y_2.

	Y_2		
Y_1	Included	Not included	Total
Included	n_{11}	n_{10}	n_{1+}
Not included	n_{01}	n_{00}	n_{0+}
Total	n_{+1}	n_{+0}	N

where we use the classic notation of hierarchical loglinear models reporting only the higher order interaction terms. Then, each additional interaction parameter with respect to (1) represents a deviation from the local independence assumption of a LCM. The term Local Dependence Model is generally used for this setting.

A necessary condition for any such model to be identifiable is that the number of parameters is not smaller than the number of degrees of freedom of the observed contingency table. As a consequence, we necessitate at least four lists. Note that it is possible to use three lists by introducing constraints on the parameters (e.g., see Biemer 2011).

To clarify things, we describe the case of two lists (previously introduced in Table 1) with the introduction of the latent variable X. The complete contingency table is illustrated in Table 2, where the generic cell is $n_{y_1y_2x}^{Y_1,Y_2,X}$ for $y_1, y_2, x \in \{0,1\}^3$, and for instance n_{101} represents the number of in-scope units captured only in list Y_1. The shadowed cells in Table 2 are the observed counts, and their sum is n_{obs}, while the structural zeros cell are n_{001} and n_{000}. The target population size is $N_1 = n_{++1}$.

Next, we address the presence of incomplete lists. As described in the introduction, in the subpopulations where the incomplete lists do not operate, cell counts are treated as if part of the capture history is missing. This is formalized by defining a stratifying random variable S, taking values in a finite set $\mathcal{S} = \{s_1, s_2, \ldots\}$, that partitions the observed population into different strata where different sets of incomplete lists do not operate. For example, if we have just an incomplete list, we have two strata: one where all lists operate, and one where the incomplete list does not operate. Note that we assume a perfect knowledge of the target population of each list; that is, we can distinguish without uncertainty whether a unit is not captured in an incomplete list by chance or because it is out of the scope of that list.

In Table 3 we continue the example of Table 2 by introducing the stratifying variables S. The two strata are s_1, where both lists operate, and s_2, where the incomplete list Y_2 does not operate. The complete contingency table is illustrated in Table 3, where the generic cell is $n_{sy_1y_2x}^{S,Y_1,Y_2,X}$ for $y_1, y_2, x \in \{0,1\}^3, s \in \{s_1, s_2\}$, for instance n_{s_1101} represents the number of in-scope units captured only in list Y_1 in strata s_1. We remark that an asterisk ($*$) in the subscript of n indicates that the corresponding list does not operate in that stratum. For example, $n_{s_21^*+}$ denotes the number of units captured in the first list in stratum s_2 where the second list does not operate. The observed counts are the ones shadowed in the table and their sum is n_{obs}. In s_2 we observe only $n_{s_21^*+} = n_{s_211+} + n_{s_210+}$.

Table 2. Example of the contingency table defined by two lists Y_1, Y_2 and a latent variable X.

		X		
Y_1	Y_2	In scope	Out of scope	Total
Included	Included	n_{111}	n_{110}	n_{11+}
	Not included	n_{101}	n_{100}	n_{10+}
Not included	Included	n_{011}	n_{010}	n_{01+}
	Not included	n_{001}	n_{000}	n_{00+}
Total		n_{++1}	n_{++0}	N

Table 3. Example of the contingency table defined by two lists Y_1, Y_2, the latent variable X, and the stratifying variable S.

			X			
S	Y_1	Y_2	In scope	Out of scope	Total	
s_1	Included	Included	n_{s1111}	n_{s1110}	n_{s111+}	
		Not included	n_{s1101}	n_{s1100}	n_{s110+}	
	Not included	Included	n_{s1011}	n_{s1010}	n_{s101+}	
		Not included	n_{s1001}	n_{s1000}	n_{s100+}	
s_2	Included	Included	n_{s2111}	n_{s2110}	n_{s211+}	
		Not included	n_{s2101}	n_{s2100}	n_{s210+}	$n_{s_2 1^*+}$
	Not included	Included	n_{s2011}	n_{s2010}	n_{s201+}	
		Not included	n_{s2001}	n_{s2000}	n_{s200+}	
Total			n_{+++1}	n_{+++0}	N	

In the presence of incomplete lists, there will be more than one structural zero cell to estimate. The number of structural zero cells varies in each stratum depending on the number of incomplete lists. More formally, let $S = s$ indicates the stratum where $Y_1, \ldots, Y_k$ do not operate. Then, for that stratum we have the following 2^k structural zero cells:

$$\left\{ n^{S,Y_1,\ldots,Y_k,Y_{k+1},\ldots,Y_f}_{sy_1\ldots y_k 0\ldots 0} \right\}_{(y_1,\ldots,y_k)\in\{0,1\}^k}. \tag{2}$$

Taking into account also the latent variable X, the number of structural zero cells for each stratum is 2^{k+1}. In the example of Table 3, we have two structural zero cells in stratum s_1 (n_{s_1001} and n_{s_1000}), and four in stratum s_2 (n_{s_2011}, n_{s_2010}, n_{s_2001} and n_{s_2000}).

By restating the problem in the general frame of inference with missing data, we can easily handle both incomplete lists and the latent variable X. We assume the existence of a complete contingency table $T^* = \left[n_{sy_1\ldots y_f x}\right]$ of which we observe the marginal counts T, and we want to estimate

$$N_1 = \sum_{\substack{y_1,\ldots,y_f\in\{0,1\}^f \\ s\in S}} n_{sy_1\ldots y_f 1} = n_{+\ldots+1}.$$

In this setting, it is not difficult to jointly estimate the cells affected by missing data (excluding structural zero cells) and the missing dimension X, conditional on T. For this, we define a loglinear model for T^* with parameters $\{\lambda\}$, and use the EM algorithm (Dempster et al. 1977) iterating over the following two steps:

Algorithm 1

E-step: compute the expected counts of cells affected by missing data in T^* conditionally on the observed marginal T and the current estimate of $\{\lambda\}$. Note that the structural zero cells are not considered in this step;

M-step: update the MLE of the parameters $\{\lambda\}$ of the loglinear model over the frequencies in the current estimate of T^* computed at the E-step.

In the M-step, we used the Iterative Proportional Fitting (IPF) algorithm to obtain the MLE of the loglinear parameters (Fienberg 1970).

Once the EM algorithm converges, the current MLE of $\{\lambda\}$ are used to estimate each cell of T^* including the structural zero cells and, consequently, N_1. The structural zero cells of interest in this context are the unobservable cells for which $X = 1$, that is, we are interested in:

$$\left\{ n^{S,Y_1,\ldots,Y_k,Y_{k+1},\ldots,Y_f,X}_{sy_1\ldots y_k 0\ldots 01} \right\}_{(y_1,\ldots,y_k)\in\{0,1\}^k}. \tag{3}$$

We remark that one cannot easily treat the estimates of the partially missing lists and of the latent variable X separately, at least not in the context of loglinear models. In fact, one could be tempted to estimate the complete table $T' = \left[n_{sy_1\ldots y_f}\right]$, and then the complete table $T^* = \left[n_{sy_1\ldots y_f x}\right]$ conditionally on T', by using two loglinear models. However, the model for T' would not be a submodel of the one for T^*. In fact, loglinear models are not "reproducible" or "collapsible", that is, if (X, Y, Z) have joint distribution described by the loglinear model with parameters $\{\lambda\}$, the joint distribution of (Y, Z) would not be readily derivable from $\{\lambda\}$. Even null interaction parameters can have non-zero values in the marginal model. So, two independent models should be selected and estimated, resulting in an unpractical procedure.

As an alternative to the EM algorithm implementation presented above, it is possible to adopt two nested EM algorithms, where the outer one initializes and updates the structural zero cells (2), while the inner one updates T^* including cells (3). This second approach would maximize the unconditional likelihood, while Algorithm 1 is based on the maximization of the conditional likelihood (see Fienberg 1972). However, we opted for the conditional likelihood approach since it is computationally much easier.

3. Simulation Study

In the simulation presented here, we use four lists, the minimum number for any Local Dependence Model to be identifiable. For the sake of a simpler notation, we denote the four lists as A, B, C, and D. The probability distribution of our model will be denoted as in the classic notation of LCM:

$$Pr(A = a, B = b, C = c, D = d, X = x) = \pi^{ABCDX}_{abcdx}$$

with $a, b, c, d, x \in \{0, 1\}^5$. The superscript of π will be omitted where the reference to the r.v.s is clear. The conditional probabilities $Pr(A = a \mid X = x)$ will be denoted as $\pi^{A|X}_{a|x}$. Note that the probability $\pi^{A|X}_{1|0}$ represents the overcoverage rate of list A, while $\pi^{A|X}_{0|1}$ represents its undercoverage.

We test our model in four different scenarios. In all scenarios, we use the following values and coverage rates: $N = 10^6$, $\pi^X_0 = 0.4$, $\pi^X_1 = 0.6$, and

$$\begin{aligned} &\pi_{1|0}^{A|X} = 0.25, \quad \pi_{1|0}^{B|X} = 0.2, \quad \pi_{1|0}^{C|X} = 0.21, \quad \pi_{1|0}^{D|X} = 0.29, \\ &\pi_{0|1}^{A|X} = 0.3, \quad \pi_{0|1}^{B|X} = 0.18, \quad \pi_{0|1}^{C|X} = 0.14, \quad \pi_{0|1}^{D|X} = 0.17 \end{aligned} \tag{4}$$

The four scenarios, presented in order of complexity are:

Scenario 1. The generating model is a simple LCM with an additional interaction parameter between C and D. C and D have a correlation of about 0.72 both under $X = 1$ and $X = 0$.

Scenario 2. We enhance Scenario 1 by adding a parameter of interaction between A and B, leaving parameters (4) substantially unchanged. A and B have a correlation of about 0.6 both under $X = 1$ and $X = 0$.

Scenario 3. List A is now incomplete. We add S indicating the subpopulation where all lists are available ($S = s_1$), and the subpopulation for which list A does not operate ($S = s_2$). S is independent of all other variables, and $\pi_{s_1}^S = \pi_{s_2}^S = 0.5$.

Scenario 4. We add a parameter of interaction between S and D indicating a different capture probability for D in the two subpopulations. S and D have a correlation of 3%.

All lists are complete in Scenarios 1 and 2, whereas in the remaining two Scenarios there is a single incomplete list A which operates just over half of the population. In particular, in Scenario 3 the missing mechanism can be considered missing completely at random (MCAR), as S does not interact with other variables, whereas the missing mechanism is MAR in Scenario 4.

For each scenario, we specified the generating model for the complete contingency table $T^* = [n_{sabcdx}]$ with fixed probabilities $Pr\{N_{sabcdx} = n_{sabcdx}\} = \pi_{sabcdx}$ and generated 200 independent realizations (samples) of T^*. For each sample, we registered the generated ("true") values of N_1 (the target population size), and of $n_{00001}^{ABCDX} = \sum_s n_{s00001}^{SABCDX}$, (the undercoverage in the target population), then derived the marginal "observed" counts T on which we fitted various models.

Table 4 describes the statistical properties of the estimated values of $\hat{N}_1$ and $\hat{n}_{00001}$ for each studied model by scenario. The bias and root mean squared error (RMSE) for each estimate are computed with respect to the corresponding true population values. So, let $N_1(i)$ and $n_{00001}(i)$ be the "true values" generated in the $i-$th sample, while $\hat{N}_1(i)$ and $\hat{n}_{00001}(i)$ are the resulting estimates. Then, bias and MSE reported in Table 4 respectively are:

$$Bias(\hat{N}_1) = \sum_{i=1}^{200} \frac{\hat{N}_1(i) - N_1(i)}{200}, \quad Bias(\hat{n}_{00001}) = \sum_{i=1}^{200} \frac{\hat{n}_{00001}(i) - n_{00001}(i)}{200},$$

$$MSE(\hat{N}_1) = \sum_{i=1}^{200} \frac{(\hat{N}_1(i) - N_1(i))^2}{200}, \quad MSE(\hat{n}_{00001}) = \sum_{i=1}^{200} \frac{(\hat{n}_{00001}(i) - n_{00001}(i))^2}{200}.$$

Relative bias (RB) and relative RMSE (RRMSE) are computed with respect to bias or error respectively of estimates specified with the generating model (indicated with an

Table 4. Results of the simulations. The asterisks indicate the cases where the estimating model coincides with the generating model, with respect to which the RB and RRMSE are calculated. [LCM] stands for the parameters of the Latent Class Model (5).

Generating Model	Estimating Model	N_1				n_{00001}^{ABCDX}			
		Bias	RB	RMSE	RRMSE	Bias	RB	RMSE	RRMSE
1) [LCM]+[CD]	[LCM]+[AB][CD]	370	1	2,683	1.5	80	1.04	411	3.2
	*[LCM]+[CD]	−361	1	1,771	1	−77	1	127	1
	[LCM]	−77,292	214.1	77,294	43.6	−5,865	76.17	5,866	45.9
2) [LCM]+[AB][CD]	*[LCM]+[AB][CD]	469	1	11,061	1	−66	1	1,447	1
	[LCM]+[CD]	−44,541	95	44,553	4	−17,664	267.6	17,665	12.2
	[LCM]	−38,074	81.2	38,092	3.4	−17.577	266.3	17,578	12.1
3) [LCM]+[AB][CD][S]	*[LCM]+[AB][CD][S]	929	1	12,437	1	112	1	1,543	1
	[LCM]+[CD][S]	−45,069	48.5	45,085	3.6	−19,356	172.8	19,357	12.6
	[LCM]+[S]	−58,455	63	60,378	4.9	−20,327	181.5	20,331	13.2
4) [LCM]+[AB][CD][SD]	*[LCM]+[AB][CD][SD]	5,784	1	9,157	1	336	1	876	1
	[LCM]+[AB][CD][S]	−9,584	1.7	12,503	1.4	−890	2.6	1,190	1.4
	[LCM]+[S]	−76,112	13.2	76,083	8.3	−12,243	36.4	12,250	14

asterisk), so that

$$RB(\hat{N}_1) = \frac{Bias(\hat{N}_1)}{Bias\left(\hat{N}_1^*\right)}, \quad RB(\hat{n}_{00001}) = \frac{Bias(\hat{n}_{00001})}{Bias\left(\hat{n}_{00001}^*\right)},$$

$$RRMSE(\hat{N}_1) = \frac{RMSE(\hat{N}_1)}{RMSE\left(\hat{N}_1^*\right)}, \quad RRMSE(\hat{n}_{00001}) = \frac{RMSE(\hat{n}_{00001})}{RMSE\left(\hat{n}_{00001}^*\right)}.$$

Thus, when the estimating and generating model coincide, both RB and RRMSE equal unity.

Note that a simple LCM in this case is equivalent to the loglinear model

$$[AX][BX][CX][DX]. \tag{5}$$

The results shown in Table 4 indicate that our estimation strategy works well both in terms of bias and variance even in presence of various interaction parameters and incomplete lists when the estimating model is the same as the model generating the simulations. On the other hand, whenever the estimating model does not coincide with the generating model, the estimates can be very biased. In particular, models with missing interaction parameters severely underestimate both N_1 and the undercount. Note that an overparameterized model (see Scenario 1) leads to a less severe deviation from the true values than the underparameterized models.

Turning to model selection, in Scenarios 1 and 4, the values of the AIC and BIC favor the correct model in all 200 samples. However, in Scenarios 2 and 3, the AIC and BIC criteria occasionally favor the second model. In detail, the second estimating model has a lower AIC in about 25% of samples and a lower BIC in 30% in Scenario 2; the percentages rise to 40% and 50% in Scenario 3. The simple LCM is never preferred in any Scenario.

To illustrate the estimation procedure, we refer to Generating Model 3 in Table 4:

$$\begin{aligned} \lg n_{sabcdx} &= \lg(N\pi_{sabcdx}) = \\ &= \lambda + \lambda_x + \lambda_a + \lambda_b + \lambda_c + \lambda_d + \lambda_s + \lambda_{ax} + \lambda_{bx} + \lambda_{cx} + \lambda_{dx} + \lambda_{ab} + \lambda_{cd} \end{aligned} \tag{6}$$

Here, A and B are independent from C and D conditionally on X and the model can be defined in terms of conditional probabilities by the following equations:

$$\pi_{sabcdx} = \pi_x \pi_{ab|x} \pi_{cd|x} \pi_s$$

$$\pi_{sabcd} = \sum_{x \in \{0,1\}} \pi_x \pi_{ab|x} \pi_{cd|x} \pi_s$$

where $\pi_{ab|x}$ and $\pi_{cd|x}$ are restricted by means of the absence of the second order loglinear interaction parameters $[ABX]$ and $[CDX]$. The observed marginal counts T correspond to $[n_{s_1abcd}] \cup [n_{s_2bcd}]$. The log-likelihood of the observed incomplete data is:

$$\sum_{a,b,c,d} n_{s_1abcd} \log \pi_{s_1abcd} + \sum_{b,c,d} n_{s_2bcd} \log \pi_{s_2bcd}$$

while Algorithm 1 is specified in the following way:

1. initialize at random an estimate of the posterior probabilities $\left\{\hat{\pi}_{x|s_1abcd}\right\}$ and $\left\{\hat{\pi}_{xa|s_2bcd}\right\}$;
2. estimate the complete contingency table $\hat{T}^* = [n_{sabcdx}]$ excluding the structural zero cells by computing

$$\begin{aligned} \hat{n}_{s_1abcdx} &= n_{s_1abcd}\hat{\pi}_{x|s_1abcd}, \quad \forall \text{ cells s.t. } (a,b,c,d) \neq (0,0,0,0) \\ \hat{n}_{s_2abcdx} &= n_{s_2bcd}\hat{\pi}_{xa|s_2bcd} \quad \forall \text{ cells s.t. } (b,c,d) \neq (0,0,0); \end{aligned}$$

3. estimate loglinear model (6) on $\hat{T}^*$ via IPF conditionally on the unobservable structural zero cells;
4. update the current value of the observed log-likelihood and of the posterior probabilities estimate;
5. repeat 2–4 until convergence.

After convergence, we have to estimate the structural zero cells $n^{SABCDX}_{s_100001}$, $n^{SABCDX}_{s_200001}$, and $n^{SABCDX}_{s_210001}$.

4. An Application to Business Statistics

In this section, we present an application of our method to estimate the number of active local units of Italian enterprises of large dimensions (50 or more employees) in 2011 using solely administrative data. We compare our model results to the official counts obtained by a yearly complete enumeration survey on local units of Italian enterprises (IULGI) conducted by Istat (Consalvi et al. 2008).

We have five administrative sources containing information on the local units of Italian enterprises that can be used to model the annual measures, namely:

A. The register of enterprise and local units owned by the Italian Chamber of Commerce, containing compulsory declarations to be submitted by anyone who wants to open a new local unit.
B. The Yellow Pages owned by SEAT, supplying all the business addresses that have at least one telephone line.
C. Territorial Insurance Position owned by the agency for the insurance against work-related injuries, consisting of information on the number of employees with insurance against accidents per local unit.
D. The Bank register owned by Bank of Italy, holding the addresses of all the bank tellers.
E. The Big Distribution Division Register, owned by NIELSEN, containing the addresses and the employees of the local units of enterprises operating in the sector of Big distribution.

In addition to these sources, we have at our disposal the results of the field survey IULGI conducted to verify the presence of the local units and to obtain selected information on their characteristics. In the 2011 edition, the survey included all the approximately 30,000 enterprises with more than 50 employees. Since we are focusing on those enterprises, IULGI constitutes a total enumeration survey. As a consequence, we use it to evaluate the single sources quality, and the model results.

Table 5 presents the results of a coverage analysis of each single administrative source with respect to the survey. These results are indicative of high coverage error in all

Table 5. Coverage of the administrative sources with respect to the survey.

List		IULGI	
		Not observed	Observed
A	0	–	38,109
	1	104,570	158,478
B	0	–	145,040
	1	6,581	51,547
C	0	–	108,394
	1	33,228	88,193
D	0	–	2,811
	1	1,190	4,754
E	0	–	3,142
	1	3,352	28,325

sources. In particular, source A has high overcoverage level, whereas sources B, C, and D are mainly affected by undercoverage errors. Note that the values of sources D and E in Table 5 are constrained to the Bank sector and to the Big distribution sector respectively.

Table 6 presents the observed counts of the capture profiles of the administrative lists. The capture profiles are classified according to the operating strata (S), and the asterisks ($*$) indicate where the lists do not operate. In the application, only three strata are observed since the lists D and E target on not overlapping populations, for instance in s_1 only the lists (A, B, C) operate. The total amount of observed units (n_{obs}) is 298,253.

The data at hand fit perfectly the range of problems addressed in this article. In fact, all five sources have both under- and overcoverage, lists D and E are incomplete lists of defined subpopulations, and the capture probabilities of an individual unit are not independent among the lists. Moreover, there is a credible benchmark given by the

Table 6. Frequencies (n_{ABCDE}) of the capture profiles of the administrative lists by operating strata S.

s_1		s_2		s_3	
$n_{001^{**}}$	24,865	$n_{0001^{*}}$	1,058	$n_{000^{*}1}$	662
$n_{010^{**}}$	2,721	$n_{0010^{*}}$	2,188	$n_{001^{*}0}$	596
$n_{011^{**}}$	2,148	$n_{0011^{*}}$	493	$n_{001^{*}1}$	166
$n_{100^{**}}$	125,459	$n_{0100^{*}}$	13	$n_{010^{*}0}$	58
$n_{101^{**}}$	40,884	$n_{0101^{*}}$	23	$n_{010^{*}1}$	62
$n_{110^{**}}$	20,200	$n_{0110^{*}}$	17	$n_{011^{*}0}$	60
$n_{111^{**}}$	29,414	$n_{0111^{*}}$	16	$n_{011^{*}1}$	59
–	–	$n_{1000^{*}}$	5,338	$n_{100^{*}0}$	3,349
–	–	$n_{1001^{*}}$	15,220	$n_{100^{*}1}$	1,491
–	–	$n_{1010^{*}}$	1,575	$n_{101^{*}0}$	767
–	–	$n_{1100^{*}}$	29	$n_{110^{*}0}$	455
–	–	$n_{1011^{*}}$	13,984	$n_{101^{*}1}$	2,030
–	–	$n_{1101^{*}}$	250	$n_{110^{*}1}$	444
–	–	$n_{1110^{*}}$	22	$n_{111^{*}0}$	474
–	–	$n_{1111^{*}}$	633	$n_{111^{*}1}$	1,030

complete enumeration survey, which allows us to compare and evaluate the results of the proposed strategy.

The first step is the identification of the model that better fits data. We restricted our choice of candidate loglinear models to the class of hierarchical model, starting with the simple LCM and adding interaction terms in an increasing order of complexity. An exhaustive search is obviously not possible because of the huge number of potential models. Out of the approximately one hundred models examined, the following models performed the best in terms of AIC and BIC:

(M1) LCM + [S][AB][AC][AE][BC][ABX][ACX][BCX]

(M2) LCM + [S][AC][AE][BC][BE][CD][ACX][BCX]

(M3) M1 + [BE]

where LCM indicates the latent class model [AX][BX][CX][DX][EX].

The estimates of undercount, overcount and in-scope population size N_1 obtained with models M1, M2, M3, and LCM are presented in Table 7. For each model, the goodness-of-fit measures (log likelihood, AIC, and BIC) are provided. For evaluation purposes, we have also included the official values given by the survey IULGI (first column of Table 7). Undercount and overcount are calculated comparing the local units captured by at least one administrative list and IULGI's list. So, a local unit captured in a list which is not in IULGI is overcount, while a local unit captured in IULGI which is not in any administrative list is undercount. Note that the estimate of N_1, obtained by summing the estimate of all cells of T^* having $X = 1$, equals n_{obs} (298,253 in this case) minus the estimate of the overcoverage plus the estimate of the undercoverage.

The simple LCM has a poor fit and the estimated counts are distant from the target values. The models with the lowest AIC and BIC are M3 and M2 respectively. These two models yield the estimates of N_1 closest to the official values. However, each model yields very different estimates of the under- and overcoverage. In particular, Model M2 entirely misses the undercount, but the errors in the under- and overcounts offset each other, resulting in a good estimate of N_1. In a real situation, we would discard Model M2 as implausible and choose Model M3. These results illustrate the sensitivity of the proposed models to differences in the parameterization. The results indicate that the main challenge in successfully applying the proposed strategy is finding a model that adequately estimates both under- and overcoverage in the studied data. In general, model selection is a critical

Table 7. Application result.

	IULGI	LCM	M1	M2	M3
Loglik		−453,024	−447,958	−447,923	−447,920
AIC		906,074	895,956	895,887	895,883
BIC		906,212	896,168	896,099	896,106
Under	28,519	6,701	20,996	0	20,139
Over	130,185	81,305	116,458	104,905	117,790
N_1	196,587	223,648	202,791	193,348	200,602

point to capture-recapture modeling as population size estimate can be sensitive to changes in the parameterization.

5. Conclusions

The article focuses on the estimation of a population size by using multisource data. The availability and usage of more information is certainly important, providing new opportunities for more timely or more detailed estimates. However, it also presents new methodological challenges. In this study, we estimate the population size by using multiple-record system methodologies, considering each source as a capturing list of the units of interest and capture-recapture techniques. The usual assumptions, such as independence of the captures of a unit in different lists ("Causal Independence") and absence of overcoverage, are not valid. Moreover, many administrative lists are "incomplete", that is, target a subset of the population of interest. We propose an estimation procedure that accounts for overcoverage, dependence of the captures in different lists, and the presence of incomplete lists. In particular, loglinear models are employed to model the dependence of the captures of a unit in different lists, overcoverage is modeled by a latent dichotomous variable that represents whether an observation belongs to the target population, and incomplete lists are addressed by means of an inferential approach developed in the context of inference with missing data.

We evaluate the proposed estimation on simulated data and on Istat business survey data. The simulation results are encouraging in that whenever the fitted model is the same as the one used for generating data, good estimates of the population size in terms of MSE are obtained. However, when the estimating model differs from the generating one, the resultant estimates may be biased. Hence, a sensible point when applying the proposed strategy is to focus on validating the selected estimating model. In our empirical evaluation, the AIC and BIC proved useful, but before implementing these models in a production setting, more research is needed in determining useful model selection procedures.

The empirical application highlighted other features of the proposed method. The algorithm provides estimates of over- and undercoverage in addition to the population estimates. The variation in levels obtained from the different models indicates that the coverage estimates may be less precise than the total population size estimates, but they are nonetheless useful. For example, they can be provided to subject matter experts to help assess the plausibility of the estimating model, given their practical knowledge or experience with the input lists. Of course, the choice of the estimating model should not only take advantage of such specialized knowledge, but should be plausible without it. For example, a model that produces viable population but, but unrealistic estimates of over- or undercoverage – such as the M2 model – is not acceptable.

Further studies will examine whether adding covariates might improve the estimating models and mitigate the impact of a wrong estimating model on the fitted estimates. In addition, we will consider a Bayesian approach that is, in general, more apt to smooth the results and to introduce prior information to help the model selection. For example, we will attempt to explicitly model the uncertainty about the model selection by utilizing model averaging techniques, placing a prior distribution over a set of possible models; see Madigan and York (1997) for an application in capture-recapture.

6. References

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

Bartolucci, F. and A. Forcina. 2001. "Analysis of Capture-Recapture Data with a Rasch Type Model Allowing for Conditional Dependence and Multidimensionality." *Biometrics* 57: 714–719. Doi: http://dx.doi.org/10.1111/j.0006-341X.2001.00714.x.

Biemer, P. 2011. *Latent Class Analysis of Survey Error*. John Wiley & Sons Inc., NY. Doi: http://dx.doi.org/10.1002/9780470891155.

Biggeri, A., E. Stanghellini, F. Merletti, and M. Marchi. 1999. "Latent Class Models for Varying Catchability and Correlation Among Sources in Capture-Recapture Estimation of the Size of a Human Population." *Statistica Applicata* 11: 563–576.

Consalvi, M., L. Costanzo, and D. Filipponi. 2008. "Evolution of Census Statistics on Enterprises in Italy 1996–2006: from the Traditional Census to a Register of Local Units." In Proceedings of the IAOS Conference on Reshaping Official Statistics, October 2008, Shanghai.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, Series B (methodological) 39: 1–38.

Fienberg, S.E. 1970. "An Iterative Procedure for Estimation in Contingency Tables." *Annals of Mathematical Statistics* 41: 907–917. Doi: http://dx.doi.org/10.1214/aoms/1177696968.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables." *Biometrika* 59: 409–439.

Hagenaars, J.A. 1988. "Latent Structure Models with Direct Effects Between Indicators Local Dependence Models." *Sociological Methods & Research* 16: 379–405. Doi: http://dx.doi.org/10.1177/0049124188016003002.

Hagenaars, J.A. 1993. *Loglinear Models with Latent Variables*. Newbury Park: CA: Sage. Doi: http://dx.doi.org/10.4135/9781412984850.

Kamen, C.S. 2005. *The 2008 Israel Integrated Census of Population and Housing Basic conception and procedure*. Central Bureau of Statistics. Available at: http://www.cbs.gov.il/mifkad/census2008e.pdf (accessed May 2017).

Madigan, D. and J.C. York. 1997. "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika* 84: 19–31. Doi: https://doi.org/10.1093/biomet/84.1.19.

Pledger, S.A. 2000. "Unified Maximum Likelihood Estimates for Closed Capture-Recapture Models Using Mixtures." *Biometrics* 56: 434–442. Doi: http://dx.doi.org/10.1111/j.0006-341X.2000.00434.x.

Stanghellini, E. and P.G. Van der Heijden. 2004. "A Multiple-Record Systems Estimation Method that Takes Observed and Unobserved Heterogeneity into Account." *Biometrics* 60: 510–516. Doi: http://dx.doi.org/10.1111/j.0006-341X.2004.00197.x.

Sutherland, J.M. 2003. *Multi-List Methods in Closed Populations with Stratified or Incomplete Information*. PhD Thesis. Simon Fraser University.

Sutherland, J.M. and C.J. Schwarz. 2005. "Multi-List Methods Using Incomplete Lists in Closed Populations." *Biometrics* 61: 134–140. Doi: http://dx.doi.org/10.1111/j.0006-341X.2005.021126.x.

Wallgren, A. and B. Wallgren. 2007. *Register-Based Statistics: Administrative Data for Statistical Purposes*. John Wiley and Sons: Chichester. Doi: http://dx.doi.org/10.1002/9780470061350.

Zhang, L.-C. 2012. "Topics of statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Zhang, L.-C. 2015. "On Modelling Register Coverage Errors." *Journal of Official Statistics* 31: 381–396. Doi: http://doi.org/10.1515/jos-2015-0023.

Zwane, E.N., K.M. van der Pal-de Bruin, and P.G. van der Heijden. 2004. "The Multiple-Record Systems Estimator when Registrations Refer to Different but Overlapping Populations." *Statistics in Medicine* 23: 2267–2281. Doi: http://dx.doi.org/10.1002/sim.1818.

Received October 2016
Revised July 2017
Accepted July 2017