Estimating Population Size from Multisource Data with Coverage and Unit Errors

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

Abstract

In recent years, the quantity and quality of administrative information available for National Statistical Institutes have been constantly increasing. In estimating a population size based on several administrative sources, the misalignment between the scope of the administrative data and that of the statistician poses several methodological challenges and sets us apart from a classical capture-recapture setting. In this paper we deal with two main aspects: misclassification of the units, leading to lists with both overcoverage and undercoverage, and lists targeting a subpopulation, which imply large sets of units having null probability of being captured. We show some results on the use of a class of capture–recapture models based on latent class analysis to address these issues.

Key Words: Multisource Integration, Capture–Recapture, Latent Class Models

1. Introduction

Traditionally, official statistics were mainly based on primary data gathered by means of survey sampling and censuses, while secondary data (namely administrative data) was used as auxiliary source of information. Nowadays, National Statistical Institutes (NSIs) are investigating the possibility of assigning to administrative data a more important role, till arriving to produce statistics solely based on them, as in register-based statistics (Wallgren and Wallgren, 2007). In this context, the main problem is that data are gathered by other organizations for their specific aims, hence, units and variables refer to specific populations and measurements that are of interest for the body collecting information, and, in general, do not fit perfectly the statistical interests. Hence, an important task is that of transforming/aligning data to fit the NSIs' research interests. The errors made in this transformation step may refer both to units and measurements (Zhang, 2011). Unit errors as such are rarely mentioned in literature of survey sampling, but their impact on the statistics is particularly important in such a context. A discussion about unit errors can be found in (Zhang, 2012), where they are described essentially as the errors made in the creation/derivation of statistical units of interest. Different types of errors fit this general definition, for instance the creation of an household based on information available in different data sources, or the erroneous classification of a unit into the target population.

In this paper we focus on population size estimation based on multiple data sources referring to different but overlapping populations. This scenario is frequently encountered in practice. In fact, in recent years, the number of available sources for NSIs has been constantly increasing, this fact on one side gave us the opportunity of using statistical methodologies that exploit the information redundancy, on the other side, this abundance is not always associated to a uniform quality of the information. In general, every additional source we are willing to include in the analysis has a lower quality, and the more lists we include, the more classical assumptions of a capture–recapture setting we could violate. In particular, the main issues (in terms of coverage) we encountered are:

- partial information, when some sources refer to a subpopulation of our target population (we will call them "incomplete sources/lists"). This problem is frequently

[1]Istat, Via Cesare Balbo, 16, Roma

encountered in practice. In fact, the large number of available sources is related to a large number of organizations collecting data, that typically target specific set of units (e.g., specific categories of workers, enterprises having certain legal form,...);

- 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 lists potentially having both under and overcoverage.

In the vast literature of capture-recapture, these two problems have been rarely addressed. The problem of incomplete lists is studied in (Sutherland 2003, Sutherland et al., 2004), and, in different terms, in (Zwane et al., 2004). They show that, in general, ignoring the incompleteness of the lists, i.e., treating the uncatchable units as sampling zeros for the incomplete lists, results in biased estimates of the population size. To address this, all cited works essentially suggest to treat 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, i.e., as if the capture history is partially missing, and an EM algorithm is presented to estimate the missing part. Note that this approach allows us to use all records even in presence of incomplete lists. This is particularly important for us since, from on side, we often deal with situations where the missing patterns are complementary and there is little/no subset where all sources operates, and, on the other side, there is need of complex models requiring several sources to be identified.

As regards to overcoverage, in the usual practice, it is preliminarily treated with clerical review of spurious events and duplicate records or with ad hoc studies to identify and remove units not belonging to the target population according to a set of deterministic rules. Afterwards, capture–recapture methodologies are utilized on the data classified as composed solely of target units. Sometimes, ad hoc surveys are deployed to estimate the overcoverage; this approach is common in census, where O-sample coverage surveys are integral part of the process. For example, in the 2008 Israeli Population Census, a 20% sample was set to provide, among other information, 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 by means of the weights, (see Kamen, 2005).

Whenever a survey is not available, the possibility of an unsupervised approach can be considered. We essentially propose to treat overcoverage as induced by misclassification, without further analysis of the source of error. Latent Class models are particularly suitable for handling misclassification errors in an unsupervised fashion.

The use of finite mixture models in capture–recapture applications to account for unobserved heterogeneity in capture probabilities is well known, in particular the use of Latent Class models dates back to (Agresti, 1994). Several extensions have been proposed to include covariates, to model individual heterogeneity, and to relax the local independence assumption of the Latent Class models (see, e.g., Bartolucci and Forcina, 2001). This last aspect is particularly interesting for us, as in our applications dealing with administrative data, the hypothesis of conditional independence of the captures is hardly tenable (just consider that registration on a list can be mandatory preliminary on others). For these reasons, we opted for a generalization of the Latent Class models that include 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, Stanghellini et al., 2004).

That being said, the use of a latent variable to directly model overcoverage has been largely ignored. To our knowledge, the sole contribution in this sense is in (Biemer 2011, chap. 6.3) who proposes the use of these models in the situation of a Triple System Estimate, for 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 paper we present some results on the use of these models to jointly estimate under and overcoverage in presence of incomplete lists.

2. Some motivating cases in Business statistics

A first example in business statistics is the case of agricultural holdings. In this context, administrative sources containing information concerning agriculture in Italy are the Integrated Administration and Control System (AGEA), the System for the Identification and Registration of Bovine Animals and other species (AA.ZZ.), lands property Incomes (Tax Agency), the Land registry, Agricultural self-employed workers (Social Security), regional agricultural Systems. There are moreover further general administrative sources that can be used for agricultural holdings: the Chambers of Commerce, the VAT declarations (Tax Agency). These sources were used to create in 2010 the list assisting the agricultural census, but they provide information that could be used to give a population size estimation in a capture-recapture approach. Nevertheless, problems concerning the classification of target units and population partially overlapping arise. As clearly illustrated in Viviano 2009,

- *"the agricultural sector is characterized by small and very small productive units, strongly aided for the realization of a minimum income.*

  Labour force is mainly based on family labour, often seasonal and part-time. Moreover the sector is strictly integrated with other activities such as

  transformation, trade, tourism, etc. These items make complex the correct identification of units as well as the estimation of their actual size and

  their principal activity;

- *At national level as defined by the Census of Agriculture and according to the definition of the FSS regulation (article2), the statistical*

  definition of Agricultural holding means a single unit, both technically and economically, constituted by lands, even not contiguous, and by

  machineries managed by a holder i.e. natural or juridical person or institution which undertakes agricultural, forestall and zootechnics activities. On

  the other side each administrative body has its own function to collect data and manage the corresponding records, under specific legislation governing

  relations between them and individuals and between them and the public administration. Thus, each source uses peculiar definitions and classifications

  that need to be translated according to a statistical framework before their usage. The main difficulty is to identify easily the statistical unit of

  reference starting from the units recorded into administrative files;

Although the administrative data sources overlap each other, some of them refer to specific population, for instance the observations in the Chamber of commerce refer only to units carrying out an agricultural economic activity.

A second motivating example is given by a project of the Italian National Statistical Institute (Istat) that aims to improve the timeliness of the Italian Business Register (BR), providing information of the structure of the business population at six months distance from the reference year. The need to get a timely representation of the BR derives from the important role conferred to the BR as a main source for the statistical analysis of the economic system, especially concerning the territorial aspects of the economic dynamics. This means that the BR is not only a picture of the business population to be used as a sampling frame for the economic surveys and for grossing up the sampling results, but it is also considered the official source for statistical information on the enterprises structure and demography.

Up to now, the set-up process of the Register for the reference year *t* started in the last quarter of the year *t+1*, when the yearly data supplied from the main sources are available. Then, after the process of standardization and integration of the administrative units and the estimation of the main structural variables, the BR is published with fifteen months of delay. The use of longitudinal information may improve the timeliness of the BR , i.e., all the administrative sources with reference year *t-1*.

The administrative sources available within the first quarter of the year *t+1* and with reference year *t* and reference year *t-1* are the followings:

- the *Tax Register*, owned by the Ministry of Economy and Finances, which records all natural and legal persons operating over the national

  territory, who are required to comply with fiscal legislation;

- the *Register of Enterprises and Local Units*, owned by the Chambers of Commerce, gathering compulsory declarations to be submitted by

  anyone who wants to start a new enterprise (excluding the self-employers);

- the *Social Security data* managed by the Social Security Authority, which record the enterprises with employees as well as the sole

  traders, subject to the payment of social security contributions.

All these sources are continuously updated, as new submissions continuously arrive in a flow. This involves the addition of new units and the updating of the values of existing records. Istat does not have a continuous access to the data, indeed it acquires the entire datasets in two distinct occasions during the year. The continuous updating of the administrative archives may cause misalignments in the reference population when the archive is downloaded in different periods of the same reference year. This means that using an earlier supply of the administrative data with respect to the timing regularly adopted for the realization of the BR, might lead to problems of completeness, both in terms of undercoverage and accuracy of the information. The administrative information available only with reference year *t-1* are the presence of an yearly turnover in the Tax Register and the payment of the annual tax for the Chamber of Commerce, all these information are essential for the identification of active enterprises.

Although the available administrative information represent overlapping lists of the same target population, to correctly identify the target statistical units (i.e., the active enterprise), we need to concurrently deal with problems of erroneous units enumeration and of partially overlapping list, since some of the lists cover only one year.

3. Model and estimation

First of all let us formalize a simple capture–recapture setting based on loglinear models: Suppose we have f lists, and let $Y_1, \ldots, Y_f$ be the binary variables associated to each list such that $Y_i = 1$ when a given unit is observed in the i–th 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 $T = \left[n_{y_1,\ldots,y_f}\right]$, where each cell represents a pattern of inclusion in the lists (also known as capture history). Cell $n_{0,\ldots,0}$, 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 $N = n_{obs} + n_{0,\ldots,0}$. The use of loglinear models is typical in situations where each list has a different capture probability and we want to explicitly model the dependencies between the r.v.s $Y_1, \ldots, Y_f$ when captures of the same unit are not independent among the lists. The estimate of the unobserved count $n_{0,\ldots,0}$ (and, in general, of any structural zero cell) is simply obtained by the ML estimates of the loglinear model conditionally on the observed data (Fienberg, 1972).

In our case, units observed in the lists do not all refer to the target population $\mathcal{P}$. In order to model this, we add to our loglinear model a dichotomous latent variable X identifying the in–scope and the out–of–scope units, that is

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

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

Problems of identifiability may arise in this context: as the number of parameters should not be less than the number of degrees of freedom of the contingency table, we necessitate at least four lists for any Local Dependence model to be identifiable, unless ad hoc constraints on the parameters are introduced (see for instance (Biemer, 2011)). In the motivating cases, it is shown that there is in practice the possibility of having such a high number of lists.

As regards to the presence of incomplete lists, first of all we assume a perfect knowledge of which units cannot be captured by the incomplete lists. This is formalized by a stratifying r.v. S, $S = \{s_1, s_2, \ldots\}$ indicating the subpopulations of units (or strata) where different sets of incomplete lists do not operate. In the subpopulations where the incomplete lists do not operate, cell counts are treated as if part of the capture history is missing. Note that in this setting we have more than one structural zero cell to estimate. Their number 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:

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

We can easily handle both incomplete lists and the latent variable X if we rephrase the problem in the general frame of inference of missing data. We assume the existence of a complete contingency table $T^* = \left[n_{x,y_1,\ldots,y_f,s}\right]$, of which we observe the marginal counts T, and we want to estimate

$$N_1 = \sum_{y_1,\ldots,y_f,s} n_{1,y_1,\ldots,y_f,s}$$

The joint estimate of the cells affected by missing data (excluding structural zero cells) and the missing dimension X conditionally on T presents no difficulties. We fix a loglinear

model for T^* having parameters $\{\lambda\}$, and then use a simple EM algorithm 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.

Once the algorithm converges, the current estimate of $\{\lambda\}$ are used to estimate each cell of T^*, the structural zero cells and, consequently, N_1. We notice that the structural zero cells of interest in this context are the unobservable cells for which $X = 1$. That is, in the stratum defined in (1), those would be:

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

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, loglinear models are not "reproducible" or "collapsible", i.e., 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 different loglinear models (one for $(Y_1, \ldots, Y_f)$ and one for $(X, Y_1, \ldots, Y_f)$) should be selected and utilized in two distinct EM algorithms, and that would result in an unusable procedure.

A second remark on the algorithm we want to point out is that, as an alternative to Algorithm 1, it is possible to adopt two nested EM algorithms where the outer one initializes and updates the structural zero cells (1) while the inner one updates T^* including cells (2) with passages identical to Algorithm 1. This second approach would refer to the maximization of the complete likelihood, while Algorithm 1 is based on the maximization of the conditional likelihood, see (Fienberg ,1972). However, we opted for the conditional likelihood approach of Algorithm 1 since it is more natural in the context of loglinear models, and is computationally much easier.

4. Simulation Study

As previously discussed, the minimum number of lists for any Local Dependence model to be identifiable is four. It may be the case that in some applications an even higher number of lists is available, however, we stick to this minimum number in all our simulations.

In this Section, for the sake of a simpler notation, we will name the four lists as A, B, C and D. We will use the classic notation of Latent Class models for which:

$$Pr(X = x, A = a, B = b, C = c, D = d, S = s) = \pi^{XABCDS}_{xabcds}$$

The superscript of π will be omitted where the reference to the r.v.s is clear. 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 set four scenarios. For each scenario we fixed a generating model for the complete contingency table $T^* = [n_{x,a,b,c,d,s}]$ with fixed probabilities $Pr\{N_{xabcds} = n_{xabcds}\} = \pi_{xabcds}$. For each scenario, we generated 100 realizations of T^*. For each sample we registered the generated ("true") values of N_1 (the target population size), and of $n_{10000}^{XABCD} = \sum_s n_{10000s}^{XABCDS}$, (the undercoverage in the target population), and derived the marginal "observed" contingency table T on which we fitted various models.

In all simulations we set as fixed the following values: $N = 10^6$, $\pi_0^X = 0.4$, $\pi_1^X = 0.6$, and the following coverage rates:

$$\begin{array}{llll} \pi_{1|0}^{A|X} = 0.25, & \pi_{1|0}^{B|X} = 0.2, & \pi_{1|0}^{C|X} = 0.21, & \pi_{1|0}^{D|X} = 0.29, \\ \pi_{0|1}^{A|X} = 0.3, & \pi_{0|1}^{B|X} = 0.18, & \pi_{0|1}^{C|X} = 0.14, & \pi_{1|0}^{D|X} = 0.17 \end{array} \tag{3}$$

Table 1 shows the results of the estimate of N_1 and of n_{10000} in terms of bias, Root Mean Square Error (RMSE), and relative RMSE wrt the case where the estimating model coincides with the generating model (estimates $\hat{N}_1^*$ and $\hat{n}_{10000}^*$), i.e.,

$$\frac{\sqrt{\sum \left(\hat{N}_1 - N_1\right)^2}}{\sqrt{\sum \left(\hat{N}_1^* - N_1\right)^2}} \qquad \text{and} \qquad \frac{\sqrt{\sum \left(\hat{n}_{10000} - n_{10000}\right)^2}}{\sqrt{\sum \left(\hat{n}_{10000}^* - n_{10000}\right)^2}}.$$

Note that a simple Latent Class Model which involves local independence can be written as a loglinear model that, in this case, takes the following form:

$$[AX][BX][CX][DX]$$

where we used the conventional notation reporting only the higher order interactions.

The following are the four scenarios in increasing order of complexity:

Scenario 1.) The generating model has a single additional interaction parameter between C and D wrt a simple Latent Class model. C and D have a correlation of 0.72 both under $X = 1$ and $X = 0$ while leaving parameters (3) unchanged.

Scenario 2.) We add a parameter of interaction between A and B: they 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.

In scenarios 1) and 2) we do not have incomplete lists, while in the remaining two we have a single incomplete list A which operates just over half of the population. In particular, in scenario 3) the missing mechanism can be considered MCAR, as r.v. S does not interact with other variables, while in Scenario 4) the missing mechanism is MAR.

As the results shown in Table 1 indicate, our estimation strategy works well both in terms of bias and variance even in presence of various interaction parameters and incomplete lists. We note however that, if the estimating model does not coincide with the generating model, the estimates can deviate significantly from the true value. This is also true when when the assumed estimating model is overparameterized (see first case in Generating model 1)). However, in almost all simulations the values of the AIC and BIC have been in favour of the correct model.

To give an insight of the passages involved in the estimation, we refer to Generating Model 3) in Table 1:

$$n_{xabcds} = N\pi_{xabcd} = \lambda_x + \lambda_a + \lambda_b + \lambda_c + \lambda_d + \lambda_s + \lambda_{ax} + \lambda_{bx} + \lambda_{cx} + \lambda_{dx} + \lambda_{ab} + \lambda_{cd} \quad (4)$$

Here, A and B are independent from C and D conditionally on X and the model can be equivalently defined by the following equations:

$$\begin{aligned} \pi_{xabcds} &= \pi_x \pi_{ab|x} \pi_{cd|x} \pi_s \\ \pi_{abcds} &= \sum_{x \in \{0,1\}} \pi_x \pi_{ab|x} \pi_{cd|x} \pi_s \end{aligned}$$

where $\pi_{ab|x}$ and $\pi_{cd|x}$ are restricted by means of the absence of second order (three variable) loglinear interaction. The observed contingency table is $T = [n_{abcds_1}] \cup [n_{bcds_2}]$. The log–likelihood of the observed incomplete data is:

$$\sum_{a,b,c,d} n_{abcds_1} \log \pi_{abcds_1} + \sum_{b,c,d} n_{bcds_2} \log \pi_{bcds_2}$$

while Algorithm 1 is specified in the following way:

---

1. initialize at random an estimate of the posterior probabilities $\{\hat{\pi}_{x|abcds_1}\}$ and $\{\hat{\pi}_{xa|bcds_2}\}$;
2. estimate the complete contingency table $\hat{T}^* = [n_{xabcds}]$ excluding the structural zero cells by computing

$$\begin{aligned} \hat{n}_{xabcds_1} &= n_{abcds_1} \hat{\pi}_{x|abcds_1}, && \forall \text{ cells s.t. } (a,b,c,d) \neq (0,0,0,0) \\ \hat{n}_{xabcds_2} &= n_{bcds_2} \hat{\pi}_{xa|bcds_2} && \forall \text{ cells s.t. } (b,c,d) \neq (0,0,0); \end{aligned}$$

3. estimate loglinear model (4) 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 that in this case are: $n^{XABCDS}_{10000s_1}$, $n^{XABCDS}_{10000s_2}$, and $n^{XABCDS}_{11000s_2}$.

Generating Model		Estimating Model	N_1 Bias	N_1 RMSE	N_1 RRMSE	$n_{0000\|1}$ Bias	$n_{0000\|1}$ RMSE	$n_{0000\|1}$ RRMSE
1)	[AX][BX][CX][DX][CD]	[AX][BX][CX][DX][AB][CD]	-1,321	13,021	3.2	-509	3,447	12.1
		*[AX][BX][CX][DX][CD]	-365	4,084	1	-66	285	1
		[AX][BX][CX][DX]	-77,356	172,978	42.4	-5,845	13,072	45.9
2)	[AX][BX][CX][DX][AB][CD]	*[AX][BX][CX][DX][AB][CD]	407	14,265	1	-155	1,974	1
		[AX][BX][CX][DX][CD]	-44,645	63,152	4.4	-17,672	24,992	12.7
		[AX][BX][CX][DX]	-37,924	53,653	3.8	-17,583	24,866	12.6
3)	[AX][BX][CX][DX][AB][CD][S]	*[AX][BX][CX][DX][AB][CD][S]	1,559	15,537	1	5	2,097	1
		[AX][BX][CX][DX][CD][S]	-45,012	45,024	2.9	-17,671	17,672	8.4
		[AX][BX][CX][DX][S]	-58,567	60,521	3.9	-17,801	17,803	8.5
4)	[AX][BX][CX][DX][AB][CD][SD]	*[AX][BX][CX][DX][AB][CD][SD]	6,035	8,551	1	72	1,005	1
		[AX][BX][CX][DX][AB][CD][S]	-10,343	13,081	1.5	-931	1,131	1,1
		[AX][BX][CX][DX][S]	-76,149	76,162	8.9	-11,168	11,169	11.1

Table 1: Results of the simulations. The asterisks indicate the cases where the estimating model coincides with the generating model, wrt which the RRMSE are calculated.

5. Conclusions

The paper focuses on the estimation of a population size by using multisource data. The interest on this topic is increasing with the increase of available administrative sources for the National Statistical Institutes. The availability and usage of more information is certainly important to enrich the analysis we may carry out, however new methodological problems need to be dealt with. In this study we estimate the population size by using multiple-record system methodologies, that is, considering each source as a capturing list of the units of interest and then by using capture-recapture techniques. The usual assumptions such as independence of the captures of a unit in different lists ("Causal Independence") and absence of overcoverage cannot be taken. Moreover, many administrative lists are "incomplete", i.e., target a subset of our population of interest. We propose an estimation procedure based on a latent model to deal with overcoverage, dependence of the captures in different lists, and the presence of incomplete lists. In particular, loglinear models are introduced 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. The results of the estimation strategy proposed are encouraging, in fact in the simulations we have performed, whenever the model is the same as the one used for generating data, good estimates in terms of MSE are obtained. On the other hand, the experiments pointed out that, when the model is not the generating one, estimates may be biased. Hence, a sensible point to apply the proposed strategy is the choice of the model. The experiments analysed in the paper have shown that AIC and BIC were able to find the right model, but more studies are needed in this direction. Further studies will be devoted to see whether covariates may help to improve the results both in terms of model selection and to reduce the impact of a wrong model. In addition, it is our interest to study a Bayesian approach that is in general more apt to smooth the results and to introduce prior information to help the model selection.

REFERENCES

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

Bartolucci, F. and Forcina, A. (2001), "Analysis of Capture-Recapture Data with a Rasch Type Model Allowing for Conditional Dependence and Multidimensionality," *Biometrics*, 57, 714–719.

Biemer, P., (2011) *Latent Class Analysis of Survey Error*, John Wiley & Sons Inc., NY.

Biggeri, A., Stanghellini, E., Merletti, F. and Marchi, M. (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.

Fienberg, S. (1972), "The multiple recapture census for closed populations and incomplete 2k contingency tables," *Biometrika*, 59, 409–439.

Hagenaars, J. A., "Latent structure models with direct effects between indicators local dependence models." *Sociological Methods & Research* 16.3 (1988): 379–405.

Hagenaars, J.A. (1993) *Loglinear models with latent variables.*, Newbury Park: CA: Sage.

Kamen, C., S. (2005). "The 2008 Israel Integrated Census of Population and Housing Basic conception and procedure." Central Bureau of statistics. (http://www.cbs.gov.il/mifkad/census2008_e.pdf)

Sutherland, J. M., Multi-list methods in closed populations with stratified or incomplete information. 2003. PhD Thesis. Simon Fraser University.

Sutherland, J. M., and Schwarz, C. J., (2005) "Multi-List Methods Using Incomplete Lists in Closed Populations". *Biometrics*, 61, 134–140.

Viviano, C. (2009) *The setting up of Farm Register: the production process to building up the prototype of the pre-census list*, 21th Meeting of the Wiesbaden Group on Business Registers - International Roundtable on Business Survey Frames "The Central Place of Business Registers in Response to Globalisation Needs" 24–27 November 2009, OECD Paris, France

Wallgren, A., and Wallgren, B. (2007), *Register-based statistics: Administrative data for statistical purposes*, John Wiley and Sons, Chichester

Zhang, L. C. (2011), "A Unit-Error Theory for Register-Based Household Statistics," *Journal of Official Statistics*, 27, 415–432.

Zhang, L. C. (2012), "Topics of statistical theory for register-based statistics and data integration," *Statistica Neerlandica*, 66, 41–63.

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

Zwane, E.N., van der Pal-de Bruin and van der Heijden P.G. (2004), "The multiple–record systems estimator when registrations refer to different but overlapping populations," *Statistics in medicine*, 23, 2267–2281.