Statistical Methods & Applications (2005) 14: 83–99
DOI: 10.1007/s10260-004-0090-6

SMA

Estimates based on preliminary data from a specific subsample and from respondents not included in the subsample*

Piero Demetrio Falorsi, Giorgio Alleva, Fabio Bacchini, and Roberto Iannaccone**

ISTAT, University of Rome "La Spapienza", Italy

Abstract. Various approaches to obtaining estimates based on preliminary data are outlined. A case is then considered which frequently arises when selecting a subsample of units, the information for which is collected within a deadline that allows preliminary estimates to be produced. At the moment when these estimates have to be produced it often occurs that, although the collection of data on subsample units is still not complete, information is available on a set of units which does not belong to the sample selected for the production of the preliminary estimates. An estimation method is proposed which allows all the data available on a given date to be used to the full - and the expression of the expectation and variance are derived. The proposal is based on two-phase sampling theory and on the hypothesis that the response mechanism is the result of random processes whose parameters can be suitably estimated. An empirical analysis of the performance of the estimator on the Italian Survey on building permits concludes the work.

Key words: Timeliness, preliminary estimates, two-phase sampling, response probabilities, expansion estimators

1. Objectives and context of reference

Over recent years the need for increased timeliness of short-term data has become more pressing at both a national and an international level. In Europe this need is particularly urgent, because the infra-annual statistics available at a community

* The Sects. 1,2,3,4 and the technical appendixes have been developed by Giorgio Alleva and Piero Demetrio Falorsi; Sect. 5 has been done by Fabio Bacchini and Roberto Iannaccone.

** Piero Demetrio Falorsi is chief statisticians at Italian National Institute of Statistics (ISTAT); Giorgio Alleva is Professor of Statistics at University "La Sapienza" of Rome, Fabio Bacchini and Roberto Iannaccone are researchers at ISTAT.

level has lower timeliness than that found in many countries outside the EU. The USA, in this respect, sets an example that the European system would be advised to follow.

Several studies have recently been initiated by the EU to identify the main issues regarding the timeliness of short-term data. These studies fall into two broad types: first, those which focus on the feasibility of constructing a sample stratified by country, to provide estimates of European retail trade (Eurostat, 2001); second, those undertaken by ESTEI (Expert group on Sampling for Timely European Indicators), which has been working on guidelines to which individuals member states ought to refer to ensure timely production of short-term indicators.

The report (Eurostat, 2002) drawn up by the ESTEI team indicates three possible approaches to reducing the time taken to deliver indicators:

I – *acceleration of national procedures*: under this approach each stage of the process leading to the production of statistical indicators would need to be redesigned. Timeliness is seen as an issue pertaining to the entire productive process - hence it has to be addressed by the overall way a survey is organized, there cannot be a simple statistical solution to it. This approach would require large financial investment in the data-collection phase;

II – *selection of a specific subsample*: this approach would involve selecting a subsample of the surveys that each European country currently carries out individually. This would require an investment in resources for collecting information on the units of the subsample, within a deadline that allows preliminary estimates to be produced;

III – *approaches based on statistic models:* statistical models can also lead to the production of timely estimates. This would be the cheapest way because it would not call for resources during the data acquisition stage, it would only require investment related to methodological research and to information processing.

As regards approach II, there are two possible ways of selecting the subsample:

- two-phase sample design, developed within the classical approach to sampling finite populations (Sarndal et al., 1992);
- optimal sampling design, developed by those using predictive inference for finite populations (Royall, 1992; Dorfman and Valliant, 2000).

As regards approach III, there are three main methods:

- methods based on techniques developed within time series analysis, which use exclusively estimates calculated on previous occasions (Box and Jenkins, 1976). In this category can also be placed methods that find specific sub-series within a time series, thereby permitting a good prediction of the overall time series to be made (Van Garderen et al., 2000; Maravalle et al., 1993; Battaglia and Fenga, 2003);
- models using both information obtained from units responding within a fixed time (timeliness respondents) and estimates made on previous occasions. These models can be further classified into (i): methods based on the imputation of the variable values of the non-respondents units. Such methods can be based on a non-parametric approach (Chen and Shao, 2000), or on a suitable superpopulation model (Little, 1986); a relevant set of mod-

els are linear dynamic models (Harvey, 1984, 1989; Tam, 1987; Bell and Hillmer, 1990); (*ii*) methods based on re-weighting techniques that correct weights assigned to timeliness respondents, so that they can also represent non-timeliness units in a suitable way. These weights can be based on superpopulation models explaining either the stochastic process that generates the timeliness response, or the values of the target variable of the non-respondents unit at the current time (Rizzo et al. 1996; Eltinge and Yansaneh, 1997); the probability of timeliness response co uld also be defined using non-parametric techniques (Giommi, 1987; Niyonsenga, 1994);
- suitable econometric methods based on the use of the relation linking the target variables to *proxi* indicators used as early index of the target variables (Bodo and Signorini, 1987; Bruno and Lupi, 2002).

The proposal being made in this paper has been developed within the classical approach to sampling on finite populations. A case is considered which frequently arises when selecting a subsample of units, the information for which is collected within a deadline that allows preliminary estimates to be produced. At the moment when these estimates have to be produced it often occurs that, although the collection of data on subsample units is still not complete, information is available on a set of units which does not belong to the sample selected for the production of the preliminary estimates.

A simple way of producing preliminary estimates could be to use exclusively the information collected from the respondents belonging to the subsample. However, this would not only go against the statistical principle that all available information should be used to the full, it would also result in inefficiency in estimates if the number of respondents not involved in the subsample was large.

Below we present an estimation method which allows all the information available on a given date to be used. Our proposal is based on the theory of two-phase sampling and on the hypothesis that the response mechanism is the result of random processes whose parameters can be appropriately estimated.

The paper is structured as follows: Sect. 2 introduces the parameter of interest and the context of observation; Sect. 3 presents the modelling of the response process; Sect. 4 describes the expansion estimator based on the inclusion probability of first and second phase samples, and on the models employed for explaining the preliminary response probability. In Sect. 5 the performance of the proposed estimator is studied with an application to the *Italian Survey on building permits*. The analysis is carried out comparing the mean absolute percentage error and the quarterly and year growth rates of the proposed estimator with two different estimators that use only subsample data. Finally in Sect. 6 the conclusion are given

2. Parameters of interest and sample observations

Let us denote by: $U_t = \{1, ..., k, ..., N_t\}$ the *target population* at the *current time* t, which consists of N_t units; and by $y_{t,k}$ the value of the variable of interest for the k-th unit at the t-th time.

We assume that the parameter of interest is the total

$$Y_t = \sum_{k \in U_t} y_{t,k}. \tag{1}$$

In order to estimate this total, a first phase sample s_{at} is selected from the population U_t, through a sampling design which attributes to the sample s_{at} the probability $p(s_{at})$ of being selected.

We denote by

$$\pi_{at,k} = \sum\nolimits_{s_{at} \supset k} p(s_{at}) \tag{2}$$

$$\pi_{at,kl} = \sum\nolimits_{s_{at} \supset (k,l)} p(s_{at}) \tag{3}$$

the inclusion probability of unit k and the inclusion probability of the pair of units (k, l). Moreover, we assume that in order to obtain a preliminary estimate of Y_t a *second phase sample* s_t is selected from s_{at} with conditional probability $p(s_t|s_{at})$. Let us further assume that data collection on the units of s_t is carried out with particular care so that almost every unit of this sample is respondent at the time preliminary estimates are produced.

The inclusion probabilities of the *second phase* design are

$$\pi_{t,k|at} = \sum\nolimits_{s_t \supset k} p(s_t|s_{at}) \tag{4}$$

$$\pi_{t,kl|at} = \sum\nolimits_{s_t \supset (k,l)} p(s_t|s_{at}). \tag{5}$$

It is worth noting that s_{at} can be divided into two subsets: s_t, and the complementary subset, $\bar{s}_t$, which contains s_{at} units not included in s_t, so that $s_{at} \equiv s_t \cup \bar{s}_t$ and $s_t \cap \bar{s}_t \equiv \emptyset$.

When preliminary estimates have to be calculated we have a subset of responding units, marked r_t, which is obtained from the union of two exclusive parts: the set r_{s_t} (with $s_t \supseteq r_{s_t}$), composed of r_t units from subset s_t; the set $r_{\bar{s}_t}$ (with $\bar{s}_t \supseteq r_{\bar{s}_t}$), composed of r_t units from subset $\bar{s}_t$, which is complementary to s_t.

3. Modelling of the response process

Under the condition that the unit $k (k \in s_{at})$ is included either in s_t, or in $\bar{s}_t$, we assume that response probability of the unit within a given deadline (defined for the production of preliminary estimates) is based on a Bernoullian stochastic process. Let us denote by

$$\begin{aligned} \varphi_{k|s_t} &= \Pr(k \in r_{s_t} | k \in s_t) \\ \varphi_{k|\bar{s}_t} &= \Pr(k \in r_{\bar{s}_t} | k \in \bar{s}_t) \end{aligned} \tag{6}$$

the unit k probabilities of being respondent (within the given deadline) under the condition of one of the two events $k \in s_t$ or $k \in \bar{s}_t$.

In order to estimate these unit k probabilities we use some special models named *Homogeneous Response within Groups* (HRG), which assume that the sample may be partitioned into subsets and that units of a given subset have an equal response probability.

First, we are going to consider the HRG model for the calculation of an estimate of the probabilities $\varphi_{k|s_t}$. This model is characterised as follows:

i. the sample s_{at} is partitioned into C_{s_t} groups or weighting cells; let us denote by $s_{at,c}$ the set of units belonging to the generic c-th group $(c = 1, ..., C_{s_t})$;
ii. the units of $s_{at,c}$, when included in s_t, have a uniform value of the conditional response probability $(\varphi_{k|s_t})$ equal to $\varphi_{s_{t,c}}$; units belonging to different groups have different response probability;
iii. each unit responds independently of all other units.

This model can be formally described as

$$\begin{cases} \Pr(k \in r_{s_t} | k \in s_t) = \varphi_{k|s_t} = \varphi_{s_{t,c}} \text{ for } k \in s_{at,c}, (c = 1, ..., C_{s_t}) \\ \Pr((k,l) \in r_{s_t} | (k,l) \in s_t) = \Pr(k \in r_{s_t} | k \in s_t) \Pr(l \in r_{s_t} | l \in s_t) = \\ \quad = \varphi_{s_{t,c}} \varphi_{s_{t,c'}} \text{ for } (k \in s_{at,c}) \cap (l \in s_{at,c'}), ((c \text{ or } c') = 1, ..., C_{s_t}). \end{cases} \tag{7}$$

In the HRG model for the calculation of the estimate of the probabilities $\varphi_{k|\bar{s}_t}$, we assume that the sample s_{at} is divided into $C_{\bar{s}_t}$ weighting cells, and that the k-th element of the generic cell $s_{at,\ddot{c}} (\ddot{c} = 1, ..., C_{\bar{s}_t})$, when included in $\bar{s}_t$, has a conditional response probability $\varphi_{k|\bar{s}_t}$ equal to $\varphi_{\bar{s}_{t,\ddot{c}}}$. As with model (7), the formal expression of the model is:

$$\begin{cases} \Pr(k \in r_{\bar{s}_t} | k \in \bar{s}_t) = \varphi_{k|\bar{s}_t} = \varphi_{\bar{s}_{t,\ddot{c}}} \text{ for } k \in s_{at,\ddot{c}}, (\ddot{c} = 1, ..., C_{\bar{s}_t}) \\ \Pr((k,l) \in r_{\bar{s}_t} | (k,l) \in \bar{s}_t) = \Pr(k \in r_{\bar{s}_t} | k \in \bar{s}_t) \Pr(l \in r_{\bar{s}_t} | l \in \bar{s}_t) = \\ \quad = \varphi_{s_{t,\ddot{c}}} \varphi_{s_{t,\ddot{c}'}} \text{ for } (k \in s_{at,\ddot{c}}) \cap (l \in s_{at,\ddot{c}'}), ((\ddot{c} \text{ or } \ddot{c}') = 1, ..., C_{\bar{s}_t}). \end{cases} \tag{8}$$

Before illustrating the subsequent developments it is worth presenting a method for calculating weighting cells which is borrowed from Eltinge and Yansaneh's paper (1997) and is based on the "response propensity scoring" technique (Rosenbaum and Rubin, 1983; Little, 1986).

First we consider the case of the partition of s_{at} into the C_{s_t} cells $s_{at,c}$. The first step in applying the method consists of estimating, using the units of s_t, the parameters of a *logit* (or *probit*) model. For each unit the model connects the expected value of the binary variable $r^{\delta}_{t,k}$ (equal to one, if the unit is responding, or to zero, if it is not) to a vector $\mathbf{z}_{t,k}$ of auxiliary variables available on the whole sample s_{at}.

Following the estimation of these parameters we can predict the response probability $\hat{\varphi}_{k|s_t}$ for each unit of s_{at}. It is worth noting that, as illustrated in Eltinge and Yansaneh's paper (1997), in practical situations the predicted probabilities $\hat{\varphi}_{k|s_t}$ are not used directly for estimating the parameters of interest, since they may lead to the creation of extreme values in the estimates and thus drastically increase the variance. The predicted probabilities are used only in order to obtain an appropriate definition of the weighting cells.

In line with the optimal stratification theory (Cochran 1977, pp. 127–134), we fix the weighting C_{s_t} cells by dividing the sample s_{at} by the quantiles of predicted probabilities $\hat{\varphi}_{k|s_t}$; as a result every weighting cell includes homogeneous units as regards the response probability. If we indicate the position quantiles $(c-1)$ and c by $\hat{\varphi}_{c-1|s_t}$ and $\hat{\varphi}_{c|s_t}$, all the units for which $\hat{\varphi}_{c-1|s_t} < \hat{\varphi}_{k|s_t} \leq \hat{\varphi}_{c|s_t}$ belong to the weighting cell c.

Similarly, for the definition of the partition of s_{at} into the $C_{\bar{s}_t}$ cells $s_{at,\ddot{c}}$, when the predicted values $\hat{\varphi}_{k|\bar{s}_t}$ are determined we assign to the weighting cell $\ddot{c}$ all the units for which $\hat{\varphi}_{\ddot{c}-1|\bar{s}_t} < \hat{\varphi}_{k|\bar{s}_t} \leq \hat{\varphi}_{\ddot{c}|\bar{s}_t}$, where $\hat{\varphi}_{\ddot{c}-1|\bar{s}_t}$ and $\hat{\varphi}_{\ddot{c}|\bar{s}_t}$ indicate, respectively, the position percentiles $(\ddot{c}-1)$ and $\ddot{c}$ of the distribution over s_{at} of the predicted values $\hat{\varphi}_{k|\bar{s}_t}$.

If we have a good set of explicative variables $\mathbf{z}_{t,k}$, most of the bias reduction is achieved by forming a rather small number of weighting cells (5 or 6), (Eltinge and Yansaneh, 1997). If the response process was strongly dependent on an explanatory variable which is not available for regression, we can not decrease bias, no matter what the number of evaluated weighting cells is.

In their 1996 paper Rizzo, Kalton and Brick present alternative ways of setting up the weighting cells using response propensity scoring from longitudinal research. In De Vitiis et al. (2002) we find an application of these methods as part of a survey on the labor force carried out by the Italian National Institute of Statistics.

4. Expansion estimator

Under the condition that the unit k belongs to the first phase sample s_{at}, the fact that this unit is a respondent one (which means that it belongs to the set r_t) can be represented as the union of the following two exclusive events:

- E_1: the unit is selected in the second phase sample, s_t, and it is respondent; we have $\Pr(\mathrm{E}_1) = \Pr((k \in s_t) \cap (k \in r_{s_t}) | k \in s_{a,t}) = \pi_{t,k|at}\varphi_{k|s_t}$;
- E_2: the unit does not belong to the second phase sample s_t and it is respondent; we have $\Pr(\mathrm{E}_2) = \Pr((k \in \bar{s}_t) \cap (k \in r_{\bar{s}_t}) | k \in s_{a,t}) = (1 - \pi_{t,k|at})\varphi_{k\bar{s}_t}$.

As a consequence we have

$$\Pr(k \in r_t | k \in s_{a,t}) = \Pr(\mathrm{E}_1 \cup \mathrm{E}_2) = (\pi_{t,k|at}\varphi_{k|s_t}) + ((1 - \pi_{t,k|at})\varphi_{k|\bar{s}_t}). \quad (9)$$

If we know the response probabilities $\varphi_{k|s_t}$ and $\varphi_{k|\bar{s}_t}$ for each individual respondent unit (that is for each unit belonging to r_t) we can obtain an unbiased estimate of the parameter of the type (1) using this kind of estimator

$$\hat{Y}_t = \sum_{k \in r_t} y_{t,k} d_{t,k} \quad (10)$$

where

$$d_{t,k} = 1/\pi_{at,k}((\pi_{t,k|at}\varphi_{k|s_t}) + ((1 - \pi_{t,k|at})\varphi_{k|\bar{s}_t})). \quad (11)$$

The proof of the unbiasedness of the estimator (10) may be easily derived.

In practical situations, however, the values of $\varphi_{k|s_t}$ and $\varphi_{k|\bar{s}_t}$ are not known and have to be estimated. These estimates are done using response models which, if they are wrongly specified, can introduce bias into the estimators (10). Consequently the specification of the response models is a fundamental step in obtaining good estimates. It can be performed using an *explicit modelling of the response probabilities*, based on the estimate of a functional link between the response probabilities and a set of auxiliary variables $\mathbf{z}_{t,k}$ ($k \in s_{at}$), which are available on both the respondent sample and the non-respondent one.

The models often used are the HRG type, illustrated in expressions (7) and (8). Assuming (7) and (8) are valid models and that the probability of not observing respondents inside any of the weighting cells is about zero, it is possible to define unbiased estimators. To this end let us indicate by $s_{t,c}(c = 1, ..., C_{s_t})$ and $\bar{s}_{t,\ddot{c}}(\ddot{c} = 1, ..., C_{\bar{s}_t})$ the sets of intersection $s_{t,c} = s_{at,c} \cap s_t$, $\bar{s}_{t,\ddot{c}} = s_{at,\ddot{c}} \cap \bar{s}_t$; where n_c and m_c are the numbers of units and the number of respondent units in $s_{t,c}$; and where $n_{\ddot{c}}$ and $m_{\ddot{c}}$ are the number of units and the number of respondent units in $\bar{s}_{t,\ddot{c}}$. Finally, let us indicate by $\mathbf{m}_s$ and $\mathbf{m}_{\bar{s}}$ the vectors containing the number of observed respondent units inside the weighting cells, where

$$\mathbf{m}_s = (m_1, ..., m_c, m_{C_{s_t}})' \text{ and } \mathbf{m}_{\bar{s}} = (m_1, ..., m_{\ddot{c}}, ...m_{C_{\bar{s}_t}})'$$

and let us further indicate by $\mathbf{m} = (\mathbf{m}'_s, \mathbf{m}'_{\bar{s}})'$ the stack vector of vectors $\mathbf{m}_s$ and $\mathbf{m}_{\bar{s}}$.

If the conditions of the model (7) hold, the response probabilities of the unit belonging to s_t, corresponding to the observed vector $\mathbf{m}_s$ are:

$$\varphi_{k|s_t,\mathbf{m}_s} = \Pr(k \in r_{s_t} | k \in s_t, \mathbf{m}_s) = \frac{m_c}{n_c} \text{ for } k \in s_{at,c} \tag{12}$$

and

$$\varphi_{k,l|s_t,\mathbf{m}_s} = \Pr((k,l) \in r_{s_t} | (k,l) \in s_t, \mathbf{m}_s) = \\ = \begin{cases} \dfrac{m_c}{n_c} \dfrac{m_c - 1}{n_c - 1} & \text{for } (k,l) \in s_{at,c} \\ \dfrac{m_c}{n_c} \dfrac{m_{c'}}{n_{c'}} & \text{for } k \in s_{at,c}, l \in s_{at,c'}, c \neq c'. \end{cases} \tag{13}$$

Likewise, under model (8), the response probabilities of the units belonging to $\bar{s}_t$, corresponding to the observed vector $\mathbf{m}_{\bar{s}}$ are:

$$\varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}} = \Pr(k \in r_{\bar{s}_t} | k \in \bar{s}_t, \mathbf{m}_{\bar{s}}) = \frac{m_{\ddot{c}}}{n_{\ddot{c}}} \text{ for } k \in s_{at,\ddot{c}} \tag{14}$$

and

$$\varphi_{k,l|\bar{s}_t,\mathbf{m}_{\bar{s}}} = \Pr((k,l) \in r_{\bar{s}_t} | (k,l) \in \bar{s}_t, \mathbf{m}_{\bar{s}}) = \\ = \begin{cases} \dfrac{m_{\ddot{c}}}{n_{\ddot{c}}} \dfrac{m_{\ddot{c}} - 1}{n_{\ddot{c}} - 1} & \text{for } (k,l) \in s_{at,\ddot{c}} \\ \dfrac{m_{\ddot{c}}}{n_{\ddot{c}}} \dfrac{m_{\ddot{c}'}}{n_{\ddot{c}'}} & \text{for } k \in s_{at,\ddot{c}}, l \in s_{at,\ddot{c}'}, \ddot{c} \neq \ddot{c}'. \end{cases} \tag{15}$$

Under the conditions of the models (7) and (8) and assuming that the probabilities of the following events is zero:

$$A_s : m_c = 0, \quad \text{for some value of } c = 1, ..., C_s$$
$$A_{\bar{s}} : m_{\ddot{c}} = 0, \quad \text{for some value of } \ddot{c} = 1, ..., C_{\bar{s}}$$

an unbiased estimator (see Appendix 1) of the total of interest is:

$$\tilde{Y}_t = \sum_{k \in r_t} y_{t,k} a_{t,k} \tag{16}$$

where

$$a_{t,k} = 1/(\pi_{at,k}(\pi_{t,k|at}\varphi_{k|s_t,\mathbf{m}_s}) + ((1 - \pi_{t,k|at})\varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}})). \tag{17}$$

The variance of estimator (16) and the estimate of this variance (which is unbiased only if models (7) and (8) are unbiased and if the probability of the events A_s and $A_{\bar{s}}$ is zero), are (see Appendix 2):

$$\mathrm{V}(\tilde{Y}_t) = \sum_{k \in U_t} \sum_{l \in U_t} (\pi_{at,kl} - \pi_{at,k}\pi_{at,l}) \frac{y_{t,k}}{\pi_{at,k}} \frac{y_{t,l}}{\pi_{at,l}} + \tag{18}$$
$$+ \mathrm{E}_{p_a} \left(\sum_{k \in s_{at}} \sum_{l \in s_{at}} \left(\mathrm{E}_{\mathbf{m}} \left(a_{t,k} a_{t,l} \varphi_{k,l|s_{at},\mathbf{m}} \right) - \frac{1}{\pi_{at,k}} \frac{1}{\pi_{at,l}} \right) y_{t,k} y_{t,l} \right),$$

$$\tilde{\mathrm{V}}(\tilde{Y}_t) = \sum_{k \in r_t} \sum_{l \in r_t} \frac{1}{\pi_{at,kl}\varphi_{k,l|s_{at},\mathbf{m}}} (\pi_{at,kl} - \pi_{at,k}\pi_{at,l}) \frac{y_{t,k}}{\pi_{at,k}} \frac{y_{t,l}}{\pi_{at,l}} + \tag{19}$$
$$+ \sum_{k \in r_t} \sum_{l \in r_t} \left(a_{t,k} a_{t,l} \varphi_{k,l|s_{at},\mathbf{m}} - \frac{1}{\pi_{at,k}} \frac{1}{\pi_{at,l}} \right) \frac{y_{t,k} y_{t,l}}{\varphi_{k,l|s_{at},\mathbf{m}}},$$

where

$$\varphi_{k,l|s_{at},\mathbf{m}} = \Pr((k,l) \in r_t | s_{a,t}, \mathbf{m});$$

for $k \neq l$

$$\varphi_{k,l|s_{at},\mathbf{m}} = (\pi_{t,kl|at}\varphi_{k,l|s_t,\mathbf{m}_s}) + ((1 + \pi_{t,kl|at} - (\pi_{t,k|at} + \pi_{t,l|at}))\varphi_{k,l|\bar{s}_t,\mathbf{m}_{\bar{s}}}) +$$
$$+ ((\pi_{t,k|at} + \pi_{t,l|at} - 2\pi_{t,kl|at})\varphi_{k|s_t,\mathbf{m}_s}\varphi_{l|\bar{s}_t,\mathbf{m}_{\bar{s}}}),$$

and for $k = l$,

$$\varphi_{k,k|s_{at},\mathbf{m}} = \Pr(k \in r_t | s_{a,t}, \mathbf{m})$$
$$= (\pi_{t,k|at}\varphi_{k|s_t,\mathbf{m}_s}) + ((1 - \pi_{t,k|at})\varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}}) = \frac{\pi_{at,k}}{a_{t,k}}.$$

5. Empirical results

5.1. Background

In May 1998 the Council Regulation n.1165/98 for short term statistics (henceforth STS Regulation) was published. Regarding statistics on industry, construction, retail trade and repair and other services, the STS regulation foresees which indicators to produce and their timeliness. Henceforth the National Statistics Institute (NSI) began to revise their statistics according to the new standards. For the Italian NSI this has implied the need of new data collection methods for statistics in the construction sector. In particular, information on building permits are required 90 days from the reference period.

Until the end of 2002, the survey on building permits was organized as a monthly census of all Italian municipalities. Each municipality had to communicate information either on building permits or null activity, meaning that no building permits have been released for that month. STS Regulation requires data on building permits for the number of new residential building broken down into one-dwelling residential buildings or two and more dwelling residential buildings. The square meters are broken down into residential or non residential.

The Italian 8,100 Municipalities may be grouped into two classes that exhibit different response patterns. The two groups are:

- *First group* including 160 large municipalities with more than 50,000 citizens representing the 36.3% of the Italian population;
- *Second group* characterised by the 7,940 non-large municipalities with less than 50,000 citizens representing the 63.7% of the total population.

The response rate within the deadline useful for the production of a preliminary estimate (90 days) was very high for the first group, greater than 87%. On the contrary, the non-response phenomenon was stronger for the second group. The rate of preliminary respondent municipalities from this group ranges from 35.0% (in 1997) to 58.5% (in 1999). The principal reason for the delay in response process is due to data collection, which was organised into two steps: in the first step the data was sent from municipalities to provinces and in the second step to the Italian NSI. Another problem is related to the difficulty of undertaking a reminder activity for all the municipalities.

For all these reasons, it has been made evident that data collection method could not make the elaboration of the index with the delay required by STS possible.

In order to improve the timeliness, and starting in 2003, the first step will be the placing of the census survey side by side with a sampling survey of 814 municipalities. The information on municipalities included in the survey is collected without the intermediate step of the provincial level; while for the other 7.286 municipalities the survey continues as a monthly census. This strategy implies that for the month t at time $t + 3$ (90 days) two different sources of information are available: respondents included in the sample and respondents not included in the sample.

In particular the sample has been selected using a stratified strategy. The municipalities of the first group are included in the sample with certainty and form a take

all stratum. The other municipalities are subdivided into strata defined by the cross-classification of two auxiliary variables: region and class of resident population. A fixed number of municipalities is selected in each stratum without replacement and with equal probabilities.

5.2. *Analysis of the result*

In this paragraph we show some empirical results obtained by simulating the selection of the sample for the years 1998–2001; for each month of the period we have subdivided the preliminary respondents into two subset r_{s_t} and $r_{\bar{s}_t}$ and then calculated the estimators. We note that the methodology proposed in the paper supposes a higher response rate for the units in the sample s_t compared to that observed for nonsample units $\bar{s}_t$. However, this hypothesis is verified in the period 1997–2001 where the response rate is equal on average to 85% for the s_t and 75% for $\bar{s}_t$; the difference in the response rates will be even bigger in the future because the introduction of the sample in 2003 will improve the response rate for s_t.

We compare the estimator (16), denoted in the following as estimator E3, with two other estimators denoted in the following as E1 and E2 based only on sample data, in the context of the Italian survey on building permits. Estimator E3 uses the information on all data set of respondents. Since after 1998 the large municipalities are always preliminary respondents, the analysis is developed only with reference to the domain of municipalities of the *Second group*.

The estimates $\tilde{Y}_{t,a}$ of the total Y_t obtained with estimator $a(a =$E1, E2, E3) are defined as

$$\tilde{Y}_{t,a} = \sum_{k \in r_{s_t}} y_{t,k} a_{t,k,a} \quad \text{for } a = \text{E1, E2, and with}$$

$$\tilde{Y}_{t,\text{E3}} = \sum_{k \in r_t} y_{t,k} a_{t,k,\text{E3}} \quad \text{for } a = \text{E3},$$

where $a_{t,k,\text{E1}} = N_h/m_h$ for $k \in U_h$; $a_{t,k,\text{E2}} = 1/\pi_{at,k}\pi_{t,k|at}\varphi_{k|s_t,\mathbf{m}_s}$; $a_{t,k,\text{E3}} = a_{t,k}$ as defined in expression (17); $\pi_{at,k} = 1$; U_h is the stratum of the sample design of size N_h; m_h denotes the number of sample respondent units of stratum h.

The idea behind the three estimators is based on three different ways to re-weight the respondent units (included or not in the sample). In estimator E1, in which only data for sample units are used, the response probability is homogeneous for all units belonging to a given stratum and this probability is estimated by the response rate in the stratum.

Estimator E2 considers formation of cells to estimate the response probabilities $\hat{\varphi}_{k|s_t}$ by a logistic regression model using only the information in the sample

$$\hat{\varphi}_{k|s_t} = f(c_{k,t-12}, c_{k,t-24}, p_{k,1999}, g_k) \tag{20}$$

where $c_{k,t-12}, c_{k,t-24}$ represent the collaboration of unit k at *month* $t-12$ and $t-24$ (0 if the unit collaborates and 1 otherwise), $p_{k,1999}$ and g_k are the population at 1999 and the geographical repartition. The model is estimated for every month

Table 1. Total number of new dwellings with different estimators:MAPE and Year growth rate – Years 1997–2001

Years	Mean absolute percentage error			Year growth rate			
	E1	E2	E3	True	E1	E2	E3
1997	10.6	11.6	4.2				
1998	9.9	10.6	3.8	−3.0	−3.8	−4.9	−3.5
1999	11.0	11.4	4.3	7.6	8.9	6.8	7.9
2000	15.0	15.4	2.1	13.7	12.3	16.4	11.1
2001	18.8	25.4	2.7	4.1	14.1	19.0	4.4

in the period 1998–2001. Following the strategy proposed in Eltinge and Yansaneh the units are grouped according to their $\hat{\varphi}_{k|s_t}$ value in 6 cells; the conditional probabilities $\varphi_{k|s_t,\mathbf{m}_s}$ are obtained as the cell response rates.

Using the estimator E3, the probabilities $\hat{\varphi}_{k|s_t}$ are estimated by a logistic regression model (20) for the sample units; the estimated parameters of model (20) based on the sample observations are used to calculate the $\hat{\varphi}_{k|s_t}$ values for the municipalities not included in the sample; then are formed 6 weighting cells. The conditional probabilities $\varphi_{k|s_t,\mathbf{m}_s}$ are obtained by cell response rates. A similar approach has been used for the calculation of the probabilities $\varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}}$. For the non-sample units, the probabilities $\hat{\varphi}_{k|\bar{s}_t}$ are estimated using logistic regression model with the same predictors of model (20); the estimated parameters of this logistic model are used to calculate the $\hat{\varphi}_{k|\bar{s}_t}$ values for the municipalities included in the sample; then the conditional probabilities $\varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}}$ are calculated for 6 cells using the Eltinge and Yansaneh strategy.

With reference to the domain of municipalities of the *Second group*, Table 1 shows the Mean Absolute Percentage Error for estimator a and for year w(w =1997, ...,2001) calculated as:

$$MAPE_{a,w} = \frac{1}{12} \sum_{t=1}^{12} \left| \frac{Y_t^{\omega} - \tilde{Y}_{t,a}^{\omega}}{Y_t^{\omega}} \right| 100.$$

The table shows also the growth rate of the annual estimates and the true values obtained as

$$\frac{\tilde{Y}_a^{\omega} - \tilde{Y}_a^{\omega-1}}{\tilde{Y}_a^{\omega-1}} 100; \quad \frac{Y^{\omega} - Y^{\omega-1}}{Y^{\omega-1}} 100$$

where $\tilde{Y}_a^{\omega} = \sum_{t=1}^{12} \tilde{Y}_{t,a}^{\omega}$ and $Y^{\omega} = \sum_{t=1}^{12} Y_t^{\omega}$.

The values of MAPE show a better performance for the E3 estimator in terms of smoothness.

The results in terms of year to year growth rate are quite similar for the period 1998–2000. For 2001, instead, the estimator E3 has a better performance compared to the other two.

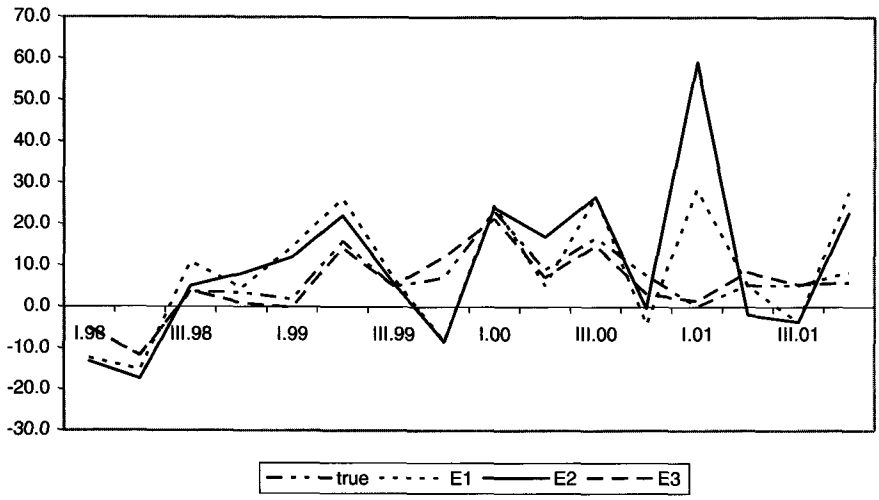

Fig. 1. Quarter to quarter year growth rate for differemt estimator – Year 1998–2001

This evidence is even more clear from quarter to quarter year growth rate for the period 1998–2001 (Fig. 1), for the domain of municipalities of the *Second group*. In particular the estimator E2 and E1 are influenced by data on the number of dwellings for one of the sampling units in the second quarter 2001. The estimator E3, using all the information available at the generic quarter t, is not influenced by this outlier. Estimator E3 gives as result a smoother series that is quite similar to the true series: the correlation between the quarter to quarter year growth rate of the true series and estimator E3 is equal to 0.96 (0.63 for estimator E1 and 0.44 for estimator E2).

To test the sensitivity of the estimator E3 to the number of respondent municipalities, and supposing that the respondents in 2001 are the same as in 2002 (with a reduction of 15%), we estimated the total number of dwellings. The absolute percentage error of the estimation ranges from a minimum of 2.3% for the fourth quarter to 4.7% for the first quarter.

6. Conclusion

Over recent year EU has given growing attention to timeliness in short-term statistics. As result it has been established the ESTEI working group (Expert group on Sampling for Timely European Indicator) that has released a guidelines on best practices for the production of timely indicators. Exploring approaches based on statistic models, in this work we have proposed an estimation method which uses all the information available.

The performance of the proposed estimator is studied with an application to the Italian Survey on building permits in comparison with two different estimators that use only subsample information. The proposed estimator gives a smoother serie quite similar to the true one in terms of mean absolute percentage error.

Technical Appendixes

Appendix 1 – Expected value of estimator $\tilde{Y}_t$

Under models (7) and (8) and supposing that the probability of the events below is zero:

$$A_s : m_c = 0, \text{ for some value of } c = 1, ..., C_s$$
$$A_{\bar{s}} : m_{\ddot{c}} = 0, \text{ for some value of } \ddot{c} = 1, ..., C_{\bar{s}}$$

the expected value of $\tilde{Y}_t$ is given by:

$$\mathrm{E}(\tilde{Y}_t) = \sum_{k \in U_t} y_{t,k} \mathrm{E}(a_{t,k} r^{\delta}_{t,k}),$$

in which

$$\mathrm{E}(a_{t,k} r^{\delta}_{t,k}) = \mathrm{E}_{p_a}(\mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k} | s_t, \mathbf{m}) | k \in s_{at}))) \text{ and}$$
$$r^{\delta}_{t,k} = \begin{cases} 1 \text{ if } k \in r_t \\ 0 \text{ otherwise} \end{cases}$$

where: $\mathrm{E}_{\mathrm{HRG}}(\cdot)$ indicates the expectation under the HRG models (7) and (8); $\mathrm{E}_p(\cdot)$ the expectation under the second phase sampling design, conditioned by the first phase sampling; $\mathrm{E}_{\mathbf{m}}(\cdot)$ denotes the expectation over different values of the vectors $\mathbf{m}_s$ and $\mathbf{m}_{\bar{s}}$; $\mathrm{E}_{p_a}(\cdot)$ the expectation under the first phase sampling design.

Under models (7) and (8) we have

$$\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k} | s_t, \mathbf{m}) = \delta_{t,k} \varphi_{k|s_t,\mathbf{m}_s} a_{t,k} + (1 - \delta_{t,k}) \varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}} a_{t,k}$$

in which $\delta_{t,k}$ is a binary variable that equals 1 if $k \in s_t$. By consequence

$$\begin{aligned} &\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k} | s_t, \mathbf{m}) | k \in s_{at}) = \\ &= (\pi_{t,k|at} \varphi_{k|s_t,\mathbf{m}_s} + (1 - \pi_{t,k|at}) \varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}}) a_{t,k} = \\ &= (\pi_{t,k|at} \varphi_{k|s_t,\mathbf{m}_s}) + ((1 - \pi_{t,k|at}) \varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}}) \cdot \\ &\quad \cdot 1/(\pi_{at,k}(\pi_{t,k|at} \varphi_{k|s_t,\mathbf{m}_s}) + ((1 - \pi_{t,k|at}) \varphi_{k|\bar{s}_t,\mathbf{m}_{\bar{s}}})) = 1/\pi_{at,k}. \end{aligned} \tag{A.1}$$

From the above we have

$$\mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k} | s_t, \mathbf{m}) | k \in s_{at})) = \mathrm{E}_{\mathbf{m}}(1/\pi_{at,k}) = 1/\pi_{at,k}, \tag{A.2}$$

and then

$$\begin{aligned} &\mathrm{E}_{p_a}(\mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k} | s_t, \mathbf{m}) | k \in s_{at}))) = \\ &= \mathrm{E}_{p_a}(\delta_{at,k} \frac{1}{\pi_{at,k}} + (1 - \delta_{at,k}) 0) = \pi_{at,k} \frac{1}{\pi_{at,k}} = 1 \end{aligned} \tag{A.3}$$

By consequence, from the result (A.3), we derive that

$$\mathrm{E}(\tilde{Y}_t) = \sum_{k \in U_t} y_{t,k} \mathrm{E}(a_{t,k} r^{\delta}_{t,k}) = \sum_{k \in U_t} y_{t,k} = Y_t.$$

Appendix 2 – Variance and estimate of variance of estimator $\tilde{Y}_t$

Under models (7) and (8) and under the adopted sampling design, the variance of $\tilde{Y}_t$ may be expressed as:

$$\mathrm{V}(\tilde{Y}_t) = \mathrm{V}_{p_a}[\mathrm{E}(\tilde{Y}_t|s_{ta})] + \mathrm{E}_{p_a}[\mathrm{V}(\tilde{Y}_t|s_{ta})], \tag{A.4}$$

in which we have denoted with: $\mathrm{V}_{p_a}(\cdot)$ the variance of the first phase sampling.

Let us consider first the expectation in the first addendum on the right hand side of the expression (A.4) $\mathrm{E}(\tilde{Y}_t|s_{ta}) = \mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(\tilde{Y}|s_t, \mathbf{m})|s_{at}))$. Using the result (A.2), we have

$$\mathrm{E}(\tilde{Y}_t|s_{ta}) = \sum_{k \in s_{at}} \frac{y_{t,k}}{\pi_{at,k}}.$$

By using the standard results on sampling from finite populations, we have that the first addendum of (A.4) is given by:

$$\mathrm{V}_{p_a}[\mathrm{E}(\tilde{Y}_t|s_{ta})] = \sum_{k \in U_t} \sum_{l \in U_t} (\pi_{at,kl} - \pi_{at,k}\pi_{at,l}) \frac{y_{t,k}}{\pi_{at,k}} \frac{y_{t,l}}{\pi_{at,l}}. \tag{A.5}$$

Let us consider now the variance defined in the second addendum on the right hand side of the expression (A.4). It is straightforward to prove the following result

$$\mathrm{V}(\tilde{Y}_t|s_{ta}) = \sum_{k \in s_{at}} \sum_{l \in s_{at}} \mathrm{Cov}(r^{\delta}_{t,k} a_{t,k}, r^{\delta}_{t,l} a_{t,l}|s_{at}) y_{t,k} y_{t,l}$$

being

$$\begin{aligned}
&\mathrm{Cov}(r^{\delta}_{t,k} a_{t,k}, r^{\delta}_{t,l} a_{t,l}|s_{at}) = \\
&= \mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k} a_{t,l} r^{\delta}_{t,l}|s_t, \mathbf{m})|(k,l) \in s_{at})) + \\
&\quad -\mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k} r^{\delta}_{t,k}|s_t, \mathbf{m})|k \in s_{at})) \cdot \\
&\quad \cdot \mathrm{E}_{\mathbf{m}}(\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{l,k} r^{\delta}_{l,k}|s_t, \mathbf{m})|l \in s_{at})) = \\
&= \mathrm{E}_{\mathbf{m}}(a_{t,k} a_{t,l} \varphi_{k,l|s_{at},\mathbf{m}}) - \frac{1}{\pi_{at,k}} \frac{1}{\pi_{at,l}}
\end{aligned} \tag{A.6}$$

Therefore the second addendum of variance (A.4) may be expressed as

$$\begin{aligned}
&\mathrm{E}_{p_a}\left[\mathrm{V}(\tilde{Y}_t|s_{ta})\right] = \\
&= \mathrm{E}_{p_a}\left(\sum_{k \in s_{at}} \sum_{l \in s_{at}} \left(\mathrm{E}_{\mathbf{m}}(a_{t,k} a_{t,l} \varphi_{k,l|s_{at},\mathbf{m}}) - \frac{1}{\pi_{at,k}} \frac{1}{\pi_{at,l}}\right) y_{t,k} y_{t,l}\right).
\end{aligned}$$

Where $\varphi_{k,l|s_{at},\mathbf{m}} = \Pr((k,l) \in r_t|s_{a,t}, \mathbf{m})$.

In order to derive explicit expressions of the probability $\varphi_{k,l|s_{at},\mathbf{m}}$ let us consider, that: if $k = l$ we have

$$\varphi_{k,k|s_{at},\mathbf{m}} = \Pr(k \in r_t|s_{a,t}, \mathbf{m}) = \varphi_{k,k|s_{at},\mathbf{m}} = \frac{\pi_{at,k}}{a_{t,k}}. \tag{A.7}$$

On the contrary, if $k \neq l$, we have

$$\begin{aligned}\varphi_{k,l|s_{at},\mathbf{m}} &= \Pr((k,l) \in r_t | k \neq l, (k,l) \in s_{a,t}, \mathbf{m}) = \\ &= \mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(r^{\delta}_{t,k} r^{\delta}_{t,l} | k \neq l, s_t, \mathbf{m}) | (k,l) \in s_{at}) = \\ &= \Pr(\mathrm{E}_1 \cup \mathrm{E}_2 \cup \mathrm{E}_3 \cup \mathrm{E}_4 | k \neq l, (k,l) \in s_{a,t}, \mathbf{m}) \end{aligned} \tag{A.8}$$

in which $\mathrm{E}_1, \mathrm{E}_2, \mathrm{E}_3, \mathrm{E}_4$ are four disjoint events described in the following table

Event	Conditional probability of the event $\Pr(E_\alpha \| k \neq l, (k,l) \in s_{a,t}, \mathbf{m})$ with $\alpha = 1, \ldots, 4$
$\mathrm{E}_1 = ((k,l) \in s_t) \cap ((k,l) \in r_t)$	$\pi_{t,kl\|at} \varphi_{k,l\|s_t,\mathbf{m}_s}$
$\mathrm{E}_2 = ((k,l) \in \bar{s}_t) \cap ((k,l) \in r_t)$	$(1 + \pi_{t,kl\|at} - (\pi_{t,k\|at} + \pi_{t,l\|at})) \varphi_{k,l\|\bar{s}_t,\mathbf{m}_{\bar{s}}}$
$\mathrm{E}_3 = ((k \in s_t) \cap (l \in \bar{s}_t)) \cap$ $\cap ((k,l) \in r_t)$	$(\pi_{t,k\|at} - \pi_{t,kl\|at})(\varphi_{k\|s_t,\mathbf{m}_s} \varphi_{l\|\bar{s}_t,\mathbf{m}_{\bar{s}}})$
$\mathrm{E}_4 = ((k \in \bar{s}_t) \cap (l \in s_t)) \cap$ $\cap ((k,l) \in r_t)$	$(\pi_{t,l\|at} - \pi_{t,kl\|at})(\varphi_{l\|s_t,\mathbf{m}_s} \varphi_{k\|\bar{s}_t,\mathbf{m}_{\bar{s}}})$

Using the above results, it is possible to prove that for $k \neq l$, we have

$$\begin{aligned}\varphi_{k,l|s_{at},\mathbf{m}} = (\pi_{t,kl|at} \varphi_{k,l|s_t \mathbf{m}_s}) + ((1 + \pi_{t,kl|at} - (\pi_{t,k|at} + \pi_{t,l|at})) \varphi_{k,l|\bar{s}_t \mathbf{m}_{\bar{s}}}) + \\ + ((\pi_{t,k|at} + \pi_{t,l|at} - 2\pi_{t,kl|at}) \varphi_{k|s_t,\mathbf{m}_s} \varphi_{l|\bar{s}_t,\mathbf{m}_{\bar{s}}}).\end{aligned}$$

Using the above results it is straightforward to prove the expression of first addendum of (3.6), indeed, given s_{at} and $\mathbf{m}$, the product $a_{t,k}\, r^{\delta}_{t,k} a_{t,l} r^{\delta}_{t,l}$ assumes value $a_{t,k} a_{t,l}$ only if $r^{\delta}_{t,k} r^{\delta}_{t,l}$ is equal to 1, otherwise it assumes value 0; by consequence:

$$\begin{aligned}&\mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(a_{t,k}\, r^{\delta}_{t,k} a_{t,l} r^{\delta}_{t,l} | s_t, \mathbf{m}) | (k,l) \in s_{at}) = \\ &\quad = a_{t,k} a_{t,l} \mathrm{E}_p(\mathrm{E}_{\mathrm{HRG}}(r^{\delta}_{t,k} r^{\delta}_{t,l} | s_t, \mathbf{m}) | (k,l) \in s_{at}) = a_{t,k} a_{t,l} \varphi_{k,l|s_{at},\mathbf{m}}.\end{aligned}$$

Estimate of the variance

Under models (7) and (8) and assuming that the probability of events A_s and $A_{\bar{s}}$ is equal to zero, it is possible to derive an unbiased estimator of the variance $\mathrm{V}(\tilde{Y}_t)$ given by:

$$\tilde{\mathrm{V}}(\tilde{Y}_t) = \tilde{\mathrm{V}}_1 + \tilde{\mathrm{V}}_2$$

where $\tilde{\mathrm{V}}_1$ and $\tilde{\mathrm{V}}_2$ represent unbiased estimators of the two addenda on the right hand side of expression (A.4) expressed by

$$\tilde{\mathrm{V}}_1 = \sum_{k \in r_t} \sum_{l \in r_t} \frac{1}{\pi_{at,kl} \varphi_{k,l|s_{at},\mathbf{m}}} (\pi_{at,kl} - \pi_{at,k} \pi_{at,l}) \frac{y_{t,k}}{\pi_{at,k}} \frac{y_{t,l}}{\pi_{at,l}},$$

$$\tilde{\mathrm{V}}_2 = \sum_{k \in r_t} \sum_{l \in r_t} \left(a_{t,k} a_{t,l} \varphi_{k,l|s_{at},\mathbf{m}} - \frac{1}{\pi_{at,k}} \frac{1}{\pi_{at,l}} \right) \frac{y_{t,k} y_{t,l}}{\varphi_{k,l|s_{at},\mathbf{m}}}.$$

The proof that $\tilde{V}_1$ is an unbiased estimator of the first addendum of (A.4), $V_{p_a}[E(\tilde{Y}_t|s_{ta})]$, derive from the following

$$E(\tilde{V}_1) = E\left(\sum_{k\in r_t}\sum_{l\in r_t}\frac{(\pi_{at,kl}-\pi_{at,k}\pi_{at,l})}{\pi_{at,kl}\varphi_{k,l|s_{at},\mathbf{m}}}\frac{y_{t,k}}{\pi_{at,k}}\frac{y_{t,l}}{\pi_{at,l}}\right) =$$

$$= E_{p_a}\left(\sum_{k\in s_{at}}\sum_{l\in s_{at}}\frac{(\pi_{at,kl}-\pi_{at,k}\pi_{at,l})}{\pi_{at,kl}\varphi_{k,l|s_{at},\mathbf{m}}}\cdot\right.$$

$$\left.\cdot\frac{y_{t,k}}{\pi_{at,k}}\frac{y_{t,l}}{\pi_{at,l}}(E_p(E_{HRG}(r^{\delta}_{t,k}\, r^{\delta}_{t,l}|s_t,\mathbf{m})|(k,l)\in s_{at}))\right) =$$

$$= E_{p_a}\left(\sum_{k\in s_{at}}\sum_{l\in s_{at}}\frac{(\pi_{at,kl}-\pi_{at,k}\pi_{at,l})\varphi_{k,l|s_{at},\mathbf{m}}}{\pi_{at,kl}\varphi_{k,l|s_{at},\mathbf{m}}}\frac{y_{t,k}}{\pi_{at,k}}\frac{y_{t,l}}{\pi_{at,l}}\right) =$$

$$= E_{p_a}\left(\sum_{k\in U_t}\sum_{l\in U_t}\frac{(\pi_{at,kl}-\pi_{at,k}\pi_{at,l})\delta_{at,k}\delta_{at,l}}{\pi_{at,kl}}\frac{y_{t,k}}{\pi_{at,k}}\frac{y_{t,l}}{\pi_{at,l}}\right) =$$

$$= \sum_{k\in U_t}\sum_{l\in U_t}(\pi_{at,kl}-\pi_{at,k}\pi_{at,l})\frac{y_{t,k}}{\pi_{at,k}}\frac{y_{t,l}}{\pi_{at,l}} = \quad V_{p_a}[E(\tilde{Y}_t|s_{ta})].$$

Under the assumption adopted in the present work, the proof that $\tilde{V}_2$ is an unbiased estimator of the second addendum of variance (A.4) is based on the following derivations

$$E_p(E_{HRG}(\tilde{V}_2|s_t,\mathbf{m},(k,l)\in s_{at})) =$$

$$= E_p\left(E_{HRG}\left(\sum_{k\in s_{at}}\sum_{l\in s_{at}}(a_{t,k}a_{t,l}\varphi_{k,l|s_{at},\mathbf{m}})\right.\right.$$

$$\left.\left.-\frac{1}{\pi_{at,k}}\frac{1}{\pi_{at,l}}\frac{r^{\delta}_{t,k}r^{\delta}_{l,k}}{\varphi_{k,l|s_{at},\mathbf{m}}}y_{t,k}y_{t,l}|s_t,\mathbf{m},(k,l)\in s_{at}\right)\right) =$$

$$= \sum_{k\in s_{at}}\sum_{l\in s_{at}}(a_{t,k}a_{t,l}\varphi_{k,l|s_{at},\mathbf{m}}) - \frac{1}{\pi_{at,k}}\frac{1}{\pi_{at,l}}y_{t,k}y_{t,l}$$

By consequence it is possible to derive the following

$$E(\tilde{V}_2) = E_{p_a}(E_{\mathbf{m}}(E_{p(E_{HRG}(\tilde{V}_2|s_t,\mathbf{m},(k,l)\in s_{at}))})) =$$

$$= E_{p_a}\left(\sum_{k\in s_{at}}\sum_{l\in s_{at}}\left(E_{\mathbf{m}}(a_{t,k}a_{t,l}\varphi_{k,l|s_{at},\mathbf{m}}) - \frac{1}{\pi_{at,k}}\frac{1}{\pi_{at,l}}\right)y_{t,k}y_{t,l}\right) =$$

$$= E_{p_a}[V(\tilde{Y}_t|s_{ta})].$$

References

Battaglia F, Fenga L (2003) Forecasting composite indicators with anticipated information: an application to the industrial production index. Applied Statistics 52: 3

Bodo G, Signorini L (1987) Short-term forecasting of the industrial production index. International Journal of Forecasting 10: 285–299

Bruno G, Lupi C (2002) Forecasting industrial production and the early detection of turning points. http://enteluigieinaudi.it/cofin98/Rt5-2000.ps

Bell WR, Hillmer SC (1990) The time series approach to estimation for repeated surveys. Survey Methodology 16: 195–215

Box GEP, Jenkins GM (1976) Time series analysis: *forecasting and control.* Holden-Day, San Francisco

Chen J, Shao J (2000) Nearest neighbor imputation for survey data. Journal of Official Statistics 16(2): 113–131

Cochran WG (1977) Sampling techniques. Wiley, New York

De Vitiis C, Falorsi PD, Falorsi S, Russo A (2002) Un analisi comparativa di alcuni metodi di trattamento della mancata risposta totale nella stima delle variazioni lorde nel campionamento ruotato. In Problemi di campionamento nella ricerca sociale, a cura di E. Aureli Cutillo. Università degli Sudi di Roma La Sapienza

Dorfman A H, Valliant R (2000) Stratification by size revisited. Journal of Official Statistics 16: 139–154

Eltinge JE, Yansaneh IS (1997) Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodology 23(1): 33–40

Eurostat (2001) Feasibility study report: country-stratified European sample for the retail trade index. Report presented at 42nd SPC meeting, Luxemburg

Eurostat (2002) Final report of the expert group on sampling for timely European indicators. Luxembourg, October 09, 2002 ESTAT/A4/JMM/ESTEI_Finalreport_EN.doc

Giommi A (1987) Nonparametric methods for estimating individual response probabilities. Survey Methodology 13: 127–134

Harvey AC (1984) Dynamic models, the prediction error decomposition and state-space. In: Hendry DF, Wallis KF (eds) Econometrics and quantitative economics, pp 37–59. Blackwell, Oxford

Little RJA (1986) Survey nonresponse adjustment for estimates of means. International Statistical Review 54: 139–157

Maravalle M, Politi M, Iafolla P (1993) Scelta di indicatori per la stima rapida di un indice provvisorio della produzione industriale. Quaderni di Ricerca. ISTAT, Roma

Niyonsenga T (1994), Nonparametric estimation of response probabilities in sampling theory. Survey Metodology 20(2): 177–184

Rizzo L, Kalton G, Brick MJ (1996) A comparison of some weighting adjustment methods for panel nonresponse. Survey Methodology 22(1), 43–53

Royall RM (1992) On finite population sampling theory under certain regression model. Biometrika 70: 41–50

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effect. *Biometrika* 70: 41–55

Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, Berlin Heidelberg New York

Tam SM (1987) Analysis of repeated surveys using a dynamic linear model. International Statistical Review 55: 63–73

Van Garderen K, Lee K, Pesaran M (2000) Cross-sectorial aggregation of non linear models. Journal of Econometrics 95: 285–331