

Sampling Strategies for Preliminary Estimates Production in Short-Term Business Surveys

*Strategie campionarie per la produzione di stime preliminari nelle indagini
congiunturali sulle imprese*

Michele D'Alò, Claudia De Vitiis, Paolo Righi, Stefano Falorsi

Istat, via Cesare Balbo 16, 00184 Roma, dalo@istat.it; devitiis@istat.it, parighi@istat.it;
stfalros@istat.it

Roberto Gismondi

Istat, via Tuscolana 1788, 00177 Roma, Italy, gismondi@istat.it

Riassunto: la tempestività nella diffusione dei principali indicatori economici congiunturali rappresenta, da tempo, uno dei principali requisiti qualitativi del sistema statistico europeo. D'altra parte, il *trade/off* tra tempestività e qualità delle stime ha animato, nel contesto della statistica ufficiale, diversi filoni di ricerca finalizzati all'identificazione di strategie di stima "anticipata" in grado di ridurre l'errore di revisione. In questo lavoro si propongono due applicazioni, riferite agli indici delle vendite al dettaglio e del fatturato del commercio all'ingrosso, nel cui contesto sono state applicate e commentate tecniche di riduzione dell'errore di stima basate sulla selezione di un campione bilanciato e l'applicazione di diversi stimatori *model-based*.

Keywords: Non-response bias, Provisional estimator, Short-term statistics, Timeliness.

1. Introduction

Among the main components on which the EU statistical definition of quality for short-term statistics is founded (EUROSTAT, 2000), *accuracy* and *timeliness* seem to be the most relevant both for producers and users of statistical data. While accuracy is normally measured by the percent difference between provisional and final estimates, timeliness is measured as the time lag between the reference time point (or the end of the reference period) and the date of data dissemination. However, in recent years, in many fields of official short-term statistics timeliness became the driving issue, both for the increasing demand of users and the need to fill the gap respect to data release standards already achieved by USA and other developed countries. For instance, the first provisional industrial production index is currently released after 12 days from the end of the reference month in the U.S. and after 42 days in the European Union (consolidated indexes of 27 countries). In particular, the Amendment Eu Regulation on Short-Term Statistics (introduced in August 2005) requests all the statistical institutes of the EU Member States to transmit to EUROSTAT provisional short-term indicators with a reduced delay with respect to the timeliness set in the original 1998 Regulation: from 60 to 30 days for the retail trade index, from 90 to 60 days for the quarterly turnover index of other services activities (not including retail trade). Given that, the goal of reducing delay in short-term data release steered theoretical and applied research in the field of official statistics, as widely commented, for instance, in ISTAT (2007).

In the following, in the context of a given sampling survey we mean as *provisional estimate* the estimation of a parameter of interest obtained on the basis of a sub-sample of *quick respondent units* that is available within a time lag Δ'_t after the reference time point t (or end of the reference period) of the survey, while the correspondent *final estimate* is based on a final sample, including both *quick* and *late respondents*, observed within a time lag Δ_t ($> \Delta'_t$). The indicators measuring the statistical quality of a *preliminary estimation method* are based on the differences, evaluated at the different times (identifying the correspondent survey occasions) between provisional estimates obtained by means of the method under study, and the corresponding final estimates. These differences are known as *revision errors*.

The main goal of this paper is discussing possible strategies that could reduce the average revision error affecting the actual provisional estimates of economic short-term indicators. The choice of techniques and of the case studies has been driven by previous works on the same field (Gismondi *et al.*, 2006). Furthermore, the aim of the work is to deal with a wide range of methods that may be utilized in different phases of the data production process. For this reason also statistical techniques that may be applied when defining the sampling plan for the selection of a *Preliminary Theoretical Sample* (PTS), have been considered even if their application is not operationally simple; to apply them Statistical Institutes need to take under control, in advance, the set of units belonging to the preliminary sample. Methods that may be applied during the estimation phase have been considered too, based both on cross-sectional data available at time t and on time series information coming from previous survey occasions. These methods may utilize either a *Planned Preliminary Observed Sample* (PPOS), in the case a PTS has been selected, or alternatively an *Unplanned Preliminary Observed Sample* (UPOS). In paragraph 2 general aspects of preliminary estimation methods are discussed. While paragraph 3 contains main results of an application of the balanced sampling theory for the selection of a PTS to the *Monthly Retail Trade Survey*, in paragraph 4 we propose and compare some preliminary estimation techniques aimed at reducing the average revision in the *Quarterly Wholesale Trade Survey*. In this context a model based approach seems the most suitable way to evaluate and reduce the possible bias due to the non random selection of quick respondents belonging to an UPOS. Some conclusions have been drawn in paragraph 5.

2. Methods working out the problem: what and why

The identification of the optimal *preliminary* estimation and the optimal *final* estimation strategies are always strictly connected problems. Generally speaking, both of them could require the availability of one or more auxiliary variables. Auxiliary variables can be used for the sampling design planning (stratification, evaluation of inclusion or response probabilities), in order to build up and test a super-population model and for carrying out estimates. In Italy, the most part of official short-term statistics are based on fixed panel of enterprises (*industrial production* and *turnover*) or rotating panels with a partial overlap from one year to another (*retail trade*, *service activities*). For all the enterprises in the population, structural variables as number of persons employed and the last year turnover are available from the business register ASIA at the *single unit* level.

Generally speaking, if the strategy used for final estimation is optimal (inside a given family of estimators and according to a sampling or a model based approach), there is not a particular reason justifying the use of a different strategy for preliminary estimation. However, when the final sample does not coincide with the planned theoretical sample – e.g. in the case of final non response – the final sample cannot be known in advance and when provisional estimates are

evaluated it could be necessary to estimate the expected composition and size of the corresponding final sample. If the strategy used for final estimation is not optimal (for instance, because originally it was difficult to identify the true model underlying observed data, or variables in this model were not available, or a final non-response bias occurred), the paradox is that also a provisional estimation technique, that minimizes the average revision error, should probably not be optimal, as well, with respect to the estimation of target parameter.

According to a super-population approach, normally the optimal estimation strategy is based on minimisation of the mean squared error with respect to the model underlying observed data. The main risk of this approach is due to the need to identify the right model; in a late work Hedlin *at al.* (2001) stressed the risk of additional bias due to a model miss-specification even when the asymptotically design unbiased *GREG* estimator is used. These remarks underline the need to test model rightness before carrying out estimates. This kind of study has been carried out for the application showed in paragraph 4.

Finally, provisional estimates could be improved if one could at least partially determine, at the moment of sampling selection phase, the composition of the provisional sample; this result may be obtained on the basis of the *balanced sampling theory* (Royall, 1992), that guarantees model unbiasedness even when the model is not correctly specified. Deville and Tillé (2004) proposed an efficient algorithm for finding a balanced sample; however in the frame of official statistics there are not many applications of this criterion. Paragraph 3 contains an application to the case of the retail trade monthly index.

3. Selection of a preliminary theoretical sample: the Italian Monthly Retail Trade Index context

This paragraph shows the results of a simulation study about the preliminary estimates based on a planned preliminary sub-sample data, hereinafter denoted by PTS, of the overall sample for the final estimate. In the PTS the variables of interest must be collected within time Δ'_t for the construction of preliminary estimates, where Δ_t ($> \Delta'_t$) is the deadline to produce the final estimates. This approach can require some changes in the productive process: the PTS needs to have a high quick response rate making the PPOS as close as possible to PTS. This is one of the main difference due to an approach using the respondents of an UPOS, that is merely formed by the units that are respondent before time $t + \Delta'_t$ without any planned selection process.

The experiment uses artificial data generated from the 2004 Italian Monthly Retail Trade Survey data (MRTS) (Gismondi and De Sandro, 2004). The current estimator of the MRTS index completes the estimation process. The MRTS is based on a monthly measurement of all goods sold by a stratified sample of retail enterprises (Division 52 of NACE nomenclature) of different types and sizes. The sample is composed by a panel and a non-panel component, drawn every year and observed for 12 months. The survey provides provisional estimates within 30 days after the reference time and final estimates within 54 days. The provisional retail trade indices are referred to the domains: *type of sold product* (food and non-food retail enterprises) *type of distribution* (large and small retail enterprises; TD) and, finally, crossing each other these two types of domains of estimate. Let t the reference month (with $t=13, 14, \dots$); the parameters of interest are defined as:

$$I_d^{t,0} = \left(\sum_{h \in d} I_h^{t-12,0} R_h^t \gamma_h \right) / \sum_{h \in d} \gamma_h \quad \text{with} \quad R_h^t = \frac{\sum_{i \in U_h^{t,t-12}} y_{hi}^t}{\sum_{i \in U_h^{t,t-12}} y_{hi}^{t-12}}$$

where: d is the generic domain of interest; h is the generic stratum defined by the cross-classification of some auxiliary variables; $I_h^{t-12,0}$ is the retail trade index of the same month t of the previous year in the stratum h ; γ_h is a stratum weight given by yearly turnover referred to the base 2000, derived from business register ASIA; y_{hi}^t and y_{hi}^{t-12} are respectively the total turnover of the month t and of the same month of the previous year observed on the unit i ; $U_h^{t,t-12}$ the longitudinal population in time period $(t, t-12)$ of stratum h . In the questionnaire of the reference month t both the values of the variables y^t and y^{t-12} are collected. The 2004 sample, considered in the simulation study, has 7,448 units (drawn from a population of 570 thousands), with about the 46,5% belonging to the panel component. Since October 2004, the calculation of the preliminary estimates considers the UPOS within $\Delta'_t=29$ days from the end of the reference month. All the non-respondents within Δ'_t are imputed to obtain the provisional estimates. For each domain of interest the preliminary estimation process is given by

$$\tilde{I}_d^{t,0} = \left(\sum_{h \in d} \tilde{I}_h^{t-12,0} \tilde{R}_h^t \gamma_h \right) / \sum_{h \in d} \gamma_h \quad \text{with} \quad \tilde{R}_h^t = \frac{\sum_{i \in s_{ah(t)}^t} y_{hi}^t + \sum_{i \in (\tilde{s}_h^t - s_{ah(t)}^t)} \tilde{y}_{hi}^t}{\sum_{i \in s_{ah(t-12)}^t} y_{hi}^{t-12} + \sum_{i \in (\tilde{s}_h^{t-12} - s_{ah(t-12)}^t)} \tilde{y}_{hi}^{t-12}}$$

where: $\tilde{I}_h^{t-12,0}$ is the estimate of $I_h^{t-12,0}$; $\tilde{y}_{hi}^t$ and $\tilde{y}_{hi}^{t-12}$ are the imputed values for the non respondents; $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ are respectively the sample units giving information about the variables y^t and y^{t-12} in stratum h in the survey of the month t , such that $s_{ah(t)}^t \subseteq \tilde{s}_h^t$ and $s_{ah(t-12)}^t \subseteq \tilde{s}_h^{t-12}$ being $\tilde{s}_h^t$ the theoretical overall sample for the final estimates in stratum h at month t , while $(\tilde{s}_h^t - s_{ah(t)}^t)$ and $(\tilde{s}_h^{t-12} - s_{ah(t-12)}^t)$ are the corresponding non respondent samples at time $t + \Delta'_t$. In the simulation we use the value y^t and y^{t-12} collected in the questionnaire and we suppose that unit i providing the value of y_{hi}^t gives information on y_{hi}^{t-12} as well; then, $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ coincide and they are indicated by s_{ah}^t . The imputation procedure is defined by two steps:

$$1) \tilde{y}_{hi}^{t-12} = a_{hi}^{t-12} \frac{\sum_{k \in s_{ag}^t} y_{hk}^{t-12}}{\sum_{k \in s_{ag}^t} a_{hk}^{t-12}}, \quad 2) \tilde{y}_{hi}^t = \frac{\sum_{k \in s_{ag}^t} y_{hk}^t}{\sum_{k \in s_{ag}^t} y_{hk}^{t-12}} \frac{a_{hi}^t}{a_{hi}^{t-12}} \tilde{y}_{hi}^{t-12}, \quad (3.1)$$

where: $s_{ag}^t = \bigcup_{h \in g} s_{ah}^t$ represents the sample of respondents belonging to the imputation cell g defined by crossing each other the variables *type of distribution* and class of persons employed (8 cells); a_{hi}^t and a_{hi}^{t-12} are the number persons employed respectively at month t and $(t-12)$ for the unit i , observed in the survey or imputed with values of the business register.

In order to carry out the simulation study, the final sample of 2004 was replaced by an artificial sample (*pseudo-sample*) created by an imputation method (Istat, 2007) for which complete monthly data are available for all the 7,448 units. The procedure of imputation is different from (3.1). The final *pseudo-estimates* based on the final pseudo-sample without non respondents are then calculated following the (3.1) process. They represent the target parameters of the simulation study and the properties of the compared provisional estimates are studied with respect to the pseudo-estimates.

3.1 Definition of the Preliminary Theoretical Sample

The aim of the simulation is to verify the impact of a planned PTS on the preliminary estimates. Therefore, we adopt the same imputation processes defined in (3.1). In particular the left part of (3.1) may be expressed by means of the following superpopulation model

$$y_{hi}^{t-12} = \beta_g a_{hi}^{t-12} + \varepsilon_i \quad (i \in g); \quad E(\varepsilon_i) = 0; V(\varepsilon_i) = \sigma^2 a_{hi}^{t-12}; \quad Cov(\varepsilon_i, \varepsilon_k) = 0. \quad (3.2)$$

On the other hand, the second imputation step can not be expressed via superpopulation model, although a reduction of the imputation error in the first step is important for a good imputation in the second one. Starting from (3.2), different balanced sampling designs for each imputation cell have been considered. The balanced equations are defined in a way such that the estimates of the preliminary sample are equal to the estimates of the final sample in each imputation cell. Adopting a balanced sample, we define in the model based approach a *bias-robust* strategy against the model failure (Valliant *et al.*, 2000), while in the model assisted framework the sampling errors can be reduced (Deville and Tillé, 2004). In particular, we study the bias property in the model based inference.

The selection of the PTS takes into account the possibility that the *working model* (3.2) for cell g is different from the *true underline model*. If that is true, balancing not only on the a_{hi}^{t-12} can preserve the estimates from bias. Then, let us consider the following general model for cell g :

$$\begin{cases} y_{hi}^{t-12} = \eta_0 \mu_{0,g} + \sum_{j \in J} \delta_j \alpha_{j,g} (a_{hi})^j + \sum_{j \in J} \lambda_j \phi_{j,g} (f_{hi})^j + \varepsilon_i, \\ E(\varepsilon_i) = 0; V(\varepsilon_i) = \sigma^2 c_i; Cov(\varepsilon_i, \varepsilon_k) = 0. \end{cases} \quad (i \in g) \quad (3.3)$$

where: a_{hi} is number of persons employed observed in October 2003 for the panel unit i or the number of persons employed in the business register (2002) for the non panel data or for panel non respondents; f_{hi} is the 2003 overall retail turnover for panel unit i and the turnover obtained from the business register (2002) for non panel data; μ , α 's and ϕ 's are regression coefficients; η_0 , δ_j and ϕ_j are zero-one indicators respectively of whether the intercept (constant variable equal to 1) and regressors $(a_{hi})^j$ and $(f_{hi})^j$ are included in the model (3.3); $J = \{0.5, 1, 2\}$. We have then tested some weighted balanced designs satisfying (or approximately satisfying) one or more equations like:

$$\sum_{i \in s_{ag}^t} c_i^{-1/2} (x_{hi})^j / n_{ag} = \sum_{i \in \tilde{s}_g^t} (x_{hi})^j / \sum_{i \in \tilde{s}_g^t} c_i^{1/2}, \quad (3.4)$$

where: x_{hi} indicates a_{hi} or f_{hi} ; c_i may assume alternatively the value a_{hi} or 1; n_{ag} is the preliminary sample size in cell g ; $\tilde{s}_g^t = \bigcup_{h \in g} \tilde{s}_h^t$. The sample selection has been carried out by means of the *cube algorithm* (Deville and Tillé, 2004).

3.2 Results of the Monte Carlo simulation

In the simulation we have considered several sampling designs for selecting a PTS. One thousand Monte Carlo samples (each one with 1,920 units as recommended by Eurostat) have been selected for each sampling design and the *Relative Bias* (RB) have been calculated as

$$RB[\tilde{I}_{d,m}^{t,0}] = \frac{\frac{1}{1000} \sum_{l=1}^{1000} (\tilde{I}_{d,m}^{t,0(l)} - \tilde{I}_{d,f}^{t,0})}{\tilde{I}_{d,f}^{t,0}} \times 100$$

where: $\tilde{I}_{d,m}^{t,0(l)}$ is the estimate of the index at the iteration l with a PTS selected with the sampling design m ; $\tilde{I}_{d,f}^{t,0}$ is the estimate of the index with the overall theoretical sample. We have considered the Simple Random Sampling WithOut Replacement design (SRSWOR), the Probability Proportional to Size (given by a_{hi}) design (PPS) and some balanced sampling designs. For the national domain, table 3.1 compares the RB obtained by a PTS with the RB calculated with the UPOS, really observed in the IMRTI survey, by means of the ratio of the RB's,

$$RRB[\tilde{I}_{d,m}^{t,0}] = \frac{|RB[\tilde{I}_{d,m}^{t,0}]|}{|RB[\tilde{I}_{d,UPOS}^{t,0}]|} - 1.$$

The table shows the results of the balanced sampling designs that have had the best performance and the results of PTS composed by the largest retail enterprises for cell g in terms of turnover as well. For the PTS the allocation in each cell g has been accomplished according to the Neyman method with respect to the variable a_{hi} . We suppose that all the units in PTS respond. In the table the balanced samplings are denoted by the zero-one indicator variables of the model (3.3) that are equal to 1. The survey months of October, November and December have been analysed for which size of the UPOS has been respectively of 2,068, 2,520 and 2,283 units.

Table 1: *Ratio of the RB of PTS with respect to the UPOS in the Monte Carlo simulation*

Designs	c_i	RRB October	RRB November	RRB December
Largest units	-	1.10	1.23	-0.85
SRSWOR	1	-0.36	1.02	-0.48
PPS	a_{hi}	1.79	-0.49	-0.86
$\eta_0, \lambda_{0.5}, \lambda_1, \lambda_2$	1	-0.65	1.42	-0.43
$\eta_0, \delta_{0.5}, \delta_1, \delta_2, \lambda_{0.5}, \lambda_1, \lambda_2$	1	-0.42	1.25	-0.48
$\eta_0, \delta_{0.5}, \delta_1, \delta_2$	a_{hi}	2.02	-0.68	-0.94
$\eta_0, \delta_{0.5}, \delta_1, \delta_2, \lambda_{0.5}, \lambda_1, \lambda_2$	a_{hi}	1.95	0.09	-0.91

The table shows that the bias of the estimates depend on term c_i of the (3.4) and, in particular, in October the designs perform better with $c_i = 1$, while in the other two months the designs using $c_i = a_{hi}$ lead to the best results. Moreover, given the best choice of c_i , generally the balancing decreases the bias of the estimates: that occurs in October balancing on the intercept and on the variables $(f_{hi})^{1/2}$, f_{hi} and $(f_{hi})^2$ (RRB=-0.65), while, replacing $(f_{hi})^{1/2}$, f_{hi} and $(f_{hi})^2$ with $(a_{hi})^{1/2}$, a_{hi} and $(a_{hi})^2$, in the other two occasions, the minimum RRB is achieved (RRB=-0.68 in November, RRB=0.94 in December). Finally, it is important to note that the RRBs have been obtained being the size of PTS always smaller than the UPOS size.

4. Use of model based estimators in the wholesale trade survey

Starting from the first quarter of 2001, ISTAT elaborates and releases 8 (referred to the 7 economic activities, according to classification NACE Rev.1, activities from 51.1 to 51.9 and the total) quarterly index numbers (with base 2000=100) concerning turnover of the wholesale trade and commission trade sector.

The survey is based on a stratified random sample including about 7.500 units with a yearly rotation of about 2.000 units, where strata are obtained combining 9 economic activities, 3 employment classes (1-5; 6-19; >19) and 3 geographic areas (North, Centre, South). A stratum index is currently calculated for each of the 81 strata. Calculations of higher order indexes – among which the total wholesale trade one – are based on weighted means of lower order indexes, where weights refer to 2000 and were based on turnover data drawn from structural business statistics. A peculiarity concerns the technique used to obtain each turnover index with base 2000=100: in order to better manage wave non-response, first the ratio between average turnovers referred to quarters t (the actual one) and $(t-4)$ is calculated; secondly, it is multiplied by the index number with base 2000=100 referred to the quarter $(t-4)$ and calculated one year before. This option is guaranteed by the presence, in each quarterly questionnaire, of questions concerning turnover of both quarters t and $(t-4)$. So, in order to guarantee quality of estimates at each occasion it is sufficient to control the turnover ratio $t/(t-4)$. At the moment, available time series of final indexes, released after 180 days from the end of the reference quarter, and provisional indexes after 60 days are available for the 14 quarters from I-2003 to II-2006. The provisional indexes are obtained using the same estimator as that utilized for the final estimate, in the following referred as *standard preliminary estimator*, *SPE*.

In order to improve the quality of preliminary estimates, some model based estimation techniques have been studied with the aim to reduce the revision error. An application of these methods has been carried out for the construction of preliminary estimates for the 14 available survey quarters. These methods can be classified into two types, A and B. *Type A* preliminary estimators are based on cross-sectional models specified on elementary ratios, obtained for each unit by dividing the value of the interest variable at time t by the corresponding value at time $(t-4)$, being t the reference quarter (with $t=5, 6, \dots$). *Type B* preliminary estimators utilize autoregressive models on the time series of final estimates and revision errors, in order to correct preliminary estimates utilising the autocorrelation structure of revision errors.

4.1. Type A estimators

In order to describe *type A* provisional estimators, let's introduce for each stratum h ($h=1, \dots, H$) the stratum ratio:

$$R_h^t = \frac{Y_h^t}{Y_h^{t-4}} = \sum_{i \in U_h^{t,t-4}} r_{hi}^t q_{hi}^t \quad (4.1)$$

where $U_h^{t,t-4}$ is the longitudinal population, of dimension N_h^t , of units belonging to stratum h in period $(t-4, t)$, Y_h^t and Y_h^{t-4} are the stratum totals of the target variable y at times t and $t-4$, $r_{hi}^t = y_{hi}^t / y_{hi}^{t-4}$ is the *elementary ratio*, being y_{hi}^t e y_{hi}^{t-4} the y -values of unit i in stratum h at

times t and $t-4$, q_{hi}^t is the weight for unit hi observed at time t with respect to stratum total at time $(t-4)$, that is $q_{hi}^t = y_{hi}^{t-4} / Y_h^{t-4}$. After the sample s_h^t of dimension n_h^t is observed, the stratum ratio can be written as sum of ${}_s R_h^t$ and ${}_{\bar{s}} R_h^t$, with the subscripts of s and $\bar{s}$ corresponding to sample and non-sample population units respectively. These subscripts will be used below to denote conformable partitions of the quantities involved in the estimation process; thus formula (4.1) is expressed conformably as

$$R_h^t = {}_s R_h^t + {}_{\bar{s}} R_h^t = \sum_{i \in s_h^t} r_{hi}^t q_{hi}^t + \sum_{i \in \bar{s}_h^t} r_{hi}^t q_{hi}^t.$$

A general expression for the Best Linear Unbiased Predictor (BLUP) of R_h^t is

$$\tilde{R}_h^t = {}_s R_h^t + {}_{\bar{s}} \tilde{R}_h^t = \sum_{i \in s_h^t} r_{hi}^t q_{hi}^t + \sum_{i \in \bar{s}_h^t} \tilde{r}_{hi}^t q_{hi}^t \quad (4.2)$$

where $\tilde{r}_{hi}^t$ is the predicted value of r_{hi}^t obtained by fitting a linear model.

Formula (4.2) refers to the final estimator and the s_h^t is the complete set of n_h^t observed units, while $\bar{s}_h^t$ is the set of $N_h^t - n_h^t$ non-observed population units. This formula can be utilized also to define preliminary estimator; in this case s_h^t should be substituted with the set s_{ah}^t composed of $n_{ah}^t \leq n_h^t$ anticipated observed units, while $\bar{s}_h^t$ may represent different sets of non-observed units. This set, in fact, may represent the $N_h^t - n_{ah}^t$ non-observed population units, leading to the definition of *non conditional estimators*, or alternatively the $n_h^t - n_{ah}^t$ late respondent units.

In the latter case, for each time t , the estimator depends on the values of y_{hi}^{t-4} referred to the late observed units and on the size of late respondent set; for these reasons the correspondent estimator can be defined as *conditional BLUP*. These last quantities are unknown at the moment of production of preliminary estimate, so that an evaluation of them is necessary when using this approach. In the case of *non conditional BLUP* for final and preliminary estimators the weights q_{hi}^t (for $i \in s_h^t$ and $i \in \bar{s}_h^t$) are unknown and depend on the stratum total Y_h^{t-4} ; they may be estimated by $\tilde{q}_{hi}^t = (y_{hi}^{t-4} / y_h^{t-4}) Q_h$, where y_h^{t-4} is the stratum sample total of y at time $(t-4)$ based on final or quick sample units and $Q_h = (X_h / X)$ is the population weight of stratum h with respect to the auxiliary variable x . For the *conditional BLUP* estimator, the weights are unknown too - depending on sample total y_h^{t-4} evaluated, as discussed below, on the n_h^t units of the final observed sample – and are calculated as $q_{hi}^t = 1 / y_h^{t-4}$.

The general form of the model, following Valliant (1999) is

$$r_{hi}^t = \beta_g^t + \varepsilon_{hi}^t, \quad h \in g \quad (g=1, \dots, G) \quad (4.3)$$

in which subscript g denotes a generic stratum aggregation, called *model group*, and the random variables ε_{hi}^t are assumed to be mutually independent with $E(\varepsilon_{hi}^t) = 0$ and

$\text{Var}(\varepsilon_{hi}^t) = \sigma^2 c_{hi}$. It is possible to define different models according to the *model group* choice and the specification of hetheroscedasticity factors c_{hi} ; so we denote the generic model as $M_\xi = M[g, c_{hi}]$. It is important to note that under model $M_0 = M[g \equiv h, c_{hi} = 1/y_{hi}^{t-4}]$ the *unconditional* BLUP $\tilde{R}_h^t$ reduces to the current estimator based on stratum ratios, used for both final and provisional estimates.

The general expression of the *conditional* BLUP of the long-term index of change $I^{t,0}$ from the base period 0 to time t based on model M_ξ is given by the following composite estimator

$$\tilde{I}_{M_\xi}^{t,0} = \sum_{g \in G} \sum_{h \in g} \hat{I}_h^{t-4,0} \left\{ f_h^t r_h^t + (1 - f_h^t) \sum_{h \in g} r_h^t \xi C_h \right\} \quad (4.4)$$

in which $\hat{I}_h^{t-4,0}$ is the final estimate of long-term index of change $I_h^{t-4,0}$ from the base period 0 to time $t-4$, referred to stratum h , while r_h^t , f_h^t and ξC_h are respectively expressed by:

$$r_h^t = \sum_{i=1}^{n_{ah}^t} y_{hi}^t / \sum_{i=1}^{n_{ah}^t} y_{hi}^{t-4}, \quad f_h^t = \sum_{i=1}^{n_{ah}^t} y_{hi}^{t-4} / \sum_{i=1}^{n_h^t} y_{hi}^{t-4} \quad \text{and} \quad \xi C_h = \sum_{i=1}^{n_{ah}^t} \xi c_{hi} / \xi C_g,$$

being $\xi C_g = \sum_{h \in g} \sum_{i \in s_{ah}^t} \xi c_{hi}$. We note that in (4.4) f_h^t (≤ 1) is the weight of direct component

of estimator while $(1 - f_h^t)$ weighs the synthetic component referred to model group g . Because f_h^t depends on the n_h^t values observed in the final sample, to evaluate the preliminary estimation we could (method-a) predict the $n_h^t - n_{ah}^t$ values y_{hi}^{t-4} for the late respondent units or (method-b) predict f_h^t , under specific hypothesis on the time series of the previous f_h^t values.

In the application, six alternative preliminary estimators, Pre1- Pre6, have been obtained on the basis of models $M_\xi = M[g, \xi c_{hi}]$ ($\xi = 1, \dots, 6$), defined by different choices of hetheroscedasticity factors c_{hi} ; in fact the *model group* g is the same for all the six models and has been defined studying the longitudinal stability of the fit of several different models on the available survey data. The chosen model group g ($g=1, \dots, G=27$) is defined combining the 9 economic activities and the 3 geographic areas.

The different factors ξc_{hi} depend on quantities known at unit level, like the enterprise turnover y_{hi}^{t-4} , and/or stratum level turnover totals, X_h^α , available from ASIA business register for reference year α . The first three choices for c_{hi} 's utilize the ASIA turnover at the base year 0 (year 2000) and are expressed as $_1c_{hi} = 1/(y_{hi}^{t-4} X_h^0)$, $_2c_{hi} = 1/X_h^0$ and $_3c_{hi} = 0.5(_2C_g - _1C_{hi} + _1C_g - _2C_{hi})/(_2C_g - _1C_g)$; furthermore, the corresponding factors using

ASIA turnover at a more recent year α (year 2004) have been defined as: ${}_4c_{hi} = 1/y_{hi}^{t-4} X_h^\alpha$, ${}_5c_{hi} = 1/X_h^\alpha$ and ${}_6c_{hi} = 0.5({}_5C_g - {}_4c_{hi} + {}_4C_g - {}_5c_{hi})/({}_5C_g - {}_4C_g)$.

In order to estimate f_h^t in (4.4), it has been adopted the method-B through $f_h^t = \bar{f}_h = \sum_{\tau=1}^{t-1} f_h^\tau$ ($\forall t$).

The empirical results of the application of the type A methods to the quarterly data on wholesale trade are showed in paragraph (4.3) together with results obtained by means of type B methods.

4.2. Type B methods

Type B methods are based on a time series approach, as proposed by Rao *et al.* (1989), using the set of final and preliminary estimates available at time t . Under the hypothesis that the final estimate and the measurement errors follow a Markov process, and writing a state space formulation of models, optimal predictors of final estimates are obtained using the Kalman Filter. In particular, let I_p^t, I^t and $^*I^t = I^t - I_p^t$ be respectively the preliminary estimate at time t , the final estimates and the preliminary estimates' errors at time t , $t = 1, \dots, T$. Ignoring sampling errors, it is possible to obtain the following final preliminary estimator (TS1) for the period $(t+1)$ by means of Kalman filter (see for instance Harvey, 1984):

$$\hat{I}^{t+1} = \alpha(\phi I^t) + (1 - \alpha)(I_{p+1}^t - \psi^* I^t) \quad \text{where } \alpha = \sigma^2 / (\sigma^2 + \sigma_0^2).$$

The preliminary estimate can be viewed as a weighted average of the final estimate of the previous period t and the preliminary estimate for time $(t+1)$, adjusted for the previous estimate error.

Modelling the true value of the parameter of interest $\tilde{I}^t$ rather than the final estimates, i.e.

considering the sampling error u_t , and putting $\sigma_t^2 = \phi^2 V(\hat{\tilde{I}}^t - \tilde{I}^t) + \sigma^2$, a second preliminary estimator (TS2) is given by:

$$\hat{I}^{t+1} = \alpha(\phi \hat{\tilde{I}}^t) + (1 - \alpha)(I_{p+1}^t - \psi^* I^t) \quad \text{where } \alpha = \sigma_0^2 / (\sigma_t^2 + \sigma_u^2 + \sigma_0^2).$$

Standard estimation methods can be used to estimate the parameters of the models, in each domains of interest or at national level.

Type B estimators have been considered either in the case in which revision errors are defined by means of standard preliminary estimator (SPE) and in the case when errors are obtained making use of *type A* methods. In the first case it is possible to verify the performance of *type B* methods conditional to the use of SPE as cross-sectional preliminary estimator (in the following denoted as *type B*/_{SPE} methods). In the second case, it is possible to evaluate the performances of *type B* provisional estimators conditional to *type A* methods as cross-sectional preliminary estimators (this methods will be named *type B*/_A preliminary estimators)

4.3. Main results of the application

The above methods have been applied to the available data of the Wholesale Trade survey referred to the 14 quarters from I-2003 to II-2006. To evaluate the performance of the proposed estimators, the absolute relative revision error (ARE) with respect to the corresponding final estimate, at aggregate and at domain level have been calculated for each quarter. A synthesis of the outcomes has been obtained by means of the mean of the AREs on the 14 quarters (MARE). In Table 2 the MARE for the national estimates are reported, for the two type B methods TS1 and TS2, applied by using as preliminary series the estimates obtained through the type A predictors Pre1-Pre6 and the SPE. TS1 and TS2 have been implemented by estimating the parameter both at the single domain level and for the total level. In the table the first column reports the performances of type A estimators only, the first row the results of *type B*_{SPE} methods; the other cells give the *type B*_A outcomes.

Table 2: Percentage revision errors of Type A and B estimators at national level

Type A estimator	Revision Error of type A estimator	Domain level parameters		General level parameters	
		Rev. Err. TS1	Rev. Err. TS2	Rev. Err. TS1	Rev Err. TS2
SPE	0,287	0,273	0,276	0,265	0,272
Pre1	0,232	0,236	0,242	0,244	0,250
Pre2	0,305	0,276	0,279	0,274	0,290
Pre3	0,230	0,251	0,254	0,244	0,247
Pre4	0,278	0,245	0,259	0,263	0,290
Pre5	0,244	0,270	0,273	0,249	0,249
Pre6	0,198	0,234	0,238	0,236	0,236

The predictor producing the lowest revision error is the Pre6 (0,198). TS1 e TS2 lead always to an improvement when using the SPE (*type B*_{SPE}, in the first row), but only in some cases when using a type A estimate as preliminary series (*type B*_A). Therefore, at the aggregate level the use of type A predictors is alternative to time series approach.

As further analysis, the MARE have been calculated at domain level averaging over the 14 quarters; the results, here not presented for shortness, show that type B methods lead in general to a reduction of revision error. Among type A methods only Pre5 and Pre6 lead to a lower revision error then SPE.

The empirical application pointed out rged that the more the series were long the more the results of the type B methods improved, due to the fact that the parameter estimation is more accurate.

5. Main conclusions

In this paper some techniques to estimate the provisional indices in short-term business surveys have been formalised and compared. Efficiency of the provisional estimation strategy should always be linked to the final estimation one, whenever in some empirical contexts the final estimation is not carried out using the true best estimator: it could be due to the not availability

of any auxiliary variable useful to implement alternative strategies, or to a too poor correlation with the variable under study.

In particular, starting from standard methods applying the same technique utilized for the final estimate, in this work advanced methods have been considered for sample selection and estimation. The empirical studies show the good performances of proposed methods when applied to a couple of relevant short-term surveys conducted by Istat.

Future enhancements may concern the extension of elementary ratios models also in the sample balanced design and the introduction of a random effect in the unit level model of type A, in order to consider longitudinal correlation between indices.

References

- Eurostat (2000), *Short-term Statistics Manual*, Luxembourg, Eurostat.
- Eurostat (2005), *Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council – Unofficial Consolidated Version*, documento non pubblicato, Eurostat, Lussemburgo.
- D'Alò M., Gismondi R., Solari F., Naccarato A. (2006), "Estimation in Repeated Business Surveys using Preliminary Sample Data", *Atti della XLIII Riunione Scientifica SIS*, 14-16 giugno, Torino.
- De Sandro L., Gismondi R., (2004) Provisional Estimation of the Italian Monthly Retail Trade Index, *Contributi-Istat*, 24/2004.
- Harvey A. C. (1984), "Dynamic Models, the Prediction Error Decomposition and State-space", in *Econometrics and Quantitative Economics*, D. F. Hendry and K. F. Wallis (eds.), pagg.37-59. Blackwell, Oxford.
- Hedlin D., Falvey H., Chambers R., Kokic P. (2001), "Does the Model Matter for GREG Estimation? A Business Survey Example", *Journal of Official Statistics*, 17, 4, 527-544.
- ISTAT (2007), Rapporto Finale del gruppo di lavoro: "Sperimentazione di stime anticipate per specifici indicatori congiunturali, finalizzata al rilascio in produzione delle relative metodologie", Direttiva Istat TRAC16, *Documento interno*.
- Rao J.N.K., Srinath K.P., Quenneville B. (1989), "Estimation of Level and Change using Current Preliminary Data", in *Panel Surveys*, (Kasprzyk, Duncan, Kalton G., Singh eds.), 457-485, John Wiley & Sons, New York.
- Royall R.M. (1992), "Robustness and Optimal Design Under Prediction Models for Finite Populations" *Survey Methodology*, 18, 179-185.
- Valliant R. (1999). Uses of Models in the Estimation of Price Indexes: a Review. In *Proceedings of the Survey Research Methods Section*, American Statistical Association.