

Estimation in Repeated Business Surveys Using Preliminary Sample Data

*La Stima Anticipata in Indagini Ripetute sulle Imprese
mediante Informazioni Campionarie Parziali*

Michele D'Alò Roberto Gismondi Fabrizio Solari
Via Cesare Balbo 16, 00184 Roma gismondi@istat.it

Alessia Naccarato
Università di RomaTre, Via Ostiense 169, 00154 Roma

Riassunto: L'esigenza di aumentare la tempestività di diffusione degli indicatori utilizzati per individuare l'andamento del ciclo economico e delle principali variabili obiettivo della politica economica, oltre a dar luogo a una domanda crescente da parte degli utilizzatori, viene posta con forza dalle istituzioni europee. Questo lavoro trae la sua origine dagli studi avviati nell'ambito di un progetto di ricerca istituito all'interno dell'ISTAT al fine di approfondire alcune metodologie statistiche utili ad ottenere stime preliminari nelle indagini congiunturali condotte dall'Istituto ed a verificarne empiricamente la loro validità.

Keywords: preliminary estimation, time series data, short-term business survey.

1. Introduction

Timeliness of statistical information, i.e. the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available, is a key issue for improving survey process quality (Rosen and Elvers, 1999). There may be, of course, an interplay and/or a trade-off between timeliness and other dimensions of data quality, such as accuracy and coherence, and finding a balance among these factors is a critical aspect for yielding accurate short-term statistics (Bier and Ahnert, 2001; Papageorgiou and Vardaki, 2002). Timeliness is a particularly critical component of quality for producing short-term business statistics at National Statistical Institutes (NSIs) of European Community, as each Member State has to meet the standard quality requirements of the *Regulation No 1165/98* – amended by the *Regulation No 1158/2005* – about terms for transmission of the results and details of the information provided on short-term statistics. More precisely, the amended regulation provides for a substantial improvement of timeliness for the production of some important indicators disseminated at present by Istat. Besides reducing from 45 to 40 days the dissemination delay of the industrial production index, relevant improvement in terms of timeliness are requested for some of the main short-term business indicators: quarterly turnover indicators concerning the “other market services” (from 90 to 60 days), the index of production in construction (from 60 to 45 days), the quarterly employment rate (from 90 to 60 days). Increasing pressure on the

EU Member States for producing timely indicators also stems from financial market analysts, governments and the European Central Bank.

In repeated surveys, optimization of data collection and processing may not be enough in order to attain the requested timeliness in terms of final estimates of the indicators of interest. Therefore, it is usual practice for NSIs to produce *preliminary* estimates before the time settled for the dissemination of the final estimates. The preliminary estimation error, that is the discrepancy between preliminary and final estimates, is generally considered an important indicator of accuracy, representing the component of total survey error due to the partial amount of information available for the releasing of preliminary estimates. In order to reduce this error component, we need to develop statistical methods allowing to control both variability and bias related to preliminary estimation process. In this way, preliminary estimates (diffused in one or more occasions) are expected not to differ substantially from final estimates. This explains the reason why NSIs are spending increasing resources to research projects addressed to the analysis of theoretical and empirical aspects related to the problem of preliminary estimation.

In particular, Istat has recently carried out some research projects aimed at studying and comparing different preliminary estimation methodologies and at the same time evaluating the empirical properties of the proposed methods in the context of the short-term business surveys.

2. Informative Context for Preliminary Estimation

For each survey, the standard process from collection to elaboration of survey data requires to be accomplished in a fixed period of time in order to disseminate final estimates of the parameters of interest at the settled time t . In this context, direct estimators of the target parameters – based on the sampling units included in the *Theoretical Sample* (TS), selected by a probabilistic sampling design – are approximately design unbiased; the sampling error depends on the variability of the phenomenon under study, the planned sample size, effectiveness of the selection procedure and calibration method defining the adopted sampling strategy. Direct estimators based on the *Observed Sample* (OS) – that is a subset of TS whose size depends on the total non response rate – can be affected by bias effects – not easily valuable – depending on the random process generating the OS. Furthermore, sampling errors increase as the planned sampling size decreases and modify as inclusion probabilities of the sampling units in OS change for the non response mechanism.

We assign the term “preliminary” to the estimates obtained on the basis of statistical information available at time t' preceding time t . More precisely, at time t' we can observe only a subset of OS denoted as *Preliminary Sample* (PS). A common practice in this situation is to apply to PS data the same estimation techniques utilised to produce the final estimates. This strategy represents, often, the most straightforward solution to obtain preliminary estimates and from now on it will be denoted as *standard*.

Let us consider the case when the adopted standard solution is the direct estimator: in comparison with the final direct estimator based on the OS data, the preliminary direct estimator is affected by higher variance since the lower sample size. Furthermore, the preliminary direct estimator may systematically depart from the final direct estimator due to the different mechanism generating non response in OS and in PS. For this

reason, direct estimates based on preliminary data may differ substantially from those available when all data have been processed and, hence, suitable estimation methodologies are to be adopted in order to obtain good preliminary estimates.

As a matter of fact, the main theoretical problem to be faced in a short-term provisional estimation context concerns the possible *self-selection* of quick respondents, that can lead to biased estimates of the unknown population mean and variances (Drudi and Filippucci, 2000). As underlined by Bolfarine and Zacks (1992), the question of robustness of predictors of population quantities could be faced using 3 approaches: 1) imposing restrictions (often unrealistic) to the possible superpopulation models adopted; 2) imposing restrictions to the samples to be selected; 3) using Bayes predictors that adaptively consider the possibility that each one out of a series of alternative models is the correct model.

In the context of short-term business surveys managed by Istat - planned for estimating parameters such as indexes and their changes over time - one common method is based on the evaluation, for each design stratum, of the direct estimator of the index on TS data, imputing the missing responses for the sampling units belonging to TS but not to OS. In this case the use of standard strategy results into imputing the sampling units belonging to TS but not to PS. Another type of procedure utilises the direct estimates of the design stratum indexes without imputation of the missing responses either in OS or in PS. The former strategy is adopted for the dissemination of preliminary and final estimates of the retail trade monthly turnover index (Retail Trade Survey), while the latter is currently applied for the wholesale quarterly turnover index (Wholesale Trade Survey).

When standard strategies are applied to the production of preliminary estimates, since they do not properly take into account the specific problem of production of good preliminary estimates, their quality may be negatively affected. Nonetheless, it is worthwhile to note that in many applications standard strategies may produce preliminary and final estimates similarly biased. In this situation, the use of alternative estimation techniques, although reducing bias, may produce a difference between final and preliminary estimates (*revision error*) larger than that observed for standard strategies.

Even though many studies about preliminary estimation have been discussed in literature especially for the solution of particular problems, it is not available a general and comprehensive account of the methods and theory of preliminary estimation. In the following we suggest a taxonomy of preliminary estimation techniques that may help to this aim.

A first criterion of classification pertains the type and the nature of information utilised for preliminary estimation. We can distinguish among two main classes:

- *survey information*, including two distinct cases: in the first case it is represented by the observed values of the variable of interest for each sampling unit at the previous survey occasions (*disaggregated survey information*); in the second case the information is constituted by the time series of preliminary and final estimates computed at the previous times (*aggregated survey information*). It is worthwhile to note that disaggregated information is available only in case of panel surveys or rotated surveys;
- *administrative register information* (the business register ASIA in the Italian case) from which we can draw auxiliary information on variables supposed to be correlated with the variable of interest either at unit level (*disaggregated*

administrative register information) or at aggregations of units level (*aggregated administrative register information*), e.g. totals computed over PS, OS, TS or the whole population.

A second classification criterion considers the phase of the statistical process of data production in which the methods are applied. According to this criterion, it is possible to identify methods acting:

- at the sampling design phase based on the selection of a preliminary sub-sample of TS;
- at the estimation phase on the units in PS; these methods include:
 - imputation techniques of missing data, that are applied to the non respondent units in TS but not in PS;
 - weighting adjustment modifying the sampling weights assigned to the units in PS in order to take into account non respondents in TS;
 - direct and indirect estimators utilising known population totals of auxiliary variables and/or time series of preliminary and final estimates of the variable of interest.

It is useful to relate each other the classes identified by the two classification criteria given above. The techniques based on the selection of a preliminary sample and the methods requiring imputation and weighting adjustment are generally based on *unit level* models. These models use disaggregated auxiliary information coming from survey data at previous times and/or administrative register data. For the methods in the last class the relation between the variable of interest and the auxiliary variables is usually formalised by *domain level* models in which the auxiliary information is expressed in terms of domain known totals or estimates.

The planning of a preliminary sub-sample is realistic only in the case in which suitable resources are available for the optimization in terms of time spending of data collection and processing on a sub-sample of TS. In this case, it seems natural to choose the sub-sample in TS by means of a sampling design as efficient as possible starting from the auxiliary information available for the units in TS. A solution comes from *balanced sampling*, which is based on the selection of a sub-sample in TS by means of a controlled probabilistic algorithm. This algorithm allows to satisfy equality constraints between preliminary estimates of one or more auxiliary variables totals and the correspondent estimates based on the sampling units in TS, so that representation with respect to final estimates of the variable of interest is guaranteed. The preliminary sample selected by means of balanced sampling techniques is denoted with *Balanced Preliminary Sample* (BPS).

The methods adopting imputation techniques are able to re-construct the values of the variable of interest for all the non respondents, while weighting adjustment is based on calibration techniques able to modify the respondents sampling weights in order to reduce bias effects due to missing responses in TS. When dealing with longitudinal surveys, information available for the same units in the previous surveys can be used in the imputation models or in the models formalizing the response probability mechanism.

Similarly to the previous classes defined above, the last class comprises a wide range of design-based and model-based methodologies. Among these, in the next section we will describe some important techniques adopted successfully by Istat to obtain preliminary estimates in many short-term business surveys.

3. Empirical Studies

The empirical studies carried out by Istat aimed to assess the opportunity of introducing methodologies alternative to standard techniques, validating standard techniques in case of no better alternative methods.

In most applications, the precision of estimates is evaluated by means of synthetic indicators allowing to compare the performances of alternative techniques for preliminary estimation. One of the most common quality indicators is the *Mean Absolute Percentage Error* (MAPE) obtained averaging over the time (and, if the case, over the number of simulated replications that could be carried out on a given set of data) the absolute relative differences between preliminary and final estimates computed on OS. In the most part of the empirical studies conducted by Istat the performances of the preliminary estimation methods have been compared in terms of MAPE; nonetheless, together with MAPE, in some experimental studies some alternative indicators have been introduced in order to assess some bias components of preliminary and final estimators. In particular, bias of preliminary estimates can be divided into three components based on the expected values of the following differences: the difference between preliminary estimates and final estimates computed on OS, the difference between final estimates computed on OS and TS and finally the difference between the final estimates computed on TS and the parameters of interest. The information arising from the ASIA register on total yearly turnover (available for each unit in the target population), highly correlated with the target variable, has been used in order to obtain the values of the final estimates on TS and the values of the corresponding parameters of interest.

An important case in which we attempted to evaluate some of the bias components introduced above is the study carried out on the Retail Trade Survey aimed at the analysis of the methods for the selection of a BPS. A simulation exercise has been undertaken in order to evaluate MAPE indicator with respect to the final estimates computed on TS, that is an estimate of the sum of the first two bias components. To this aim, data derived from the business register ASIA and survey data available for previous times have been utilised in order to create a *pseudo*-TS by means of re-sampling techniques. First results showed that recourse to a BPS could reduce MAPE even in a context when the actual provisional estimation errors are yet quite low.

Another similar study is given by the application of missing response adjustment methods to the Wholesale Trade Survey. In this context, final definitive estimations are released after 180 days from the end of the reference quarter, while provisional estimations are normally calculated after 60 and 90 days. In the study, a comparative analysis among different methods for preliminary estimation has been carried out. In this case for each method has been evaluated, also, the last component of bias averaging the differences between the final estimates on TS and the corresponding parameters of interest. Moreover, an undergoing study – related to 2004 and 2005 quarterly data – is evaluating the different effect played on the average revision error by three main factors: 1) the type of quick estimator used and the link between *preliminary* and *final* estimators; 2) the treatment of outlier observations (inclusion or exclusion from calculations); 3) the estimation of the probability that a unit not included in the PS will be part of the final OS.

A detailed description of the experimental studies is given in ISTAT (2005), while Table 1 displays synthetically which preliminary estimation methods have been adopted for each of the different short-term surveys considered in the studies.

Table 1. Preliminary estimation methods applied in the ISTAT short-term surveys.

<u>Methods</u>	<u>Survey</u>					
	<u>Wholesale Trade</u>	Retail Trade	Industrial Turnover	Maintenance and Repair of Motor Vehicle Turnover	OROS	Construction
Methods for the selection of a preliminary subsample		X				X
Methods for imputation of missing data	X	X	X	X	X	
Methods for weighting adjustment	X			X		X
Direct or in direct utilising aggregated information	X	X	X			

Methods using aggregated information have been used for the preliminary estimation of the industrial turnover index for the Industrial Turnover Survey. Two different contexts have been adopted: the former is related to panel survey data, for which information for each sampling units is available for all the surveys carried out in the previous periods of time; the latter refers to a more general context concerning situations where sampling units are not necessarily the same over the time. In the first case the natural choice is to use a unit level model with all the information coming from the previous periods of time. Unfortunately, in the second case the unit level model is not feasible since auxiliary information available for the units not always in the surveys is not coherent with the model. One possible solution is to use a model using aggregating information, on the basis of one or more auxiliary variables correlated with the phenomenon of interest.

In the case of panel survey data, a dynamic linear model approach adapting the model proposed by Tam (1989) has been applied to obtain preliminary estimates (Lamberti *et al.*, 2004). Early and late respondents at previous surveys and early respondents at the current survey are used as auxiliary information to obtain preliminary estimates. In general, the model gives good results in terms of estimation of the monthly industrial turnover; moreover, the application of the model requires that each sampling unit (early or late respondent) is observed all over the time. Therefore, disaggregated information level models implies a substantial loss of the monthly information.

In order to use all the information available at each month, sampling units have been aggregated according to the NACE Rev.1 two digits-economic activity which have been considered as level specification of the model. Preliminary model based estimates have

been computed by means of the methods proposed in Rao *et al.* (1989). The authors suggest to adjust standard preliminary estimates by three different time series methods. Two different standard preliminary estimation methods have been used: naïve estimation and a method based on modelling the auxiliary information by means of Preliminary estimates have been obtained by means of Seemingly Unrelated Regression model. The results obtained are satisfying, particularly when for standard preliminary estimation SUR methods have been applied.

From that we can deduce that better results are achieved when preliminary estimation is computed integrating different techniques, showing, in this way, the future activity of research.

References

- Bolfarine H., Zacks S. (1992), *Prediction Theory for Finite Populations*, Springer-Verlag.
- Drudi I., Filippucci C. (2000), “Inferenza da campioni longitudinali affetti da selezione non casuale”, In: Filippucci C. (ed), *Tecnologie informatiche e fonti amministrative nella produzione di dati*, 415-432, Franco Angeli, Milano.
- ISTAT (2005), Rapporto Finale del Progetto Interarea finalizzato ad approfondire e a verificare sperimentalmente alcune metodologie statistiche utili per ottenere stime preliminari nelle indagini congiunturali condotte dall'Istat, *Documento Interno di Ricerca*.
- Lamberti, A., Naccarato, A., Pallara, A. (2004). Improving timeliness of short-term business statistics through state-space modelling of preliminary survey data, *Proceedings of the European Conference on Quality and Methodology in Official Statistics*, http://q2004.destatis.de/download/Table-of-contents_Programme.pdf.
- 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; Wiley, New York.
- Tam, S. M. (1987), Analysis of repeated surveys using a dynamic linear model, *International Statistical Review*, 55: 63-73.