

An Application of Small Area Estimation Methods for the Agricultural Census Undercount

Un'Applicazione dei Metodi di Stima per Piccole Aree per la Copertura del Censimento dell'Agricoltura

Nicoletta Cibella, Michele D'Alò, Tiziana Tuoto
ISTAT, tuoto@istat.it

Riassunto: Nel presente lavoro vengono fornite stime per la copertura del 5° Censimento dell'agricoltura per sotto-domini non pianificati: Regioni e domini incrociati per Regione e ampiezza dell'azienda. L'indagine di copertura è stata infatti progettata con l'intento di valutare il tasso di copertura del censimento a livello nazionale e per ripartizione geografica; per ottenere stime affidabili a livello più disaggregato è necessario avvalersi di specifiche tecniche di stima per piccole aree. In questa sede vengono applicati i metodi tradizionali di stima per piccole aree.

Keywords: Coverage Rate Estimation; Petersen Model; Small Area Methods.

1. Introduction

Censuses are aimed at complete enumeration of the total population, but many factors affect census coverage, causing omissions, duplications and misclassifications. In order to measure the completeness of the 5° Agricultural Census, ISTAT conducts an independent coverage survey. The main target of this Post-Enumeration Survey (PES) is to evaluate the ratio between the census count and the true size of the population, defined as the coverage rate. In order to estimate the true number of farms is applied the Petersen model, one of the most widespread coverage error models.

The PES provides reliable estimates of the true population size for large domain (national and main geographic levels); to obtain estimates for regional level and for geographic-size farm combinations domains, it is necessary to apply specific small area estimation techniques. The aim of this paper is to estimate the Census undercount for such domains using the traditional small area methods.

2. Post Enumeration Survey and Petersen Model

In order to obtain an independent estimate of total farm count in a dual-system estimation model (the Petersen one), ISTAT constructs an area frame (the cadastral maps), selects a sample and conducts a field post-enumeration survey. It's used a two stage stratified sample, where the primary sampling unit is the Municipality and the second-stage unit is the cadastral map (not the farm).

The farm universe total N can be estimated as an application of the Petersen coverage error model. This model uses the independent post enumeration survey in conjunction with the Census to provide an estimate of N , assuming (Wolter, 1986):

1. both the Census and the survey attempt to measure the same universe for the same reference period (the Census one);
2. Census and survey enumerations are independent of one another, i. e., the probability of a farm being on the census list (*capture probability*) is independent of the probability of a farm being enumerated by the survey and vice versa;
3. it is possible to match the survey sample results to the census results without error;
4. spurious events, e.g. duplicates, non-existent cases, out-of-scope cases, have been eliminated.

Under the assumptions, the capture probabilities are modelled by a multinomial distribution:

		PES		
Census		In	Out	
	In	p_{11}	p_{12}	p_{1+}
	Out	p_{21}	p_{22}	p_{2+}
		p_{+1}	p_{+2}	1

and the maximum likelihood estimators are

$$\hat{p}_{1+} = N_{11}/N_{+1} \quad \hat{p}_{+1} = N_{11}/N_{1+} \quad \hat{N} = N_{+1} \cdot N_{+1}/N_{11} = N_{1+}/\hat{p}_{1+}$$

where N_{11} are the farms enumerated in both lists and N_{1+} (N_{+1}) the ones enumerated in the Census (PES).

Note that the Petersen model assumes capture probabilities are homogeneous across units: this assumption is unlikely to hold for the whole population. Therefore, homogeneous groups within each of which the assumption holds are create.

3. Small area estimation

The survey is designed to produce reliable direct estimates for large areas and for large domains, such as farm size or type of organisation at national and main geographic area levels. Otherwise the direct estimates for smaller areas or domains, such as regional level and geographic-size farm combinations, may have unacceptable high standard errors due to insufficient domain sample size. In order to obtain reliable estimates for these small areas and domains, small area techniques are needed.

Let d denote a particular small area or domain of interest, the generic estimator of the true population size in d , ${}_dN$, is:

$${}_d\hat{N} = {}_dN_{1+}/\tilde{p}_{1+}$$

where ${}_dN_{1+}$ is the known Census number of farms in the small area d and $\tilde{p}_{1+}$ is an estimator of the capture probability based upon the available data.

Different estimators $\tilde{p}_{1+}$ are available: for example, if $\tilde{p}_{1+} = {}_d\hat{N}_{11}/{}_d\hat{N}_{+1}$ the direct estimator for ${}_dN$ is obtained.

Otherwise a synthetic estimator can be obtained under the assumption that the coverage rate in some small areas is in average equal to the rate of a larger domain which includes several areas. So the estimator $\tilde{p}_{1+}$ is given by:

$$\tilde{p}_{1+} = {}_L\hat{N}_{11} / {}_L\hat{N}_{+1}$$

where the subscript L denotes a large area which includes small area d .

The synthetic estimate of true number of farms, ${}_d\hat{N}$, has a smaller variability than the direct one, but, if the previous assumption doesn't hold, it could be strongly biased. It is possible to find a balance between bias and variance by considering a weighted average of the two previous estimators: so, the composite estimator of ${}_d\hat{N}$ is obtained with $\tilde{p}_{1+}$ given in the following form:

$$\tilde{p}_{1+} = {}_dw({}_d\hat{N}_{11} / {}_d\hat{N}_{+1}) + (1-{}_dw)({}_L\hat{N}_{11} / {}_L\hat{N}_{+1})$$

where the weights ${}_dw$ are suitably chosen in the range $[0,1]$. Optimal weights can be obtained by minimizing the MSE of the composite estimator with respect to ${}_dw$; assuming realistically that direct and synthetic estimators are not correlated, Ghosh and Rao (1994) suggest:

$${}_dw = \text{MSE}({}_d\hat{N}_{\text{syn}}) / (\text{MSE}({}_d\hat{N}_{\text{syn}}) + v({}_d\hat{N}_{\text{dir}}))$$

An EBLUP estimator of true population size is obtained by modelling the adjustment factor, given by $\theta_d = {}_dN / {}_dN_{1+}$, using an area level mixed model (Dick, 1995):

$$\hat{\theta}_d = \theta_d + e_d \quad \text{and} \quad \theta_d = X_d\beta + u_d$$

where $\hat{\theta}_d$ are the direct estimates, e_d are the sampling errors, u_d are the area level random effects, X_d represent the area level covariates and β the regression parameters. The chosen covariates are the farms' size, the type of organization and the main geographic areas. The parameters involved in the model can be estimated by the Restricted Maximum Likelihood (REML) method. Finally, the true number of farms in each small area is given by:

$${}_d\hat{N} = \hat{\theta}_d^{\text{EBLUP}} \cdot {}_dN_{1+}.$$

4. Results and conclusion

In figure 1 the estimates for the true number of farms in the 20 Regions using the direct estimator, the synthetic estimator with L = 'main geographic areas', the composite estimator and the EBLUP one are showed. The EBLUP estimator is quite closed to the direct estimator, but it has a lower standard error, as shown in figure 2, where the standard errors are reported. Note that, generally, the composite estimator seems to be

more reliable than the others in terms of standard errors, while the direct estimator is the worse.

Figure 1: *Number fo farms- Census count and estimates*

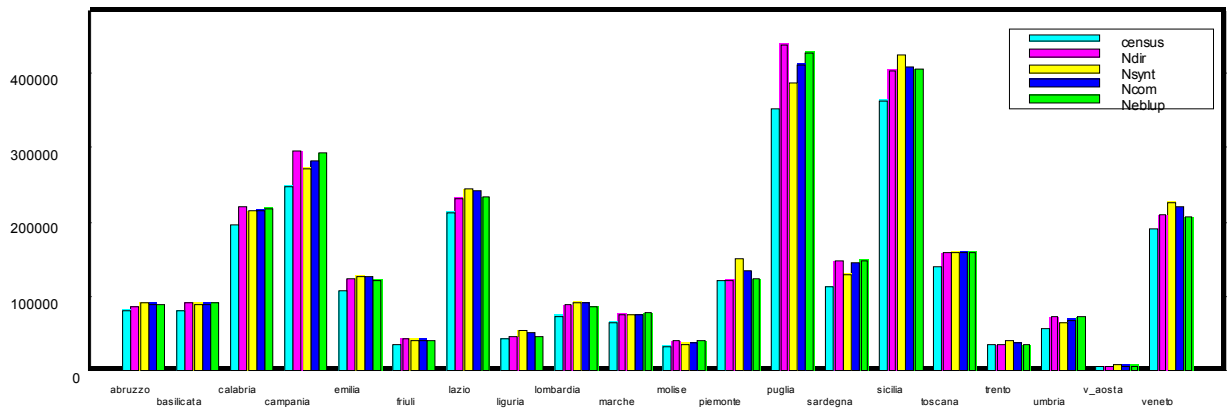
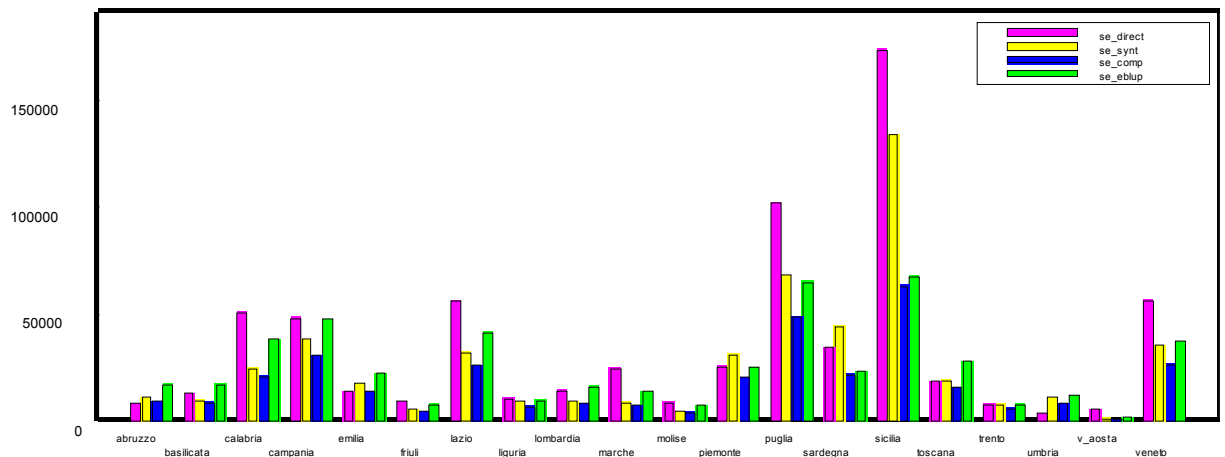


Figure 2: *Standard errors of estimates*



References

- Dick P. (1995) Modelling net undercoverage in the 1991 Canadian Census, *Survey Methodology*, vol.21, 1, 45-54.
- Ghosh M. and Rao J.N.K.(1994), Small area estimation: an appraisal (with discussion), *Statistical Science*, 9,55-93.
- Rao J.N.K, (2003) *Small area estimation*, John Wiley & Sons, Hoboken, New Jersey.
- Wolter K. (1986) Some coverage error models for Census data, *Journal of the American Statistical Association*, 81, 338-346.