

Measuring Response Errors in 2000 Census of Agriculture

La Valutazione degli Errori di Misura del 5° Censimento Generale dell'Agricoltura

Antonia Boggia, Matteo Mazziotta, Alessandro Pallara, Monica Russo

ISTAT, Via Depretis, 74/b

Rome, Italy

e-mail: boggia@istat.it

e-mail: mazziott@istat.it

e-mail: pallara@istat.it

e-mail: morusso@istat.it

Riassunto: In questo lavoro si presentano i primi risultati dell'analisi dei dati dell'indagine di controllo per la valutazione degli *errori di risposta* connessi ad alcune delle variabili più significative rilevate nell'ambito del 5° Censimento Generale dell'Agricoltura. L'indagine è stata condotta mediante reintervista di un campione di circa 7800 aziende agricole già rilevate al Censimento. Il metodo della *reintervista* consente di ottenere, sotto opportune condizioni sul processo di misurazione ripetuta, una stima della *varianza semplice* di risposta e della *distorsione*. In particolare, nel lavoro si affronta il problema della costruzione di tali stime quando non si può assumere che le due misurazioni siano ottenute nelle medesime condizioni, come nel caso considerato, in cui il dato censuario è stato ottenuto mediante intervista diretta, mentre l'indagine di controllo è stata effettuata con tecnica di rilevazione telefonica (CATI).

Keywords: Measurement Error, Reinterview, Response bias, Response variance, Reconciliation.

1. Introduction

Reinterviews have been extensively used as a tool for estimating and reducing response errors in censuses and sample surveys. By *response error* it is meant any error occurring at the data collection stage for a variety of reasons. Errors may be due to the respondent, to the interviewer, or to both. Such errors can never be completely avoided and, therefore, in the practice of censuses and large surveys some different techniques (cf., e.g., Forsman and Schreiner, 1991) have been proposed in order to measure response error components. One such method relies on replicated measurements on the same units. That is, for a sample of units a set of questions from the original interview is asked once again (the *reinterview*), and the two answers given by the same units to the same question are then matched. When the responses obtained during the reinterview differ from those obtained in the original interview, the differences can be evaluated through the so-called *reconciliation*.

The standard response error model based on reinterview data was developed at the U.S. Bureau of the Census (Hansen, Hurwitz and Pritzker, 1964). Using this model unbiased estimates of the response errors components can be obtained under fairly restrictive

assumptions. Biemer and Forsman (1992) discuss various problems when the basic assumptions of the model are not met, proposing a general model for response errors, of which the Bureau of the Census' model can be viewed as a special case. In Falorsi *et al.* (2001) it is shown how to yield unbiased estimates of the error variance components when the assumption of identical measurement conditions for the two occasions fails to hold. This seems definitely to be case with the application presented herein, relative to studying the accuracy of the 2000 Italian Census of Agriculture. Indeed, the reinterview survey carried on after census field operation has been based on computer-assisted telephone interviewing (CATI), so that the two survey measures cannot be considered as being collected under the same conditions.

2. The statistical model for the analysis of reinterview data

The standard approach to statistical modeling of measurement errors based on reinterview data (Hansen, Hurwitz and Pritzker, 1964) assumes that a measurement model m holds so that the response y_k to any given question can be represented as the sum of two terms: a true value q_k and an error e_k , as follows:

$$y_k = q_k + e_k \quad (1)$$

Consider then the following conditions: (i) a complete enumeration of the population U is carried on (without coverage errors) so that for each element k of U a measurement of the variable y of interest is observed, which will be denoted in what follows with y_{k1} ; (ii) a sample s is drawn from U and for each element $k \in s$ the variable y is observed a second time, thus obtaining the repeated value, denoted as y_{k2} ; (iii) the measurement conditions for the two occasions have been identical, or as close to identical as possible.

With the above conditions, the expected values $E_m(\cdot)$ with respect to the model (1) can be specified as follows:

$$E_m(y_{kt}|U, s) = m_k \quad (\text{for } k \in U \text{ if } t=1 \text{ and } k \in s \subset U \text{ if } t=2) \quad (2)$$

$$E_m[(y_{kt} - m_k)^2|U, s] = s_k^2 \quad (\text{for } k \in U \text{ if } t=1 \text{ and } k \in s \subset U \text{ if } t=2) \quad (3)$$

$$E_m[(y_{kt} - m_k)(y_{lt} - m_l)|U, s] = s_{kl} \quad (\text{for } k \in U \text{ if } t=1 \text{ and } k \in s \text{ if } t=2) \quad (4)$$

$$E_m[(y_{k1} - m_k)(y_{l2} - m_l)|U, s] = 0 \quad (\text{for } (k \in U) \cap (l \in s)) \quad (5)$$

We assume that the objective is to estimate the population total of the *true unknown values* q_k : $t_y = \sum_{k \in U} q_k$. An estimate of the parameter t which can be calculated after

census data collection is: $Y = \sum_{k \in U} y_{k1}$. Under the model (1) – (5) the mean square error

of Y can be represented as follows (cf. Särndal *et al.*, 1992, pp.608-609):

$$E_m[(Y - t_y)^2|U, s] = V_1 + B^2 \quad (6)$$

where the two components

$$V_1 = \sum_{k \in U} s_k^2 + \sum_{k \in U} \sum_{\substack{l \neq k \\ l \in U}} s_{kl} \quad (7)$$

$$B^2 = \left[\sum_{k \in U} (m_k - q_k) \right]^2 \quad (8)$$

denote the response variance and the (squared) measurement bias, respectively. Note that the bias component arises when the expected measurement values do not agree with the true values.

Under the (2)-(5) an unbiased estimate of the response variance V_I can be obtained as:

$$\tilde{V}_I = \frac{1}{2} \left[\sum_{k \in s} \frac{1}{p_k} (y_{k1} - y_{k2})^2 + \sum_{\substack{l \in s \\ l \neq k}} \frac{1}{p_{kl}} (y_{k1} - y_{k2})(y_{l1} - y_{l2}) \right], \quad (9)$$

where p_k and p_{kl} denote, respectively, first and second order probabilities of inclusion in sample s .

An estimate of the bias can be obtained assuming that design of reinterview survey will yield the most correct responses. In the practice of reinterview programs this involves usually that discrepancies between the reinterview responses and the original interview responses are reconciled by asking the respondent to determine what is the correct information between the two interviews. Therefore, after reconciliation, a third response y_{k3} is available, whose expected value under the model is assumed to be the true value,

$$E_m(y_{k3}|U, s) = q_k \quad (\text{for } k \in s). \quad (10)$$

Under this hypothesis an unbiased estimate of the measurement bias will be:

$$\tilde{B}^2 = \sum_{k \in s} \frac{1}{p_k} \left(y_{k3} - \frac{1}{2}(y_{k1} + y_{k2}) \right)^2 + \sum_{\substack{l \in s \\ l \neq k}} \frac{1}{p_{kl}} \left(y_{k3} - \frac{1}{2}(y_{k1} + y_{k2}) \right) \left(y_{l3} - \frac{1}{2}(y_{l1} + y_{l2}) \right) \quad (11).$$

The measurement error model presented above relies on a number of assumptions that in practice hardly ever hold. In particular, for the application presented in this paper the two most critical are that: *i)* the measurement conditions are identical for the two occasions and *ii)* the reconciled reinterview process yields the true response.

3. The Census of Agriculture reinterview program

Data collection for the 5th Census of Agriculture in Italy began at the end of November 2000, through face-to-face interviewing of some 2,75 million names, from individual or family operations to very large corporations as well as publicly managed farms or woodlands included in the census farm list. Field operations were concluded by the end of March 2001, with a few exceptions that were made primarily for small towns that had been affected by floods during the data collection period.

A reinterview survey has been planned at the end of data collection to evaluate accuracy of the information recorded during census field work. The sample design is two-stage with farms selected within municipalities (primary sampling units) to control geographical spread of the sample. The choice of this sampling plan has been motivated essentially from budget and time constraints, since survey design involved a number of preliminary operations related to special processing of census questionnaires of farms

selected in the reinterview sample. The selected sample included some 8200 farms in about 200 municipalities from all regions of the country.

As mentioned before, interviewing was conducted through Computer Assisted Telephone Interviewing (CATI). Using CATI had a number of advantages with respect to field reinterview for obtaining uncontaminated estimates of response errors (cf. Forsman and Schreiner; 1991), although in this case it was the only option available due to limited budget allocated to the reinterview program. Primarily, for the measurement of response bias there were specific advantages concerning reconciliation of differences because: a) the reinterviewer had non access to the original interview data, unless the difference between responses to the original interview and the reinterview for a given question did not cross certain tolerance limits which had been programmed in the software for the reinterviews; b) it was not possible for the reinterviewer to alter the reinterview response once the reinterview had been completed. Conversely, in measuring response variance, the assumption that the reinterview response, although independent from the original face to face interview, has been obtained under identical or as close to identical measurement conditions is clearly violated when using telephone for reinterview. The estimates of response variance from Census of Agriculture reinterview data have been obtained using the result from Falorsi *et al.* (2001), who show how to obtain an unbiased estimate of the response variance when the assumption of identical second-order moment of the two measurements is violated.

Because of complexity of the original census questionnaire, it was decided that for successful realization of telephone interviewing, only a subset of the original questions was included in the reinterview pertaining to: i) major crops, flowers or tree-growing and vineyards, ii) cattle and poultry raising and iii) family and other personnel employed in the farm. The reinterview included some 50 questions, 30 of which involving reconciliation as part of the reinterview process.

References

- Biemer P.P., Forsman, G. (1992) On the Quality of Reinterview Data with Application to the Current Population Survey, *Journal of the American Statistical Association*, 87, 915-923.
- Forsman G., Schreiner, I. (1991) The Design and Analysis of Reinterview: an Overview, in: *Measurement Errors in Surveys*, Biemer, P. Groves, R.M. Lyberg, L. Mathiowetz, N. and Sudman, S. (Eds.), New York, Wiley, 279-302.
- Falorsi P.D., Fortini M., Pallara A., Russo, A. (2001) Estimation of Simple and Correlated Measurement Variance when First and Second Interview use Different Survey Modes, *Proceedings of Statistics Canada Symposium 2001 – Achieving Data Quality in a Statistical Agency: a Methodological Perspective* (forthcoming).
- Hansen M. H., Hurwitz W.N., Pritzker, L. (1964) The Estimation and Interpretation of Gross Differences and the Simple Response Variance, in: *Contribution to Statistics*, Rao, C.R. (Ed.), Calcutta, Statistical Publishing Society, 111-136.
- Särndal C.E., Swensson B., Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.