Provisional Estimation of the Italian Monthly Retail Trade Index

Roberto Gismondi (*)
Laura De Sandro (**)

(*) ISTAT – Ufficio del Direttore delle Statistiche Economiche Congiunturali
(**) ISTAT - Servizio SCO (Statistiche Congiunturali)

Keywords

Balanced sampling; Cluster analysis; Predictor; Provisional estimation; Retail trade.

Riassunto

Il lavoro affronta il problema della selezione di un *panel* di rispondenti "rapidi", rappresentativo dell'intera popolazione di riferimento, da cui si possano derivare stime anticipate utili per l'analisi congiunturale. A tale fine è stato proposto un particolare adattamento della teoria del campionamento bilanciato, con proposte operative circa l'algoritmo di selezione delle unità ed una applicazione empirica a dati reali estratti dell'indagine mensile sulle vendite al dettaglio dell'ISTAT. Sulla base della metodologia descritta l'ISTAT, a partire dal mese di riferimento di gennaio 2004, è in grado di trasmettere ad EUROSTAT le stime anticipate a 30 giorni dell'indice mensile sulle vendite al dettaglio, utili per il calcolo di un indicatore anticipato del commercio al dettaglio a livello UE. La metodologia può essere applicata senza modifiche sostanziali ad altri contesti d'indagine

Abstract

The problem faced concerns the selection of a panel of "advance" respondents, representative of the whole target population, from which advance provisional estimates useful for short-term analyses can be derived. For this goal, we have proposed a particular adaptation of the theory of balanced sampling, with operative proposals concerning the algorithm for selecting sample units and an empirical application to data drawn from the monthly retail trade survey currently carried out by ISTAT. In this way ISTAT can supply EUROSTAT with monthly provisional estimates of the retail trade indexes available after 30 days from the end of the reference month, as requested in a specific task force aimed at building up an EU monthly retail trade provisional index.

1. Introduction[1]

Timeliness is considered one of the various aspects on which quality of statistical data should be founded in addition to exhaustiveness, coherence, comparability, low cost and degree of discrepancy with true data (EUROSTAT, 2000). However, there is an obvious trade/off between timeliness and precision of estimates, that is commonly intended as the most relevant qualitative issue of sampling estimates.

This problem, even though widely discussed in literature, should be carefully monitored in each specific contest concerning short-term indicators production. One solution consists in building up a statistical system able to calculate and spread out both definitive and provisional *quick* data.

The retail trade index represents the only monthly indicator on the economic activity of the service sector currently calculated and spread out by ISTAT, with a delay of about 54 days from the end of the reference month, considered too large by many users.

No provisional advance data referred to month *m* are available for publication at *m*+30, delay that would be considered rather satisfactory for short-term economic analysis. However, provisional retail trade index estimations could be carried out on the basis of three main strategies:

1) use of data on retail trade dynamics related to previous months, through time series forecasts based on ARIMA models. However, in such a way only historical data will be used, without any additional information on the trend related to the month of reference.
2) Use of regression models based on delayed values of the variable of interest (as in case 1), but adding other auxiliary variables available *for the period object of estimation*; though better than the previous one, also this strategy doesn't use actual survey data at all.
3) Use of data related to the reference month *m* and to a part of the units included in the sample (a *panel*, a natural sub-sample of advance respondents, a whatever sub-set of units), whose data related to month *m* are available within a short time.

Various experiences exist related to the third methodology, that seems to be the fittest for a national statistical institute. In particular, a study referred to the first half of '90 years (Gismondi, 1996) stressed how, generally speaking, there is a structural statistically significant difference between the average value of retail sales concerning respondent and non-respondent units: on the average, the latter had a higher retail trade turnover, even though this evidence is only due to non-food products, while food products showed an opposite profile. In that context there was a first tentative proposal for the calculation of a provisional retail trade index at *m*+30. Calculation of provisional indexes was simply based on all and only the questionnaires filled in and received within 30 days from the end of the reference month. Limits of this approach concerned: 1) differences between the average profile of these advance respondents and the whole sample; 2) unsteadiness of advance respondents in following waves of response, so that advance respondents at month *m* could be quite different in composition from those at month (*m*+1).

Problems to be solved further concern: a) how to identify an optimal sub-sample of advance respondents and b) how to convince them to respond quickly.

In details, the Italian retail trade monthly sample survey is aimed at estimating monthly retail trade turnover indexes (Division 52 of the NACE nomenclature). Sales in non fixed outlets, out of shops and second hand goods are not observed.

In 2003 the sample was based on 7.122 enterprises - drawn from a population of about

---

[1] The opinions herein expressed don't involve ISTAT and must be addressed to the authors only (gismondi@istat.it; ldesandr@istat.it), as well as possible errors or omissions.

570 thousands - object of a partial yearly rotation involving about 2.000 enterprises.

Until the end of 2002, the most part of questionnaires, which enterprises were requested to send back no later than 15 days from the end of the reference month, were received by ordinary mail. Moreover, no postal reminders were used, but sensitive firms (about 400, the most part of which are very big and self-representative) were contacted by telephone at *m*+30 to speed up data collection[2].

After the calculation of 150 elementary indexes, obtained crossing each other 15 main groups of product sold, 5 classes of persons employed (1-2, 3-5, 6-9, 10-19, >19) and 2 main geographic areas (North and Centre/South), higher level indexes are obtained on the basis of the Laspeyres formula, where the weight of each stratum is given by the yearly turnover referred to the base year 2000, derived from structural business statistics.

Monthly indexes are calculated as ratios between the average turnover of each month *m* and the average of the base year 2000. By the way, let's note that even though under given super-population models the recourse to more detailed estimators than the sample mean could improve quality of estimates - as described in paragraph 3 -, the need to guarantee comparability between final and provisional estimates obliges to use sample means for calculations of provisional estimates as well.

For what concerns timeliness, indexes currently calculated are based on responses got within 52 days from the end of the reference month, are released by a monthly press release after 54 days[3] and at the same time are sent to EUROSTAT. Delay mostly depends on response burden, need to use external accountants for filling in questionnaires and delay occurred when using ordinary mail. Reduction of this delay within 30 days for the whole sample will be achieved only along some years, while the possibility to calculate provisional estimates at *m*+30 represents a goal that can be reached into a short time.

In the next paragraph we'll recall the main features of the Country-stratified European Sample project, while in paragraph 3 the main principles of balanced sampling are summarised and in paragraph 4 we propose a technique to find balanced samples useful for advance provisional estimates. In paragraph 5 main technological and operational innovations introduced in the production process are presented, while paragraph 6 shows main results concerning provisional estimates for all the 12 months of 2003, with some perspective conclusions.

2. European Union needs: the Country-stratified European Sample for retail trade

EUROSTAT currently calculates and releases an overall EU retail trade monthly index, based on a weighted arithmetic mean of the single EU countries indexes. The delay of publication, that is about 60 days from the end of the reference month, is considered too large by operators, researchers and decision makers. For this reason, since 2001 a task force managed by EUROSTAT has been planning a statistical strategy aimed at selecting, in each EU country, a particular sub-sample from the national samples currently used, contributing to the calculation of a provisional quick index at the EU level to be released within about 30 days.

The basic idea is that for defining overall size and breakdown by country of a European sample able to produce reliable advance estimates at the EU level, it can be possible to consider each country as a single stratum and to split an overall advance sample size by

[2] More detailed methodological and operational issues can be found in ISTAT (1998; 2002).

[3] The average delay in publication of the Italian monthly retail trade index has been shortened from 56,9 days in 2001 to 55,6 in 2002 and to 54,8 in 2003.

country according to the Neyman allocation and to a given expected sampling error for the monthly average turnover at the EU level[4]. In this way a relatively small EU sample – obtained summing up all national sub-samples – could guarantee, on the average, small estimate errors. The optimal allocation of the sub-sample size is based on the common Neyman formula:

$$n_c^* = n^* (\hat{\sigma}_c w_c) \left(\sum_{c=1}^{15} \hat{\sigma}_c w_c \right)^{-1} \quad (2.1)$$

where n^* is the optimal overall "advance" sample size (guaranteeing an error level not higher than 1% with the 99% of probability), c indicates a country, w_c is the relative weight of the number of retail trade enterprises in the universe operating in that country and $\hat{\sigma}_c$ is the estimate of its average monthly standard deviation[5]. Let's note that, in practice, application of Neyman allocation at the EU level was more detailed than in formula (3.1), since:

1. EUROSTAT asked for advance estimates not only for the total of sales, but separately for food and non-food sales as well;
2. stratifications currently used in several EU countries suggested that an additional breakdown of formula (2.1) by firm size (evaluated on the basis of five classes of persons employed) could have improved final precision, so that this formula was applied separately in each of the ten strata got crossing stratification by type of product sold and firm size class.

According to the optimal Neyman allocation, EUROSTAT calculated that Italy, starting from 2003, should use for advance estimates a sub-sample of 1.929 retail enterprises, to be selected from the whole sample composed – as yet said – by 7.122 units (EUROSTAT, 2001).

A not trivial problem to be faced has been the choice of the technique for the sub-sample selection, topic on which EUROSTAT didn't give any particular recommendation. The problem consists in the selection of a sub-sample which longitudinal monthly profile is "similar" to the corresponding one evaluated on the overall sample. This choice, that should guarantee a good quality of provisional indexes, is not easy, also because:

1. retail trade enterprises are very heterogeneous, even in the same stratum;
2. the retail trade turnover distribution is far from normal, so that use of simple random sampling in each stratum could not often lead to satisfactory results;
3. even if an optimal advance sample can be identified, we are not sure that all enterprises belonging to it will respond, or will respond within 30 days so that, in addition to technical evaluations, an efficient system of reminders must be used as well.

On the other hand, common estimators (or *predictors* in a model-based context) can be improved using additional information available on the whole sample units, as historical monthly data yet picked up and registered along year 2002, as shown in paragraph 4. In the following table 2.1 we have resumed the main figures concerning the optimal Neyman allocation for Italy and the EU. In Italy size of the universe is quite large and about 3,5 times greater than the EU countries average. However, sample size is not much larger than the EU average (5.868 units), while the optimal sub-sample size is about the double with respect to

[4] For simplicity, other relevant sources of (non-sampling) errors – as business longitudinal changes, measurement errors, non responses, under coverage of the list – were not taken into account. Moreover, let's note that fixing an estimate error for average turnover doesn't guarantee the same error level for index numbers.
[5] For Italy, this estimation was carried out on the basis of turnover data, referred to year 2001, available for each enterprise from the ISTAT business register (ASIA).

the EU average of 987 units. As a consequence, the percent ratio between sizes of the optimal sub-sample and the total sample is 27,1%, in line respect to the EU average, equal to 25,6%.

Table 2.1: *Optimal Neyman allocation for each EU country (EU sampling error level:1%)*

Country	Universe	Total sample	Neyman sub-sample	% ratio sub-sample/sample
Italy	**570.379**	**7.122**	**1.929**	**27,1**
Total EU	2.414.465	76.279	13.868	18,2
Average EU (*)	160.694	5.868	987	25,6

(*) It's based on 13 countries among the 15 belonging to the EU, because sample sizes for Luxembourg and France were not available. *Source*: elaboration on EUROSTAT and ISTAT data.

From table 2.2 we can see how the most part of Italian retail trade firms are very small (almost 81 retail trade enterprises on 100 have 1 or 2 persons employed) and the actual whole sample is only the 1,25% of the universe; in particular, it includes only the 0,55% of enterprises with 1 or 2 persons employed. Since enterprises selling food products are more heterogeneous with respect to those selling non-food products, the relative weight of the former in the advance sub-sample is higher than in the whole sample. So, almost the fifty percent of enterprises in the whole sample selling food products belong to the advance sample as well. This ratio raises up to 77,07% for firms with more than 19 persons employed.

Table 2.2: *Universe, sample and optimal EU Neyman sub-sample for the Italian retail trade sector (universe referred to 2001, samples to year 2003)*

Persons employed	Universe			Total sample			Neyman sub-sample		
	Total	Food	Non food	Total	Food	Non food	Total	Food	Non food
1-2	461.574	144.842	461.574	2.537	657	2.030	442	306	136
3-5	81.274	22.452	81.274	1.089	264	825	169	74	95
6-9	18.182	4.492	18.182	868	290	578	65	27	38
10-19	6.587	2.250	6.587	723	148	575	49	23	26
>19	2.762	1.386	2.762	1.905	750	1.155	1.204	578	626
Total	**570.379**	**175.422**	**570.379**	**7.122**	**2.109**	**5.163**	**1.929**	**1.008**	**921**
Persons employed	**Sample/universe**			**Neyman/universe**			**Neyman/sample**		
	Total	**Food**	**Non food**	**Total**	**Food**	**Non food**	**Total**	**Food**	**Non food**
1-2	0,55	0,45	0,44	0,10	0,21	0,03	17,42	46,58	6,70
3-5	1,34	1,18	1,02	0,21	0,33	0,12	15,52	28,03	11,52
6-9	4,77	6,46	3,18	0,36	0,60	0,21	7,49	9,31	6,57
10-19	10,98	6,58	8,73	0,74	1,02	0,39	6,78	15,54	4,52
>19	68,97	54,11	41,82	43,59	41,70	22,66	63,20	77,07	54,20
Total	**1,25**	**1,20**	**0,91**	**0,34**	**0,57**	**0,16**	**27,09**	**47,80**	**17,84**

Source: elaboration on EUROSTAT and ISTAT data.

3. Theoretical background: superpopulation model and balanced sampling

From now on we'll suppose to refer to the Italian current sample and a whatever sample stratum among the ten considered by EUROSTAT in the stratification frame adopted for the Neyman allocation. Strata have been obtained combining two main kinds of products sold – food and non-food – and five class of persons employed: 1-2, 3-5, 6-9, 10-19, >19. Symbol N will indicate sample size in each stratum and n the size of the (optimal) sub-sample to be selected.

The target consists in estimating the turnover population mean $\bar{y}$ (that, in practice, will refer to each single month in a year) on the basis of a sample survey. In each stratum we'll suppose as true the following simple regression model (R), defined as:

$$y_i = \beta\, x_i + \varepsilon_i \qquad \text{where} \qquad \begin{cases} E(\varepsilon_i) = 0 & \forall i \\ VAR(\varepsilon_i) = \sigma^2 v_i & \forall i \\ COV(\varepsilon_i, \varepsilon_j) = 0 \quad if & i \neq j \end{cases} \tag{3.1}$$

where expected values, variances and covariances are referred to the model (and not to any sampling design), y is turnover, x is an additional variable strongly correlated with y and to be specified, as well as the function v_i, with β and σ^2 given, but generally unknown parameters.

If only one single auxiliary x-variable is taken into account, a sample s of size n drawn from a population U of size N is said *balanced with respect to the weights root(v)* if it satisfies the condition:

$$\sum_s x_i / n\sqrt{v_i} = \sum_U x_i \Big/ \sum_U \sqrt{v_i} \tag{3.2}$$

The balanced sample could be chosen among all the possible samples of size n using various algorithms, as those proposed by Droesbeke, Fichet and Tassi (1987), Rose (1996) and Valliant, Dorfman and Royall (2000). Royall (1992) showed that if the previous linear model R holds and a balanced sample *can be found*, then the best linear unbiased predictor under the model (3.1), given the vector **x**, is:

$$\hat{T}_{bal,v} = n^{-1}\Big(\sum_U \sqrt{v_i}/N\Big)\Big(\sum_s y_i/\sqrt{v_i}\Big) \tag{3.3}$$

Two common statements are v=1 and v=x. The first one (v=1) translates into a model-based context the common hypothesis of homoschedasticity that justifies the frequent recourse to a (SRSWR[6]) design, while v=x implies the more realistic hypothesis of a lower relative (model) variability for largest units. The optimal predictors derived from (3.3) will be, respectively:

$$\hat{T}_{bal,1} = n^{-1}\sum_s y_i \quad \text{and} \quad \hat{T}_{bal,x} = n^{-1}\Big(\sum_s y_i/\sqrt{x_i}\Big)\Big(\sum_U \sqrt{x_i}/N\Big) \tag{3.4}$$

so that if the sample is balanced the sample mean is still optimal even when x≠1. When v=1 and v=x the minimum (*among all samples satisfying* (3.2)) model mean squared errors are given by, respectively:

$$\big[N(N n^{-1} - 1)\big]\frac{\sigma^2}{N^2} \quad \text{and} \quad \left[\left(\sum_U \sqrt{v_i}\right)^2 n^{-1} - \sum_U v_i\right]\frac{\sigma^2}{N^2}. \tag{3.5}$$

[6] Simple Random Sampling Without Replacement.

There are some relevant points in favour of the use of balanced sampling:

1. it preserves from a bias when the model (3.1) is wrong, so that misspecification of the model does not lead to bias under a broader model. This is an important advantage with respect to the more common regression estimator (Park, 2002), that is optimal under model (3.1) when the sample is not balanced.
2. Under model (3.1), the best choice among all possible samples is in favour of the sample including the n biggest units. However, the model variance depends on the relative weight of the sample units on the overall x-amount in the population U: generally speaking, when this weight is lower than 50% other estimators and/or sample selection rules could perform better.
3. Both from a theoretical (given the condition 3.2) and an heuristic point of view, search for balanced samples leads to the selection of a *representative panel*, often chosen otherwise in a deterministic subjective way, not always justified by objective considerations.

Properties of balanced samples represent a useful theoretical result up to now not very exploited in current sampling practice, but that can be adapted to our context, as resumed hereafter. However, serious problems can occur concerning the availability of an algorithm to select, *if it exists*, a balanced sample defined as before.

In the following paragraph a simple methodology to get quickly to a *quasi*-balanced sample is proposed, avoiding the risk to try and reject all the possible samples until a balanced one is found. For simplicity, the balance conditions will be limited to the cases v=1 (sample and population x-means must be equal) and v=x.

4. *Quasi-balanced* sample selection

4.1 Theoretical aspects

Let's suppose again to refer to a given stratum, including N units, and to know for each unit i the values x_i and v_i.

We can divide the stratum population into n sub-strata including N_h units each. From every sub-stratum h a single unit i is drawn, e.g. the one minimising a loss function – to be defined further – so that, remembering (3.2), this identity can be considered approximately true:

$$x_{hi}\Big/\sqrt{v_{hi}} \approx \sum_{j=1}^{N_h} x_{hj} \Big/ \sum_{j=1}^{N_h} \sqrt{v_{hj}} \tag{4.1}$$

The idea is that a *quasi-balanced* one-unit sample in each sub-stratum should lead to a *quasi-balanced* sample of size n for the stratum considered, where the final predictor for the population mean of the stratum taken into account is given by:

$$\hat{T} = N^{-1}\sum_{h=1}^{n} \hat{T}_h N_h \qquad \text{where} \qquad \hat{T}_h = \Big(\sum_{j=1}^{N_h} \sqrt{v_{hj}}\Big/N_h\Big)\, y_{hi}\Big/\sqrt{v_{hi}} \tag{4.2}$$

It is not difficult to verify that, if (4.1) holds for each of the n sub-strata, then (3.2) is approximately true for the whole sample as well.

Main problems are: 1) how defining the n sub-strata; 2) how to choose the "optimal" unit i in each sub-stratum.

Concerning point 1), the problem of which is and how can be obtained an optimal (sub)stratification is still unsolved, the choice depending on the concentration of x in the population, the sampling technique and the kind of estimator used.

As a premise, we remark that all the following considerations hold when v=1, otherwise we can write again (4.1) using this reasonable approximation:

$$\sum_{j=1}^{N_h} x_{hj} \Bigg/ \sum_{j=1}^{N_h} \sqrt{v_{hj}} \approx N_h^{-1} \sum_{j=1}^{N_h} x_{hj} \Big/ \sqrt{v_{hj}} \tag{4.3}$$

so that unit i in (4.1) must have a value of $z_{hj} = x_{hj} \big/ \sqrt{v_{hj}}$ as similar as possible to the sub-stratum mean.

Cochran (1977) proposed to order the N units according to their not decreasing z-values and to calculate for each unit i the cumulative of $\sqrt{z_i}$. Boundaries of n sub-strata can be obtained imposing that each sub-stratum must cover the same cumulative value $\sum_{j=1}^{N} \sqrt{z_j} / n$.

In practice, the *Cochran-root* method could be satisfactory only if strata are numerous and small.

An alternative idea is driven by the fact that the risk of the choice of only one unit from each sub-stratum will be as much lower as the sub-stratification used guarantees a high ratio Var(B)/Var(T), being the two variances respectively equal to the “Between strata” and the “Total” variance evaluated on variable z and calculated on the whole sample (*Max(VarB) method*).

In this case sub-strata can be obtained using any univariate hierarchical cluster analysis algorithm based on the Ward optimisation, easily available on common statistical packages as SAS or SPSS. In practice, when N>3.000 it could be helpful to use alternative not hierarchical time-saving cluster analysis algorithms, as suggested by Fraire (1984).

On the second point, in each sub-stratum h we can select the unit i satisfying the condition:

$$\left|z_{hi} - \bar{z}_h\right| = \underset{j \in U_h}{MIN} \left(\left|z_{hj} - \bar{z}_h\right|\right) \tag{4.4}$$

Even though samples found as described above could be not exactly balanced – as a consequence of positions (4.1) and (4.3) – they present the following advantages:

a) since each unit selection is carried into a restricted sub-stratum, the hypothesis v=1 – that on the basis of approximation (4.3) should guarantee more precise results and lead to the use of a simpler estimator – is more realistic than when referred to the whole overall stratum, for which empirical evidence shows that often the best model is based on v=x, so that balance could turn out to be more imprecise.
b) Quasi-balanced samples are selected on the basis of functions considering *for each unit* the degree of distance with respect to the mean.
c) The selection rule is simple and quick, since only N_h attempts are needed in each sub-stratum. This is a fundamental advantage in comparisons with other proposed procedures, based on mathematical optimisation (Khan *et al.*, 1999).
d) In general, a (sub)optimal result is always guaranteed whatever is n - often fixed in advance because of budget constraints - according to optimal predictors defined before.

Moreover, if a stratum contains a sub-group of very big units, among all the n sub-strata probably the algorithm will create some clusters composed by one unit only, that will be included in the optimal sub-sample by definition. In other words, if x-distribution is very positively skewed (as it often happens in practice), a part of the sub-sample will include all the largest units, in agreement with the optimal strategy mentioned in point 2) of paragraph 3.

If K x-variables reasonably linked with the observed y-values are available, we can calculate the corresponding z-values, standardise them in order to deal with K new variables comparable in magnitude and variability and calculate for each unit the synthetic function:

$$Z_i = \sum_{k=1}^{K} z_{ki}^{*} / K \qquad (4.5)$$

where z^* indicate values standardised with respect to mean and standard deviation. Then the (*Max(VarB) method*) can be applied to the new variable Z in the same way as described above.

For the choice of the optimal unit i in each sub-stratum h we can use again formula (4.4) applied to Z, and this selection method will be defined *z-univariate*.

Otherwise, it could be guessed that better results can be achieved if the unit i satisfies this condition:

$$\sum_{k=1}^{K} \left| z_{khi}^{*} - \bar{z}_{kh}^{*} \right| = \underset{j \in U_h}{MIN} \left(\sum_{k=1}^{K} \left| z_{khj}^{*} - \bar{z}_{kh}^{*} \right| \right) \qquad (4.6)$$

of which (4.4) is a particular case for K=1[7]. This second selection method will be defined *z-multivariate*.

4.2 Practical aspects

According to the technical decisions taken by the task force, ISTAT selected from the overall retail trade sample composed by 7.122 enterprises a sub-sample which retail trade average evolution was, along year 2002, very similar to that of the overall sample according to a "balanced-sampling" procedure as follows.

Let's note that the sub-sample size fixed by EUROSTAT guarantees a 1% error at the aggregate EU level, but *not necessary at the Italian level*. For this reason, ISTAT decided to increase the advance sub-sample size, adding to the initial 1.929 units other 811, chosen taking into account operative aspects as the availability of an e-mail address or an easy access to the Web. So, the final Italian balanced sample was based on 2.740 enterprises.

These technological factors should render easier an advance response from enterprises. However, in this way the gain due to a larger amount of effective advance respondents is paid in terms of the inclusion in the sub-sample of about the 30% of enterprises which overall profile could not be balanced according to the rule (3.2).

When this preliminary analysis was carried out[8], available definitive retail trade data covered the period January-November 2002. From the ISTAT retail trade monthly survey sample we extracted 4.616 enterprises, that are those for which monthly data for 2002 were available (since, as yet said in paragraph 1, a part of the sample is rotated each year). Moreover, units that didn't respond for at least three of the eleven months were also excluded. Since some wave non-responses affected available historical data, about the 10% of monthly micro-data were estimated using the same techniques adopted in the current survey, in order

[7] Let's note that even though each z^* has mean equal to zero, generally *in each sub-stratum* the z^* mean will be different from zero.

[8] January 2003.

to deal with a complete dataset.

Available sample data were broken down in 10 "universes" of reference obtained, as yet said, combining two main kinds of products (food and non-food) and five size classes (1-2, 3-5, 6-9, 10-19, >19). The aim was drawing optimal sub-samples from each universe, that is itself a part of the current overall retail trade sample.

For each of these ten populations we looked for balanced sub-samples on the basis of data referred to the first six available months (January-June), which turnovers were considered as x-variables. The purpose consisted in estimating *ex-post* the (known) average monthly turnover referred to each of the five months from July to November and to verify the size of error in estimates. The underlying idea is that good balances for old data should produce, on the average, good provisional estimates in the future as well, limiting the analysis to monthly turnover means, even though index numbers are the true object of estimates.

In our context, in formula (4.5) we have K=6 and each x-variable is turnover for each month from January to June 2002, so that k=1,2,…6. The breakdown of each stratum in n sub-strata is based on a function (4.5) given by the average standardised turnover for the first half of the year, without (case v=1) or with (case v=x) the correction $z=x/v^{0,5}= x^{0,5}$ to be used for balance as in (4.1), where x is the y-variable lagged.

The better performance of balanced sampling in comparison with more used and traditional estimation strategies (as stratified random sampling and systematic sampling), at least for what concerns retail trade, has been proved in Gismondi (2003). In that context the correctness of model (3.1) has been tested as well. Moreover, it was also proved that, according to empirical results, the best strategy for selecting a *quasi*-balanced sample was the one based on the (*Max(VarB) method*). This is the criterion on which results presented in the following paragraph 6 are based.

We can also underline since now that empirical evidence showed how the model position v=x, according to the *z-multivariate* option can be considered the most suitable for retail trade data, in terms both of percent average balance error[9] for the first six months of 2002 and percent average estimate error for months from July to November. As a consequence, that is the overall strategy used to calculate provisional estimates from January 2003 ahead.

5. Operational aspects

The additional problem to be still faced concerns tools to be used for increasing the response rate and timeliness of responses for all enterprises included in the advance sample.

For what concerns the retail trade monthly survey, starting from the last months of 2002 ISTAT tried to build up an internal system able to manage the sending of questionnaires based on telefax and e-mails and the realisation of an electronic questionnaire available on the Web, in order to make easier and quicker responses from enterprises. More in general, the changes gradually introduced in the operational process were the following ones:

1) sending of forms with the "priority mail" to all the enterprises in the sub-sample;
2) sending of forms using a fax server to the enterprises that didn't receive the postal questionnaire, or that didn't respond within 18 days;
3) building up of an electronic questionnaire on the Web and optical reading of forms received on the Internet with the *software* Teleform[10], avoiding to register data manually;

[9] The balance error is given by the percent discrepancy between the left and right terms of formula (3.2).

[10] Teleform is a particular software useful for data capturing and direct video-revision of paper and electronic questionnaires received by Web or through a fax server, by which the whole process becomes quicker. In

4) interactive check of these data with Teleform;
5) a more efficient way of managing data in the statistical database.

At the beginning, the evaluation of the possibility to manage the most important innovations as those concerning previous points 2), 3) and 4) was experimented on a small sub-sample concerning 194 enterprises, representing the 10% of enterprises belonging to the optimal advance EUROSTAT sub-sample and selected in order to be representative of that.

Further, on the basis of good results obtained with this first small sample, the number of enterprises engaged in the Teleform experiment was enlarged. Starting from the whole theoretical sub-sample of 2.740 enterprises, from the reference month of April 2003 we decided to add to the Teleform experiment other 1.635 enterprises that could be reached and, of course, could also respond by fax.

At the end of May, since the fax server system was not able to manage the large amount of questionnaires, a definitive choice was undertaken in favour of the use of priority mail for all enterprises in the sub-sample.

In this way, we had also the possibility to compare precision of early estimates concerning the first three months of 2003 – evaluated, on the average, on less than one thousand of advance respondents – with that concerning the following months, whose early estimates were based on a larger number of advance respondents, many of which using fax instead of ordinary mail.

Summarising the results got for the months of 2003, questionnaires *concerning the sub-sample of advance respondents* are received mainly by fax (74,8%), then by ordinary mail (21,2%) and web (4,0%). For what concerns the *whole effective final sample*, at the end of 2003 the share of questionnaires received by fax was 58,8%. At the end of 2002 the same percent rate was significantly lower (33%).

Due to the quickness of responses got by fax, within 30 days from the end of the reference month the average number of advance respondents in the months of 2003 was equal to 1.126 (Graph 5.1), that is about the double of the number of advance responses available until the end of 2002. This amount also represents more than one third of the effective overall sample used for the calculation of final retail trade indexes, released after about 54 days from the end of the reference month.

On the other hand, monthly variability of the number of advance respondents is still quite high: after the first two months characterised by less than nine hundreds units, in the first half of 2003 the highest amounts were got in March (1.200) and April (1.253). In August and, especially, in September there was a significant increase in the number of advance respondents that reached the top of 1.552. That was due to an improvement in the process of reminders started from the end of Summer.

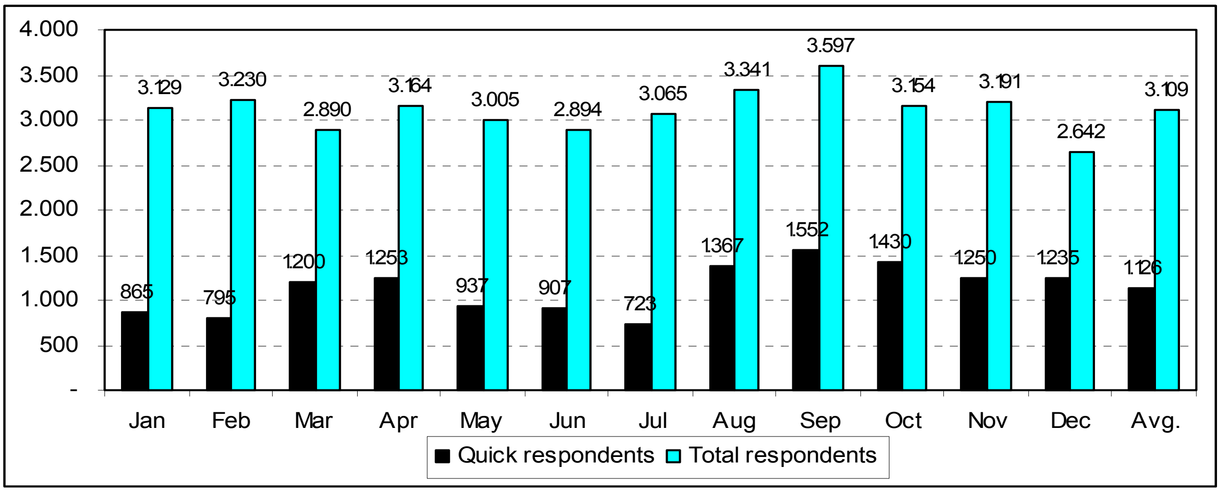
Moreover, at the moment the effective rate of advance responses is rather low in comparison with the theoretical sample established by the EUROSTAT task force. In particular, we must remark that in the overall theoretical sample of 7.122 enterprises, 712 are classified as not specialised enterprises with prevalence of food products. Among them, 291 belong to the theoretical optimal sub-sample. At the moment, we can observe among the advance respondents only 77 enterprises on 291. This under-coverage is currently under observation, as it can determine less precise advance estimates than for non-food products,

---

particular, ISTAT used this software for reading electronic questionnaires. In this way, enterprises do not wait anymore for receiving the electronic form, but they only need to have the access to the Internet for filling it in and to submit it; then, through Teleform the filled form is received and visualised by ISTAT within some minutes. The reviser opens the form on his personal computer without waiting for receiving it by fax (that sometimes could not be readable or busy), makes corrections and submits the form toward Oracle databases.

also because in Italy sales of food and beverages are strongly increasing just in not specialised stores.

Graph 5.1: *Advance and total respondents in the retail trade monthly survey (months from January to December 2003)*



6. Main results and perspectives

First, observation of balance errors – given by the percent discrepancies between the left and right terms of formula (3.2) – is useful to evaluate the overall quality of the advance sample selection procedure. Results refer to the use of the (*Max(VarB) method*) and to the Neyman optimal sub-sample composed by 1.929 enterprises[11], have been resumed in table 6.1. For what concerns balance errors these evidences raise:

- for total turnover, no definitive indication in favour of univariate or multivariate procedures raised, nor when *v*=1 or *v*=*x*: *z-univariate* is better when *v*=*x* (the average of the percent errors for the six months is equal to 1,98% against the 2,83% of *z-multivariate*), but worst when *v*=1 (5,82% against 1,29%).
- If we consider separately food and non-food products, we still note that univariate is worst than multivariate when *v*=1 (for food and non-food we have, respectively, 5,07% and 7,33% with univariate and 1,12% and 1,61% with multivariate), but also, for non-food products, when *v*=*x* (3,44% against 2,91%), so that it should be preferred only for food products when *v*=*x* (1,22% against 2,75%).

In short, for food, non-food and total the best performances of *z-multivariate* were got when *v*=1, while *z-univariate* should be preferred when *v*=*x*. Anyway, *z-multivariate* produces, on the average, more steady estimates (lower average errors).

The effective precision of estimates was assessed evaluating precision of advance estimates in comparisons with definitive estimates for months from July to November 2002. Main results are the following ones:

- On the average, for "total" the forecast error was higher than the corresponding balance error when *v*=1 (the average percent forecast errors were 7,36% with *z-univariate* and

[11] For a better coherence respect to all the methodological issues stressed by the EUROSTAT task force, results in table 6.1 are based on the original 1.929 EU advance sample units only, without the 811 added to guarantee a good level of estimates at the Italian level as well.

1,79% with *z-multivariate*), but substantially equal or lower when *v*=*x* (respectively 2,08% and 1,69%, when the corresponding balance errors were respectively 1,98% and 2,83%). Moreover, it's clear that forecasts show the better results of *v*=*x* both for univariate and multivariate (differently from balances, when *v*=*x* was better only with *z-univariate*) and the best performance of *z-multivariate*.

- Also when considering separately food and non-food products we have that univariate is worst than (or at most equal to) multivariate: when *v*=1 forecasts errors with *z-univariate* are equal to 7,69% and 6,71% respectively for food and non-food, while the corresponding errors for *z-multivariate* are 1,30% and 2,73%. When *v*=*x* we have 1,21% and 3,71% on one hand and 1,24% and 2,50% on the other. Moreover, *v*=*x* is quite always to be preferred to *v*=1.

In short, forecast analysis results stress that, for food, non-food and total, the best performances of *z-multivariate* are got when *v*=*x*. This seems to be the most suitable strategy to carry out along the available months of year 2003.

Table 6.1: *Balance and forecast results got using the z-univariate and the z-multivariate methods – year 2002*

Domain	N	n	n/N	Balance (Jan-Jun 02)		Forecasts (Jul-Nov 02)	
				z-univariate	z-multivariate	z-univariate	z-multivariate
FOOD	1.703	1.008	59,2				
Average turnover (Euro)				1.630.502	1.630.502	1.675.078	1.675.078
Average % error (*v*=1)				5,07	1,12	7,69	1,30
Average % error (*v*=*x*)				1,22	2,75	1,21	1,24
NON FOOD	2.913	921	31,6				
Average turnover (Euro)				346.494	346.494	365.477	365.477
Average % error (*v*=1)				7,33	1,61	6,71	2,73
Average % error (*v*=*x*)				3,44	2,91	3,71	2,50
TOTAL	4.616	1.929	41,8				
Average turnover (Euro)				729.249	729.249	755.860	755.860
Average % error (*v*=1)				5,82	1,29	7,36	1,79
Average % error (*v*=*x*)				1,98	2,83	2,08	1,69

Since one relevant purpose was not only the selection of the optimal sub-sample to be used for EUROSTAT's purposes, but the assessment of the precision of advance estimates for the Italian retail trade index as well, ISTAT verified precision of early estimates at the national level for the twelve months of 2003[12].

In table 6.2 we have resumed the main results of early estimates on the basis of the *quasi*-balanced sub-sample finally chosen[13].

For the total, the percent differences between definitive and provisional indexes are quite low until September, ranging from 0,1 in March up to 0,6 in April, May and June. October and November were affected by higher percent errors (respectively, -1,2% and -1,3%), mainly due to non-food products. The presence of 7 negative and 5 positive errors, and the small value (equal to 0,26) for the average of the twelve estimate errors seem to confirm the absence of significant structural biases and the randomness of errors. However, the presence of four consecutive negative errors in the last four months, all due to over-

[12] Starting from the end of February 2004 (having with reference month January 2004), ISTAT began to send to EUROSTAT official early estimates at *t*+30, in agreement with the task force decisions.

[13] In this case estimations for 2003 are based on the added 811"quick" units as well.

estimations of the non-food sector, could depend on "quick" large enterprises which non-food trend, generally speaking, is better than that of the overall final sample[14].

On the average, it's a bit more difficult to estimate indexes for food (the average of absolute errors is equal to 0,97) that non-food (0,82). That can be due to the higher variability of individual data related to enterprises selling food products and to their more sprightly longitudinal dynamic when compared with the non-food sector[15]. On the whole, the average of the errors taken in absolute value is equal to 0,58, that represents a satisfactory result.

Table 6.2: *Indexes and differences between definitive and advance provisional retail trade indexes for the twelve months of 2003, by main kind of product sold*

Domain	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg.	Avg. absol.
							Definitive indexes with base 2000=100							
Food	102,3	103,3	110,6	114,7	107,4	110,0	104,2	97,5	108,9	124,0	112,9	160,6	113,0	
Non-food	88,5	88,7	95,7	102,8	112,9	99,9	106,1	78,8	100,8	109,0	110,6	148,4	103,5	
Total	94,0	94,4	101,6	107,5	110,8	103,9	105,4	86,2	104,0	114,9	111,5	153,2	107,3	
							Differences between definitive and provisional indexes							
Food	-0,9	+1,8	-0,3	+1,0	-0,3	+1,0	-0,8	+1,1	+0,7	-0,5	+2,4	-0,8	+0,37	+0,97
Non-food	+0,0	-0,5	-0,2	+0,3	-0,8	+0,4	+0,2	-0,2	-1,4	-1,6	-3,6	-0,6	-0,66	+0,82
Total	-0,4	+0,3	+0,1	+0,6	-0,6	+0,6	-0,2	+0,3	-0,6	-1,2	-1,3	-0,7	-0,26	+0,58

Note: "Avg. absol." is the average of the nine months calculated on absolute values.

Results got up to now seem to confirm how the recourse to a quasi-balanced panel of advance respondents, composed in theory by *the same enterprises* each month[16], should improve the quality of preliminary estimates, both at the EU and at the national level.

That is even more relevant if we remember that retail trade indexes are currently calculated on the basis of ratios between simple sample means, which use is optimal under model (3.1) and a *quasi-balanced* sub-sample only when v=1: being in this empirical study v=x, the use of an alternative estimator as the second one in (3.5) could further improve precision of advance estimates.

Nevertheless, we stress again the importance to verify, at least for another complete year, the degree of discrepancy between preliminary estimates at m+30 and final indexes at m+54.

This represents the most relevant quality measure on the basis of which drawing right conclusions on the opportunity to start a monthly release of advance estimates of the retail trade index at the national level as well.

[14] This aspect will be investigated further along 2004.

[15] In 2003 the Italian retail trade index increased of 2,0% respect to 2002, but while sales of food products raised of 4,6%, the increase for non-food products was only the 0,2%.

[16] In practice, in 2003 the 55,2% of enterprises responded *effectively quickly* for at least 7 months on 12.

References

BATTAGLIA F., MARAVALLE M. (2003), "Data Analysis and Forecasting Techniques for Deriving Anticipated Estimates", proceedings of the *Italian Statistical Society Conference on Multivariate Statistical Analysis*, 9-11 June, 201-211, Naples, Italy.

BREWER K.R.W. (1995), "Combining Design-Based and Model-Based Inference", *Business Survey Methods*, 589-606, John Wiley & Sons, New York.

CICCHITELLI G., HERZEL A., MONTANARI G.E. (1992), *Statistical Sampling Theory*, Il Mulino, Bologna.

COCHRAN W.G. (1977), *Sampling Techniques*, John Wiley & Sons, New York.

DEVILLE J.C., GROSBRAS J.M. (1988), "Efficient Sampling Algorithm and Balanced Sample", *COMPSTAT proceedings in computational statistics*, Springer Verlag.

DEVILLE J.C., SÄRNDAL C.E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, vol.87, 376-382.

DORFMAN A.H., VALLIANT R. (2000), "Stratification by Size Revised", *Journal of Official Statistics*, Vol.16, 2, 139-154.

DROESBEKE J.J., FICHET B., TASSI P. (1987), *Les sondages*, Economica, Paris.

EUROSTAT (2000), *Short-term Statistics Manual*, Eurostat, Luxembourg.

EUROSTAT (2001), *Country Stratified European Sample for Retail Trade – National Methodologies*, Final Report, Eurostat, Luxembourg.

FRAIRE M. (1994), *Methods for Multidimensional Analysis of Data*, CISU, Rome.

GISMONDI R. (1996), "Effects of Non-responses in the Retail Trade Monthly Survey", *Research Copybooks*, 4, 199-236, Istat, Rome.

GISMONDI R. (2003), "Optimal Provisional Estimation of Monthly Retail Trade Data", Proceedings of the *Annual Meeting of the Statistical Society of Canada – Survey Methods Session* (fore coming print), June 8-11 2003, Halifax, Nova Scotia, Canada.

GISMONDI R., DE SANDRO L. (2003), "The Monthly Retail Trade Survey. Methodological Aspects and the Need of Quick Estimates: Late and Future Experiences", working paper discussed in the *EUROSTAT task force on the Country-stratified European Sample*, 28 April 2003, Luxembourg.

GISMONDI R., SUCCI R. (2003), "Optimal Sample Selection Using Auxiliary Data", paper presented at the *Italian Statistical Society Conference on Multivariate Statistical Analysis*, 9-11 June 2003, Naples, Italy.

ISTAT (1998), *La nuova indagine sulle vendite al dettaglio: aspetti metodologici e contenuti innovativi,* Metodi e norme, 3, Istat, Roma.

ISTAT (2000), *Gli indici delle vendite al dettaglio per area geografica: metodologia e risultati*, Argomenti, 19, Istat, Roma.

ISTAT (2002), *Gli indici delle vendite al dettaglio nel 2001*, Informazioni, 48, Istat, Roma.

KHAN E. A., KHAN M. G. M., KHAN M. J. (1999), "Optimum Stratification: a Mathematical Programming Approach", *Proceedings of the Sixth Islamic Countries Conference on Statistical Sciences*, 207-217, Lahore, Pakistan.

LUCEV D. (1997), *Typologies and Non-response Error Control for Quality of Economic Data*, Rocco Curto Editor, Naples.

PARK M. (2002), "Regression Estimation of the Mean in Survey Sampling", *PHD thesis*, Iowa State University, Ames, Iowa.

ROSE D.M. (1996), "A Network Approach to Balanced Sampling", *Congressus Numerantium*, 118, 33-47.

ROYALL R.M. (1992), "Robustness and Optimal Design Under Prediction Models for Finite Populations", *Survey Methodology*, 18, 179-185.

ROYALL R.M., HERSON J. (1973), "Robust Estimation in Finite Population", *Journal of the*

American Statistical Association, vol.68, 880-893.
SÄRNDAL C.E., SWENSSON B., WRETMAN J. (1993), *Model Assisted Survey Sampling*, Springer Verlag.
VALLIANT R., DORFMAN A.H., ROYALL R.M. (2000), *Finite Population Sampling and Inference – A Prediction Approach*, John Wiley & Sons, New York.
SAS INSTITUTE INC. - SAS CAMPUS DRIVE (1996), *SAS/FSP Software* Version 6, Cary, North Carolina.