Integration Among Statistical Sources: Some Methodological Proposals

Roberto Gismondi (*)

(*) ISTAT - Servizio SCO

Sommario

Il problema della integrazione tra più fonti statistiche costitutisce una delle tematiche metodologiche di maggiore attualità. Ciò è dovuto soprattutto alla necessità di ridurre il fastidio statistico sui rispondenti, in un contesto internazionale fortemente caratterizzato dalla necessità di disporre di un maggior numero di informazioni statistiche su famiglie ed imprese. In questo contesto, il documento descrive alcune strategie operative finalizzate alla raccolta di informazioni utilizzabili a fini statistici. Successivamente, partendo da uno stimatore generalizzato basato su due fonti statistiche relative ad una variabile di interesse, vengono proposti alcuni criteri per determinare i pesi da assegnare alle due fonti. Una applicazione empirica relativa al contesto nazionale conclude il documento.

Abstract

While statistics are quickly moving towards a more and more wide and integrated International Statistical System, the need to contain the statistical nuisance is forcing many countries in developing strategies aimed at using all the existing statistical data and at limiting the recourse to new survey if not strictly necessary. In this paper, starting from a general weighted estimator based on two statistical sources concerning the same variable of interest, we'll describe and compare possible choices of the weights to be assigned to each source, mostly based on qualitative, "optimal" and probabilistic approaches. A practical "case study" concerning the Italian context ends the article.

1. Premise[1]

While statistics are quickly moving towards a more and more wide and integrated International Statistical System, in many practical circumstances unsustainable problems are met when addressing to households or enterprises new statistical questionnaires, which compilation is often paid in terms of monetary costs and/or waste of time. In short, the need to contain the statistical nuisance is forcing many countries in developing strategies aimed at using all the existing statistical data and at limiting the recourse to new survey if not strictly necessary (no available source on the phenomenon of interest). Since in many cases more than one source concerning the same variable exists, what we often need is the skill in how merging these sources in some optimal way, in order to obtain a composed estimator which efficiency could reveal to be rather high.

Starting from this last need, this paper summarises and compares some statistical techniques – some of which are original proposals – for integrating two (or more) quantitative statistical sources concerning the same variable of interest. Before proceeding, we must underline that, in a wide sense, the need to integrate data coming from different sources is in any case very common because of the wide set of possible uses of data that could be done, and on this point some examples could be helpful.

First of all, in some circumstances there is a clear trust in one particular source and the redundant use of statistics aims to confirm or update some information: let's think to the case of variables requested in a questionnaire but yet known because included into an original archive from which a sample has been drawn (employment, for instance). This is a situation where there is a strong *a priori* feeling that one source (the archive) should be very reliable, and the recourse to the second source (the sample survey) is just a tool to increase the informative content and the degree of precision of the archive itself.

If we exclude this case, the other situations in which problems related to the complex process of comparison among sources are present could be reduced to three main cases:

1. integration;
2. synthesis;
3. weighting,

[1] Obviously the opinions herein expressed are not necessarily the same of ISTAT and can be attributed to the author only. Even if in this frame a stronger attention is going to be paid in favour of enterprise statistics, some of the technical and qualitative considerations that will follow could be successfully applied for household statistics as well.

where only the third case is strictly concerned with the implicit need that all sources should refer *to the same object of interest* (an amount, a change, a proportion), that is the particular problem to which this paper is addressed to.

In fact in the first situation what we do in practice consists in using more than one source in order to achieve at the estimation of an unknown parameter. A first case, very common in practice, consists in aggregating, with opportune weights, data concerning partial estimates in order to estimate the variable of interest for the whole domain of reference. For instance, we could have separate estimates of value added for the enterprises with less than 20 persons employed (S_1) and with more than 19 (S_2), finally synthesised by means of deterministic weights (P_1 and P_2) in a unique estimate $S_1P_1+S_2P_2$. Another example concerns a survey carried out using both paper and electronic questionnaires, depending on the choice of the respondents based on their degree of informatisation. But in practice these processes are equivalent to consider both sources as biased estimators of the overall unknown parameter of interest, transformed in a more efficient and probably unbiased estimator by a simple weighted arithmetic mean. A further case is concerned with many panel surveys aimed at calculating indexes of variation respect to a base year where the index is given by the product of the index concerning the data observed in the panel and the index related to the "demography" of the operators active in the domain under study, often derived from a different source[2].

The second situation is normally faced with multivariate techniques: if several variables are considered useful to describe a certain latent feature of a population (competitiveness, quality, welfare, etc.), the main aim is to pass from a matrix composed by *k* variables to a new one composed by 2 or 3 variables at most, generally got using factor analysis.

Finally the third situation is particularly devoted to the estimation of optimal weights to be assigned to each source: the weight, even when not introduced in an explicit way, is a tool to fix our degree of confidence on a certain source, depending or not on objective qualitative evaluations. All the paper refers mostly to this situation, even if some results could be useful for the situation 1 as well.

The relevance of this topic is strongly confirmed if we analyse how the EU member states are able to fulfil all the requests of the Structural Business Statistics Regulation (EUROSTAT, 1997), mostly satisfied using both administrative and sampling data. For what concerns the Short-term Regulation's needs (EUROSTAT, 1998) the EU member states make use of data collected in

[2] Given that the overall index could be written as the product of two sources, this case could be reconnected to the above mentioned weighted mean of two sources if we put $P_1=(S_1S_2-S_2)/(S_1-S_2)$.

more than one survey: in particular (EUROSTAT, 1999), 3 member states for what concerns turnover (France, Luxembourg and Finland) and 4 member states for what concerns employment (Denmark, France, Netherlands and Finland), while Italy will probably follow a mixed approach as well, yet to be better defined. These cases could be reconnected to the situations 1 and 3 above mentioned.

The previous cases are not always simple to be detected and faced with the necessary care. Also for this reason in the following paragraph we'll describe other practical frames all characterised - even if in different ways - by a process of integration among statistical sources, sometimes in implicit form. Then in paragraph 3 the general weighted estimator based on the synthesis of two sources will be deeply commented. After paragraph 4 - where the case of a qualitative approach is also introduced - paragraph 5 contains an empirical "case study" referred to the Italian frame. Original methodological proposals mainly concern the second part of paragraph 2 and the sub-paragraphs 3.3 and 3.4.

2. Sources of statistical information and statistical strategies

Developing and running surveys is not the only option for capturing enterprise data. There are other sources of information for many kinds of industries, and some of them are administrative sources, even if not so many practical cases of real use of these kind of data are available at the moment.

The real problem consists in the transition from a situation characterised by the simple *coexistence* of administrative and statistical sources to a new one, where several attempts are spent in order to integrate both sources into a common informative system able to rationalise what is existing in the field of statistical information (Monducci, 1997).

In many cases owners of administrative files cannot evaluate the informative content they could manage, so it's first necessary to change their way of thinking for what concerns statistics and their use in the society and then to elaborate technical tools to guarantee the best use of the existing statistical sources.

Some of the possible strategies useful to pick up enterprise data are shortly resumed in the next sub-paragraphs; among them, the case discussed in sub-paragraph 2.4 is the basis for all the following arguments.

2.1 Recourse to surveys

Notwithstanding direct surveys are the most traditional way used by statisticians to capture data, increasing difficulties in carrying out surveys are pushing many National Statistical Institutes towards the recourse to other statistical source. In other words, if some data yet exist we could avoid to build up a new *ad hoc* survey concerning the field of interest, also because hard problems to solve concerning surveys are:

1. missing values and editing procedures;
2. recourse to archives not really so updated;
3. in many cases, too heavy costs of response for enterprises.

Even the EU Regulations are sometimes concerned with problems of response burden and comparability among similar sources. For instance, in the Short-term Business Statistics Regulation it's assessed that "...*It is necessary to have reliable and rapid statistics*..." but, on the other hand, for what concerns quality from Article 10 we derive that "...*The quality of the variables shall be measured by each Member State according to common criteria...The quality of the variables shall be tested by comparing them with other statistical information...Quality evaluation shall be carried out comparing the benefits of the availability of the data with the costs of collection and the burden on business, especially on small enterprises*...".

So a great elasticity seems to rise up from these sentences, as well as the implicit request to compare data when possible in order to better use all the statistical information yet existing.

2.2 Use of VAT data

The VAT fiscal data are normally available in all the most developed countries, with different degrees of coverage and delay.

Among the main problems concerning the recourse to administrative data, we mention: different and often not linkable classifications, delays and not exhaustiveness of the observable domain.

It seems realistic to imagine a statistical infra-annual production process based on the integration of administrative data by statistical models able to reduce or cancel bias and to link more directly administrative data to the real trends we want to observe.

In any case, as also suggested by Thomsen (1988), a strategy finalised to a systematic use of administrative sources generally should be based on these three main components:

1. a system of well co-ordinated administrative records; this system should primarily guarantee that public administration is more effective, and in any case should be co-ordinated by the central statistical office;
2. the link among administrative files and a main statistical archive of reference, that represents one of the most difficult aims to reach and maintain along time;
3. a deeper knowledge of what exists in administrative archives and how it could be used for statistical purposes.

2.3 Proxy variables

This field is strongly concerned with short-term data (monthly or quarterly). When available, sub-annual employment sources (administrative and/or derived from surveys) could be used to describe the sub-annual dynamic of output variables not directly observed. In this case the ratio of labour input to gross output is assumed to be constant in the short term, which allows hours worked to be used as a projector for most of industries, mainly if belonging to the service sector.

In practice, recent experiences carried out in some highly developed countries, like Canada (see Mc Mechan and Marcil, 1998) and New Zealand (Mc Kenzie, 1998) show how, generally, labour input bears only a faint connection with output, in particular when indexes of output at constant prices are needed; the only cases of good approximation concern highly concentrated sectors (for instance, air transports) or activities for which the under-coverage of employment as derived from administrative declarations is low (for instance, financial services).

Other proxy variables of infra-annual turnover indexes could be given by the corresponding sub-annual provisional quantitative indicators, when existing. For instance, at the moment in Italy no sub-annual survey on hotels and other receptive accommodations exist, but quantity indicators on arrivals and nights spent are currently available with about 3 months of delay from the last month of reference, so that they could be useful to satisfy the Short-term Business Statistics Regulation needs.

The recourse to sector quantity indicators (available also for road, air and maritime transports) could be more useful than employment data because reflects more directly the output infra-annual seasonally pattern, but could produce misleading results, also because it doesn’t take into account the dynamic of prices. In this field it’s possible to include the use of indicators currently produced by other bodies besides the National Statistical Institute (D'Alessandro and Pisani, 1996).

The conclusion is that, before starting to produce output indexes using available proxy variables, comparisons should be done for those sectors for which both data coming from proxies and current surveys exist, in order to evaluate the degree of discrepancy between them.

2.4 Synthesis of two (or more) sources

From the previous synthetic resume, we should accept since now that many of the future statistics will be based on more than one source, and probably the best solution for a certain economic sector could be quite different from the optimal solution for another.

In any case the underlying philosophy actually on among many European Union countries is in favour of the maximum and reasonable use of all existing data, coming out or not from direct surveys, which cost in terms of monetary budgets, response burden and need of human resources in many circumstances seems to be unsustainable.

So the most useful strategy for many statistical agencies should be probably based on these two actions:

- picking up all data concerning a certain variable of interest;
- harmonising these data with the recourse to particular statistical processes.

Given these premises, in the following paragraphs we'll propose a resume of techniques and procedures aimed at obtaining more efficient estimators when more than one statistical source is available concerning a given *Y* variable of interest, while in the follow the symbol Y will mean the corresponding amount.

Just to put in evidence the real practical significance of the concept of *coherence* - e.g. comparability of sources concerning the same topic of interest - this example could be helpful, from which is evident how the idea of integration among sources is strongly taken into account for solving problems apparently concerning other matters. We could suppose to be interested to a quantitative variable *Y*, to refer to a certain year A and to observe at time *t* a statistical source $S^1_{\mathrm{A}t}$, that is an estimate of the true unknown value $\mathrm{Y}_{\mathrm{A}t}$, with *t*=1,...,*k* (for instance, with *k*=4 we have quarterly observations). After the end of year A we suppose to know the true value Y_{A} referred to the whole year A, and got from administrative data or a structural yearly survey, supposed to be more precise than the infra-annual statistics. If all the above mentioned figures are yet known for the previous year (A-1), and if the global yearly amount $\mathrm{Y}_{(\text{A-1})}$ is always obtainable as a weighted arithmetic mean of the *k* infra-annual estimates with weights w_t (each weight could be equal to 1 or to 1/*k*) we could find, for year (A-1), the new infra-annual estimates $S^2_{(\text{A-1})t}$, trying to minimise the average distance from the original estimates $S^1_{(\text{A-1})t}$, with the constraint that the new infra-annual estimates must reproduce the yearly true amount $\mathrm{Y}_{(\text{A-1})}$. So, if $\mathrm{Y}^1_{(\text{A-1})} = \sum_{t=1}^{k} S^1_{(\text{A-1})t}\, w_t$, minimising this Lagrange's function:

$$\sum_{t=1}^{k}[S^2_{(\text{A}-1)t} - S^1_{(\text{A}-1)t}]^2 + \lambda\left[\sum_{t=1}^{k} S^2_{(\text{A}-1)t}\, w_t - \mathrm{Y}_{(\text{A}-1)}\right]$$

we can easily obtain the optimal solution:

$$S^2_{(\text{A}-1)t} = S^1_{(\text{A}-1)t} + \left\{ w_t[\mathrm{Y}^1_{(\text{A}-1)} - \mathrm{Y}_{(\text{A}-1)}] \Big/ \sum_{t=1}^{k} w_t^2 \right\}. \qquad (2.4.1)$$

Now, if during year A we need to correct the infra-annual estimates $S_{1\text{A}t}$ on the basis of the previous formula we could put:

$$S^2_{\text{A}T} = S^1_{\text{A}T} + [S^2_{(\text{A-1})T} - S^1_{(\text{A-1})T}]\left[\sum_{t=1}^{T} v_t\, S^1_{\text{A}t} \Big/ \sum_{t=1}^{T} v_t\, S^1_{(\text{A-1})t}\right] \qquad (2.4.2)$$

for a given time *T*, where the coefficients v_t are seasonal parameters useful to give a sense to the ratio into the square brackets. This procedure is very simple and could be useful for adjusting provisional short-term data in order to render them more similar to the definitive ones when a benchmark value $\mathrm{Y}_{(\text{A-1})}$ is known.

This is the typical situation occurring when an integration between the Short-term and the Structural Business Regulations is needed, even if very different approaches could be used to solve this problem.

3. Linear combination of two statistical sources

As said before, the general formula we'll consider in the follow will be given by the source:

$$S = S_1 P_1 + S_2 (1 - P_1) \tag{3.1}$$

obtained as a weighted arithmetic mean of two sources P_1 and P_2. These weights can be assigned on the basis of different approaches, herein briefly resumed.

3.1 Qualitative and heuristic selection of weights

The simplest situation concerns the case in which we assign subjective weights in the previous formula: for instance, when we use only one source (one of the two weights is equal to zero) or when the two weights are equal to 0.5. These options are common in practice, but could be dangerous especially because of the lack of optimality in the case where both sources are considered as random variables. In other words, it could happen that the average of two sources it's less efficient than one single source in terms of variance of the final estimate.

In a wide definition, the general formula to determine the first source's weight could be written as:

$$P_1 = \frac{f(S_1; S_2)}{g(S_1; S_2)} \tag{3.1.1}$$

where *f* and *g* are two functions to be determined, concerning particular features of the two sources.

A first particular case derived from (3.1.1) occurs when we are able to evaluate, but often in a subjective way only, the overall quality of each source, considering quality as the result of more than one performance desired for an estimator[3]. The weight assigned to each source could be directly proportional to its quality (indicated with the symbol *Q*), following the simple formula:

$$P_1 = \frac{Q_1}{(Q_1 + Q_2)} \tag{3.1.2}$$

and this formal structure will be common also in the next paragraphs. For instance, in sub-paragraph 3.2 the quality of each source is inversely correlated with its variance.

A second situation occurs when we can evaluate the cost of each source in terms, for instance, of monetary and/or human resources, fixed capital, use of time to elaborate data in order to render them useful for our purposes, etc.. If the costs of the two sources are given, respectively, by C_1 and C_2, we could assign a lower weight to the source costing less, putting:

$$P_1 = \frac{C_2}{(C_1 + C_2)}. \tag{3.1.3}$$

If both the sources come from sampling surveys, an approximation of the previous formula could be given by the ratio $n_2/(n_1+n_2)$, where n indicates the number of sampling units, that in this case are supposed to have the same statistical cost.

3.2 Some optimal choices

A family of methods for estimating the weights in (3.1.1) derives from the stochastic properties of the estimators S_1 and S_2 concerned with the two sources. As well known on the basis of the latest literature on this topic (For instance, Falorsi and Russo, 1998), the first situation derived from the general estimator (3.1) concerns the case in which both the statistical sources provide unbiased estimators of the unknown amount Y so that, as a consequence, also S will be unbiased if P_1 is not itself the result of a random variable, as we'll see at the end of this paragraph. If the symbols V and COV indicate the operators variance and covariance referred to the model[4], with the symbols 1 and 2 referring respectively to S_1 and S_2 it's easy to get the formula:

$$V(S) = V_1 P_1^2 + V_2 (1 - P_1)^2 + 2 P_1 (1 - P_1) COV_{12}.$$

[3] Following the EUROSTAT's definitions (EUROSTAT, 1996), some features of quality are given by pertinence, accuracy, opportuneness, consistency, dissemination, cost, comparability.

[4] In other words the variability of estimates will be evaluated referring to the superpopulation model underlying the observations, and not in relation with a sampling design. This choice is mostly due to the fact that we could deal with non-sampling sources (administrative files, census data affected by non-sampling errors, etc.). Clearly if we deal with two sampling sources we could directly refer to sampling variances, as we'll see in paragraph 5.

As previously mentioned, the use of whatever weight P_1 could be dangerous: in fact, supposing for simplifying that the term of covariance is zero, we could prove that the mixed estimator (3.1) is not less efficient than each single estimator S_1 and S_2 only if:

$$\left(\frac{1-P_1}{1+P_1}\right) \le \frac{V_1}{V_2} \le \left(\frac{2-P_1}{P_1}\right);$$

for instance, this condition is satisfied, fixed V_1=70 and V_2=30, when P_1=0.5 but not when P_1=0,75. From the previous formula of the variance, deriving respect to P_1 we obtain the optimal solutions:

$$P_1 = \frac{V_2 - COV_{12}}{V_1 + V_2 - 2\,COV_{12}} \qquad \text{and} \qquad 1 - P_1 = \frac{V_1 - COV_{12}}{V_1 + V_2 - 2\,COV_{12}}, \qquad (3.2.1)$$

which reduce to the simpler ratios between each variance and the sum of variances if the covariance between the two estimators is (approximately) equal to zero[5]. This fundamental property is not necessarily satisfied even if just one of the two estimators is biased, as it's very common in practice when one of the sources derives from administrative archives or is referred to a domain wider than the one under observation (for instance, a geographical area greater than the one object of estimation by the formula 3.2.1). If the squared bias of the second source is given by $(BIAS_2)^2 = (E_2 - \mathrm{Y})^2$, where E means stochastic expectation as regards the model, the mean squared error (*MSE*) of (3.1.1) will be given by:

$$MSE(S) = V(S) + [BIAS\,(S)]^2 = V(S) + [(1-P_1)\,BIAS_2]^2.$$

and these optimal solutions will follow:

$$P_1 = \frac{V_2 - COV_{12} + (BIAS_2)^2}{V_1 + V_2 - 2\,COV_{12} + (BIAS_2)^2} \quad \text{and} \quad 1 - P_1 = \frac{V_1 - COV_{12}}{V_1 + V_2 - 2\,COV_{12} + (BIAS_2)^2}, \qquad (3.2.2)$$

from which it's clear that, *coeteris paribus*, the higher is the bias of the second estimator, the lower will be its relative weight.

A simple generalisation of the formula (3.1) useful in presence of more than two sources is

[5] Depotout and Arondel (1998).

given by:

$$S = \sum_{h=1}^{k} S_h \, P_h \qquad (3.2.3)$$

where we suppose to have *k* different statistical sources S_h. For simplicity, we'll suppose absence of correlation among the single sources, because of the difficulty to get explicit solutions of the various minimisation problems, so that $COV(S_h;S_r)=0$ for *h≠r*. We can simply find the optimal weights P_h with the constraint that their sum is equal to one, in a case similar to the first one seen in paragraph (3.2). If V_h is the variance of the *h-th* source, it's easy to obtain the formula:

$$P_h = \frac{1}{V_h} \Bigg/ \left(\sum_{h=1}^{k} \frac{1}{V_h} \right) \qquad (3.2.4)$$

which reduces to (3.2.1) if *k*=2.

In practice, it's difficult to have more than three or, in some cases, four (possibly unbiased) sources for the same amount, so that the use of the previous formula can be considered not usual. Moreover, in the case of more than two sources it could be possible first to mix couples of them until we reduce their number to two, and then is possible to use formulas introduced in paragraph 3.2.

An approach in some way comparable to the one derived from the formula (3.2.3) is based on well known multivariate techniques as factorial analysis which, reduced to its essential scope, aims at calculating new *underlying* variables given by the averages of *k* sources of information concerning the same latent variable not directly observable.

Clearly the use of the optimal weights (3.2.1) and (3.2.2) is possible if we could know - or at least evaluate - all the parameters in the corresponding formulas. In practice we must estimate these parameters, mainly using time series concerning both the estimators and, in the case of (3.2.2), the true *Y*-values as well[6].

3.3 Least squares techniques

A particular aspect concerned with the estimator (3.1) is easily obtainable, as also suggested in Falorsi and Falorsi (1994) and Thomsen (1988), if time series referred to the true *Y*-amounts are available. In other words, we suppose to know the *Y*-values for at least one period after the one

[6] On this matter see Gismondi (1996).

selected as reference period for the estimate of P_1 (and obviously of P_2 too). Otherwise, we could suppose to know the Y-values for the reference period but concerning a particular domain (for instance, other economic sectors, a particular subset of enterprises, or some specific geographic areas).

If the symbol n is used to indicate the length of a time series (in this case n=T) or the size of a particular domain for which the true amounts Y are available, we could find the optimal P_1 as the weight able to minimise the sum of the squared differences between true and estimate data, given by the function:

$$\sum_{i=1}^{n}\left[S_{1i}P_1+S_{2i}(1-P_1)-\mathrm{Y}_i\right]^2 .$$

In practice we want to estimate P_1 using the usual least squares technique, supposing implicitly both S_1 and S_2 as explanatory variables for Y with the constraint given by the sum of the two unknown parameters equal to one. In particular, the recourse to least squares is justified by the hypotheses of disturbance terms identically and independently distributed and characterised by homoschedasticity, and this last assumption seems realistic for phenomena that don't present strong changes in size along time and – as it's reasonable – with the two sources similar in level. We can show that the optimal value of the first weight is given by

$$P_1=\left[\sum_{i=1}^{n}(S_{1i}-S_{2i})(\mathrm{Y}_i-S_{2i})\right]\Big/\left[\sum_{i=1}^{n}(S_{1i}-S_{2i})^2\right]. \qquad (3.3.1)$$

from which is evident how could be relevant the influence of even few high differences between the two sources or the second source and the true value, and these high differences could be mainly due to the increase of the Y-values along time and not necessarily to a real increase of unprecision of estimates.

Let's note that, mainly when the parameter of interest is given by a change instead of an amount, we could prefer to minimise the sum of the "relative" estimate errors, given by:

$$\sum_{i=1}^{n}\left[\frac{S_{1i}P_1+S_{2i}(1-P_1)}{\mathrm{Y}_i}-1\right]^2$$

and in this case we obtain the optimal choice:

$$P_1 = \left[\sum_{i=1}^{n} \frac{(S_{1i} - S_{2i})(\mathrm{Y}_i - S_{2i})}{\mathrm{Y}_i^2} \right] \bigg/ \left[\sum_{i=1}^{n} \frac{(S_{1i} - S_{2i})^2}{\mathrm{Y}_i^2} \right]. \qquad (3.3.2)$$

Both the solutions satisfy the condition that the sum of the estimated values is equal to the sum of the true ones, because of the peculiar properties of all the least square estimates. As an example, at the moment in Italy two official sources concerning tourism statistics are available: the first concerns the supply side (nights spent in hotels and complementary accommodations) and the second the demand side (nights spent by residents in Italy). These sources can provide quarterly data on nights spent by tourists at the regional level (20 regions), but the definitive data are available only after 6 months from the end of the quarter of reference. Really these data are requested by users with a delay of about 3 months (following the deadline imposed by the EU Directive on Tourism Statistics), when the only definitive data are available at the level of big geographic area (North/West, North/East, Centre and South/Island). Since 1997 both the survey produce comparable data, so at the moment we have at least 8 quarterly couples of regional and big area data on tourist flows. In order to estimate regional flows for a new quarter after 3 months (when only provisional regional estimates are available), we commonly use the formula (3.3.1) with *n* equal to the number of previous quarters for which both the true regional and big area data are known[7]. After having estimated P_1 we can use the estimator *S* for estimating flows at the regional level, putting in the estimator’s formula each couple of provisional regional flows.

3.4 Probabilistic approaches

A very intuitive selection of the weight in the general formula (3.1) concerning the first source could be based on the “probability” that this source is more reliable than the second one. On this field, the following techniques are relatively easy to be implemented.

Approach 1

A first, very simple case is when we consider P_1 equal to the percent share of cases in the recent past in which the first source satisfied the following condition:

[7] In this example quarterly data should be previously seasonally adjusted, in order to represent comparable data in the formula (3.3.1).

$$(S_{1t}-\mathrm{Y}_t)^2 \leq (S_{2t}-\mathrm{Y}_t)^2 \qquad \text{for } t=1,...,T \tag{3.4.1}$$

where *T* is the period of reference. So in this case we need a time series rather long to hope in a good estimate of P_1, but the risk is the use of a sub-optimal estimator and, moreover, in practice a solution that could reveal to be better consists in putting P_1=1 if the number of cases in which the condition (3.4.1) is satisfied is higher than *T*/2.

Approach 2

The availability of time series concerning the performances of two sources in comparison with the true amounts could be used in a better way imagining to have a database with *T* records, where at the first position of the *t-th* record we have the dummy dependent variable G_t, equal to 1 if at time *t* the best estimate was S_1 (in terms of formula 3.4.1) and 0 otherwise. Further we have as independent variables the only 2 ones given by the estimates S_1 and S_2.

As well known, discriminant analysis is based on the estimate of the *ex post* conditioned probability PR that G_t=1 (that is the first source is better than the second), when S_{1t} and S_{2t} have been observed. The Bayes formula (3.4.2) is:

$$\mathrm{D_t} = \mathrm{PR}(G_\mathrm{t}=1/S_{1\mathrm{t}},S_{2\mathrm{t}}) = \frac{\mathrm{PR}(S_{1\mathrm{t}},S_{2\mathrm{t}}/G_\mathrm{t}=1)\mathrm{PR}(G_\mathrm{t}=1)}{\mathrm{PR}(S_{1\mathrm{t}},S_{2\mathrm{t}}/G_\mathrm{t}=1)\mathrm{PR}(G_\mathrm{t}=1)+\mathrm{PR}(S_{1\mathrm{t}},S_{2\mathrm{t}}/G_\mathrm{t}=0)\mathrm{PR}(G_\mathrm{t}=0)}$$

and gives us the possibility to put P_1=D or, more roughly, P_1=1 if D>0.5.

Obviously in this case some additional hypotheses on the density form of the estimates are needed, and we should analyse the share of cases that in the past would have been correctly classified using the discriminant function.

Approach 3

In a third circumstance we can suppose normal density forms for both the sources, in a situation where the availability of a time series is not needed.

We start again from the mixed estimator $S_1 P_1 + S_2(1-P_1)$ and then we can choose the probability P_1 as:

$$P_1 = \text{Prob}\left\{(S_1 - \vartheta)^2 \le (S_2 - \vartheta)^2\right\}, \qquad (3.4.3)$$

where the hypothesis that both S_1 and S_2 are unbiased estimators of θ holds. If they are also characterised by variances given, respectively, by σ_1^2 and σ_2^2 we can write the previous probability as:

$$P_1 = \text{Prob}\left\{\left(\frac{S_1 - \vartheta}{\sigma_1}\right)^2 \sigma_1^2 \le \left(\frac{S_2 - \vartheta}{\sigma_2}\right)^2 \sigma_2^2\right\} = \text{Prob}\left\{\left(\frac{S_1 - \vartheta}{\sigma_1}\right)^2 \Big/ \left(\frac{S_2 - \vartheta}{\sigma_2}\right)^2 \le \frac{\sigma_2^2}{\sigma_1^2}\right\}.$$

Then, if S_1 and S_2 are supposed to have a normal distribution, i.e. $S_1 \approx N(\theta, \sigma_1^2)$ and $S_2 \approx N(\theta, \sigma_2^2)$, and they are *independent in probability* we have that:

$$\begin{cases} \left(\dfrac{S_1 - \vartheta}{\sigma_1}\right)^2 \approx \chi^2_{(1)} & \text{being} \quad \left(\dfrac{S_1 - \vartheta}{\sigma_1}\right) \approx N(0,1) \\ \left(\dfrac{S_2 - \vartheta}{\sigma_2}\right)^2 \approx \chi^2_{(1)} & \text{being} \quad \left(\dfrac{S_2 - \vartheta}{\sigma_2}\right) \approx N(0,1) \end{cases}$$

and the main consequence is that, because of the fact that the ratio between two independent $\chi^2_{(1)}$ is given by a Fisher's F random variable with (1,1) degrees of freedom, the previous probability is simply given by:

$$P_1 = \text{Prob}\left\{F_{(1,1)} \le \frac{\sigma_2^2}{\sigma_1^2}\right\} \qquad (3.4.4)$$

and can be easily obtained using the F tables commonly available. For instance, if $\sigma_1^2 = 1$ and $\sigma_2^2 \cong 39$ we have that $P_1 \approx 0.9$

Obviously it's recommended a preliminary study about the stochastic independence between the two compared estimators. In practice, variances too different each other could cause a very high or low weight for one of the two sources.

4. A qualitative approach

Starting from the general formula (3.2.3), a methodology to choice how to assign the weights P_h could be based on some of the decision making methods proposed by Naumann (1998), introducing the slight modifications necessary to adapt the main results to our context.

Let's suppose to assign to each source a score (ranking, for example, from 1 to 10) concerning (k-1) qualitative features, where the *k-th* feature is given by the cost to sustain for getting the statistical source available. Some features could be given by timeliness, degree of "tuning" with the domain object of interest, comparability, coherence and other qualitative characteristic as the ones reported in EUROSTAT (1996) and yet mentioned. We could also suppose (but it's not necessary) to assign a weight w_i to each feature f_i, whose score on the *h-th* source is given by f_{ih}. All these (k-1) features are positively correlated with the good performance of a statistical source, while the last feature (cost of the source) must be transformed in order to be higher when the cost decreases. Moreover, all the indicators must be in any case manipulated in order to be comparable, so that these new scores can be used, according to the SAW (Simple Additive Weighting) method:

$$\begin{cases} v_{ih} = \dfrac{f_{ih} - f_i^{\min}}{f_i^{\max} - f_i^{\min}} & \textit{for} \quad i = 1,2,\ldots,(l-1) \\ v_{ih} = \dfrac{f_i^{\max} - f_{ih}}{f_i^{\max} - f_i^{\min}} & \textit{for} \qquad i = l \end{cases}$$

and the overall score for the *h-th* statistical source will be given by

$$F_h = \sum_{i=1}^{l} w_i v_{ih} \,.$$

Finally, the choice of the weights P_h could be based on these two possible rules:

$$P_h = \begin{cases} 1 & \textit{if} \qquad F_h = F_h^{\max} \\ 0 & \textit{otherwise} \end{cases} \tag{4.1}$$

$$P_h = F_h \Bigg/ \left(\sum_{h=1}^{k} F_h \right) . \tag{4.2}$$

These rules can be easily reduced to the case of only two sources analysed in paragraph 3. In any case the use of both the alternative rules should be based on a deeper investigation on their statistical properties, because in this context the decision method is more derived from a descriptive analysis of the intrinsic characteristics of the sources rather than from the application of a certain optimisation criterion[8].

5. An application to the Italian case

In this paragraph we'll present a simple application to the estimate of the overall yearly expenses for consumption of goods in Italy in the period 1991-1998, for which no "true" historical data can be considered available at the moment. On this field, ISTAT carries out two surveys[9].

The first one is given by the households' consumption survey, carried out on a monthly basis. It's based on a yearly sample of about 35,000 households, it's able to produce data for each quarter and has as domain of reference all the expenses for goods and services done in each month.

The second one is given by the monthly retail trade survey - mainly finalised to the calculation of the monthly retail trade indexes - able to produce monthly index numbers concerning the value of retail trade sales compared with the average value calculated for the base year 1995. The sample is based on about 5,000 enterprises each month.

The two surveys are characterised by various differences, as the following ones:

- the observation unit (it's the household for the first source and the enterprise for the second one).
- The interviewing technique (based on a questionnaire directly submitted by an interviewer for the first source, sent and received back by mail for the second one).
- As above mentioned, the sampling size.
- The smaller domain of reference covered by the second source: in practice the monthly retail trade survey refers to the main part of the NACE Division 52, with the exception of sales of goods outside shops and repairs of consumption goods, while the first source includes these items as well and, in addition, expenses for goods as motorvehicles and cars (including

[8] A more complicated, but also more precise methodology to assign weights P_h is given by the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), based a different transformation v_{ih} , but for more details on that we address again to Naumann's article.

[9] A third source is obviously given by estimates produced in the National Accounts Department, but this information is partially based on the two sources analysed in this context, so that we preferred not to take into consideration national accounts' data.

accessories) and fuel, and all the expenses for services; moreover, expenses realised in wholesales markets and self consumption of food products are taken into account as well.

- ❑ The main aims of the two surveys: estimation of a global amount in the first case and of a month-to-month change in the second.

From table 1 – where figures are expressed in thousands of billion lire – we see how, on the average, the first source (labour forces) produces lower estimates than the second one, with the only exception in 1995. In any case, we can see how the two sources don't show so highly differences each other, and this evidence is also due to the fact that, to render them comparable, we considered only the part of consumption in common with the domain observable with the retail trade survey. So we can conclude that both sources could be considered as estimates of the same unknown amount, and our aim is to obtain a new estimate based on a weighted arithmetic mean of the two previous ones.

Remembering formula (3.1) and starting from all the above mentioned techniques for estimating P_1, we considered these possibilities:

- formula (3.1) where each weight is equal to 0.5 (simple arithmetic mean);
- method (3.1.3), where we have supposed to approximate costs with the number of sampling units multiplied by the average cost for each interview (this cost is roughly three times higher for the first source, because of the use of an interviewer);
- method (3.2.1a) derived from the general formula (3.2.1), supposing to ignore the covariance term and estimating (for each source) a unique variance using the average of the sampling variances calculated on all the eight years;
- method (3.2.1b), for which estimates of variances change from year to year, because they are given by the sampling variances for each year;
- method (3.4.4), based on the F distribution.

It's easy to see how the only method leading to significant differences between P_1 and P_2 is (3.1.3), because of the higher estimated costs related to the first source. On the other hand, the method (3.2.1b) – leading to weights different from year to year – shows significant differences only before 1994, due to the higher variability of data derived from the first source in those years (on the whole, P_1 ranges from 0.406 to 0.516).

As a consequence, the estimates obtained with the previous methods are not very different each other and/or along time – even with the recourse to method (3.1.3) – so that we could conclude that both sources can produce good estimates of households' goods consumption[10].

Table 1 – Comparison among five different methods

Year	S_1	S_2	3.1	3.1.3	3.2.1a	3.2.1b	3.4.4
			Integration of sources				
P_1			0.500	0.340	0.516	See note	0.505
1991	508	516	512	513	512	513	512
1992	534	537	536	536	536	536	536
1993	541	546	544	544	543	544	544
1994	559	565	562	563	562	562	562
1995	582	582	582	582	582	582	582
1996	592	597	594	595	594	595	594
1997	607	617	612	613	611	612	612
1998	623	633	628	630	628	628	628
AVG	568	574	571	572	571	571	571
CV	6.46	6.61	6.53	6.55	6.53	6.49	6.53

Source: elaboration on ISTAT data. CV=Coefficient of Variation.
Starting from 1991, the weights got with (3.2.1b) are: 0.406, 0.406, 0.448, 0.481, 0.457, 0.481, 0.505, 0.516.

For the next future, we recall that procedures like those seen in the previous application are going to be better adapted to real data as soon as the planning of permanent short-term statistics for all the main service activities – mostly due to the Short-term Regulation's needs and going into force in one year time – will give us the possibility to add sources to the ones actually available, not only for employment but for turnover as well.

References

DENTON F. (1971), "Adjustment of Monthly or Quarterly Series to Annual Totales: an Approach Based on Quadratic Minimization", *Journal of the American Statistical Association*, Vol.66, 333, 99-102.

D'ALESSANDRO P. – PISANI S. (1996), "Short-term Supply Indicators for the Italian Service Sector", document prepared for the *Meeting of Service Statistics Experts*, 28-29 March, OCSE, Paris.

DEPOTOUT R. – ARONDEL P. (1998), International Comparability and Quality of Statistics, in Biffignardi S. (ed.), *Micro and Macrodata of Firms*, 125-156, Physica-Verlag, New York.

[10] Having supposed both sources able to produce unbiased estimators as regards the domain under observation.

EUROSTAT (1995), “Council Directive 95/57/EC Concerning Tourism Statistics”, Luxembourg.

EUROSTAT (1996), *Manuel Methodologique sur la Statistique d’Entreprises (ver.2.0)*, Luxembourg.

EUROSTAT (1997), “Council Regulation Concerning Structural Business Statistics (58/97)”, *Official Journal of the European Communities*, L 14/1.

EUROSTAT (1998), “Council Regulation Concerning Short-term Business Statistics (1165/98)”, *Official Journal of the European Communities*, L 162.

EUROSTAT (1999), “Industrial Trends: National Methods – Retail Trade”, document discussed during the *Working Group on Trade Statistics*, 22/23 March, Luxembourg.

FALORSI S. – FALORSI P.D. (1994), “Quarterly Estimates at Provincial Level for the Labour Forces Survey”, *Istat Research Copybooks*, 3, Rome.

GISMONDI R. (1996), “Some Considerations on the Estimation of Level and Change Using Preliminary Information”, *Italian Journal of Applied Statistics*, 2, 6-31, Rocco Curto Editor, Naples.

GISMONDI R. (1998), “The Impact of the Short-term Business Statistics Regulation”, paper presented at the *13-th Voorburg Group Meeting*, September, Rome, Italy.

LAAKSONEN S. (1996), “Notes on International Comparability of Sample Based Business Survey Data”, paper presented at the *CAED Conference*, Helsinki, 17-19 June 1996.

MC KENZIE R. (1998), “Short Term Economic Indicators using Administrative Data”, paper presented at the *13th Voorburg Group Meeting*, Rome.

MC MECHAN J. – MARCIL J.E.R. (1998), “Short Term Service Indicators”, paper presented at the *13th Voorburg Group Meeting*, Rome.

MONDUCCI R. (1995), “Businesses’ Statistical Information: Problems Concerning the Integration Among Sources”, Proceedings of the *Second National Statistical Conference*, 614-622, ISTAT, Rome.

NAUMANN F. (1998), “Data Fusion and Data Quality”, paper presented at the seminar on *New Techniques and Technologies for Statistics*, Sorrento, Italy, 4/6 November.

THOMSEN I. (1988), “Estimating Change in a Proportion by Combining Measurements from a True and a Fallible Classifier”, *Scandinavian Journal of Statistics*, 15, 139-145.