OPTIMAL PROVISIONAL ESTIMATION OF MONTHLY RETAIL TRADE DATA

Roberto Gismondi[1]

ABSTRACT

The problem faced concerns the selection of a panel of "quick" respondents, representative of the whole target population, from which quick 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.

KEY WORDS: Balanced sampling; Cluster analysis; Predictor; Provisional estimation; Retail trade.

RÉSUMÉ

Le problème ici étudié regarde la sélection d'un panel des répondants «rapides», qui rèpresent tous la population étudiée, de la quelle on peut dériver des estimations anticipâtes nécessaires pour l'analyse conjoncturelle. Dans ce but on était proposé une particulière adaptation de la théorie du échantillon équilibré, avec des propositions opératives, qui regardent la technique de sélection des unités et une application empirique à des données réelles tirées de l'enquête mensuelle sur le commerce au détail de l'ISTAT.

MOTS CLÉS : Analyse par grappes; commerce au détail; échantillons équilibrés; prévision; provisoire estimation.

1. PROVISIONAL SHORT-TERM ESTIMATES: NEEDS FOR RETAIL TRADE

Timeliness is maybe the most relevant feature of short-term statistics. However, there is an obvious trade/off between timeliness and overall quality, when intended, as commonly done in practice, in terms of precision of estimates. This problem, even though widely faced in literature, should be carefully monitored in each specific contest concerning calculation and diffusion of short-term indicators. One solution could consist in building up a statistical system able to calculate and spread out both definitive and provisional (quick) data.

At the moment, the retail trade turnover index represents the only *monthly* economic indicator on 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. Up to now, no provisional quick data are available for publication at *m*+30, delay that is considered rather satisfactory for short-term economic analysis.

The retail trade sample survey, referred to Division 52 of the NACE nomenclature, was based in 2003 on a sample of 7.122 enterprises, drawn from a population of about 570 thousands. Each year about 2.000 enterprises are rotated. Questionnaires are received mainly by telefax (50%), then by ordinary mail (45%), e-mail or web (5%). The final average rate of response is about 70%. 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 yearly turnover referred to the base year 2000.

Delay in publication mostly depends on response burden, need to use external accountants for filling in questionnaires and delay occurred when using ordinary mail. At the moment, reduction of this delay within 30 days for the whole sample is not a realistic goal, while the possibility to calculate *provisional* estimates at *t*+30 can be satisfied within a short time. Generally, provisional retail trade index estimations could be carried out on the basis of two main strategies:

[1] Roberto Gismondi (gismondi@istat.it), ISTAT, Viale Liegi 13, Roma, Italy, 00198

1) using forecasts based on ARIMA models. However, in such way only historical data will be used, without any information on the trend of the month of reference based on the "quick" observation of a part of the whole sample.
2) Using data related to the reference month *m* and to a part of the units included in the sample (a *panel*, or a natural sub-sample of quick respondents), whose data related to month *m* are available within 30 days.

While no relevant empirical attempts have been carried out, up to now, for what attains the first methodology, various experiences exist related to the second one. However, a relevant problem is still unsolved, that is how to identify an optimal sub-sample of quick respondents. In particular, Gismondi (1996) showed that the use of whatever quick respondent generally produces a structural bias for the retail trade provisional index.
According to EUROSTAT's needs (paragraph 2), under a very general super-population model we'll recall the main definitions of balanced sampling (paragraph 3), while in paragraph 4 we'll propose a technique to find a balanced sub-sample including quick respondents. Main results of empirical attempts are resumed in paragraph 5.

2. THE EUROPEAN COUNTRY-STRATIFIED SAMPLE FOR RETAIL TRADE

EUROSTAT, the European Union statistical office, actually calculates and spreads out an overall EU retail trade monthly index based on a weighted arithmetic mean of the single EU countries indexes. The delay of publication is about 60 days from the end of the reference month and is considered too large by 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, on the basis of which provisional quick indexes at the EU level could be calculated within a delay of about 30 days.
The basic idea is that for defining overall size and breakdown by country of a European sample able to produce quite precise quick provisional estimates at the EU level, it can be possible to think each country as a single stratum and to split an overall quick sample size – fixed according to a 1% level of sampling error – by country according to the Neyman allocation. In this way a relatively small EU sample – obtained summing up all national sub-samples – could guarantee, on the average, small estimate errors.
According to the optimal Neyman allocation, EUROSTAT calculated that Italy, starting from 2003, should use for quick estimates a sub-sample of 1.929 retail enterprises, to be drawn from the whole sample of 7.122 units. The most part of Italian retail trade firms are very small, the actual whole sample is only the 1,25% of the universe (and only the 0,55% of enterprise with 1 or 2 persons employed) and the Neyman quick sample is the 27,09% of the whole sample. Since enterprises selling food products are more heterogeneous respect to those selling non-food products, the relative weight of the formers in the quick sample is higher than in the whole sample, so that almost the fifty percent of enterprises in the whole sample selling food products are in the quick sample as well.
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 basic idea consists in selecting a sub-sample which average 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 very far from normality, so that use of simple random sampling in each stratum – especially for small strata – could not lead to satisfactory results.
3) Even if an optimal quick sample can be identified, not 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.

Leaving outside the third aspect, common estimators (or *predictors* in a model-based context) can be improved using additional information on sample units, as historical monthly data available for year 2002 (paragraph 5).

3. SUPERPOPULATION MODEL AND BALANCED SAMPLING

We'll refer to a whatever sample stratum of the ISTAT retail trade survey, among the ten considered by EUROSTAT for the Neyman allocation, obtained crossing 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* is the size of the (optimal) sub-sample to be selected. In each stratum we'll suppose as true the following regression model (*R*):

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

where expected values, variances and covariances are referred to the model, y is turnover, x is an additional variable correlated with y and to be specified, as well as the function v_i, with β and σ^2 given, but generally unknown parameters. In general, if a sample s of size n is drawn from a population U of size N, under model (3.1) the optimal linear predictor of the (monthly) population mean $\bar{y}$ – e.g. the linear estimator minimising the model *MSE* – is given by (Park, 2002):

$$T^* = \bar{y}_s\left(\frac{n}{N}\right) + \bar{x}_{\bar{s}}\hat{\beta}\left(\frac{N-n}{N}\right) \quad \text{where} \quad \hat{\beta} = \frac{\sum_s x_i y_i / v_i}{\sum_s x_i^2 / v_i} \tag{3.2}$$

and its variance respect to the model will be equal to:

$$E(T-\bar{Y})^2 = \left[\left(\sum_{\bar{s}} x_i\right)^2 \Big/ \left(\sum_s (x_i^2/v_i)\right) + \sum_{\bar{s}} v_i\right]\frac{\sigma^2}{N^2} \tag{3.3}$$

Relevant particular cases are got if v=1 – when (3.2) reduces to the regression estimator – and v=x – when (3.2) is the common ratio estimator. Let's note that under model (3.1) the sample mean is optimal if and only if x=v=1. Formula (3.3) suggests that the best choice of the (sub)sample simply consists in selecting the n units in the universe with the biggest x-values. However, a strategy based on estimator (3.2) and the n biggest units could be dangerous, for at least two reasons:

1. quality of estimates strongly depends on the validity of all assumptions in model (3.1).
2. The choice of the n biggest units doesn't assure a low variance, because it depends on the relative weight of these units on the overall x-amount in the universe.

Given these premises, the recourse to *balanced sampling* could significantly improve quality of estimates. 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:

$$\sum_s x_i / n\sqrt{v_i} = \sum_U x_i \Big/ \sum_U \sqrt{v_i} \tag{3.4}$$

It could be chosen among all the possible samples of size n using various algorithms, as those proposed by Droesbeke, Fichet and Tassi (1987), Deville and Grosbas (1988), 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 is given by:

$$\hat{T} = n^{-1}\left(\sum_U \sqrt{v_i}/N\right)\left(\sum_s y_i/\sqrt{v_i}\right) \tag{3.5}$$

Under the common statements v=1 and v=x the optimal predictors derived from (3.2) will reduce, respectively, to:

$$\hat{T}_0 = n^{-1}\sum_s y_i \quad \text{and} \quad \hat{T}_1 = n^{-1}\left(\sum_s y_i/\sqrt{x_i}\right)\left(\sum_U \sqrt{x_i}/N\right) \tag{3.6}$$

so that if the sample is balanced the sample mean is still optimal even when $x \neq 1$. Points in favour of the use of balanced sampling are the following ones:

1. it preserves from a bias when the model (3.1) is wrong.
2. It drives the optimal choice of the sample, reducing the search to the subset including only balanced samples.
3. In practice, search for balanced samples means to find what is commonly said a *representative panel*, often chosen in a deterministic subjective way.

Let's note that while balance as in (3.4) tries to calibrate *ex-ante* the choice of the sample, the recourse to *calibration estimators* (Deville and Särndal, 1992) is aimed at calibrating *ex-post* sample estimates respect to some marginal constraints. That's the procedure used, for instance, in many of the business structural surveys carried out by ISTAT.
Even though properties of balanced sampling represent a theoretical tool useful for our context, serious problems occur concerning both the model correct identification and the use of an algorithm to select, *if it exists*, a balanced sample.

4. *QUASI*-BALANCED SAMPLE SELECTION

Let's suppose to refer to a given stratum including N units, and to know for each unit i the values x_i and v_i. Instead of drawing a simple random sampling or a systematic sampling, we can divide the stratum population into n sub-strata

including each N_h units. From each sub-stratum h a single unit i is drawn, e.g. the one minimising a loss function – to be defined further – so that, remembering (3.4), this identity can be considered approximately true:

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

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:

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

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 how can be obtained an optimal (sub)stratification is still unsolved, depending the choice on the concentration of x in the population, the sampling technique and the kind of estimator used. As a premise, we remark how all the following considerations hold when v=1, otherwise we can write (4.1) using this approximation:

$$\sum_{j=1}^{N_h} x_{hj} \Big/ \sum_{j=1}^{N_h} \sqrt{v_{hj}} \approx N_h^{-1} \sum_{j=1}^{N_h} x_{hj} \Big/ \sqrt{v_{hj}} \qquad (4.3)$$

so that unit i in (4.1) must have a value of $z_{hj} = x_{hj} / v_{hj}^{0,5}$ as much 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 $z_i^{0,5}$. Boundaries of n sub-strata can be obtained imposing that each sub-stratum must cover the same cumulative value Z/n, where Z is the global amount of z in the stratum. In practice, the goodness of this method could be satisfactory only if strata are numerous and narrow. Moreover, if more than one z-value is available it should be applied separately for each z, leading to *different* optimal samples (Dorfman and Valliant, 2000).

An alternative idea is driven by the fact that the dangerousness 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.

Concerning point 2), in each sub-stratum h we can select the unit i satisfying the condition:

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

Even though samples selected as above could not be exactly balanced, they present these advantages:

a) *quasi*-balanced samples are selected considering *for each unit* the degree of distance respect to the mean.
b) 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).
c) A (sub)optimal result should be always guaranteed whatever is n according to optimal predictors defined before.

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

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

where z^* indicate values standardised 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, 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. This second selection method will be defined *z-multivariate*.

5. MAIN EMPIRICAL RESULTS

From the ISTAT retail trade monthly survey sample we extracted 4.616 enterprises, that are those for which historical monthly data for 2002 were available. These data cover the period January-November 2003. Units not responding for at least three of the eleven months were excluded. Available sample data were broken down in 10 "universes" got crossing two main kinds of products sold (food and non-food) and five classes of persons employed (1-2, 3-5, 6-9, 10-19, >19). The aim consists in drawing an optimal sub-sample from each universe, itself a part of the overall retail trade sample.

For each of these ten populations we decided to look 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 as defined in paragraph 4. The purpose consists in estimating *ex-post* the (known) average monthly turnover referred to each of the five months from July to November and to verify the degree of error in estimates. So, we have K=6 and each x-variable is turnover for each month from January to June 2002, so that $x_k=y_{m-k}$ for k=1,2,…6 and m=7,…11. 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 $z=x/v^{0,5}$ for the balance condition (4.1).

As a preliminary step, efficiency of balanced sampling was compared with more used and traditional procedures as well. Comparisons among percent errors calculated as an average of estimate errors referred to months from July to November showed how the sample mean was significantly more precise when used under a quasi-balanced sampling instead of simple random sampling (error equal to 8,25%) or systematic sampling (10,56%). In particular, on the average the *z-multivariate* technique produced better results both for v=1 (the average error is equal to 1,79%) and $v=x$ (1,69%), while balance didn't produce good results when v=1 and the *z-univariate* procedure is used (7,36%), while good results were got putting $v=x$ (2,08%). In details (see table 1), for what concerns balance errors we have that:

- for the type of product "total" 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%). However, the monthly variability of estimate precision – calculated as the coefficient of variation (C.v.) of monthly estimate errors – is lower for multivariate than for univariate for whatever v.
- 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%). Also for food and non-food monthly variability of estimate errors got with multivariate is lower than *z-univariate*.

In short, for food, non-food and total the best performances of *z-multivariate* are 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).

Table 1: Balance and forecast results using the *z*-univariate and the *z*-multivariate methods (*) C.v. = coefficient of variation

				Balance (Jan-Jun 02)		Forecasts (Jul-Nov 02)		Forecasts (03)
Domain	*N*	*n*	*n/N*	*z*-univariate	*z*-multivariate	*z*-univariate	*z*-multivariate	*z*-multivariate
FOOD	1.703	1.008	59,2					
Average turnover (Euro)				1.630.502	1.630.502	1.675.078	1.675.078	1.131.754
Average % error (v=1)				5,07	1,12	7,69	1,30	
Average % error ($v=x$)				1,22	2,75	1,21	1,24	0,99
Error C.v. (v=1) (*)				0,47	0,29	0,52	0,39	
Error C.v. ($v=x$) (*)				0,44	0,37	0,45	0,43	0,64
NON FOOD	2.913	921	31,6					
Average turnover (Euro)				346.494	346.494	365.477	365.477	246.105
Average % error (v=1)				7,33	1,61	6,71	2,73	
Average % error ($v=x$)				3,44	2,91	3,71	2,50	0,23
Error C.v. (v=1) (*)				0,69	0,51	0,74	0,62	
Error C.v. ($v=x$) (*)				0,61	0,27	0,68	0,39	0,88
TOTAL	4.616	1.929	41,8					
Average turnover (Euro)				729.249	729.249	755.860	755.860	531.229
Average % error (v=1)				5,82	1,29	7,36	1,79	
Average % error ($v=x$)				1,98	2,83	2,08	1,69	0,29
Error C.v. (v=1) (*)				0,54	0,37	0,48	0,39	
Error C.v. ($v=x$) (*)				0,52	0,26	0,50	0,32	0,42

The effective precision of estimates was assessed evaluating precision of quick estimates in comparisons with definitive estimates for months from July to November. 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 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 stress 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*. Moreover, the coefficient of variation of forecast errors is quite lower with *z-multivariate*, without particular differences between v=1 and v=x, and this is true for both food and non-food.
- 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. Also for food and non-food, the monthly variability of estimate errors is quite always lower with *z-multivariate*.

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.
By the way, the *z-multivariate* approach under the hypothesis v=x was just applied to get provisional estimates for the first three months of 2003. The average quarterly estimate error was equal to 0,29%, and it was more difficult to estimate indexes for food (0,99%) that non-food (0,23%). That can be due to the higher variability of individual data and to a more sprightly longitudinal dynamic related to enterprises selling food rather than non-food products.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques* (3rd. ed.). New York: John Wiley.

Deville, J.C., Grosbas, 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.* Paris: Economica.

Gismondi, R. (1996). Effects of non-responses in the retail trade monthly survey. *Research Copybooks*, **4**, 199-236, Rome: Istat.

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.

Park, M. (2002). Regression estimation of the mean in survey sampling. *PHD thesis*, Ames, Iowa: Iowa State University.

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.

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*. New York: John Wiley.