rivista di statistica ufficiale

n. 3
2006

Temi trattati

Interventi

Istat
Istituto nazionale di Statistica

rivista di statistica ufficiale

n. 3
2006

Temi trattati

Interventi

Istat

Comitato di redazione della Rivista di statistica ufficiale
c/o Dipartimento per la Produzione Statistica e il Coordinamento Tecnico Scientifico,
Via Cesare Balbo, 16 - 00184 Roma
tel.: 06.46732774 – fax: 06.47888069
e-mail: rivista@istat.it, cadeli@istat.it

rivista di statistica ufficiale

n. 3/2006

Periodico quadrimestrale
ISSN 1828-1982

Registrazione presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Servizio Produzione editoriale
Via Cesare Balbo, 16 - Roma

Videoimpaginazione:
Raffaella Rose

Copertina:
Maurizio Bonsignori

Stampa:
Istat – Produzione libraria e centro stampa
Via Tuscolana 1776 - Roma
Ottobre 2008 - Copie 400

More Rapid Short-term Statistics Using Auxiliary Variables

Roberto Gismondi[1]

Abstract

Timeliness has more and more becoming a driving feature of short-term statistics, that normally are released according to a first provisional estimate, followed by one or more revisions. In this context, we propose and compare some quick estimation methods aimed at improving timeliness and quality of provisional estimates. The particular field of interest is the ISTAT monthly survey on arrivals and nights spent in the Italian tourist establishments. A particular attention is paid to the potential self-selection bias affecting natural quick respondent units. An empirical application – referred to the period 2002-2004 – has been carried out, based on random replications of theoretical quick respondents.

Keywords: Provisional estimate, Late respondent, Nights spent, Self-selection bias, Tourism

1. The trade-off between timeliness and precision of estimates

Among the main components on which the EU statistical definition of quality for short-term statistics is founded (EUROSTAT, 2000), accuracy and timeliness seem to be the most relevant both for producers and users of statistical data. While accuracy is normally measured by the percent difference between provisional and final estimates, timeliness is measured as the time lag between the reference time point (or the end of the reference period) and the date of data dissemination.

However, along the last years, in many fields of short-term official statistics timeliness became the driving issue, both for the increasing demand of users and the need to fill the gap respect to data release standards as those already achieved by USA and other developed countries

In particular, the EU Council Directive 95/57/EC on tourism statistics (Council of the European Union, 1995) requests all the statistical institutes of the EU Member States to collect and transmit to EUROSTAT monthly data concerning internal tourism. A very sensitive variable is the number of nights spent in tourist accommodations located inside the national territory, broken down by nationality (Italians and foreigners) and kind of establishment (hotels and other collective accommodations, "o.c.a.").

In this context, source of data is the ISTAT census monthly survey on tourist establishments (ISTAT, *Anni vari*). Data are not picked up directly by ISTAT, but by local agencies for tourist promotion; further, they are summed up by the 103 Italian provinces and then sent to ISTAT in order to obtain national figures. Even though provisional estimates are spread out after 3 months (according to requests of the Directive), complete

[1] Senior Researcher (Istat);e-mail: gismondi@istat.it

definitive data are available only after 6 months. Delay depends from the extremely heterogeneous sensitiveness of tourist accommodations with respect to short-term tourism analysis, different territorial organisations and data transmission tools. Moreover, national and international users - asking for quicker data in order to better analyse short-term developments - identified in 45 days a reasonable time benchmark for provisional estimates for the tourism sector.

As a matter of fact, currently monthly data concerning some provinces are available in a relatively short time: the basic idea is that quick complete data of some units can be used to get, with a certain degree of error, quick estimates for national amounts as well. On the other hand, the main theoretical problem faced in the paper concerns the possible *self-selection* of quick units, so that the use of whatever quick respondent for estimating late responses could lead to seriously biased provisional estimates (Royall, 1988; Drudi and Filippucci, 2000).

As underlined by Bolfarine and Zacks (1992), the issue of robustness of predictors of population quantities can be faced using 3 strategies: 1) imposing restrictions to the possible super-population models adopted; 2) imposing restrictions to the samples to be selected; 3) using Bayes predictors that adaptively consider the possibility that each one out of a series of alternative models is the correct model. In this context, we propose and compare various quick estimation methods (paragraphs 2, 3, 4 and 5), mostly based on strategy 1) and, on a lesser extent, on strategy 2). In particular, in paragraph 3 we propose an approach aimed at evaluating and reducing the possible bias due to the non random selection of quick respondents, while paragraphs 6 contains an application to true tourism data, where some *model-based* estimation techniques are compared with the ratio estimator actually used in the survey in terms of *MAPE* (mean of the absolute percent errors, assessing empirical efficiency).

A serious technical constraint, also influencing the choice of methodological proposals, is the shortness of available time series, since monthly data at the province level have been officially diffused by ISTAT only from year 2002; at the moment, time series are only 36 months long, so that for each province only 3 observations related to the same month are available. That is the main reason why we did not use a time series approach, that assumes long time series and regularity along time of the error profile. Useful theoretical suggestions are available in Tam (1987), Rao *et al*. (1989), Yansaneh and Fuller (1998), Battaglia and Fenga (2003).

Moreover, at the moment the sub-sample of quick respondent provinces cannot be defined *a priori* or *driven* in some way, so that also methods acting on properties of particular quick samples cannot be used[2].

In the following, statistical units will be given by the 103 Italian provinces: use of aggregated data leads to some loss of information, but gives the possibility to manage less changeable data and renders easier the identification of a subset of quick respondents quite steady along time. We mean as “provisional quick estimate” the estimation of a parameter of interest – in the frame of a given statistical survey – obtained on the basis of a quick sub-sample available at a time t’ before time t correspondent to the “final estimate”, that will be based on a final sample including both quick and late respondents. Revisions can be calculated by the difference between provisional and final estimates. In the particular case

[2] A relevant example is given by balanced sampling (Royall, 1992), that will be considered in paragraph 2.

of tourism, final estimate refers to the complete population (Italy as a whole). Two main methodological approaches could be used:

approach based on the sampling design: the efficiency of a quick estimation strategy depends on the probability distribution derived from the particular sampling design adopted both for final and (if any) preliminary estimates. Of course, this approach cannot be considered in this particular context.

Approach based on a super-population model (Cassel et al., 1983), on the basis of which quality evaluations and the choice of the units belonging to the quick sub-sample are carried out on the basis of the mean squared error respect to the particular model underlying observed data.

For both approaches, the availability of additional information external to the survey - or related to the survey in terms of historical micro-data, as it happens in many longitudinal surveys - can be extremely helpful.

The need to explore tourist data is further steered by clear evidence: even though literature concerning forecasting tourist demand is very wide, problems related to the use of quick respondents to predict final estimations including late respondents as well, are quite uncommon in the tourist field. In particular, a resume of methodologies for forecasting tourism demand is given in Song and Witt (2000), while traditional linear regression models are proposed by Costa and Manente (2000, 129-202) and Divisekera (2003). Recourse to ARIMA models and exponential smoothing (widely discussed in Harvey, 1984) is available in Lim and Mc Aleer (2001), while late attempts to apply genetic algorithms to real tourist populations for a better decision making are given by Hernández-López (2004). Further sampling technique applied to tourist data are also available in Gismondi (2007).

On the other hand, applications concerning other economic sectors, but strictly linked with the problem herein discussed can be found in Maravalle *et al*. (1993), Aelen (2003), Ullberg (2003), Falorsi *et al*. (2005).

2. Super-population model with a single auxiliary variable

From now on U will indicate the target population with size N, n is the size of a sample S and the main purpose of the sample survey is the estimation of the population mean $\bar{y}_U$. S can indicate both the provisional quick sample and the final sample including late respondents as well. For each population unit we'll suppose as true the following regression model, defined as:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{where:} \quad \begin{cases} E(\varepsilon_i) = 0 & \forall i \\ VAR(\varepsilon_i) = \sigma^2 v_i & \forall i \\ COV(\varepsilon_i, \varepsilon_j) = 0 \quad if \quad i \neq j \end{cases} \tag{2.1}$$

where expected values, variances and covariances are referred to the model and not to any sampling design, x is an additional variable strongly correlated with y and to be specified, as well as the function v_i, with α, β and σ^2 given, but generally unknown parameters. Even though a specific model (2.1) should be defined for each reference period

t, at the moment no time labels are used. Under model (2.1), a sample $S \subseteq U$ is supposed to be available, with $S \cup \overline{S} = U$.

2.1 The case $\alpha = 0$

If in model (2.1) we put $\alpha = 0$**,** we get the widely used regression model *R*. As well known, (Cicchitelli *et al*., 1992), the optimal linear predictor – e.g. the one minimising *MSE* with respect to the model, $E(T - \overline{y}_U)^2$ – is given by:

$$T^* = \overline{y}_S\left(\frac{n}{N}\right) + \overline{x}_{\overline{S}}\,\hat{\beta}^*\left(\frac{N-n}{N}\right) \qquad \text{where:} \qquad \hat{\beta}^* = \left(\sum_S x_i y_i / v_i\right)\left(\sum_S x_i^2 / v_i\right)^{-1} \tag{2.2}$$

and its variance (equal to the *MSE* under model (2.1)) respect to the model will be equal to:

$$MSE(T^*) = \left[\left(\sum_{\overline{S}} x_i\right)^2 \Big/ \left(\sum_S (x_i^2 / v_i)\right) + \sum_{\overline{S}} v_i\right]\frac{\sigma^2}{N^2}\,. \tag{2.3}$$

Relevant particular cases are obtained if v=1 – when (2.2) reduces to the regression estimator through the origin – and v=x – when (2.2) is the common ratio estimator (the one currently used in the survey on tourist establishments for calculating provisional estimates within 90 days) and the corresponding model variances will follow straightforwardly. Moreover, under model (2.1) the sample mean is optimal *if and only if* x=v=1. Let's note how the first case translates in a model-based context the common hypothesis of homoschedasticity, while the second one implies the more realistic hypothesis of a lower relative (model) variability for largest units. If (2.2) formalises optimal estimator formula, (2.3) suggests that the best choice of the sample simply consists, *when it is possible*, in selecting the *n* units in the universe having the largest *x*-values.

A consequence of (2.3) is that whatever sample *S* is available – in particular, when S=S_P is the sample including the *provisional* quick respondents – the best strategy consists, according to predictor (2.2), in using all the n available units. However, this strategy could be dangerous, for these two main reasons:

the quality of estimates strongly depends on the validity of all assumptions in model (2.1). In particular, an estimator as (2.2) could be seriously biased when model (2.1) is wrong.

The choice of the *n* biggest units does not guarantee a low variance, because it depends on the relative weight of the sample on the overall *x*-amount: generally speaking, when this weight is lower than 50%, other estimators and/or sample selection rules could perform better.

A way to reduce potential bias mentioned in the above point 1 consists in using *balanced samples*. Under model (2.1) – e.g., when only one auxiliary *x*-variable is taken into account – a sample *S* of size *n* is *balanced with respect to the weights root(v)* if it satisfies the condition:

$$\sum_S x_i \Big/ n\sqrt{v_i} = \sum_U x_i \Big/ \sum_U \sqrt{v_i} \,. \tag{2.4}$$

It could be chosen among all the possible samples of size *n* using various algorithms, as those proposed by Valliant et al. (2000), Gismondi (2002) and Deville and Tillé (2004). Royall (1992) showed that, if the previous linear model *R* holds and a balanced sample *can be found*, the best linear unbiased predictor under the model is:

$$\hat{T}_{bal,v} = n^{-1}\Big(\sum_U \sqrt{v_i}/N\Big)\Big(\sum_S y_i/\sqrt{v_i}\Big) \tag{2.5}$$

having, *among all samples satisfying* (2.4), the lowest mean squared error given by:

$$MSE(\hat{T}_{bal,v}) = \left[\left(\sum_U \sqrt{v_i}\right)^2 n^{-1} + \sum_U v_i - 2\,n^{-1}\sum_S \sqrt{v_i}\sum_U \sqrt{v_i}\right]\frac{\sigma^2}{N^2} \,. \tag{2.6}$$

Under the statement v=1, the balance condition (2.4) becomes $\bar{x}_S = \bar{x}_U$ (sample and population means must be equal), while the optimal predictor derived from (2.5) is the sample mean $\hat{T}_{bal,1} = \bar{y}_S$, which *MSE* is $\left(N\,n^{-1}-1\right)\sigma^2$. So, if the sample is balanced the sample mean is still optimal even when $x \neq 1$. The great advantage in using a balanced sample is that it preserves from bias if model (2.1) is wrong[3]. More in general, it reduces the negative effects of respondents self-selection process on parameters estimation (as for β estimation in (2.2) and σ^2 estimation in (2.3)), mainly if the true model formalisation is unknown (Drudi and Ferrante, 2003).

Even though characterised by relevant theoretical properties, until now balanced sampling did not have a wide practical use, with some exceptions given by ISTAT (2005) and Gismondi (2003; 2007). As a matter of fact, the use of an estimator as (2.5) does not necessarily reproduce the true population mean whenever all units were available. Moreover, given *U* and model (2.1), balanced samples could not exist, or anyway a quick balanced sample cannot be planned in advance, as in the case of tourism statistics. In the most part of practical situations when provisional estimates are needed, the available natural sub-sample of quick respondents must be considered as *given* and it is generally *not balanced*, because of the mentioned self-selection bias. However, a simple *ex-post* strategy consists in selecting from the whole available quick sample a sub-sample – as much larger as possible – that minimises the *unbalancing ratio*.

In symbols, when v=1 instead of (2.4) one could have $\bar{x}_S = k\,\bar{x}_U$, where $k = \bar{x}_S / \bar{x}_U$ is the unbalancing ratio. Then the predictor to be used would turn out to be $\bar{y}_S / k$, that using the whole sample *S* is equivalent to the ratio estimator derived from (2.2), but when v=x.

However, levels of *k* more near to one (e.g., sub-samples *almost* balanced) could be

[3] For instance, if $y_i=\theta+\beta x_i+\varepsilon_i$, then the mean squared error (2.3) increases for a constant equal to θ^2.

obtained using for provisional estimates only a sub-sample $s \subseteq S$ with size $n_s \leq n$. The (quick) estimator will be given by:

$$\bar{y}_s\left(\frac{\bar{x}_U}{\bar{x}_s}\right) \qquad \text{where:} \quad s \subseteq S\,. \tag{2.7}$$

In principle, one could choose a very small sub-sample *s* in order to have $\bar{x}_s \approx \bar{x}_U$ and to reduce near to zero the original bias - so that (2.7) becomes the simple sub-sample mean - but on the other hand a too small sub-sample size n_s could dangerously increase the sub-sample variance. Then, an empirical rule could consist in imposing *a priori* that $n_s \geq (1-\gamma)n$, where γ=0,05 or γ=0,10. Finally, it is worth underlining that this strategy leads to a *modified ratio estimator*, given by:

$$\bar{y}_S\left(\frac{\bar{x}_U}{\bar{x}_S}\right)\left[\frac{\bar{y}_s}{\bar{y}_S}\cdot\frac{\bar{x}_S}{\bar{x}_s}\right] \tag{2.8}$$

where the term in squared brackets is the coefficient that modifies the original ratio estimator based on the whole sample *S* with the aim to reduce its original bias.

It must be remarked that even though this *ex post* strategy could be applied in the most part of empirical cases, the correct use of balanced sampling strictly refers to the *ex ante* planning of a given theoretical sample of respondents.

2.2 The case α ≠ 0

If we consider the only case with the common statement v_i=1 for each unit *i*, we get the usual linear homoschedastic regression model with one auxiliary variable. As a particular case of the general solution (8.2), we have the optimal predictor given by:

$$T^* = \bar{y}_S\left(\frac{n}{N}\right) + \left(\hat{\alpha}^* + \bar{x}_{\bar{S}}\,\hat{\beta}^*\right)\left(\frac{N-n}{N}\right) \quad \Leftrightarrow \quad T^* = \bar{y}_S + \hat{\beta}^*\,(\bar{x}_U - \bar{x}_S) \tag{2.9}$$

where:

$$\hat{\beta}^* = \left(\sum_S (x_i - \bar{x}_S)\,y_i\right)\left(\sum_S (x_i - \bar{x}_S)^2\right)^{-1} \quad \text{and} \quad \hat{\alpha}^* = \bar{y}_S - \hat{\beta}^*\,\bar{x}_S \tag{2.10}$$

and its mean squared error will be (see general formula (8.4)):

$$MSE(T^*) = \left[N(\bar{x}_U - \bar{x}_S)^2 \left(\sum_S (x_i - \bar{x}_S)^2 \right)^{-1} + \left(\frac{N-n}{n} \right) \right] \frac{\sigma^2}{N} . \qquad (2.11)$$

It is worthwhile to note that under a model-based approach the optimal sampling strategy will be purposive, namely, the one that selects with probability one the sample *S* consisting of those units whose *x*-values minimise the first quantity in squared brackets in (2.11). In a provisional estimation context, one should use a provisional quick sample as much balanced as possible with respect to variable *x* (the quick sample and the population *x*-means should be approximately the same), or anyway a quick sample characterised by a very large *x*-variance (Cassel *et al*., 1977, 128). This strategy does not correspond to the one which minimises (2.3) when a regression model through the origin is supposed to be true (choice of the *n* biggest units).

3. A model for evaluating self-selection bias

It is possible to model potential structural differences between provisional and late respondents. We can suppose that population *U* can be split into 2 separate sub-populations U_P and U_L, including respectively N_P units (those that are all potential *Provisional* quick respondents) and N_L units (those that are all potential *Late* respondents), with $U = U_P \cup U_L$ and $N=N_P+N_L$. These sub-populations do not derive from any preliminary stratification, but depend on some latent factor underlying units under observation. For each of the 2 sub-populations (labelled with *h,* where $h=P,L$) this model is assumed to be:

$$y_{hi} = \beta_h x_i + \varepsilon_{hi} \qquad \text{where:} \quad \begin{cases} E(\varepsilon_{hi}) = 0 & \forall h,i \\ VAR(\varepsilon_{hi}) = \sigma_h^2 v_i & \forall h,i \\ COV(\varepsilon_{hi}, \varepsilon_{hj}) = 0 \quad if \quad i \neq j \end{cases} \quad \text{for } h=P,L \qquad (3.1)$$

where all symbols keep the same meaning as in model (2.1). The main difference is that we suppose a priori that provisional and late respondents could have different model means and variances[4]. Supposing no final non-response, provisional and late samples S_P and S_L will include respectively n_P and $n_L=(n-n_P)$ units, where $S = S_P \cup S_L$, $U_P = S_P \cup \overline{S}_P$, $U_L = S_L \cup \overline{S}_L$, $\overline{S} = \overline{S}_P \cup \overline{S}_L$. The number of non-observed units will be, respectively, N_P-n_P and N_L-n_L. Table 3.1 gives an overall resuming scheme.

If the main purpose is the estimation of the overall total y_U, if x_U is the *x*-total in the whole population *U* the unknown total to be estimated will be given by:

[4] In this context we also suppose that the belonging to one of the two sub-populations is a *deterministic* (even though often unknown) feature of each unit and does not depend on any *probabilistic* mechanism.

$$y_U = \sum_{i=1}^{N_P} y_{Pi} + \sum_{i=1}^{N_L} y_{Li} = y_P + y_L \qquad \text{where:} \qquad E(y_U) = \beta_P x_P + \beta_L x_L \,. \tag{3.2}$$

Table 3.1: Different patterns for preliminary and late respondents

	STRUCTURE			SIZE		
	Total units	Provisional units	Late units	Total units	Provisional units	Late units
Population	U	U_P	U_L	N	N_P	N_L
Sample	S	S_P	S_L	N	n_P	n_L
Not observed population	$\overline{S}$	$\overline{S}_P$	$\overline{S}_L$	N-n	$N_P - n_P$	$N_L - n_L$

3.1 Optimal final prediction under model (3.1)

If y_S is the sample y-total, a linear predictor of y_U can be written as:

$$T_{(PL)} = y_S + \hat{y}_{(PL)\overline{S}} = y_S + \sum_{S_P} c_{Pi} y_{Pi} + \sum_{S_L} c_{Li} y_{Li} \tag{3.3}$$

where $\hat{y}_{(PL)\overline{S}}$ is the predictor of the unknown amount (y_U-y_S) and coefficients c_{Pi} and c_{Li} must be determined. Under model (3.1), the unbiasedness condition is equivalent to $E(T_{(PL)} - y_U) = 0$. We can show that under model (3.1) the *final* best linear unbiased predictor will be given by[5]:

$$T^*_{(PL)} = \left[y_{S_P} + x_{\overline{S}_P} \left(\sum_{S_P} \frac{x_{Pi} y_{Pi}}{v_{Pi}} \right) \left(\sum_{S_P} \frac{x_{Pi}^2}{v_{Pi}} \right)^{-1} \right] + \left[y_{S_L} + x_{\overline{S}_L} \left(\sum_{S_L} \frac{x_{Li} y_{Li}}{v_{Li}} \right) \left(\sum_{S_L} \frac{x_{Li}^2}{v_{Li}} \right)^{-1} \right] \tag{3.4}$$

with a mean squared error given by:

[5] See appendix 8.2. Formulas (3.4) and (3.5) become analogous to (2.2) and (2.3) if model (3.1) reduces to (2.1).

$$MSE(T^{*}_{(PL)}) = \sigma^2_P \left[x^2_{\bar{S}_P} \left(\sum_{S_P} \frac{x^2_{Pi}}{v_{Pi}} \right)^{-1} + \sum_{\bar{S}_P} v_{Pi} \right] + \sigma^2_L \left[x^2_{\bar{S}_L} \left(\sum_{S_L} \frac{x^2_{Li}}{v_{Li}} \right)^{-1} + \sum_{\bar{S}_L} v_{Li} \right]. \qquad (3.5)$$

One can verify consequences of the assumption of model (2.1) instead of the *true* model (3.1). In this case no distinction between provisional and late respondents would occur in the final estimation process, so that the predictor used would have the form $T = y_S + \sum_{S} c_i y_i$ and, in particular, would be given by predictor T^* in (2.2) unless the multiplying factor N. Under the *true* model (3.1), NT^* would still be unbiased if:

$$\beta_P \left[\left(\sum_{S_P} \frac{x_i^2}{v_i} \right) \left(\sum_{S} \frac{x_i^2}{v_i} \right)^{-1} x_{\bar{S}} - x_{\bar{S}_P} \right] + \beta_L \left[\left(\sum_{S_L} \frac{x_i^2}{v_i} \right) \left(\sum_{S} \frac{x_i^2}{v_i} \right)^{-1} x_{\bar{S}} - x_{\bar{S}_L} \right] = 0 . \qquad (3.6)$$

If $v=x$, condition (3.6) implies that we should have:

$$\frac{\sum_{S_P} x_i}{\sum_{\bar{S}_P} x_i} = \frac{\sum_{S_L} x_i}{\sum_{\bar{S}_L} x_i} \quad \Leftrightarrow \quad \frac{\sum_{S_P} x_i}{\sum_{U_P} x_i} = \frac{\sum_{S_L} x_i}{\sum_{U_L} x_i} \qquad (3.7)$$

and in this case the predictor (3.4) would be equal to NT^*. Identity (3.7) is satisfied if provisional respondents determine a share of the x-total in the provisional respondents' sub-population equal to the one concerning late respondents. However, in general the recourse to predictor NT^* under a true model as (3.1) leads to a not null bias, given by:

$$Bias = E(NT^* - y_U) = x_{\bar{S}} \left(\sum_{S} \frac{x_i^2}{v_i} \right)^{-1} \left[\beta_P \left(\sum_{S_P} \frac{x_i^2}{v_i} \right) + \beta_L \left(\sum_{S_L} \frac{x_i^2}{v_i} \right) \right] - \left(\beta_P \sum_{\bar{S}_P} x_i + \beta_L \sum_{\bar{S}_L} x_i \right). \qquad (3.8)$$

Moreover, its mean squared error will be given by[6]:

$$MSE(NT^*) = E(NT^* - y_U)^2 = VAR\left(\sum_{S} c_i^* y_i \right) + VAR(y_{\bar{S}}) + Bias^2 \qquad (3.9)$$

where:

[6] We can compare formula (8.7), formally similar to (3.9) when adding the squared bias term.

$$VAR\left(\sum_S c_i^* y_i\right) = x_{\bar{S}}^2 \left(\sum_S \frac{x_i^2}{v_i}\right)^{-2} \left(\sigma_P^2 \sum_{S_P} \frac{x_i^2}{v_i} + \sigma_L^2 \sum_{S_L} \frac{x_i^2}{v_i}\right) \quad \text{and:}$$

$$VAR(y_{\bar{S}}) = \sigma_P^2 \sum_{\bar{S}_P} v_i + \sigma_L^2 \sum_{\bar{S}_L} v_i \tag{3.10}$$

3.2 Provisional estimation under model (3.1)

If only provisional quick respondents can be used for estimation, we can define the general linear predictor of the unknown amount y_U, given by:

$$T_{(P)} = y_{S_P} + \hat{y}_{\bar{S}_P} = y_{S_P} + \sum_{S_P} c_{Pi} y_{Pi} \,. \tag{3.11}$$

The unbiasedness condition under model (3.1) implies:

$$E(T_{(P)} - y_U) = E\left[(y_{S_P} + \hat{y}_{\bar{S}_P}) - (y_{S_P} + y_{S_L} + y_{\bar{S}_P} + y_{\bar{S}_L})\right] = 0 \tag{3.12}$$

that, after some passages, leads to:

$$\sum_{S_P} c_{Pi} x_{Pi} = \left(\frac{\beta_L}{\beta_P}\right) x_L + x_{\bar{S}_P} \tag{3.13}$$

where $x_L = x_{S_L} + x_{\bar{S}_L}$. We will also have:

$$MSE(T_{(P)}) = E\,(T_{(P)} - y_U)^2 = VAR\left(\sum_{S_P} c_{Pi} y_{Pi}\right) + VAR(y_{\bar{S}_P} + y_L) \tag{3.14}$$

where the main difference from (8.7) is that the second variance term refers to the *y*-amount concerning *all* the non observable units (those belonging to U_P but not included in the sample and those belonging to U_L). Under model (3.1) we easily obtain:

$$MSE(T_{(P)}) = \sigma_P^2 \sum_{S_P} c_{Pi}^2 v_{Pi} + \sigma_P^2 \sum_{\bar{S}_P} v_{Pi} + \sigma_L^2 \sum_{U_L} v_{Li} \tag{3.15}$$

and minimisation of (3.15) under constraint (3.13) leads to optimal solutions:

$$c_{Pi}^* = \frac{x_{Pi}}{v_{Pi}} \left(\frac{\beta_L}{\beta_P} X_L + x_{\bar{S}_P}\right) \left(\sum_{S_P} \frac{x_{Pi}^2}{v_{Pi}}\right)^{-1} \tag{3.16}$$

$$T^{*}_{(P)} = y_{S_P} + \left(\frac{\beta_L}{\beta_P} x_L + x_{\bar{S}_P}\right)\left(\sum_{S_P} \frac{x_{Pi} y_{Pi}}{v_{Pi}}\right)\left(\sum_{S_P} \frac{x^2_{Pi}}{v_{Pi}}\right)^{-1} . \quad (3.17)$$

We could verify that predictor (3.17) is also the optimal predictor under the constraint $E(T_{(P)} - T^{*}_{(PL)}) = 0$ instead of (3.12), so that in a provisional estimate context minimising mean squared error with respect to the overall amount y_U is equivalent to minimising the expected difference between provisional and final estimates. Finally, under the *true* model (3.1) the predictor (2.2) – optimal under the *false* model (2.1) – can be written as:

$$T_{(P)} = y_{S_P} + (x_P + x_{\bar{S}_P})\left(\sum_{S_P} \frac{x_{Pi} y_{Pi}}{v_{Pi}}\right)\left(\sum_{S_P} \frac{x^2_{Pi}}{v_{Pi}}\right)^{-1} \quad (3.18)$$

that turns out to be equal to (3.17) if $\beta_L = \beta_P$.

3.3 Implementation of provisional prediction

In order to implement the optimal solution (3.17) one should know the true values β_P and β_L. Since they are generally unknown, we can use a sub-optimal provisional prediction derived from (3.17), that for the estimation of the mean can be written as:

$$\frac{\hat{T}^{*}_{(P)}}{N} = \frac{1}{N}\left[y_{S_P} + \left(\frac{\hat{\beta}_L}{\hat{\beta}_P} \hat{x}_L + \hat{x}_{\bar{S}_P}\right)\beta^{*}_P \right] \quad (3.19)$$

where β^{*}_P is given by the second relation (2.2) when the only provisional sample S_P is considered, while $\hat{\beta}_L, \hat{\beta}_P, \hat{x}_L$ and $\hat{x}_{\bar{S}_P}$ are estimates for β_L, β_P, x_L and $x_{\bar{S}_P}$. One can reasonably put:

$$\hat{x}_L = \bar{x}_{S_L} \hat{N}_L = \bar{x}_{S_L}\left(\frac{n_L}{n}\right) N \quad (3.20)$$

$$\hat{x}_{\bar{S}_P} = \bar{x}_{S_P}(\hat{N}_P - n_P) = \bar{x}_{S_P}\left(\frac{n_P}{n} N - n_P\right) = \bar{x}_{S_P} n_P \left(\frac{N}{n} - 1\right). \quad (3.21)$$

For what concerns β_P e β_L, there are 2 options. In the first case, one could use the optimal estimates derived from the second relation (2.2) applied separately to provisional and late respondents, both available with reference to a period (t-1) before time of reference

t[7], so that:

$$\hat{\beta}_L = \beta^*_{L(t-1)} \quad \text{and} \quad \hat{\beta}_P = \beta^*_{P(t-1)} . \tag{3.22}$$

In the second case, one could put:

$$\hat{\beta}_L = \beta^*_{L(t-1)} \quad \text{and} \quad \hat{\beta}_P = \beta^*_{P(t)} \tag{3.23}$$

where the theoretical advantage respect to (3.22) is that in this case we use the *actual* optimal estimate of β_P, e.g. that based on provisional quick respondents available with reference to time t. In both cases, provided that in (3.19) we can substitute $\beta^*_P = \beta^*_{P(t)}$, the final formula of the sub-optimal predictor of the population mean is given by:

$$\frac{\hat{T}^*_{(P)}}{N} = \frac{1}{N}\left[\bar{y}_{S_P} n_P + \left(\frac{\hat{\beta}_L}{\hat{\beta}_P} \bar{x}_{S_L} (n - n_P) \left(\frac{N}{n} \right) + \bar{x}_{S_P} n_P \left(\frac{N-n}{n} \right) \right) \beta^*_P \right] . \tag{3.24}$$

If we estimate the optimal coefficients (3.16) on the basis of (3.20), (3.21) and one between (3.22) and (3.23), and estimates for model variances are available as well, an estimate of the mean squared error of predictor (3.24) will be given by:

$$\hat{MSE}\left(\frac{\hat{T}^*_{(P)}}{N} \right) = \frac{1}{N^2}\left[\hat{\sigma}^2_P \sum_{S_P} (\hat{c}^*_{Pi})^2 v_{Pi} + n_P \left(\frac{N}{n} - 1 \right) \hat{\sigma}^2_P \bar{v}_{S_P} + n_L \left(\frac{N}{n} \right) \hat{\sigma}^2_L \bar{v}_{S_L} \right] . \tag{3.25}$$

4. The model proposed by Fuller

Fuller (1990) analysed the general form of the *BLU* predictor of the population mean in a generalised least squared context. Supposing the simple case of one only auxiliary variable x, the supposed underlying model is given by:

[7] Generally speaking, in a short-time survey context, if we refer to a month t of the year A, period (t-1) is given by the same month t referred to the previous year (A-1), in order to properly take into account seasonality of coefficients in the model (3.1).

$$y_i = \mu_y + \theta(x_i - \mu_x) + \varepsilon_i \qquad \text{where:} \qquad \begin{cases} E(\varepsilon_i) = 0 & \forall i \\ VAR(\varepsilon_i) = \sigma^2 & \forall i \\ COV(\varepsilon_i;\varepsilon_j) = 0 \quad if \quad i \neq j \end{cases} \tag{4.1}$$

where in this case the super-population means μ_y and μ_x are explicitly formalised into the linear model and represent, as well as θ and σ^2, unknown parameters to be estimated.

Supposing a simple random sampling design, and denoting as $\hat{\mu}_x$ the GLS estimator for μ_x and as $\hat{\theta}$ the regression coefficient obtained in the regression of y on x using *the complete set S of n observations*, then the estimator for the mean is given by:

$$\hat{\mu}_y = \bar{y}_S + \hat{\theta}(\hat{\mu}_x - \bar{x}_S) \tag{4.2}$$

with variance given approximately by:

$$VAR(\hat{\mu}_y) = n^{-1}\sigma^2 + \theta^2 VAR(\hat{\mu}_x)\,. \tag{4.3}$$

The use of result (4.2) in a provisional estimation context follows straightforwardly. The optimal prediction of the *effective* unknown population mean $\bar{y}_U$ can be based on (4.2) as well, taking into account that for provisional estimation only the provisional sample S_P is available; according to the notation introduced in paragraph 3 we can write (4.2) as:

$$\hat{\bar{y}}_U = \bar{y}_{S_P} + \hat{\theta}(\bar{x}_U - \bar{x}_{S_P}) \tag{4.4}$$

where the population x-mean is assumed known. The main difference between (4.4) and the standard regression estimator (2.9) is that in (4.4) estimation should be based on all the n units belonging to S and not only to the n_P units belonging to the provisional quick sample S_P. A particular, relevant case is when $x=y_{(t-1)}$, because the procedure (4.2) is normally very efficient when correlation between y and x is very high (Fuller, 1990, 173). It follows that (4.4) becomes:

$$\hat{\bar{y}}_{U(t)} = \bar{y}_{S_P(t)} + \hat{\theta}\left(\bar{y}_{U(t-1)} - \bar{y}_{S_P(t-1)}\right)\,. \tag{4.5}$$

Formula (4.5) shows that, at time t, the mean calculated on provisional quick respondents must be added to the difference between true and estimated y-means at time (t-1), weighted on the basis of a regression coefficient estimated using *both provisional and*

late respondents. That can be done, for instance, estimating this coefficient with reference to the previous time (*t*-1), in a way similar to section 3.3. As a matter of fact, procedure (4.5) can be another way to reduce self-selection bias effects.

A generalisation of formula (4.5) simply consists in supposing a multiplicative instead of an additive super-population model. In this case the new predictor will be:

$$\hat{\bar{y}}_{U(t)} = \bar{y}_{S_P(t)} \left[\frac{\bar{y}_{U(t-1)}}{\bar{y}_{S_P(t-1)}} \right]^{\hat{\theta}} \tag{4.6}$$

that can be viewed as another form of modified ratio estimator (compare (2.8)).

5. A model with two auxiliary variables

At the moment, provisional estimates of tourism nights spent are based on one auxiliary variable, given by nights spent in the same month of the previous year. As we will see in paragraph 6, provisional estimates could be based on two auxiliary variables. In this case the theoretical reference is given by the *general multi-regression model* G_{MR}, briefly commented in the appendix 1. The general idea is that the use of more than one auxiliary variable could increase precision of provisional estimates, that on the other hand could be still affected by a self-selection bias. In this context the symbol *S* will still indicate (as in paragraph 2) a general available sample (final or provisional). The bivariate regression model – that is a particular case of model (8.1) – is given by:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad \text{where:} \qquad \begin{cases} E(\varepsilon_i) = 0 \\ V(\varepsilon_i) = \sigma^2 v_i \\ C(\varepsilon_i, \varepsilon_j) = 0 \quad if \quad i \neq j \end{cases} \tag{5.1}$$

with α, β_1, β_2 and σ^2 given unknown parameters.

5.1 The case α = 0

If in model (2.1) we put $\alpha = 0$**,** we get a bivariate regression model through the origin. Addressing to appendix 8.1 for details, we can obtain these explicit formulas for the optimal unbiased linear predictor:

$$T^* = \left(\frac{n}{N} \right) \bar{y}_S + \left(1 - \frac{n}{N} \right) \left(\bar{x}_{1\bar{S}} \hat{\beta}_1^* + \bar{x}_{2\bar{S}} \hat{\beta}_2^* \right) \tag{5.2}$$

where:

$$\hat{\beta}_1^* = \left(\sum_S \frac{x_{2i}^2}{v_i} \sum_S \frac{x_{1i} y_i}{v_i} - \sum_S \frac{x_{1i} x_{2i}}{v_i} \sum_S \frac{x_{2i} y_i}{v_i} \right) \left[\sum_S \frac{x_{1i}^2}{v_i} \sum_S \frac{x_{2i}^2}{v_i} - \left(\sum_S \frac{x_{1i} x_{2i}}{v_i} \right)^2 \right]^{-1} \qquad (5.3a)$$

$$\hat{\beta}_2^* = \left(\sum_S \frac{x_{1i}^2}{v_i} \sum_S \frac{x_{2i} y_i}{v_i} - \sum_S \frac{x_{1i} x_{2i}}{v_i} \sum_S \frac{x_{1i} y_i}{v_i} \right) \left[\sum_S \frac{x_{1i}^2}{v_i} \sum_S \frac{x_{2i}^2}{v_i} - \left(\sum_S \frac{x_{1i} x_{2i}}{v_i} \right)^2 \right]^{-1} \qquad (5.3b)$$

and the *MSE* formula derives directly from (8.4). In particular, if v_i=1 for each unit i (5.2) reduces to the common regression estimator without constant term; in the case of heteroschedasticity, one can put $v_i = x_{1i}\, x_{2i}$ for each unit i, so that (5.2) leads to a *double ratio estimator* - an extension of the univariate ratio estimator derived from (2.1) when v=x – where:

$$\hat{\beta}_1^* = \left(\sum_S \frac{x_{2i}}{x_{1i}} \sum_S \frac{y_i}{x_{2i}} - n \sum_S \frac{y_i}{x_{1i}} \right) \left(\sum_S \frac{x_{1i}}{x_{2i}} \sum_S \frac{x_{2i}}{x_{1i}} - n^2 \right)^{-1} \qquad (5.4a)$$

$$\hat{\beta}_2^* = \left(\sum_S \frac{x_{1i}}{x_{2i}} \sum_S \frac{y_i}{x_{1i}} - n \sum_S \frac{y_i}{x_{2i}} \right) \left(\sum_S \frac{x_{2i}}{x_{1i}} \sum_S \frac{x_{1i}}{x_{2i}} - n^2 \right)^{-1} . \qquad (5.4b)$$

Empirical results related to tourism data showed that the predictor (5.2) performed better than the bivariate *ratio-cum-product type estimators* proposed by Singh (1965) and renewed by Perri (2005), whose precision can be seriously affected by the presence of some anomalous ratios.

5.2 The case α ≠ 0

As well known**,** under an ordinary bivariate regression model optimal solutions can be written as:

$$T^* = \left(\frac{n}{N} \right) \bar{y}_S + \left(1 - \frac{n}{N} \right) \left(\hat{\alpha}^* + \bar{x}_{1\bar{S}}\, \hat{\beta}_1^* + \bar{x}_{2\bar{S}}\, \hat{\beta}_2^* \right) \qquad (5.5)$$

where – supposing that v_i=1 for each unit i and putting r_{Szw} as the sample correlation coefficient between variables z and w and Std_{Sz} as the sample standard deviation concerning variable z:

$$\hat{\beta}_1^* = \left(\frac{r_{Sy x_1} - r_{Sy x_2} r_{S x_1 x_2}}{1 - r_{S x_1 x_2}^2} \right) \left(\frac{Std_{Sy}}{Std_{S x_1}} \right); \qquad \hat{\beta}_2^* = \left(\frac{r_{Sy x_2} - r_{Sy x_1} r_{Sy x_2}}{1 - r_{S x_1 x_2}^2} \right) \left(\frac{Std_{Sy}}{Std_{S x_2}} \right);$$

$$\hat{\alpha}^* = \bar{y}_S - \hat{\beta}_1^* \bar{x}_{1S} - \hat{\beta}_2^* \bar{x}_{2S} \qquad (5.6)$$

and also in this case the *MSE* formula derives directly from (8.4). In particular, one could verify that, as in the case $\alpha = 0$ and the univariate case $\alpha \neq 0$, the best choice of the sample does not necessarily lead to the selection of the *n* largest units.

Recently Dalabehera and Sahoo (1999) analysed conditions under which a linear estimator based on two auxiliary variables and a stratified sampling should perform better than a single variable regression or ratio estimator. A more general theoretical proposal for large samples was also given by Montanari (1987).

A bivariate regression model was recently applied in the tourism statistics context by Gismondi *et al*. (2003), where estimation of each late respondent *y*-value was based on a constant term, the late respondent's *y*-value in the same period of the previous year (x_1) and the current *y*-mean calculated on quick provisional respondents (x_2)[8].

6. Empirical results

Late experience in the Italian monthly survey internal tourism showed that available quick responses - normally supplied within 45 days from the end of reference month - can be used to estimate late responses (non responses at the moment of the quick estimation) of the missing provinces. The estimator of late responses currently used in the survey is a simple ratio estimator (see formula (2.2) when *x* indicates nights spent in month (*m*-12) – where *m* is the reference month – and *v*=*x*).

However, assessment of quality of compared estimators must be based on a wider platform, so that an additional simulation applied to true data has been carried out as well. Year of reference was 2004, while data concerning 2002 and 2003 were used as source of auxiliary information[9].

In details, 4 separate domains have been taken into account: Italians in hotels, Italians in the other collective accommodations ("o.c.a."), foreigners in hotels and foreigners in "o.c.a.". Then, we selected at random the 20% of the 103 provinces (10 provinces) and, in a second step, the 50% of provinces (51 provinces). We repeated this procedure 100 times and, in each of the 100 cases, a particular simulated subset of provisional quick respondents has been supposed to be available, on the basis of which a provisional estimation of the simulated non respondents was carried out. In each replication, the same subset of quick respondents was adopted in each domain and for each month. This kind of experiment can give a quite clear idea of the robustness of 13 alternative provisional estimation techniques, resumed in table 6.3.

The frequency distribution of the 103 Italian provinces by class of total nights spent is quite asymmetric, as the most part of output distributions concerning the Italian service activities. Tourism in hotels represents almost 70% of total internal tourism in Italy, so that in order to guarantee good overall provisional estimates a predictor should perform well especially for the first 2 domains. Even though coverage of random replications referred to the number of provinces and not to the amount of nights spent, final coverage of the *y*-variable of interest was very similar and equal, on the average, to 19,6% for the first simulations and to 51,8% for the second one (table 6.1).

[8] The same approach was applied on the current monthly dataset used for the empirical attempt described in the next paragraph. Since quality of preliminary estimates turned out to be quite always worst than that obtained using other compared methods, these results have been omitted.

[9] Monthly tourism data at the province level have been officially released by ISTAT only for these years. When simulations were carried out, the last available year was 2004, since the delay of their publication is normally about one year. Older data were used by Gismondi (2007), but they were partially based on unofficial *ad hoc* extrapolations.

Table 6.1: Main features of tourism in Italy and % nights spent coverage - 2004

Domain	103 Provinces			10 top provinces		Coverage=20%		Coverage=50%	
	Nights spent (million)	Nights spent %	% change 2004/2003	Nights spent %	C.v.	Nights spent %	C.v.	Nights spent %	C.v.
Italians hotels	136,6	39,6	1,0	39,3	-	20,4	0,27	51,4	0,14
Foreigners hotels	97,2	28,1	3,4	61,5	-	21,3	0,47	51,3	0,25
Italians "o.c.a."	67,5	19,6	-2,9	26,7	-	18,0	0,27	52,4	0,12
Foreigners "o.c.a."	44,0	12,7	-3,8	56,6	-	17,6	0,52	52,8	0,27
Total	**345,3**	**100,0**	**0,3**	**45,3**	**-**	**19,6**	**0,30**	**51,8**	**0,15**

C.v.: coefficient of variation of nights spent coverage measured on 100 random replications.

Because of the high seasonality of tourism, nights spent in a given month identify a specific variable characterised by levels and dynamics that could be very different month by month. According to that and to preliminary analyses, the variable x used for univariate predictors was the number of nights spent in the same month m of the previous year 2001, that is $x_m=y_{(m-12)}$, so that the y-values can be estimated according to an autoregressive first order linear model (Fuller, 1990).

A preliminary empirical validation of a model as (2.1) was carried out according to tests yet applied in similar contexts by Gismondi (2002; 2007). For what concerns expected values, a simple technique consists in evaluating results of the regression model applied to the province i:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{6.1}$$

verifying the level of the correct-R^2 and the statistical significance of the usual test t.

As concerns heteroschedasticity, it can be tested using the known White test (1980). We estimate the parameters of the following regression model:

$$e_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \tau_i \tag{6.2}$$

where e_i are the residuals got from regression (6.1) and τ_i is a common random error, and then verify significance of the statistic nR^2 – where n is the number of observations (n=103) – that in this case is distributed as a χ^2 with 1 degree of freedom. If nR^2 is lower than the *chi-square* threshold for a given probability level, model (6.2) is not significant and the hypothesis of heteroschedasticity can be rejected, so that position v=1 can be accepted[10].

As showed in table 6.2, a regression through the origin leads to good results[11]. The

[10] Thresholds of *chi square* with 1 degree of freedom leaving on the right the 5% and the 1% of probability are, respectively, 3,84 and 6,63. Yearly results are confirmed for the most part of monthly data as well.

[11] In particular, the application to yearly data referred to 2003 (x) and 2004 (y) showed that in the model (2.1) the constant parameter is always not significant.

White test was always significant at the 99% level, with the exception of the domain "Italians in hotels" (95% only). As a consequence, model (6.1) should be affected by heteroschedasticity, so that $v \neq 1$.

Table 6.2: Model validation using tests (6.1) and (6.2)

	Model mean – Test (6.1)			Model variance – Test (6.2)			
Domain	*Correct* R^2	β estimate	Sign. T_β	Correct R^2	White test (nR^2)	Sign. 95%	Sign. 99%
Italians in hotels	0,991	1,011	0,000	0,045	6,562	Yes	No
Foreigners in hotels	0,990	1,036	0,000	0,275	29,811	Yes	Yes
Italians in "o.c.a."	0,989	0,971	0,000	0,067	8,773	Yes	Yes
Foreigners in "o.c.a."	0,990	0,964	0,000	0,804	83,218	Yes	Yes

In each of the 4 domains, for estimation of level in a given month (y_m or $\bar{y}_m$) the main synthetic quality measure used is the percent difference between estimated and true values, taken in absolute value. The overall yearly percent error is then obtained averaging monthly errors, using simple or weighted means, where weights are given by the percent incidence of tourism in each month on the whole yearly amount in the domain. While simple means are more related to the *general* precision of estimates, weighted means more strictly refer to the *effective* impact of errors in the overall estimate. In order to synthesize results derived from the 100 random replications, we considered the mean of absolute percent errors (*MAPE*), calculated as mean of 100 percent errors: as well known, *MAPE* is an unbiased estimation of the mean squared error of a quick predictor conditioned to the final data; then, final synthesis of results is obtained averaging *MAPE* for the 4 domains using as weights the relative incidence of tourism in each domain on the whole yearly amount.

For estimation of percent changes (namely, $(y_m/y_{(m-12)}-1)\cdot 100$), the procedure was similar to that for levels, but the main synthetic quality measure used is the difference (*not percent*) between estimated and true changes, taken in absolute value.

Moreover, the role played by the bias component in the overall mean square error of each estimator can be estimated on the basis of the empirical random replications carried out. The limit of this approach is that it is not possible to separate bias due to the particular considered estimator (that under the unknown underlying model could be or not be biased) from bias properly due to a self-selection effect. If h indicates a random replication (h=1,2,…,100), the empirical mean of predictor T for month m will be given by:

$$\bar{T}_m = \sum_{h=1}^{100} (T_{mh})/100 \tag{6.3}$$

while, denoting as $\bar{y}_m$ the true mean and using label E to indicate "empirical" variances, biases and mean squared errors, we can put:

$$MSE_{Em} = \sum_{h=1}^{100} (T_{mh} - \bar{y}_m)^2 / 100 = \sum_{h=1}^{100} (T_{mh} - \bar{T}_m)^2 / 100 + (\bar{T}_m - \bar{y}_m)^2 = VAR_{Em} + BIAS_{Em}^2 \qquad (6.4)$$

from which, putting w_m as the relative y-weight in the month m on the whole year, we can derive the percent incidence of squared bias on the whole *MSE*:

$$100 \sum_{m=1}^{12} \left(\frac{BIAS_{Em}^2}{MSE_{Em}} \right) w_m \,. \qquad (6.5)$$

All the 13 predictors used and compared in the empirical attempt have been yet defined in paragraphs from 2 to 5 and have been resumed in table 6.3[12]. They can be divided in 4 groups. The first group includes classical estimators based on one auxiliary variable:

- ratio (2.2 with $v=x$). It is the estimator currently used in the survey for non-responses imputation. It derives from the linear regression model (2.1) without constant and heteroschedasticity, with $x_m=y_{(m-12)}$.
- Regression (2.2 with $v=1$). It derives from the linear regression model (2.1) without constant and homoschedasticity, with $x_m=y_{(m-12)}$.
- Regression with constant (2.9). It derives from the linear regression model (2.1) with constant and homoschedasticity, with $x_m=y_{(m-12)}$.

A second group includes estimators that, by definition, tend to reduce the possible self-selection bias:

- ratio based on a balanced sub-sample (2.8). It is a sort of modified ratio estimator based on the after sampling selection of a sub-sample almost balanced; same hypotheses of ratio estimator 1), with $x_m=y_{(m-12)}$.
- Separate regression (3.24 and 3.22). Correction of self-selection bias supposing separate populations for provisional and late respondents, with $x_m=y_{(m-12)}$.
- Separate regression (3.24 and 3.23). The same features of estimator 5), with $x_m=y_{(m-12)}$.
- Modified Fuller (4.6). Correction of self-selection bias supposing a log-linear model[13] where the unobserved y-mean is modelled through the product between the mean of provisional respondents and the estimation error in a previous period, with $x_m=y_{(m-12)}$.

A third group includes 3 estimators that, even though not built up in order to reduce bias, exploit information of 2 auxiliary variables instead of 1:

- Ratio bivariate (5.2 with $v=x_1x_2$). Linear bivariate regression model without constant and heteroschedasticity, with $x_1=y_{(m-12)}$, $x_2=y_{(m-24)}$.
- Regression bivariate (5.2 with $v=1$). Linear bivariate regression model without constant and homoschedasticity, with $x_1=y_{(m-12)}$, $x_2=y_{(m-24)}$.
- Regression bivariate with constant (5.5). Linear bivariate regression model with constant and homoschedasticity, with $x_1=y_{(m-12)}$, $x_2=y_{(m-24)}$.

[12] Recourse – instead of imputation of non-responses – to estimation techniques based on re-weighting of respondents as calibration (Rizzo *et al*., 1996) – produced worst estimates, so that these techniques and relative results have been dropped.

[13] Preliminary analyses showed that a log-linear model based estimator as (4.6) lead to better results than a linear model based estimator as (4.5).

A fourth group includes the same estimators of group 3, but with a difference in the second auxiliary variable, that in this case is given by the *y*-value 6 months before. As a matter of fact, quite definitive data are available, *for all the provinces*, just after about 6 months from the end of the reference month. The idea is that, even though affected by a different seasonal pattern, more late data (respect to data referred to 2 years before) could improve estimates. We have:

- Ratio bivariate (5.2 with $v=x_1x_2$). As estimator 8), but with $x_1=y_{(m-12)}$, $x_2=y_{(m-6)}$.
- Regression bivariate (5.2 with $v=1$). As estimator 9), but with $x_1=y_{(m-12)}$, $x_2=y_{(m-6)}$.
- Regression bivariate with constant (5.5). As estimator 10), but with $x_1=y_{(m-12)}$, $x_2=y_{(m-6)}$.

Even though all the previous predictors have been introduced according to a super-population model, in the following, the terms "predictor" and "estimator" will be both used without ambiguity.

Table 6.3 - Provisional predictors for the number of monthly nights spent in tourist establishments

Code	Definition	General remarks
1	Ratio (2.2 with $v=x$)	Linear regression model without constant and heteroschedasticity ($x_m=y_{(m-12)}$)
2	Regression (2.2 with $v=1$)	Linear regression model without constant and homoschedasticity ($x_m= y_{(m-12)}$)
3	Regression with constant (2.9)	Linear regression model with constant and homoschedasticity ($x_m=y_{(m-12)}$)
4	Balanced sub-sample (2.8)	Modified ratio estimator based on a sub-sample almost balanced; same hypotheses of ratio estimator ($x_m=y_{(m-12)}$)
5	Separate regression (3.24 and 3.22)	Correction of self-selection bias supposing separate populations for provisional and late respondents ($x_m=y_{(m-12)}$)
6	Separate regression (3.24 and 3.23)	Correction of self-selection bias supposing separate populations for provisional and late respondents ($x_m=y_{(m-12)}$)
7	Modified Fuller (4.6)	Correction of self-selection bias supposing a log-linear model ($x_m=y_{(m-12)}$)
8	Ratio bivariate (5.2 with $v=x_1x_2$)	Linear bivariate regression model without constant and heteroschedasticity ($x_{1m}=y_{(m-12)}$, $x_{2m}=y_{(m-24)}$)
9	Regression bivariate (5.2 with $v=1$)	Linear bivariate regression model without constant and homoschedasticity ($x_{1m}=y_{(m-12)}$, $x_{2m}=y_{(m-24)}$)
10	Regression bivariate with constant (5.5)	Linear bivariate regression model with constant and homoschedasticity ($x_{1m}=y_{(m-12)}$, $x_{2m}=y_{(m-24)}$)
11	Ratio bivariate (5.2 with $v=x_1x_2$)	Linear bivariate regression model without constant and heteroschedasticity ($x_{1m}=y_{(m-12)}$, $x_{2m}=y_{(m-6)}$)
12	Regression bivariate (5.2 with $v=1$)	Linear bivariate regression model without constant and homoschedasticity ($x_{1m}=y_{(m-12)}$, $x_{2m}=y_{(m-6)}$)
13	Regression bivariate with constant (5.5)	Linear bivariate regression model and homoschedasticity ($x_{1m}=y_{(m-12)}$, $x_{2m}=y_{(m-6)}$)

Results have been summarized in tables from 6.4 to 6.8, where figures in bold indicate the best performance and figures underlined indicate the second best.

The main result deriving from the 100 replications guaranteeing a 20% coverage (table 6.4) is that, in a situation characterised by a low level of coverage (that is about the same coverage available when provisional estimations are carried after 45 days form the

reference month), the actual ratio estimator can be always improved: on the average, for estimation of levels there are 2 better estimators using simple means and 5 better estimators using weighted means; for estimation of changes, there are 1 better estimator using simple means and 3 better estimators using weighted means.

The overall best estimator is always given by the modified Fuller 7); on the average, gains respect to ratio estimator 1) are relevant: for weighted means error is 4,35% with 7) against 4,72%) with 1) for levels, while is 5,57% with 7) against 6,53% with 1) for changes.

The second best estimator is clearly estimator 10) for levels (regression bivariate with constant), improving the ratio bivariate estimator 8) that was the best estimator based on 2 auxiliary variables in the 8 provinces case seen in table 6.4. For changes, the second best is the ratio estimator using simple means and the separate regression estimator 5) using weighted means.

Moreover, while for changes the modified Fuller estimator 7) is almost always the best for each domain (with the exception of "Italians in hotels", for which it is the third best after the 2 ratio bivariate estimators 8) and 11)), for levels the recourse to alternative estimators for "Italians in hotels" (estimator 11) and estimators 5) or 8) for "Foreigners in hotels") can improve estimator 7).

In synthesis, when coverage of quick respondents is low, a self-selection bias can often be present, so that methods aiming at reducing this bias (in particular, estimator 7), but also the balanced sub-sample technique 4) for "Italians in o.c.a.") should be used at least for what concerns nights spent in the other collective accommodations, while for "Italians in hotels" a ratio bivariate strategy can be helpful (using as second auxiliary variable nights spent in month (m-6) or, on a second extent, nights spent in (m-24)).

When coverage of quick respondents is about 50% (table 6.5), results put in evidence a clear trade-off between estimation of levels or changes.

It is well known that optimal strategies for estimating levels or changes could be different (Rao *et al*., 1989, 458-460); in this case for estimating changes the modified Fuller estimator 7) is always the best in each domain and lead to an average error equal to 4,67% using a simple mean and to 3,87% using weighted means, and gains respect to the ratio estimator 1) are quite high, since ratio's average errors are respectively equal to 6,38% and 5,06%. This means that for estimation of change a 50% coverage is not enough to guarantee a significant reduction of self-selection bias in the available quick sample.

On the other hand, for estimation of levels the recourse to the actual ratio estimator 1) seems justified, even though lightly better results could be achieved using a regression bivariate with constant estimator 10) – as for "Foreigners in hotels" and "Italians in o.c.a.", even though the gain in precision is very much higher in the first case than in the second – or a ratio bivariate estimator 8) – as for "Italians in hotels".

From the point of view of a "minimax" strategy, table 6.6 confirms that, for level estimation, the ratio estimator 1) should be preferred to the modified Fuller estimator 7) only with a 50% coverage (and mostly because of the bad performance of estimator 7) for the domain "Foreigners in o.c.a."), while the vice-versa holds when coverage is low (20%): in the first case the highest estimate error is 3,91% for 1) and 5,05% for 7); in the second is, respectively, 9,82% against 9,21%.

Let's note that when coverage is 50%, several estimators based on 2 auxiliary variables (mostly estimator 12)) tend to contain the highest percent estimate error more than univariate estimators.

Finally, results in the second part of the table confirm that, when level estimation is considered, the availability of the largest provinces (in this case, the first ten) implies

efficiency of ratio estimator 1), according to the *MSE* formula (2.3). Of course, this situation represents only one of the $\binom{103}{10}$ quick samples that could be effectively available.

Moreover, even though in this case coverage in terms of nights spent is 45,3%[14] and, so, not very far from 50%, error obtained with ratio estimator 1) is quite higher than the average error achieved with the same estimator with 100 random replications as in table 6.6, where this level of coverage was guaranteed using a quite larger number of provinces (51 instead of 10). This means that it is better to base quick estimations on a large number of units (even though their coverage in terms of *y*-variable is not particularly high), rather than trying to induce *only a small* sub-set of large units to respond in advance.

The empirical percent incidence of squared bias on the global *MSE*, evaluated according to (6.3), reaches the lowest level just with the ratio estimator only for nights spent in hotels and a coverage equal to 20% (table 6.7). In all other cases, it can be reduced using mostly estimators belonging to the second group. On the average, the lowest incidence, for a coverage equal to 20%, is 8,01%, got using estimator 4) based on a balanced sub-sample, while for a coverage equal to 50% it is 14,62%, got with estimator 9) (regression bivariate). It is worthwhile to note that bias incidence is quite always higher with a higher coverage, meaning that increase of coverage leads to a reduction of estimators' variance more than proportional than reduction of bias.

Empirical evidence suggests the possibility to introduce in the survey some *mixed* quick estimation strategies, in order to achieve to a higher efficiency (table 6.8). Use of estimation strategies based on one or two estimators can be supposed realistic if one considers separately the 4 domains, without evaluating different *monthly* performances[15]. In details:

- With a 20% coverage (low coverage), it is convenient to use always predictor 7) – modified Fuller – except for "Italians in hotels", for which estimator 11) – ratio bivariate with $x_2=y_{(m-6)}$ – should be used. This strategy would produce a gain respect to the ratio estimator in each case except level estimation of "Foreigners in hotels". The average gain is 1,09 percent points for levels and 1,15 points for change.
- The most controversial situation concerns a 50% coverage. In this case there is a contrast between optimality of estimator 7) for changes and its clear sub-optimality for levels, so that a conservative option could be in favour of retaining ratio estimator. However, an improvement could still be reached using an alternative strategy based on the ratio bivariate estimator 8) for "Italians in hotels" and estimator 10) (regression bivariate with constant) in all the other domains. The average gain would be 0,25 percent points for levels and 0,02 for changes.

[14] The main 10 provinces in terms of overall nights spent in 2004 are: Venezia, Bolzano-Bozen, Roma, Rimini, Trento, Milano, Verona, Napoli, Firenze, Salerno.

[15] The possibility to use different quick estimators *for different months* is not realistic at the moment and actually not used in any ISTAT survey.

Table 6.4 - Provisional estimate errors with a 20% coverage (average of 12 months and 100 random replications)

Predictor	Italians in hotels		Foreigners in hotels		Italians in "o.c.a."		Foreigners in "o.c.a."		Total	
	S	W	S	W	S	W	S	W	S	W
	Percent errors (Levels) – MAPE									
Ratio (2.2 with $v=x$)	5,57	4,72	5,42	5,94	3,96	2,84	6,73	4,86	5,42	4,72
Regression (2.2 with $v=1$)	7,68	6,26	8,29	9,40	4,95	3,67	9,23	6,56	7,54	6,68
Regression with constant (2.9)	5,42	4,71	6,15	6,97	3,99	3,00	7,37	5,74	5,73	5,15
Balanced sub-sample (2.8)	5,54	4,65	5,43	5,84	**3,85**	**2,66**	8,63	4,98	5,86	4,64
Separate regression (3.24 with 3.22)	6,18	5,14	**4,48**	4,47	4,80	3,37	9,38	5,79	6,21	4,69
Separate regression (3.24 with 3.23)	8,20	6,55	6,46	6,61	6,16	4,15	12,02	6,53	8,21	6,09
Modified Fuller (4.6)	5,67	4,74	5,51	6,00	3,97	2,75	**2,88**	**1,95**	**4,51**	**4,35**
Ratio bivariate (5.2 with $v=x_1x_2$)	3,52	3,33	4,72	**4,41**	10,02	7,59	10,69	7,66	7,24	5,02
Regression bivariate (5.2 with $v=1$)	7,57	6,21	6,23	6,63	4,92	3,75	7,92	5,80	6,66	5,79
Regression bivariate with constant (5.5)	5,37	4,71	4,83	5,08	4,13	3,19	6,76	5,38	5,27	4,60
Ratio bivariate (5.2 with $v=x_1x_2$) (*)	**2,92**	**2,92**	5,15	5,36	11,30	9,55	14,87	10,66	8,56	5,89
Regression bivariate (5.2 with $v=1$) (*)	7,83	6,37	7,03	7,59	4,62	3,19	8,09	5,77	6,89	6,01
Regression bivariate with constant (5.5) (*)	5,35	4,52	5,83	6,42	4,21	2,93	7,44	5,24	5,71	4,84
	Absolute differences (Changes)									
Ratio (2.2 with $v=x$)	6,47	5,49	7,39	7,61	5,89	3,89	12,59	11,47	8,09	6,53
Regression (2.2 with $v=1$)	8,41	6,82	9,38	10,16	6,82	4,44	14,33	12,88	9,73	8,07
Regression with constant (2.9)	6,13	5,31	8,55	8,95	5,94	4,00	13,56	12,01	8,55	6,93
Balanced sub-sample (2.8)	6,85	5,78	7,61	7,78	6,26	3,78	14,31	11,50	8,76	6,68
Separate regression (3.24 with 3.22)	6,72	5,56	6,47	6,23	7,06	4,45	13,80	10,76	8,51	6,20
Separate regression (3.24 with 3.23)	8,80	7,14	7,76	7,60	8,27	5,00	16,45	11,90	10,32	7,46
Modified Fuller (4.6)	5,78	4,81	**5,69**	**6,17**	**3,97**	**2,71**	**11,95**	10,97	**6,85**	**5,57**
Ratio bivariate (5.2 with $v=x_1x_2$)	5,47	4,65	7,00	6,21	10,82	8,67	15,84	13,65	9,78	7,02
Regression bivariate (5.2 with $v=1$)	8,41	6,93	7,26	7,38	6,83	4,54	13,21	12,05	8,93	7,24
Regression bivariate with constant (5.5)	6,40	5,51	7,01	6,93	6,17	4,12	12,80	11,61	8,10	6,41
Ratio bivariate (5.2 with $v=x_1x_2$) (*)	**4,71**	**4,36**	8,04	7,43	11,63	10,29	15,77	13,02	10,04	7,48
Regression bivariate (5.2 with $v=1$) (*)	9,05	7,15	8,41	8,85	6,48	4,15	13,06	10,83	9,25	7,51
Regression bivariate with constant (5.5) (*)	6,36	5,29	8,36	8,66	6,08	4,00	13,09	**10,72**	8,47	6,68

(*) The variable x_2 is the variable $y_{(m-6)}$. *Note*: S = Simple mean; W = Weighted mean.

Table 6.5 - Provisional estimate errors with a 50% coverage (average of 12 months and 100 random replications)

Predictor	Italians in hotels		Foreigners in hotels		Italians in "o.c.a."		Foreigners in "o.c.a."		Total	
	S	W	S	W	S	W	S	W	S	W
	Percent errors (Levels) – MAPE									
Ratio (2.2 with v=x)	2,38	1,98	2,32	2,46	1,79	1,31	**3,11**	**2,12**	**2,40**	2,00
Regression (2.2 with v=1)	2,89	2,33	2,64	2,74	2,14	1,57	3,73	2,69	2,85	2,34
Regression with constant (2.9)	2,25	1,91	2,26	2,39	1,86	1,34	3,58	2,52	2,49	2,01
Balanced sub-sample (2.8)	2,56	2,19	2,61	2,72	2,07	1,43	3,83	2,63	2,77	2,25
Separate regression (3.24 with 3.22)	2,66	2,22	2,20	2,20	2,17	1,41	4,19	2,56	2,80	2,10
Separate regression (3.24 with 3.23)	3,24	2,68	2,45	2,52	2,67	1,70	4,47	3,13	3,21	2,50
Modified Fuller (4.6)	2,41	1,98	2,37	2,50	1,98	1,58	4,56	4,85	2,83	2,42
Ratio bivariate (5.2 with v=x1x2)	1,80	**1,56**	2,74	2,37	6,33	4,34	6,87	4,72	4,44	2,74
Regression bivariate (5.2 with v=1)	2,92	2,41	2,35	2,40	2,03	1,47	3,81	2,72	2,78	2,26
Regression bivariate with constant (5.5)	2,35	2,01	**2,10**	**2,12**	**1,74**	**1,23**	3,42	2,33	2,41	**1,93**
Ratio bivariate (5.2 with v=x1x2) (*)	**1,78**	1,62	2,89	2,67	4,31	3,87	8,25	5,82	4,31	2,89
Regression bivariate (5.2 with v=1) (*)	2,68	2,16	2,69	2,80	1,98	1,89	3,65	2,89	2,75	2,38
Regression bivariate with constant (5.5) (*)	4,72	4,15	2,39	2,51	2,15	1,53	3,60	3,01	3,21	3,03
	Absolute errors (Changes)									
Ratio (2.2 with v=x)	4,63	3,82	6,67	6,18	4,57	2,80	9,67	9,85	6,38	5,06
Regression (2.2 with v=1)	4,61	3,77	6,31	5,92	5,31	2,98	9,08	9,68	6,33	4,97
Regression with constant (2.9)	4,69	3,89	6,70	6,18	4,53	2,86	10,38	9,99	6,57	5,11
Balanced sub-sample (2.8)	4,88	4,00	7,01	6,54	4,63	2,93	10,52	9,86	6,76	5,25
Separate regression (3.24 with 3.22)	4,73	3,93	6,49	5,96	4,92	2,89	10,46	9,78	6,65	5,04
Separate regression (3.24 with 3.23)	4,61	3,85	6,29	5,90	5,45	3,03	10,84	10,45	6,80	5,11
Modified Fuller (4.6)	**3,46**	**3,02**	**4,46**	**4,58**	**2,99**	**2,56**	**7,77**	**6,97**	**4,67**	**3,87**
Ratio bivariate (5.2 with v=x1x2)	4,79	3,94	6,30	5,34	6,81	5,43	64,82	19,07	20,68	6,55
Regression bivariate (5.2 with v=1)	4,70	3,88	6,33	5,91	5,20	2,96	9,25	9,98	6,37	5,05
Regression bivariate with constant (5.5)	4,72	3,91	6,66	6,07	4,51	2,81	9,60	9,67	6,37	5,03
Ratio bivariate (5.2 with v=x1x2) (*)	4,66	3,85	6,66	5,65	4,43	4,15	10,01	10,09	6,44	5,21
Regression bivariate (5.2 with v=1) (*)	4,55	3,71	6,42	6,01	4,77	2,60	9,63	9,85	6,34	4,92
Regression bivariate with constant (5.5) (*)	6,33	5,39	6,73	6,17	4,81	3,13	9,71	9,86	6,89	5,74

Table 6.6 - Provisional estimate highest errors and when using the 10 largest provinces (average of 12 months)

Predictor	Italians in hotels		Foreigners in hotels		Italians in "o.c.a."		Foreigners in "o.c.a."		Total	
	The highest percent error (Levels) on 100 random replications									
	20%	50%	20%	50%	20%	50%	20%	50%	20% (**)	50% (**)
Ratio (2.2 with $v=x$)	10,20	4,06	8,52	4,34	7,55	**2,05**	13,00	5,21	9,82	3,91
Regression (2.2 with $v=1$)	19,29	4,37	13,55	4,11	8,46	2,92	18,74	5,13	15,01	4,13
Regression with constant (2.9)	8,08	4,12	12,48	4,31	6,86	3,18	16,05	5,26	10,87	4,22
Balanced sub-sample (2.8)	9,85	3,75	9,17	4,41	8,13	3,15	16,17	5,14	10,83	4,12
Separate regression (3.24 with 3.22)	9,41	3,87	8,17	4,24	8,13	2,95	22,39	6,93	12,03	4,50
Separate regression (3.24 with 3.23)	19,46	5,77	9,33	4,91	11,16	3,84	32,61	7,76	18,14	5,57
Modified Fuller (4.6)	10,25	4,10	8,40	4,35	**6,60**	2,54	11,60	9,21	**9,21**	5,05
Ratio bivariate (5.2 with $v=x_1x_2$)	**5,35**	**3,08**	8,58	5,19	21,97	9,50	18,82	10,71	13,68	7,12
Regression bivariate (5.2 with $v=1$)	17,61	4,31	9,43	3,73	8,64	2,92	17,70	6,61	13,34	4,39
Regression bivariate with constant (5.5)	7,52	4,00	8,90	**3,57**	7,22	3,14	15,81	5,70	9,86	4,11
Ratio bivariate (5.2 with $v=x_1x_2$) (*)	6,34	3,36	**6,67**	6,15	23,07	4,78	33,13	10,25	17,30	6,13
Regression bivariate (5.2 with $v=1$) (*)	16,59	4,22	9,76	4,04	9,01	2,77	9,01	**4,35**	11,09	**3,85**
Regression bivariate with constant (5.5) (*)	13,19	11,45	10,42	4,34	8,26	10,30	**8,26**	4,44	10,03	7,63
	Percent errors (Levels) using the 10 main provinces – MAPE									
		W	S	W	S	W	S	W	S	W
Ratio (2.2 with $v=x$)	4,74	3,84	2,97	3,41	3,66	**2,36**	**2,50**	2,00	**3,47**	3,19
Regression (2.2 with $v=1$)	4,56	3,21	2,80	2,48	5,44	4,22	5,29	3,22	4,52	3,20
Regression with constant (2.9)	5,52	5,50	3,63	3,11	5,28	5,12	5,56	4,57	5,00	4,63
Balanced sub-sample (2.8)	8,87	7,71	2,43	2,31	8,13	3,96	15,98	15,21	8,85	6,41
Separate regression (3.24 with 3.22)	5,10	4,44	**2,39**	2,57	5,19	3,71	4,91	4,18	4,40	3,74
Separate regression (3.24 with 3.23)	5,08	4,77	2,75	2,63	6,66	4,78	5,71	4,65	5,05	4,15
Modified Fuller (4.6)	5,46	4,42	3,03	3,68	**3,45**	2,45	2,88	**1,64**	3,70	3,47
Ratio bivariate (5.2 with $v=x_1x_2$)	4,52	4,34	2,53	**2,30**	5,54	3,55	3,93	3,07	4,13	3,45
Regression bivariate (5.2 with $v=1$)	5,83	4,00	2,86	2,56	5,95	4,48	5,32	2,79	4,99	3,53
Regression bivariate with constant (5.5)	15,41	15,36	9,59	9,62	7,86	7,36	14,46	13,68	11,83	11,96
Ratio bivariate (5.2 with $v=x_1x_2$) (*)	**3,84**	3,19	2,77	2,46	4,90	3,41	4,69	3,75	4,05	**3,10**
Regression bivariate (5.2 with $v=1$) (*)	4,42	**3,11**	2,97	2,74	4,71	3,91	4,56	3,06	4,17	3,16
Regression bivariate with constant (5.5) (*)	16,30	16,54	11,35	10,1	6,65	6,02	14,11	13,79	12,10	12,46

(*) The variable x_2 is the variable $y_{(m-6)}$. (**) Weighted means.

Table 6.7 - Empirical percent incidence of squared bias on the global MSE (average of 12 months and 100 random replications)

Predictor	Italians in hotels		Foreigners in hotels		Italians in "o.c.a."		Foreigners in "o.c.a."		Total	
	20%	50%	20%	50%	20%	50%	20%	50%	20% (*)	50% (*)
(2.2) with v=x	**5,17**	19,42	**4,78**	34,94	18,01	3,42	9,82	8,98	8,17	19,32
(2.2) with v=1	8,18	18,54	9,20	24,04	16,86	16,78	7,75	5,73	10,11	18,11
(2.9)	5,30	19,04	6,32	34,86	17,24	3,06	8,98	17,89	8,39	20,21
(2.8)	7,05	19,14	9,45	38,37	10,38	6,28	**4,17**	4,76	**8,01**	20,20
(3.24) with (3.22)	6,72	**13,01**	8,81	32,07	13,91	3,06	11,11	**4,73**	9,27	15,36
(3.24) with (3.23)	15,29	16,19	12,77	26,74	17,16	17,22	7,18	8,03	13,92	18,32
(4.6)	6,81	15,00	6,30	23,87	**8,03**	23,88	18,51	9,67	8,39	18,56
(5.2) with v=x1x2	11,77	27,08	18,50	19,06	19,86	20,29	21,16	25,48	16,44	23,29
(5.2) with v=1	8,93	15,14	12,68	16,77	16,68	14,81	9,00	7,92	11,51	**14,62**
(5.5)	5,24	18,59	11,41	30,28	17,59	**1,75**	11,01	12,34	10,13	17,78
(5.2) with v=x1x2 (**)	20,52	23,97	15,88	22,99	27,85	28,15	22,17	24,26	20,86	24,55
(5.2) with v=1 (**)	20,49	17,24	10,94	**15,02**	19,43	18,22	17,25	14,23	17,19	16,43
(5.5) (**)	11,88	17,46	15,17	19,04	20,45	9,34	19,26	20,04	15,42	16,64

(*) All errors have been obtained using weighted means. (**) The variable x_2 is the variable $y_{(m-6)}$.

Table 6.8 - Some optimal strategies for provisional estimates (at most 2 estimators are used)

Domain	Optimal strategy (*)			Gain in precision *vs* ratio estimator 1) (*)		
	Estimator	Error on levels	Error on changes	Estimator	Error on levels	Error on changes
			Coverage=20%			
Italians in hotels	11	2,92	4,36	1	1,80	1,13
Foreigners in hotels	7	6,00	6,17	1	-0,06	1,44
Italians in “o.c.a.”	7	2,75	2,71	1	0,09	1,18
Foreigners in “o.c.a.”	7	1,95	10,97	1	2,91	0,50
Total		**3,63**	**5,38**		**1,09**	**1,15**
			Coverage=50%			
Italians in hotels	8	1,56	3,94	1	0,42	-0,12
Foreigners in hotels	10	2,12	6,07	1	0,34	0,11
Italians in “o.c.a.”	10	1,23	2,81	1	0,08	-0,01
Foreigners in “o.c.a.”	10	2,33	9,67	1	-0,21	0,18
Total		**1,75**	**5,04**		**0,25**	**0,02**

(*) Weighted average errors are used.

7. Concluding remarks

The availability of historical data - playing the role of auxiliary information in a model based prediction context where the sample of quick respondents is given - led to an in-depth comparison among different prediction techniques, where the need of an *ex-post* correction for *self-selection* bias was strongly emphasized.

All methods compared can be applied also in those contexts (as, for instance, in the industrial production case) where statistical units are defined according to the original available microdata.

Even though shortness of time series did not allow for definitive conclusions on robustness of results, some final remarks resume the clearest evidences raised from the application:

- the ratio predictor used in the current survey on tourist nights spent, even though quite efficient, can be improved by other estimators, especially when coverage is low. It could be still used when coverage is quite higher (at least 50%).
- Non random *self-selection* of quick respondents can be reduced using particular regression techniques as those given by formulas (3.24) and (4.6). On the other hand, post-sampling balancing procedures do not produce significant bias reductions.
- The use of estimations techniques based on 2 auxiliary variables can be helpful only with a coverage equal to 50%, but efficiency gains respect to the ratio estimator are low.
- A bias reduction can be achieved using estimators different from ratio when coverage is 50% and mainly for nights spent in the other collective accommodations.
- Generally speaking, for the four domains taken into account (Italians in hotels, foreigners in hotels, Italians in other collective accommodations, foreigners in other collective accommodations) a mixture of *different* quick predictors could be used.

8. Appendix

8.1 Appendix 1: Optimal prediction under model G_{MR}

The G_{MR} model can be defined as follows:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \qquad \text{where:} \qquad \begin{cases} E(\varepsilon_i) = 0 \\ VAR(\varepsilon_i) = \sigma^2 v_i \\ COV(\varepsilon_i, \varepsilon_j) = 0 \quad if \quad i \neq j \end{cases} \tag{8.1}$$

where $\mathbf{x}_i' = (x_{1i}, x_{2i}, ..., x_{ki})$ and $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_k)$. If $\mathbf{y}' = (y_1, y_2, ..., y_n)$, $\mathbf{X}$ is the (*n*x*k*) matrix containing the *n* row-vectors $\mathbf{x}_i'$ (the first column could contain *n* 1*s*) and $\boldsymbol{\Sigma}$ is the (*n*x*n*) diagonal matrix containing the general *i-th* term v_i it will also follow that: $E(\mathbf{y}_S) = \mathbf{X}_S\boldsymbol{\beta}$, $E(\mathbf{y}_{\bar{S}}) = \mathbf{X}_{\bar{S}}\boldsymbol{\beta}$, $V(\mathbf{y}_S) = \sigma^2\boldsymbol{\Sigma}_S$, $V(\mathbf{y}_{\bar{S}}) = \sigma^2\boldsymbol{\Sigma}_{\bar{S}}$, $C(\mathbf{y}_S, \mathbf{y}_{\bar{S}}) = 0$, where S and $\bar{S}$ refer, respectively, to the sample and non sample units. For any given sampling design, the BLU predictor of the unknown mean $\bar{y}$ is given by (Cassel et al., 1977, 127):

$$T_{BLU} = \left(\frac{n}{N}\right)\bar{y}_S + \left(1 - \frac{n}{N}\right)\bar{\mathbf{x}}_{\bar{S}}^{'}\hat{\boldsymbol{\beta}}_{BLU} \tag{8.2}$$

where $\bar{\mathbf{x}}_{\bar{S}}^{'} = (\bar{x}_{1\bar{S}}, \bar{x}_{2\bar{S}}, ..., \bar{x}_{k\bar{S}})$ and $\hat{\boldsymbol{\beta}}^{'} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k)$. In particular, we will have:

$$\hat{\boldsymbol{\beta}}_{BLU} = \left(\mathbf{X}_S^{'}\boldsymbol{\Sigma}_S^{-1}\mathbf{X}_S\right)^{-1}\left(\mathbf{X}_S^{'}\boldsymbol{\Sigma}_S^{-1}\mathbf{y}_S\right) \tag{8.3}$$

$$E\left(T_{BLU} - \bar{y}\right)^2 = \frac{\sigma^2}{N^2}\left\{\sum_{\bar{S}} v_i + (N-n)^2\left[\bar{\mathbf{x}}_{\bar{S}}^{'}\left(\mathbf{X}_S^{'}\boldsymbol{\Sigma}_S^{-1}\mathbf{X}_S\right)^{-1}\bar{\mathbf{x}}_{\bar{S}}\right]\right\}. \tag{8.4}$$

This statement can be obtained as follows: the desired predictor must be of the form: $N^{-1}[n\,\bar{y}_S + (N-n)U]$, where $E(U) = \bar{\mathbf{x}}_{\bar{S}}^{'}\boldsymbol{\beta}$. The (model) unbiased estimator $\hat{\boldsymbol{\beta}}_{BLU}$ of $\boldsymbol{\beta}$ is the well known generalised least squared estimator (Johnston, 1972, 233). The expression (8.3) follows easily.

8.2 Appendix 2: Optimal prediction under model (3.1) (self-selection bias)

The unbiasedness condition $E(T_{(AL)} - y_U) = 0$ under model (3.1) leads to:

$$\beta_P\left(\sum_{S_P} c_{Pi}\,x_{Pi} - x_{\bar{S}_P}\right) + \beta_L\left(\sum_{S_L} c_{Li}\,x_{Li} - x_{\bar{S}_L}\right) = 0 \tag{8.5}$$

where $x_{\bar{S}_P} = \sum_{\bar{S}_P} x_{Pi}$ and $x_{\bar{S}_L} = \sum_{\bar{S}_L} x_{Li}$. A general condition that guarantees (8.5) is:

$$\sum_{S_P} c_{Pi}\,x_{Pi} = x_{\bar{S}_P} \qquad \sum_{S_L} c_{Li}\,x_{Li} = x_{\bar{S}_L}. \tag{8.6}$$

Following Cicchitelli *et al*. (1992, 394) and on the basis of (3.3) we can also write:

$$MSE(T_{(PL)}) = E\left(T_{(PL)} - y_U\right)^2 = VAR\left(\sum_{S_P} c_{Pi}\,y_{Pi} + \sum_{S_L} c_{Li}\,y_{Li}\right) + VAR(y_{\bar{S}}) \tag{8.7}$$

where:

$$VAR\left(\sum_{S_P} c_{Pi}\, y_{Pi} + \sum_{S_L} c_{Li}\, y_{Li}\right) = \sigma_P^2 \sum_{S_P} c_{Pi}^2\, v_{Pi} + \sigma_L^2 \sum_{S_L} c_{Li}^2\, v_{Li} \quad \text{and}$$

$$VAR(y_{\bar{S}}) = \sigma_O^2 \sum_{\bar{S}_O} v_{Oi} + \sigma_L^2 \sum_{\bar{S}_L} v_{Li}\,. \tag{8.8}$$

Minimisation of (8.8) under constraints given by (8.6) leads straightforwardly to (3.4), from which we can also derive (3.5). If we write (3.3) as: $T_{(PL)} = y_S + \hat{y}_{(PL)\bar{S}} = (y_{S_P} + \hat{y}_{\bar{S}_P}) + (y_{S_L} + \hat{y}_{\bar{S}_L})$, we can easily verify that the optimal predictor (3.4) can be also written as $T^*_{(PL)} = \left(y_{S_P} + x_{\bar{S}_P}\hat{\beta}_P\right) + \left(y_{S_L} + x_{\bar{S}_L}\hat{\beta}_L\right)$, where $x_{\bar{S}_P}\hat{\beta}_P$ is the optimal linear predictor of $y_{\bar{S}_P} = (y_P - y_{S_P})$ and $x_{\bar{S}_L}\hat{\beta}_L$ is the optimal linear predictor of $y_{\bar{S}_L} = (y_L - y_{S_L})$.

References

AELEN F. (2003), "Improving Timeliness of Industrial Short-term Statistics using Time Series Analysis" Statistics Netherlands Working Paper, available on www.oecd.org /dataoecd/23/12/30044343.pdf.

BATTAGLIA F., FENGA L. (2003), "Forecasting Composite Indicators with Anticipated Information: an Application to the Industrial Production Index", Statistica Applicata, 52, 3.

BOLFARINE H., ZACKS S. (1992), Prediction Theory for Finite Populations, Springer-Verlag.

CAPPUCCIO N., ORSI R. (1991), Econometria, Il Mulino, Bologna.

CASSEL C., SÄRNDAL C.E., WRETMAN J. (1977), Foundations of Inference in Survey Sampling, J.Wiley & Sons, New York.

CASSEL C., SÄRNDAL C.E., WRETMAN J. (1983), "Some Uses of Statistical Models in Connection with the Nonresponse Problem", in: Madow W.G., Olkin I., Rubin D. (eds.), Incomplete Data in Sample Surveys, vol.3, 143-160, Academic Press, New York.

CICCHITELLI G., HERZEL A., MONTANARI G.E. (1992), Il campionamento statistico, Il Mulino, Bologna.

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

COSTA P., MANENTE M. (2000), Economia del turismo, Touring Club Italiano, Milano.

COUNCIL OF THE EUROPEAN UNION (1995), Council Directive 95/57/CE on the Collection of Statistical Information in the Field of Tourism, 23th November, Bruxelles.

DALABEHERA M., SAHOO L.N. (1999), "A New Estimator with Two Auxiliary Variables for Stratified Sampling", Statistica, anno LIX, 1, 101-107, Clueb, Bologna.

DEVILLE J.C., TILLÉ Y. (2004), "Efficient Balanced Sampling: the Cube Method", Biometrika, 91, 4, 893-912.

DIVISEKERA S. (2003), "A Model of Demand for International Tourism", Annals of Tourism Research, 30, 31-49.

DRUDI I., FILIPPUCCI C. (2000), "Inferenza da campioni longitudinali affetti da selezione non casuale", in: Filippucci C. (ed.), Tecnologie informatiche e fonti amministrative nella produzione di dati, 415-432, Franco Angeli, Milano.

DRUDI I., FERRANTE M.R. (2003), "Stima da fonti amministrative longitudinali con parziale sovrapposizione delle unità", in: Falorsi P.D., Pallara A., Russo A. (eds.), Temi di ricerca ed esperienze sull'utilizzo a fini statistici di dati di fonte amministrativa, 115-132, Franco Angeli, Milano.

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

FALORSI P., ALLEVA G., BACCHINI F., IANNACCONE R. (2005), "Estimates Based on Preliminary Data from a Specific Subsample and from Respondents not Included in the Subsample", Statistical methods and applications, 14, 1, 83-99, Physica-Verlag.

FULLER W. (1990), Analysis of Repeated Surveys, Survey Methodology, 16,167-180.

GISMONDI R. (2002), "Model Based Sample Selection using Balanced Sampling", Rivista di Statistica Ufficiale, 3, 81-109, Franco Angeli, Milano.

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. June 8-11, Halifax, Nova Scotia, Canada.

GISMONDI R. (2007), "Quick Estimation of Tourist Nights Spent in Italy", Statistical Methods and Applications, on-line publication: http://dx.doi.org/10.1007/s10260-006-0035-3, Springer & Verlag.

GISMONDI R., MIRTO A.P.M., SALAMONE N. (2003), "Una stima "rapida" delle presenze turistiche in Italia: un approccio multivariato", Proceedings of the conference CLADAG 2003, 185-188.

HARVEY A.C. (1984), "A Unified View of Statistical Forecasting Procedures", Journal of Forecasting, 3, 245-275.

HERNÁNDEZ-LÓPEZ M. (2004), "Future Tourists' Characteristics and Decisions: the Use of Genetic Algorithms as a Forecasting Method", Tourism Economics, vol.10, 3, 245-262.

ISTAT (2005), Rapporto sulle stime anticipate, report finale del progetto "Stime anticipate per le indagini congiunturali sulle imprese" (a cura di Falorsi S. e Gismondi R.), Istat, Roma.

ISTAT (Anni vari), Statistiche del turismo, Collana Informazioni, Istat, Roma.

JOHNSTON J. (1983), Econometria, Franco Angeli, Milano.

LIM C., MC ALEER M. (2001), "Forecasting Tourist Arrivals", Annals of Tourism Research, Vol.28, 4, 965-977.

MARAVALLE M., POLITI M., IAFOLLA P. (1993), "Scelta di indicatori per la stima rapida di un indice provvisorio della produzione industriale", Quaderni di ricerca Istat, 6.

MONTANARI G.E. (1987), "Post-sampling Efficient QR-Prediction in Large Sample Surveys", International Statistical Review, 55, 191-202.

PERRI P.F.(2005), "Improved Ratio-cum-product Type Estimators in Simple Random Sampling using Two Auxiliary Variables", Atti del convegno Cladag 2005, 473-476, MUP editore, Parma.

RAO J.N.K., SRINATH K.P., QUENNEVILLE B. (1989), "Estimation of Level and Change Using Current Preliminary Data", in: Kasprzyk D, Duncan G, Kalton G, Singh MP (eds.), Panel Surveys, 457-485, J.Wiley & Sons, New York.

RIZZO L., KALTON G., BRICK M.J. (1996), "A Comparison of some Weighting Adjustment Methods for Panel Non-response", Survey Methodology, 22, 1, 43-53.

ROYALL R.M. (1988), The Prediction Approach to Sampling Theory, Handbook of Statistics vol. 6. North Holland.

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.

SINGH M.P. (1965), "On the Estimation of Ratio and Product of the Population Parameters", Sankhya, B, 27, 321-328.

SONG H., WITT S.F. (2000), Tourism Demand Modelling and Forecasting: Modern Econometric Approaches, Pergamon, Oxford.

TAM S.M. (1987), "Analysis of Repeated Surveys Using a Dynamic Linear Model", International Statistical Review, 55, 1, 63-73.

ULLBERG A. (2003), "More Rapid Retail Trade Statistics in Sweden", Statistics Sweden Working Paper, available on www.oecd.org/dataoecd/2/62/2956932.pdf.

VALLIANT R., DORFMAN A.H., ROYALL R.M. (2000), Finite Population Sampling and Inference – A Prediction Approach, J.Wiley & Sons, New York.

WHITE H. (1980), "A Heteroschedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroschedasticity", Econometrica, 48, 817-838.

YANSANEH I.S., FULLER W.A. (1998), "Optimal Recursive Estimation for Repeated Surveys", Survey Methodology, Vol.24, 1, 31-40.

Un nuovo approccio all'analisi delle componenti locali e strutturali

Alessandro Faramondi[1]

Sommario

Nell'ambito degli studi socio-economici, uno dei metodi più diffusi per l'analisi delle componenti locali e strutturali è l'analisi shift-share. Tale metodo consente di scomporre una variazione assoluta del parametro oggetto di studio, o del suo tasso di variazione, in tre componenti: componente tendenziale, componente strutturale e componente locale. Note le potenzialità, le criticità e le eventuali soluzioni, sembrava opinione comune che tale tecnica, almeno dal punto di vista metodologico, avesse poco da dire. E così è stato, almeno fino al 2003, quando Nazara e. Hewings, adottano un nuovo approccio che tiene conto della dipendenza spaziale tra le aree oggetto di studio. Nel presente lavoro viene presentato il nuovo approccio e una sua applicazione al valore aggiunto a livello regionale. In particolare la scomposizione del valore aggiunto regionale viene ottenuta considerando la specializzazione produttiva dei Sistemi locali del lavoro.

Abstract

The goal of this paper is to present the new approach to the shift-share analysis, proposed by Nazara and Hewings (2003). Besides the traditional components, the new approach considers the spatial component to decompose the regional growth rate.
The analysis of the Italian data is carried out. In particular, the regional value added growth rate is decomposed on the base of the Local Labour Systems productive specialization rate.

Parole chiave: shift-share, dipendenza spaziale

1. Introduzione

Nell'ambito degli studi socio-economici, uno dei metodi più diffusi per l'analisi delle componenti locali e strutturali è l'analisi shift-share. Tale metodo, che molti fanno risalire al lavoro di E.S. Dunn (1960), consente di scomporre una variazione assoluta o un tasso di variazione, in tre componenti: componente tendenziale, componente strutturale e componente locale. Gli ambiti di maggior diffusione riguardano la dinamica del mercato del lavoro e della produttività, anche se non mancano esempi applicativi in altri contesti.

Stevens e More (1980), individuano due fattori principali di successo. Il primo è la semplicità di impostazione e l'immediatezza nelle elaborazioni. Il secondo, fa riferimento alla solidità dei risultati. Infatti, nonostante le critiche, le analisi empiriche non hanno ancora evidenziato criticità tali da sconsigliarne l'utilizzo.

[1] Ricercatore (Istat), e-mail: faramond@istat.it

Per quanto concerne le critiche, Zaccomer ricorda i problemi legati all'ipotesi di additività del modello ed alla robustezza dei risultati: "il modello essendo di tipo additivo non tiene conto delle possibili interazioni tra le diverse componenti, in particolare tra quella strutturale e quella locale. Non vi è poi robustezza dei risultati rispetto al livello di classificazione delle attività economiche utilizzato per delimitare i limiti settoriali: più il livello è analitico, maggiore è l'importanza della componente strutturale a dispetto di quella locale".

Nazara e Hewings (Towards regional growth decomposition with neighbor's effect: a new perspective on shift-share analysis - 2003) invece, si concentrano sui problemi che riguardano la relazione di dipendenza tra le varie componenti, secondo cui livelli territoriali minori, ad esempio le regioni, dipendono esclusivamente dalle aree gerarchicamente superiori, ad esempio le nazioni, senza tener conto della relazione di dipendenza che sussiste tra aree contigue (dipendenza orizzontale). Nel lavoro citato, Nazara e Hewings non si limitano all'individuazione del problema ma forniscono una soluzione metodologica. La caratteristica del nuovo approccio è di considerare nel modello una relazione di dipendenza spaziale tra le unità territoriali oggetto di analisi. L'ipotesi è che le performance di un territorio non sono indipendenti dalle aree circostanti. A differenza del modello classico, che presuppone l'esistenza di una struttura gerarchica della dipendenza (ad esempio, la nazione influenza la regione), il nuovo approccio prevede anche l'esistenza di una dipendenza orizzontale (ad esempio, tra regione e regione).

L'obiettivo del presente lavoro è di presentare l'analisi shift-share con struttura spaziale proposta da Nazara e Hewings, sia nella formalizzazione analitica, presentando la metodologia proposta dagli autori, sia verificando empiricamente le potenzialità del nuovo approccio, applicando l'analisi shift-share con struttura spaziale ai dati italiani, a partire dal livello territoriale dei sistemi locali del lavoro (Sll definiti in base ai dati del Censimento della Popolazione del 2001), considerando le singole specializzazioni e quindi aggregando i dati per regione.

La definizione metodologica del nuovo approccio di Nazara e Hewings è presentata nel paragrafo 3, mentre nel paragrafo 2 viene presentato il metodo secondo l'impostazione classica.

Nel Paragrafo 4 viene proposta una variante del nuovo approccio con struttura spaziale. In particolare la relazione di dipendenza spaziale viene ipotizzata solo per la componente locale.

Quindi, nel paragrafo 5 viene presentata l'applicazione ai dati italiani, a partire dalle specializzazioni produttive dei sistemi locali del lavoro. In particolare l'applicazione fa riferimento alla variazione del valore aggiunto nel periodo 1996-2002.

Per finire, il Paragrafo 6 è dedicato alle conclusioni.

2. Analisi *Shift-Share* classica

La base metodologica dell'analisi shift-share classica è la scomposizione in più termini della variazione relativa della variabile d'interesse (ad esempio l'occupazione) per il generico territorio r (ad esempio la regione r-esima) tra il tempo (t) e (t-1). L'analisi distingue le seguenti componenti (per comodità di trattazione, d'ora in avanti si parlerà di regione e nazione):

– componente tendenziale: rappresenta la variazione percentuale nell'intero paese;

– componente strutturale: misura l'effetto della maggiore/minore presenza nella regione di settori che nel complesso del paese sono risultati a più rapida crescita;

– componente locale: misura i differenziali di crescita, per i diversi settori tra il livello regionale ed il livello nazionale.

In particolare, la variazione assoluta della variabile d'interesse, nel settore i-esimo della regione r-esima è:

$$\Delta x_{ri} = \left[g_{..} + (g_{.i} - g..) + (g_{ri} - g_{.i})\right] x_{ri} \qquad (1)$$

dove:

x è la variabile oggetto d'interesse (ad esempio gli occupati oppure il valore aggiunto);

r indica la generica regione;

i indica il generico settore;

$g_{ri} = \frac{\Delta x_{ri}}{x_{ri}^{(t-1)}}$, è il tasso di variazione della variabile oggetto d'interesse, tra il tempo (t-1) ed il tempo t, nella regione r-esima, nel settore i-esimo;

$g_{.i} =$ variazione relativa della variabile d'interesse tra il tempo (t-1) e t nel generico settore i-esimo;

$g_{..} =$ variazione relativa media.

Dividendo a destra e a sinistra dell'uguale per x_{ri}, sommando rispetto a tutti i settori della regione r-esima, si ha la classica espressione dell'analisi *shift-share*:

$$g_{r.} = g_{..} + \sum_i (g_{.i} - g_{..}) \frac{x_{ri}}{x_{r.}} + \sum_h (g_{ri} - g_{.i}) \frac{x_{ri}}{x_{r.}} \qquad (2)$$

dove

$g_{r.}$ = variazione relativa della variabile d'interesse tra il tempo (t-1) e t nella generica regione r-esima;

$g_{..}$ = variazione relativa media (componente tendenziale);

$\sum_i (g_{.i} - g_{..}) \frac{x_{ri}}{x_{r.}}$ = media ponderata dei differenziali tra la variazione relativa nel settore i-esimo e la variazione media nel complesso dei settori, in cui i pesi sono le percentuali della variabile d'interesse (ad esempio occupati) nella regione r-esima del settore i-esimo (componente strutturale);

$\sum_i (g_{ri} - g_{.i}) \frac{x_{ri}}{x_{r.}}$ = media ponderata dei differenziali tra la variazione di settore nella regione r-esima e la variazione di settore sull'intero territorio (componente locale).

3. Analisi *Shift-Share* con struttura spaziale

L'analisi shift-share classica si basa sulla definizione di tre componenti e sulla relazione gerarchica che esiste tra di esse, assumendo che il livello territoriale maggiore, ad esempio la nazione, influenzi le aree sottostanti, ad esempio le regioni. In tale contesto, la possibilità

di inter-relazione tra le aree non viene presa in considerazione. Tale assunto, utile per semplificare la definizione dei modelli, sembra poco coerente con le dinamiche dei fenomeni socio-economici, che per loro natura tendono ad essere correlati spazialmente.

Partendo da tali considerazioni, Nazara e Hewings (2003) hanno proposto un nuovo metodo di analisi, che tenga conto della dipendenza spaziale tra le aree oggetto di studio.

Prima di entrare nel merito del metodo, è opportuno affrontare il problema di come trattare la relazione tra le aree. Seguendo una logica ormai consolidata, Nazara e Hewings propongono di considerare una matrice di pesi W, di dimensione (R x R), dove R rappresenta il numero di aree territoriali e il generico elemento della matrice, w_{rs}, indica il grado di interazione tra la regione r-esima ed s-esima. Ovviamente, nel caso di w_{rs} uguale a zero si ha assenza di interazione tra le regioni r-esima ed s-esima. Nel caso di interazione diversa da zero è possibile prevedere diverse opzioni, che misurano il grado di interazione tra le aree. La soluzione più semplice è di discriminare tra aree contigue e non contigue. Nel primo caso si avrà presenza di interazione (w_{rs} assume il valore 1), nel secondo assenza di interazione (w_{rs} assume il valore 0). Una generalizzazione della contiguità binaria, è stata proposta da Cliff e Ord (1981), dove viene considerata come misura dell'interazione la parte di confine in comune tra le aree.

Oltre la contiguità fisica è possibile considerare altre caratteristiche che sono connesse al grado di interazione tra due aree. Ad esempio, laddove la variabile discriminante non sia lo spazio bensì il tempo è possibile definire i valori di w_{rs} a partire dai tempi di percorrenza effettivi. A tale proposito, nella recente sperimentazione dell'Istat sulle stime di "Occupati residenti e persone in cerca di occupazione" a livello di sistema locale del lavoro, è stata considerata tra le altre, la matrice di pesi definita in base ai tempi di percorrenza tra i centroidi dei sistemi locali.

Un'ulteriore possibilità è la matrice di connessione, che non considera né lo spazio né il tempo, bensì i flussi tra le aree. Un tale approccio, a differenza dei precedenti, pone l'accento sul fenomeno oggetto di studio. Potremmo essere interessati ad una misura che esprima la relazione tra aree in termini di import-export oppure in termini di flussi finanziari, ecc. A tale proposito è esplicativo l'esempio di Zaccomer (2005), in base al quale, nell'ambito della UE, l'Italia potrebbe risultare maggiormente connessa alla Spagna rispetto alla Slovenia, se il sistema di pesi fosse definito a partire dai flussi commerciali tra le unità territoriali prese in esame.

Dopo aver scelto la matrice dei pesi W, si passa al modello proposto da Nazara e Hewings:

$$g_{ri} = \left[g_{..} + (\ddot{g}_{ri} - g_{..}) + (g_{ri} - \ddot{g}_{ri}) \right] \qquad (3)$$

dove

$\ddot{g}_{ri} = \dfrac{\sum_{s=1}^{R} \breve{w}_{rs} x_{si}^{(t)} - \sum_{s=1}^{R} \breve{w}_{rs} x_{si}^{(t-1)}}{\sum_{s=1}^{R} \breve{w}_{rs} x_{si}^{(t-1)}}$, è il tasso di variazione del settore i-esimo del vicinato della regione r-esima.

Quindi l'espressione finale del modello shift-share con struttura spaziale, si ottiene

sostituendo nel modello classico, il tasso settoriale nazionale ($g_{.i}$) con il tasso settoriale della macroarea della regione r-esima ($\ddot{g}_{ri}$):

$$g_{r.} = g_{..} + \sum_i (\ddot{g}_{ri} - g_{..}) \frac{x_{ri}}{x_{r.}} + \sum_h (g_{ri} - \ddot{g}_{ri}) \frac{x_{ri}}{x_{r.}} \qquad (4)$$

I tre effetti della (5) hanno il seguente significato:

- $g_{..}$ ha il medesimo significato del metodo classico (*componente tendenziale*);
- $\sum_i (\ddot{g}_{ri} - g_{..}) \frac{x_{ri}}{x_{r.}}$, è un effetto misto, che dipende sia dal mix settoriale sia dalle performance del vicinato;
- $\sum_h (g_{ri} - \ddot{g}_{ri}) \frac{x_{ri}}{x_{r.}}$, è un effetto semplice, che dipende dal confronto tra le performance della regione r-esima e la macroarea a cui appartiene la regione stessa.

A differenza del modello classico, nel modello con struttura spaziale sono presenti sia effetti semplici sia effetti combinati. In particolare la presenza di questi ultimi, cioè di effetti che dipendono da più fattori, complica l'interpretazione dei risultati in quanto non è possibile effettuare un'attribuzione univoca delle risultanze dell'analisi.

Per superare tale problema gli autori hanno proposto di esprimere gli effetti combinati in termini di sequenze di effetti semplici, adottando la tecnica di decomposizione *step-by-step*. In base a tale tecnica, la seconda componente diviene una somma di effetti semplici:

$$\sum_i (\ddot{g}_{ri} - g_{..}) \frac{x_{ri}}{x_{r.}} = \sum_i \left[(\ddot{g}_{ri} - \ddot{g}_{r.}) + (\ddot{g}_{r.} - g_{r.}) + (g_{r.} - g_{ri}) + (g_{ri} - g_{.i}) + (g_{.i} - g..) \right] \frac{x_{ri}}{x_{r.}} \qquad (5)$$

Tale scomposizione da luogo a cinque effetti semplici:

1. l'effetto del mix settoriale a livello del vicinato della regione r-esima
2. l'effetto del differenziale di sviluppo tra il vicinato della regione r-esima e il livello nazionale
3. l'effetto del mix settoriale a livello della regione r-esima
4. l'effetto del differenziale di sviluppo tra la regione r-esima e il livello nazionale
5. l'effetto del mix settoriale a livello nazionale.

La logica seguita da Nazara e Hewings è quella di scomporre il tasso di variazione introducendo nel modello classico dell'analisi shift-share un nuovo livello intermedio tra l'unità territoriale locale ed il livello di aggregazione massimo. Tale operazione viene ottenuta sostituendo il tasso di variazione del settore i-esimo al massimo livello di aggregazione con il tasso di variazione i-esimo al livello intermedio, modificando quindi sia la componente strutturale sia la componente locale. Questo modo di procedere, coerente con il metodo di decomposizione step-by-step, implica l'esistenza di due nuovi effetti, uno combinato e uno semplice, con la conseguenza che l'effetto combinato per essere interpretato deve essere ulteriormente scomposto in somma di effetti semplici. Il risultato ultimo è un modello composto da sette effetti semplici, che dopo alcune semplificazioni

può essere ridotto ad una somma di cinque effetti semplici. La caratteristica di tale modello è di assumere una logica di dipendenza spaziale in tutte le componenti del modello classico, sia in quella strutturale, sia in quella locale. Una tale opzione se da un lato garantisce la relazione di coerenza del modello, dall'altro introduce un elemento dubbio dal punto di vista teorico. E cioè, ha senso considerare una relazione di dipendenza spaziale anche nella componente strutturale? Non sarebbe più logico considerare la dipendenza spaziale solo per la componente locale del modello classico, visto che è la componente che misura l'effetto della variazione attribuibile al contesto territoriale.

Inoltre, considerare la dipendenza spaziale solo per la componente locale del modello classico, comporterebbe una riduzione delle componenti, con un considerevole guadagno in capacità di interpretazione.

4. Una nuova proposta di analisi shift-share con struttura spaziale

A partire dal lavoro di Nazara e Hewings, è stato sviluppato un nuovo modello di analisi *shift-share* con struttura spaziale, che considera la relazione di dipendenza spaziale solo per la componente locale del modello classico. Infatti, seguendo l'impostazione logica dell'analisi *shift-share* classica, solo la componente locale è interessata da effetti riconducibili al livello territoriale. In tal senso è ipotizzabile un diverso effetto sulla componente locale a seconda dell'influenza esercitata dalle aree contigue. Per contro la componente media e quella strutturale non hanno nulla a che vedere con eventuali componenti territoriali.

In base a tali considerazioni il nuovo modello di analisi *shift-share* è il seguente:

$$g_{r.} = g_{..} + \sum_i (g_{.i} - g_{..}) \frac{x_{ri}}{x_{r.}} + (nuova \;\; componente \;\; locale) \qquad (7)$$

dove le prime due componenti sono le stesse del modello classico.

Per quanto riguarda la *"nuova componente locale"*, è stata suddivisa in due parti:

$$componente\ locale = (\ddot{g}_{ri} - g_{.i}) + (g_{ri} - \ddot{g}_{ri}),$$

dove

$\ddot{g}_{ri}$ è la componente definita nel modello di Nazara e Hewings.

Quindi, la formulazione finale è:

$$g_{r.} = g_{..} + \sum_i (g_{.i} - g_{..}) \frac{x_{ri}}{x_{r.}} + \sum_i (\ddot{g}_{ri} - g_{.i}) \frac{x_{ri}}{x_{r.}} + \sum_i (g_{ri} - \ddot{g}_{ri}) \frac{x_{ri}}{x_{r.}} \qquad (8)$$

dove

$\sum_i (\ddot{g}_{ri} - g_{.i}) \frac{x_{ri}}{x_{r.}}$ è un effetto semplice, che misura il differenziale di crescita tra il vicinato dell'area r-esima ed il massimo livello di aggregazione (*effetto di macroarea*);

$\sum_i (g_{ri} - \ddot{g}_{ri}) \frac{x_{ri}}{x_{r.}}$ è un effetto semplice, che misura il differenziale di crescita tra l'area r-esima ed il vicinato della stessa (*effetto della componente locale*).

In questo modo la componente locale del modello classico è stata scomposta in due componenti, la prima consente di isolare l'effetto dovuto all'auto correlazione spaziale, la seconda consente di isolare l'effetto dovuto esclusivamente all'area oggetto di studio.

5. Applicazione: analisi della variazione del valore aggiunto a livello regionale nel periodo 1996-2002

L'applicazione riguarda la variazione del valore aggiunto regionale nel periodo 1996-2002. Una possibile chiave di lettura può essere fornita separando il contributo strutturale delle diverse specializzazioni produttive, dai rimanenti fattori locali di sviluppo. Infatti la presenza di specializzazioni produttive favorevoli, espressioni delle realtà più dinamiche dell'economia, costituisce un fattore di progresso regionale autonomo e, almeno concettualmente, separabile dai fattori localizzativi e di competitività. Il ruolo delle diverse specializzazioni produttive in ambito regionale, può essere proficuamente analizzato a partire dalle specializzazioni dei singoli Sistemi locali del lavoro[2] che appartengono ad una data regione, al fine di definire una mappa regionale delle specializzazioni su base locale. La classificazione dei 784 Sistemi locali è stata realizzata a partire dai dati del censimento intermedio relativi alle unità locali e agli addetti alle unità locali. Il risultato delle operazioni di sintesi e classificazione, è stato la suddivisione dei Sistemi locali del lavoro in 11 gruppi omogenei di specializzazione produttiva (Rapporto Annuale – La situazione del Paese nel 1999) (Tavola 4).

L'analisi delle componenti strutturali e regionali è stata condotta seguendo il metodo proposto nel Paragrafo 4. Trattandosi della prima analisi shift-share sulle regioni italiane, secondo un'ipotesi di dipendenza spaziale, si è scelto di adottare un criterio di vicinato semplificato, che tenga conto esclusivamente della contiguità fisica, rinviando ad ulteriori approfondimenti l'utilizzo di matrici che tengano conto anche di altri elementi. Quindi nella presente applicazione è stata adottata una matrice simmetrica 20x20, di zero e uno. Zero nel caso di non contiguità fisica e uno nel caso opposto. Sulla diagonale principale sono stati considerati tutti 1, al fine di considerare nelle macroaree anche le regioni prese in esame. Tale scelta è coerente con la struttura gerarchica del modello shift-share, in quanto il livello nazionale implementa tutte le regioni, compresa quella in esame e allo stesso modo le macroaree devono contenere le regioni di volta in volta considerate.

Il metodo si basa sulla scomposizione della variazione relativa, nelle seguenti quattro componenti:

- *componente tendenziale*: rappresenta la variazione percentuale della variabile d'interesse nell'intero paese;
- *componente strutturale*: misura l'effetto della maggiore/minore presenza nella regione, di specializzazioni che nel complesso del paese hanno fatto registrare un incremento maggiore della componente tendenziale;

[2] I Sistemi locali del lavoro sono un'aggregazione, di due o più comuni contigui, definiti sulla base dell'autocontenimento dei flussi di pendolarismo giornaliero tra luogo di residenza e luogo di lavoro e rilevati in occasione del Censimento. Si tratta di un concetto geografico che denota un territorio dove offerta di lavoro e domanda di lavoro si incontrano. Nella presente applicazione sono stati considerati i sistemi locali del Censimento 1991.

- *componente di macroarea o di vicinato*: misura i differenziali di crescita, tra il livello della macroarea a cui la regione appartiene (in questo caso le regioni confinanti, più la regione stessa) ed il livello nazionale, per gruppi di specializzazione produttiva;
- *componente locale*: misura i differenziali di crescita, tra il livello regionale ed il livello della macroarea di riferimento, per gruppi di specializzazione produttiva.

In particolare, si considera la seguente espressione:

$$g_{r.} = g_{..} + \sum_i (g_{.i} - g_{..}) \frac{x_{ri}}{x_{r.}} + \sum_i (\ddot{g}_{ri} - g_{.i}) \frac{x_{ri}}{x_{r.}} + \sum_i (g_{ri} - \ddot{g}_{ri}) \frac{x_{ri}}{x_{r.}}$$

dove:

r = indice di regione;

i = indice di specializzazione produttiva;

x = valore assoluto del valore aggiunto nel 2002;

g = variazione relativa, nel periodo 1996-2002, del valore aggiunto;

quindi, le componenti sono:

$g_{r.}$ = variazione relativa della regione r-esima, tra il tempo (t-1) ed t;

$g_{..}$ = componente tendenziale;

$\sum_i (g_{.i} - g_{..}) \frac{x_{ri}}{x_{r.}} =$ componente strutturale della regione r-esima;

$\sum_i (\ddot{g}_{ri} - g_{.i}) \frac{x_{ri}}{x_{r.}} =$ componente di macroarea della regione r-esima.

$\sum_i (g_{ri} - \ddot{g}_{ri}) \frac{x_{rh}}{x_{r.}} =$ componente locale della regione r-esima.

Nel periodo 1996-2002, il valore aggiunto è aumentato, a livello nazionale, del 27,3% (componente tendenziale). La Regione che ha fatto registrare la performance migliore è la Campania, con un incremento di valore aggiunto del 33,4%, seguita dalla Calabria con un incremento del 30,4%. Per contro, le regioni che hanno fatto registrare gli incrementi minori sono la Valle D'Aosta (17,6%) ed il Piemonte (23,6%).

Nella Tavola 1 sono riportati i risultati dell'analisi shift-share con dipendenza spaziale (il metodo è quello del Paragrafo 4), effettuata sul valore aggiunto delle regioni italiane dal 1996 al 2002, articolata in 11 gruppi di specializzazione produttiva.

Come si può vedere, la Liguria, il Lazio e la Campania sono le Regioni che, più delle altre, risentono favorevolmente di specializzazioni maggiormente in crescita (valori della componente strutturale: +1,6% nel Lazio, + 1,3% in Liguria e +1,2% in Campania). Sia in Liguria, sia nel Lazio e sia in Campania il risultato dipende quasi esclusivamente dalla concentrazione del valore aggiunto nei Sll urbani (Tavola 3), che nel periodo considerato hanno fatto registrare un incremento superiore alla media nazionale (Tavola 2) .

Per contro, la struttura delle specializzazioni produttive ha penalizzato in modo particolare il Piemonte (-1,1%). Tale risultato è causato dalla concentrazione del valore aggiunto, circa il 70%, in specializzazioni che sono cresciute meno della componente

tendenziale (il 26,7% del valore aggiunto è prodotto dai sistemi del "made in italy" e il 43,8% dai sistemi dei "mezzi di trasporto" – Tavole 2 e 3. Altre regioni che presentano specializzazioni produttive che hanno influenzato negativamente le performance di crescita sono l'Umbria e il Molise. In entrambe, si è avuto l'effetto negativo della forte concentrazione del valore aggiunto nel settore del "made in italy" . Per contro, in altre regioni, quali la Lombardia, il Veneto ed il Friuli Venezia Giulia, la forte concentrazione della produzione nel "made in italy" è compensata da una presenza significativa di sistemi locali "urbani", che hanno fatto registrare nel periodo preso in esame incrementi superiori alla componente tendenziale (Tavola 2), portando così ad un valore della componente strutturale trascurabile.

Nel complesso, tuttavia, la componente strutturale presenta un peso poco rilevante, a dimostrazione che la composizione della specializzazione produttiva influenza solo in parte lo sviluppo economico regionale, che invece risulta maggiormente caratterizzato da fattori legati alla competitività interna.

A dimostrazione di quanto detto, è possibile osservare i valori della quarta e della quinta colonna della Tavola 1, rispettivamente la componente di macroarea e la componente locale, che presentano valori mediamente più elevati della componente strutturale.

Per quanto riguarda la componente di macroarea, emerge una netta dicotomia tra Nord e Centro-Sud: sempre negativa nelle regioni settentrionali e sempre positiva nelle regioni del Centro-Sud.

Tale andamento è il frutto del diverso ritmo di crescita delle aree del nostro Paese nel periodo considerato. Infatti le regioni del Centro-Sud hanno fatto registrare mediamente, ritmi di crescita superiori a quelli delle regioni del Nord[3]. La componente di macroarea penalizza in modo particolare la Valle d'Aosta (-8,8%) che, confinando solamente con il Piemonte risente della debolezza di entrambe le regioni. Per contro, le regioni che risentono di un effetto di macroarea particolarmente positivo sono la Basilicata e il Molise. I vantaggi di competitività delle macroaree di entrambe le regioni si riscontrano su quasi tutte le specializzazioni produttive, anche se per il tipo di distribuzione produttiva, risultano determinanti le performance dei "*sistemi senza specializzazione*" e dei "*sistemi urbani*", ambiti nei quali si concentra più del 70% del valore aggiunto di entrambe le regioni.

La quarta componente è sicuramente la più importante, in quanto consente di quantificare il contributo in termini di competitività locale, al netto dell'effetto del vicinato. Nel Nord i vantaggi dovuti alla competitività interna si riscontrano soprattutto in due regioni: Trentino Alto Adige (2,1%) e Friuli Venezia Giulia (1,8%). Nel Trentino, il guadagno di competitività si ha soprattutto in tre settori: turismo, "made in italy" e lavorazione del cuoio e della pelletteria. Mentre, nel Friuli la competitività interna si manifesta in modo significativo nei sevizi a carattere urbano e nel turismo. Sempre nel Nord, si ha la difficile situazione di Piemonte e Valle d'Aosta. In tutte e due le regioni si registrano valori negativi, sia della componente di macroarea sia della componente locale. A dimostrazione che agisce in modo negativo sia il contesto di macroarea sia i fattori riconducibili alla competitività interna.

Per quanto riguarda le regioni del Centro-Sud, è interessante il caso della Campania. Questa regione, come si può vedere dalla Tavola 1, pur godendo di un contesto di

[3] La crescita superiore nel Sud rispetto al Nord è durata fino al 2003. "Nel 2004, per la prima volta dopo diversi ani, l'economia meridionale ha fatto segnare un tasso di crescita inferiore a quello del Centro-Nord" (Rapporto Svimez, anno 2005).

macroarea favorevole (1,9%), fa registrare un ulteriore vantaggio dovuto al contesto locale (3,1%), a dimostrazione che gli elementi di competitività interna giocano un ruolo determinante nella crescita di questa regione. I settori a maggiore crescita risultano quello dei servizi urbani, del turismo, della manifattura specializzata nei materiali da costruzione.

Per contro, situazioni di specificità locali penalizzano in modo particolare due regioni del Sud, Molise (-3,5%) e Basilicata (-5,3%) . Nel Molise, la perdita di competitività è pressoché distribuita equamente in tutti i settori. Invece in Basilicata le situazioni più difficili si riscontrano soprattutto in due settori: la manifattura specializzata nei materiali da costruzione e la fabbricazione dei mezzi di trasporto.

6. Conclusioni

L'analisi delle componenti strutturali e locali, nota in letteratura come analisi *shift-share*, è stata oggetto di un interesse più applicativo che metodologico. Il motivo è riconducibile alla semplicità ed immediatezza, sia nell'applicazione sia nell'interpretazione. Inoltre, lo scarso interesse metodologico è proprio dovuto alla natura stessa di tale tecnica, essenzialmente un metodo di scomposizione di una variazione, senza particolari implicazioni di carattere statistico.

In tal senso, l'innovazione di S. Nazara e G.J.D. Hewings ha creato nuovo interesse, non solo a fini applicativi, ma anche metodologici. La possibilità di considerare una componente strutturata spazialmente è senza dubbio un elemento che arricchisce in modo profondo le caratteristiche dell'analisi *shift-share*, e apre nuovi spazi di analisi, anche e soprattutto in considerazione delle diverse ipotesi di dipendenza a livello territoriale.

Nel presente lavoro è stato proposto un metodo che si presenta come una via di mezzo tra il metodo classico e quello proposto da Nazara e Hewings. Tale scelta dipende dal fatto che la proposta di Nazara e Hewings implementa la componente spaziale in tutte le componenti del metodo classico, mentre chi scrive ritiene più idoneo considerare la componente spaziale solo nel caso della componente relativa alla competitività interna, in quanto questa misura l'effetto locale, che in questo modo viene scomposto in un effetto di vicinato e un effetto specifico. L'introduzione della dipendenza spaziale, anche nella componente strutturale, sembrerebbe avere più una giustificazione algebrica che teorica, in quanto gli elementi del mix settoriale hanno poco a che vedere con gli elementi dello spazio.

L'applicazione, effettuata con il metodo proposto in questo lavoro ha messo in evidenza importanti risultati dal punto di vista delle "nuove" componenti, pur conservando la principale caratteristica dell'analisi shift-share di essere facilmente interpretabile. Si è potuto osservare, come alcune regioni del Nord abbiano un vantaggio di competitività strettamente locale, oppure come le difficoltà di Piemonte e Valle d'Aosta sia non solo di contesto d'area, ma anche delle stesse regioni. Qualora l'analisi fosse stata condotta seguendo l'impostazione classica alcune importanti risultanze non sarebbero emerse. Ad esempio, seguendo l'approccio classico, la Liguria sarebbe risultata una regione con componente locale negativa (Tavola 1). Invece, applicando il nuovo metodo, è emerso che una parte significativa dell'effetto negativo è da attribuire alla componente di macroarea, mentre l'apporto specifico della regione è addirittura positivo. Un caso analogo è la Lombardia, che in base al metodo classico avrebbe fatto registrare una componente locale negativa, mentre con il nuovo metodo, si ha un valore della componete locale positivo, il che significa che in un contesto d'area sfavorevole (-0,7%), si registra un vantaggio, seppur minimo (0,3%), dovuto alle specificità interne.

Questi risultati forniscono interessanti informazioni derivanti dalle componenti che tengono conto della struttura spaziale delle unità territoriali oggetto di studio. Così, è possibile quantificare il contributo in termini di competitività locale, scorporando l'influenza delle aree circostanti.

In conclusione, si può ritenere che il metodo con struttura spaziale rappresenti un miglioramento rispetto al modello tradizionale, in quanto permette di considerare l'interdipendenza tra le aree. In tal senso gli spazi per ulteriori approfondimenti si hanno in modo particolare in tale ambito e cioè sulla possibilità di considerare diverse ipotesi di dipendenza spaziale, a partire dalla matrice che identifica il vicinato.

Tavola 1 – Componenti della variazione percentuale del valore aggiunto nelle regioni italiane nel periodo 1996-2002

Regioni	Tendenziale	Strutturale	Macroarea	Locale	Totale
Piemonte	27,3	-1,1	-1,5	-1,1	23,6
Valle D'Aosta	27,3	0,6	-7,7	-2,4	17,6
Lombardia	27,3	-0,3	-2,2	1,8	26,5
Trentino - Alto Adige	27,3	-0,3	-2,2	3,9	28,7
Veneto	27,3	-0,3	-0,1	-1,4	25,5
Friuli - Venezia Giulia	27,3	-0,7	-1,4	2,2	27,4
Liguria	27,3	1,3	-5,2	2,9	26,2
Emilia - Romagna	27,3	-0,6	-1,5	1,2	26,3
Toscana	27,3	-0,1	0,7	0,2	28,0
Umbria	27,3	-0,8	-0,4	1,9	27,9
Marche	27,3	-0,4	2,4	-1,4	27,9
Lazio	27,3	1,6	0,5	-1,0	28,4
Abruzzo	27,3	-0,3	0,5	-1,3	26,2
Molise	27,3	-0,8	2,7	-3,7	25,5
Campania	27,3	1,2	-1,1	6,1	33,4
Puglia	27,3	-0,1	1,9	-0,3	28,8
Basilicata	27,3	0,1	4,8	-6,3	25,9
Calabria	27,3	0,4	-0,7	3,4	30,4
Sicilia	27,3	0,9	5,1	-5,9	27,4
Sardegna	27,3	-0,6	0,0	2,8	29,5

Tavola 2 - Variazione del valore aggiunto per tipologia di specializzazione produttiva, nel periodo 1996-2002

Gruppi di specializzazione produttiva	Variazione percentuale del valore aggiunto
Sistemi senza specializzazione	27,2
Sistemi urbani	29,2
Sistemi estrattivi	1,2
Sistemi turistici	26,8
Sistemi del "Made in Italy"	25,1
Sistemi del Tessile	22,3
Sistemi del cuoio e della pelletteria	28,5
Sistemi dell'Occhialeria	27,0
Sistemi dei materiali da costruzione	26,0
Sistemi dei mezzi di trasporto	27,1
Sistemi degli apparecchi radiotelevisivi	31,4

Tavola 3 - Distribuzione del valore aggiunto regionale per specializzazione produttiva - Anno 2002 (valori percentuali)

Regioni	Sistemi senza specializzazione	Sistemi urbani	Sistemi estrattivi	Sistemi turistici	Sistemi del "made in Italy"
Piemonte	4,9	3,0	0,0	0,3	26,7
Valle D'Aosta	63,4	0,0	0,0	20,1	0,0
Lombardia	2,1	44,2	0,0	0,4	44,6
Trentino - Alto Adige	58,2	0,0	0,0	25,2	10,0
Veneto	5,7	26,8	0,0	0,7	43,5
Friuli - Venezia Giulia	0,0	30,9	0,0	4,8	53,5
Liguria	18,8	71,9	0,0	7,0	0,0
Emilia - Romagna	6,5	20,0	0,0	7,1	16,6
Toscana	11,7	11,1	0,0	1,9	14,9
Umbria	47,7	8,5	0,0	0,0	37,5
Marche	4,9	13,8	0,0	0,0	46,9
Lazio	9,5	76,3	0,0	0,2	0,9
Abruzzo	25,3	0,0	0,0	1,0	36,6
Molise	45,3	0,0	0,0	0,0	31,1
Campania	21,8	44,5	0,0	4,1	5,8
Puglia	55,1	16,5	0,0	0,4	21,8
Basilicata	61,5	17,1	0,0	0,0	4,6
Calabria	70,1	26,1	0,0	1,0	2,1
Sicilia	44,3	50,9	0,0	1,5	1,0
Sardegna	27,8	54,0	5,9	6,9	2,4

Regioni	Sistemi del cuoio e della pelletteria	Sistemi dell'occhialeria	Sistemi dei materiali da costruzione	Sistemi dei mezzi di trasporto	Sistemi degli apparecchi radiotelevisivi
Piemonte	0,0	0,0	16,2	43,8	0,0
Valle D'Aosta	0,0	0,0	0,0	0,0	16,5
Lombardia	0,1	0,0	7,4	0,0	0,0
Trentino - Alto Adige	6,6	0,0	0,0	0,0	0,0
Veneto	16,7	3,5	3,1	0,0	0,0
Friuli - Venezia Giulia	3,2	0,0	7,7	0,0	0,0
Liguria	0,0	0,0	2,4	0,0	0,0
Emilia - Romagna	3,8	0,0	45,9	0,0	0,0
Toscana	41,6	0,0	11,2	0,0	0,0
Umbria	0,0	0,0	6,3	0,0	0,0
Marche	33,8	0,0	0,6	0,0	0,0
Lazio	0,0	0,0	5,1	2,0	6,0
Abruzzo	0,0	0,0	10,6	9,8	16,7
Molise	0,0	0,0	0,0	23,6	0,0
Campania	3,6	0,0	2,8	5,7	11,6
Puglia	6,2	0,0	0,0	0,0	0,0
Basilicata	0,0	0,0	6,0	10,8	0,0
Calabria	0,0	0,0	0,0	0,0	0,7
Sicilia	0,1	0,0	0,9	1,3	0,0
Sardegna	1,0	0,0	2,0	0,0	0,0

Tavola 4 – Classificazione dei SIl, per tipologia di specializzazione produttiva

GRUPPI DI SPECIALIZZAZIONE	DESCRIZIONE
Sistemi senza specializzazione	Sistemi locali che risultano privi di fattori di localizzazione specifici e che non sono stati investiti da processi significativi di sviluppo
Sistemi urbani	Sistemi locali caratterizzati dalla specializzazione nelle attività di trasporto e di servizio che definiscono le funzioni di rango urbano superiore e raggruppa prevalentemente SIl delle città più grandi
Sistemi estrattivi	Sistemi locali di piccole dimensioni, fortemente specializzati in alcune attività estrattive
Sistemi turistici	Sistemi locali caratterizzati dalla concentrazione di addetti alle attività di ricettività e ristorazione (alberghi, campeggi, ristoranti, bar)
Sistemi del "made in Italy"	Sistemi della manifattura leggera, si caratterizza per la concentrazione degli addetti nelle attività legate alla fabbricazione di prodotti in metallo, il mobilio, l'abbigliamento, il cuoio, le calzature, prodotti alimentari
Sistemi del tessile	Sistemi della manifattura leggera, caratterizzati dalla specializzazione esclusiva nel settore tessile
Sistemi del cuoio e della pelletteria	Sistemi della manifattura leggera, fortemente specializzati nelle industrie conciarie, che riconprendono al loro interno la preparazione e la concia del cuoio; la fabbricazione di articoli da viaggio, borse, cinture e sellerie; e le calzature
Sistemi dell'occhialeria	Sistemi della manifattura leggera, presentano caratteristiche uniche per grado di concentrazione. Sono cinque sistemi, tutti della Provincia di Belluno, specializzati nella fabbricazione di apparecchi medicali, apparecchi di precisione, strumenti ottici e orologi, occhiali e montature
Sistemi dei materiali da costruzione	Sistemi locali manifatturieri, specializzati nei materiali da costruzione
Sistemi dei mezzi di trasporto	Sistemi locali manifatturieri, specializzati nella fabbricazione dei mezzi di trasporto. I sistemi locali che ne fanno parte sono soltanto 13, ma in essi si concentra la produzione nazionale dei mezzi di trasporto
Sistemi degli apparecchi radiotelevisivi	Sistemi locali manifatturieri, specializzati nella fabbricazione di apparecchiature televisive. I sistemi che appartengono a questo gruppo sono soltanto 9.

Riferimenti bibliografici

Barbieri G., Pellegrini G. (2000) *I sistemi locali del lavoro: uno strumento per la politica economica in Italia e in Europa*, Atti Convegno UVAL-DPS, Roma.

Becattini, G. (1996) *I sistemi locali nello sviluppo economico italiano e nella sua interpretazione*, Sviluppo locale, n. 2-3, 5-25.

Biffignardi S. (1993) *Aspetti metodologici ed interpretativi della tecnica shift-share*, CEDAM, Padova

Dunn E.S. (1960) A *statistical and analytical technique for regional analysis*, Paper and Proceedings of the Regional Science Association, n. 6, pp. 97 - 112

Esteban-Marquillas J.M. (2000) *Regional Convergence in Europe and the industry mix: a shift-share analysis*, Regional Science and Urban Economics, n. 30, pp. 253-364

Faramondi A., Piras M.G. (2003) "*Le nuove stime di aggregati socio-economici per i sistemi locali del lavoro*" in Sviluppo locale, n. 20 (2002), Rosenberg & Selier, Torino.

Istat (1997) *I sistemi locali del lavoro*:199, Argomenti, n.10.

Istat (1999) *Rapporto annuale: la situazione del paese nel 2000*, Roma: Istat.

Istat (2000) *Rapporto annuale: la situazione del paese nel 2001*, Roma: Istat.

Marbach G. (1998) *Statistica economica*, UTET, Torino.

Nazara S., Hewings G.J.D. (2003) *Towards regional growth decomposition with neighbor's effect: a new perspective on shift-share analysis*, Regional Economics Applications Laboratory, REAL 03-T-21.

Pascarella C. (2003) *Le stime territoriali nell'ambito dei Conti Nazionali: nuovi prodotti e sviluppi futuri*, Atti VI Conferenza Nazionale di Statistica, Roma

Sforzi, F. (1991) *I distretti industriali marshalliani nell'economia italiana, Distretti industriali e cooperazione fra imprese in Italia*, a cura di F.Pyke, G.Becattini e W. Sengernberger, Firenze, Banca Toscana, 91-117.

Sforzi, F. (1995) Sistemi *locali di impresa e cambiamento industriale in Italia*, Geotema, 2, 42-54.

Zaccomer, G.P. (2005) *La scomposizione della contrazione distrettuale: un'analisi shift-share con struttura spaziale sui dati del Registro delle Imprese*, nota di ricerca n. 4, Dipartimento di Scienze Statistiche dell'Università di Udine.

Zani, S. (a cura di) (1993*) Metodi statistici per le analisi territoriali*, Franco Angeli, Milano

Stima congiunturale dell'occupazione con l'utilizzo di fonti amministrative: metodologia, risultati e prospettive della Rilevazione Oros[1]

Ciro Baldi[2], Francesca Ceccato[3], Silvia Pacini[4], Donatella Tuzi[5]

Sommario

La rilevazione Oros produce, a titolo sperimentale, indicatori dell'occupazione dipendente. Tuttavia sempre più urgente diviene l'esigenza di diffondere dati ufficiali congiunturali sull'occupazione rilevata presso tutte le imprese. Il loro rilascio impone requisiti di qualità particolarmente stringenti delle stime provvisorie. A tal fine, è in corso un progetto per rivedere la metodologia di stima anticipata valutando metodi alternativi. Le analisi effettuate hanno consentito di scomporre l'errore di stima in due fattori: la sovracopertura del registro anagrafico utilizzato per la stima della popolazione corrente e il modello di riporto del campione alla popolazione di riferimento. Nel presente documento si illustrano i principali risultati delle sperimentazioni condotte. Gli studi effettuati hanno consentito di apportare notevoli progressi sul problema della sovracopertura. In riferimento alla natura non casuale del campione sono state individuate alcune aree critiche, su cui occorrerà effettuare ulteriori sperimentazioni.

Abstract

The Oros Survey produces data on the number of employees at the experimental stage. In a short time period this new index should be officially disseminated, demanding higher levels of data quality. In the last months the methodology currently employed to produce the preliminary estimates has been analyzed in depth. Various innovations have been designed and experimented. Two sources of causes have been isolated as possible responsible for the non-negligible error produced by the actual methodology: the over-coverage of the list for the estimation of figures referring to the current population and the non-randomness of the sample coupled with a growing but unstable sample size along time. In this paper the main results of several methodological experimentations are presented. The analysis have allowed considerable progress in terms of reduction of the over-coverage problem. As concerning the non-randomness of the sample, some critical areas of the estimation methodology have been detected and on this aspect further experimentations have been outlined.

[1] Il gruppo di ricerca costituito dagli autori è stato coordinato da Ciro Baldi nell'ambito delle attività del Servizio OCC. Gli autori ringraziano Fabio Massimo Rapiti, Leonello Tronti e Gian Paolo Oneto per i preziosi suggerimenti.

[2] Ricercatore (Istat), email: baldi@istat.it

[3] Collaboratore statistico (Istat), email: ceccato@istat.it

[4] Tecnologo (Istat), email: pacini@istat.it

[5] Ricercatore (Istat), email: tuzi@istat.it

1. Introduzione

L'Istat ha in programma la diffusione di indicatori dell'occupazione dipendente nell'ambito della rilevazione Oros basata sulle dichiarazioni contributive delle imprese all'INPS (modelli DM10). Il rilascio di questi indicatori ha reso necessaria una fase, ancora in corso, di verifiche ed approfondimenti della metodologia di stima con lo scopo di ridurre l'entità della revisione tra la stima preliminare e quella definitiva. Nel corso di questa fase, la base informativa disponibile in tempi rapidi è cambiata significativamente per effetto di modifiche normative, imponendo un considerevole ripensamento sull'intera metodologia, tuttora in via di miglioramento.

Questo lavoro documenta le attività di sperimentazione svolte prima di questo cambiamento della base dati. L'esperienza acquisita in quel contesto informativo rimane, tuttavia, un patrimonio per comprendere le potenzialità e i problemi che possono nascere dall'utilizzo di fonti amministrative per la produzione di statistiche congiunturali e per contribuire allo sviluppo di metodologie adatte ad affrontare tali problemi. Il principio sotteso alle sperimentazioni svolte è quello di conciliare i requisiti di qualità di indicatori congiunturali soggetti a revisione con il miglior sfruttamento possibile delle informazioni amministrative disponibili.

Nell'ambito delle indagini congiunturali, la rilevazione Oros ha caratteristiche di unicità legate alla ricchezza e alle peculiarità della base dati amministrativa. In primo luogo, con un certo ritardo si dispone dell'informazione sulla popolazione di riferimento, consentendo di rilasciare una stima finale non affetta da errori campionari. In secondo luogo e in conseguenza di quanto appena detto, la stima preliminare, dovendo predire nel modo più accurato possibile quella finale, deve porsi come popolazione obiettivo quella corrente del trimestre di riferimento. La rilevazione Oros, quindi, compie un passo in avanti rispetto alla gran parte delle indagini congiunturali "classiche" che invece, esplicitamente o implicitamente, rappresentano popolazioni fisse nel tempo e precedenti al periodo corrente. In altri termini, il tentativo di misurare variazioni dovute alle evoluzioni demografiche della popolazione corrente è probabilmente l'obiettivo più innovativo della rilevazione. In terzo luogo, proprio per cogliere queste evoluzioni, la rilevazione Oros usa come rappresentazione della popolazione corrente il registro anagrafico dell'INPS. Esso però è affetto da errori di copertura e il metodo di stima usato deve, quindi, contemplare la loro correzione. In quarto luogo, il campione di rispondenti disponibile in tempi rapidi, formato dalle dichiarazioni fornite su supporto elettronico, è autoselezionato e caratterizzato da un trend crescente nella sua dimensione. Anche in questo caso è compito della metodologia di stima provvedere a limitare distorsioni dovute alle caratteristiche del campione.

L'analisi svolta parte dallo studio dell'errore della stima preliminare nella metodologia esistente per configurare una serie di correttivi tesi a ridurne la dimensione. Il contributo principale di questo lavoro è una scomposizione dell'errore in una componente dovuta alla copertura della lista e una dovuta al riporto all'universo. Ciò ha permesso di individuare meglio le aree di intervento per i miglioramenti metodologici. Un altro merito dell'analisi consiste nell'aver misurato l'errore sia rispetto al livello che alle variazioni dell'occupazione. Da questa doppia chiave di lettura è emersa la difficoltà di proporre modifiche alla metodologia che assicurino una riduzione dell'errore in entrambe le direzioni. Infine, uno spunto aggiuntivo riguarda la discussione sulla natura dell'errore di riporto all'universo dovuto alla non ignorabilità del processo di selezione del campione e alla necessità di valutare accuratamente la capacità di metodologie che fanno uso di variabili ausiliarie nel ridurre tali errori.

Il lavoro è strutturato come segue. Nel paragrafo 2 vengono presentate le principali caratteristiche della rilevazione Oros e nel paragrafo 3 viene illustrata la metodologia di stima. Nel paragrafo 4 si definisce l'errore di stima preliminare, mentre nel paragrafo 5 ne viene fornita una misura generale sul totale delle imprese. Nel paragrafo 6 l'errore viene esaminato distinguendo nel dettaglio la componente dovuta alla definizione della lista di stima e quella causata dal modello di riporto alla popolazione di riferimento. Relativamente alle due fonti di errore, nei paragrafi 7 e 8 si propongono alcune innovazioni metodologiche volte a contenerne l'ampiezza e gli effetti distorsivi. Nel paragrafo 9, dai risultati delle sperimentazioni effettuate si traggono alcune conclusioni e si discutono alcune prospettive di avanzamento del progetto suggerite, per un verso, dal mutare delle caratteristiche del set informativo su cui si basa la rilevazione e, per l'altro, dalle necessità di riduzione dei tempi di diffusione degli indicatori, in accordo con gli obiettivi fissati dal nuovo regolamento comunitario sulle statistiche congiunturali.

2. Variabili, popolazione e fonti

La rilevazione Oros attualmente rilascia indicatori relativi a retribuzioni lorde, oneri sociali e costo del lavoro, per unità di lavoro equivalenti a tempo pieno (Ula), espressi in termini di indici e di variazioni (variazione congiunturale e variazione tendenziale). La popolazione obiettivo è rappresentata dalle imprese attive con almeno un dipendente, nei settori di attività economica dell'industria e dei servizi privati (vengono escluse la pubblica amministrazione e i servizi sociali e personali). La principale fonte di dati è costituita dalle denunce (modelli DM10) presentate mensilmente dalle imprese all'INPS per esporre i contributi di previdenza e assistenza sociale per i propri dipendenti. L'impresa poteva presentare il DM10 su supporto cartaceo o telematico fino all'inizio del 2004 quando, per modifiche normative, l'invio telematico è divenuto l'unico mezzo accettato[6]. Per la generalità delle imprese il termine massimo di presentazione del DM10 è fissato nell'ultimo giorno del mese successivo a quello cui si riferisce il periodo di paga[7]. L'unità di rilevazione amministrativa è la posizione contributiva, che può corrispondere ad un'impresa o a parte di essa.

Per ogni trimestre (t) vengono prodotte una stima preliminare e una stima definitiva relativa a t-5[8]. Le stime finali, quindi, vengono rilasciate dopo quindici mesi dalla diffusione delle stime preliminari.

Gli archivi da cui vengono tratte le informazioni per la rilevazione Oros sono tre: il registro anagrafico (o anagrafe), disponibile alla fine di ciascun trimestre di stima, da cui si perviene alla definizione della popolazione corrente (o lista di stima); il campione totale di DM10, a partire dal quale viene prodotta la stima preliminare, costituito dai modelli relativi ai mesi del trimestre di riferimento inviati dalle imprese per via telematica e, quindi,

[6] L'INPS, in seguito ad una disposizione, ha stabilito che banche e poste, presso cui le imprese effettuano il pagamento dei contributi, non accettino le dichiarazioni cartacee a partire dal secondo trimestre 2004, anticipando i tempi di legge che prevedevano l'invio dei DM10 per via telematica a partire dal mese di competenza di gennaio 2005.

[7] Più precisamente fino al 2004 i DM10 cartacei dovevano essere presentati entro il 16 del mese successivo a quello di paga, mentre per i DM10 telematici tale termine era posticipato all'ultimo giorno del mese successivo. Con le modifiche normative accennate il termine di consegna è diventato unico e uguale a quello previsto per i DM10 telematici.

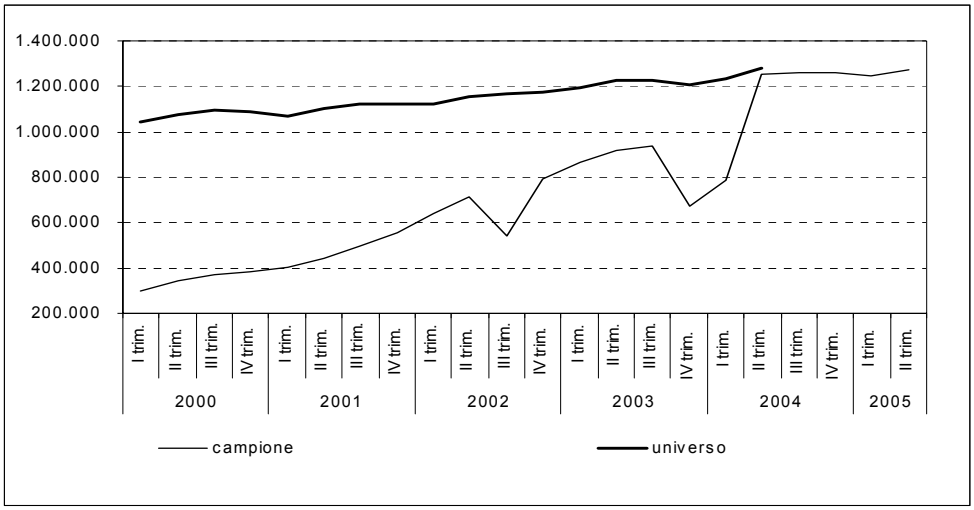
[8] In realtà, a t già si dispone dell'universo riferito a t-4. Come sarà chiaro in seguito, tuttavia, le informazioni relative a t-4 sono usate solo come variabili ausiliarie per la stima preliminare e come sussidio per l'imputazione delle mancate risposte. Tali informazioni vengono considerate definitive non prima di un ulteriore trimestre.

disponibili in tempi rapidi; infine, l'universo di DM10, che rappresenta la popolazione di riferimento, costituito dai moduli elettronici e cartacei pervenuti all'INPS, forniti all'Istat con dodici mesi di ritardo rispetto a quello di riferimento, su cui viene prodotta la stima definitiva.

I dati di fonte INPS vengono integrati con altre fonti quali la rilevazione mensile Istat su Lavoro e retribuzioni nelle grandi imprese (GI), limitatamente alle imprese con almeno 500 dipendenti[9] (Istat, 2005), e l'Archivio Statistico delle Imprese Attive (ASIA) che fornisce alcune informazioni strutturali (Abbate, Garofalo, 1997).

Per effetto di una progressiva diffusione dell'utilizzo della modalità di trasmissione elettronica dei modelli all'INPS, si è osservato nel tempo un notevole aumento della dimensione del campione.

Fig. 1 – Evoluzione temporale della dimensione e della copertura del campione INPS. Periodo I trimestre 2001 - II trimestre 2005 (numero di posizioni contributive*)*



Fonte: Elaborazioni su dati della rilevazione Oros

La crescita è avvenuta in maniera pressoché continua, con rari episodi di caduta dovuti a difficoltà organizzative e amministrative dell'INPS, fino all'improvviso *shift* a partire dal secondo trimestre del 2004, in seguito alle modifiche normative di cui sopra (Figura 1[10]).

La disponibilità recente di un campione che può essere considerato un "quasi-universo" ha cambiato sostanzialmente il quadro informativo della rilevazione inducendo delle modifiche alla metodologia di stima preliminare. Questo lavoro focalizza l'attenzione sul metodo di stima utilizzato fino al cambiamento informativo.

[9] I dati della Rilevazione GI sono usati a sostituzione ed integrazione dei dati INPS principalmente a causa della sottorappresentazione delle imprese di grandi dimensioni nel campione.

[10] I *gap* informativi che si notano nei trimestri III:2002, IV:2003 e I:2004 sono esemplificazioni dell'effetto di caduta della dimensione del campione dovuto a motivi amministrativi.

3. Stime finali e stime anticipate

Nelle procedure di stima la popolazione di Oros viene suddivisa in quattro diverse sottopopolazioni: le piccole e medie imprese (PMI); le grandi imprese che non rientrano nel dominio della rilevazione GI (GI-INPS); le grandi imprese che rientrano nel dominio della rilevazione GI (GI-RIL); le imprese che forniscono lavoro interinale (INTER). Poiché lo scopo di questo lavoro è quello di discutere i problemi di stima delle PMI, nel seguito ci concentreremo sulle procedure che riguardano questa sottopopolazione che occupa oltre il 75% dei dipendenti dei settori dell'industria e dei servizi privati.

La produzione delle stime finali, che si basa sull'universo delle dichiarazioni contributive, si riferisce ad un set informativo preliminarmente sottoposto ad imputazione delle mancate risposte totali, al fine di individuare e correggere i dati economici delle unità assenti per motivi non giustificati da eventi di demografia delle imprese (ovvero unità non stagionali, non cessate né sospese)[11]. Le stime definitive sono calcolate come somma, su tutte le unità, delle variabili di interesse (occupazione, retribuzioni e oneri sociali) provenienti dagli archivi INPS.

La stima preliminare delle PMI viene effettuata come somma ponderata sulle unità del campione, in cui i pesi derivano da una procedura di ponderazione vincolata, o calibrazione, della quale di seguito si fornisce una breve descrizione (Baldi, Falorsi, Pallara, Succi, Russo, 2001; Falorsi, Pallara, Russo, Succi, 2003).

Sia Y_t il parametro di interesse ossia il numero dei dipendenti al tempo *t* calcolabile teoricamente come:

$$Y_t = \sum_{i \in P_t} y_{it} \qquad (1)$$

dove y_{it} è il numero dei dipendenti relativo all'unità *i*-esima al tempo *t* e P_t è la popolazione teorica di riferimento al tempo *t*.

Nella metodologia di stima per ponderazione vincolata, la stima del parametro d'interesse $\hat{Y}_t$ è ottenuta come:

$$\hat{Y}_t = \sum_{i \in C_t} k_{it} y_{it} \qquad (2)$$

Nella [2] C_t è il campione del tempo *t*[12] e k_{it} è il peso della posizione *i*-esima al tempo

[11] L'imputazione nel set degli universi delle dichiarazioni contributive prevede la ricostruzione dei dati a livello micro, sfruttando prevalentemente informazioni longitudinali sulle unità da imputare. Nel corso del tempo, con l'aumento della tempestività nell'invio dei dati da parte delle imprese, la quota di DM10 da imputare si è andata via via riducendo, coinvolgendo circa lo 0,1%-0,2% delle dichiarazioni disponibili (5.000 - 6.000 DM10). L'imputazione ha un impatto poco rilevante sul livello medio di oneri e retribuzioni, mentre implica una revisione del livello dell'occupazione di circa 1 punto percentuale.

[12] Il campione a cui si fa riferimento nelle stime non è l'intero campione pervenuto dall'INPS (quello denominato campione totale o CT), ma un suo sottoinsieme (che verrà di seguito denominato campione ridotto o C). La selezione delle

t, calcolato come soluzione del seguente problema di minimo vincolato:

$$\begin{cases} \underset{\{k_{it}\}}{Min}\left[\sum_{i\in C_t} c_{it}\left(k_{it}-1\right)^2\right] \\ \sum_{i\in C_t} k_{it}\mathbf{x}_{it} = \mathbf{X}_t \end{cases} \qquad (3)$$

Nella funzione obiettivo, c_{it} è una costante che misura la dimensione della posizione e il valore 1 è il peso diretto attribuito alle unità del campione, per l'assenza di un disegno campionario. Il vincolo è un sistema di equazioni dove $\mathbf{x}_{it}$ è il vettore colonna delle variabili ausiliarie riferite alle unità del campione e $\mathbf{X}_t$ è il vettore dei totali di calibrazione relativi alle variabili ausiliarie sulla lista di stima della popolazione corrente. Su questi aspetti si tornerà nei paragrafi 4 e 7.

La disponibilità di variabili ausiliarie è diversa a seconda dei sottoinsiemi di unità della lista. In particolare, si distinguono tre gruppi di unità: le posizioni con età superiore o uguale ad un anno con informazioni economiche nell'universo di *t*-4 (unità panel con informazione ausiliaria), le posizioni con età superiore o uguale ad un anno senza informazioni economiche nell'universo di *t*-4 (unità panel senza informazione ausiliaria), le posizioni con età inferiore ad un anno (unità nuove nate). Per le unità panel con informazione ausiliaria, la disponibilità di informazioni è massima: per esse, si usano il numero di posizioni contributive secondo la lista della popolazione corrente a *t*, il numero di dipendenti a *t*-4, il monte retributivo a *t*-4, il monte oneri a *t*-4. Per le unità panel senza informazione ausiliaria, l'unica variabile disponibile è il numero di posizioni secondo la lista della popolazione corrente. Per le nuove nate le variabili ausiliarie sono: il numero di posizioni secondo la lista e il numero di dipendenti dichiarati all'iscrizione.

Il problema di minimo vincolato espresso nella (3) è risolto nell'ambito di gruppi omogenei di unità (*model groups*)[13], definiti sulla base delle informazioni relative ad attività economica, localizzazione geografica e dimensione (quest'ultima solo sulle panel con variabili ausiliarie).

I totali di calibrazione sono calcolati come somma delle variabili ausiliarie sulla lista della popolazione corrente stimata. Tuttavia, una somma semplice porterebbe ad una sovrastima dei totali di calibrazione, a causa del problema della sovracopertura della lista, derivante dal fenomeno delle cessazioni non registrate, di cui si parlerà diffusamente nei paragrafi 6 e 7. I totali di calibrazione vengono, dunque, corretti per tenere conto del fatto che non tutte le posizioni della lista sono effettivamente attive. Il metodo di correzione prevede l'applicazione di una probabilità di essere attiva ad ogni unità della lista di stima del trimestre *t*.

unità che confluiranno alle stime si rende necessaria per attenuare gli effetti di distorsione causati dalla presenza di mancate risposte nel campione di origine. Questo concetto verrà ampiamente argomentato nel paragrafo 8 dedicato all'analisi del modello di riporto all'universo.

[13] Attualmente i *model groups* sono 546.

$$\boldsymbol{X}_t = \sum_{i \in A_t} p_{it} \boldsymbol{x}_{it} \tag{4}$$

Dove A_t è la lista di stima che è la migliore rappresentazione della popolazione di riferimento al tempo *t* e p_{it} è la probabilità della unità *i*-esima di essere attiva al tempo *t*.

Considerato che la popolazione teorica di riferimento è compresa per approssimazione nella lista di stima ($P_t \subseteq A_t$) (cfr. figura 2, paragrafo 7), si vuole stimare la probabilità che ciascuna unità *i*-esima presente nella lista sia effettivamente attiva al tempo *t*:

$$p_{it} = pr(i \in P_t / i \in A_t) \tag{5}$$

Tale probabilità è calcolata per gruppi omogenei di unità (*register error groups* che non coincidono con i *model groups* definiti poco sopra)[14], ottenuti stratificando le posizioni contributive rispetto alle variabili età, ripartizione geografica, classe dimensionale, classificazione di attività economica e presenza delle variabili ausiliarie in *t*-4.

Nel metodo di stima attuale, la probabilità di essere attive è calcolata su tutte le unità del trimestre *t*-4 come rapporto tra il numero di unità appartenenti alla popolazione di riferimento di *t*-4, in simboli $n(P_{t-4})$, e il numero di unità definite attive dalla corrispondente lista stimata, $n(A_{t-4})$. Assumendo l'ipotesi di invarianza temporale tra *t*-4 e *t,* la probabilità viene applicata, nel trimestre corrente *t*, solo alle unità che non appartengono al campione C_t. Alle unità campionarie, invece, viene attribuita ex-post la probabilità pari ad uno in quanto, avendo presentato almeno un DM10 nel trimestre, sono ritenute certamente attive in *t* (tale metodo viene abbreviato con la sigla M0). In formule:

$$\begin{aligned} &\hat{p}_{it} = 1 && i \in C_t \\ &\hat{p}_{it} = \frac{n(P_{t-4})}{n(A_{t-4})} && i \notin C_t \end{aligned} \tag{6}$$

4. L'errore nelle stime preliminari: definizione

L'obiettivo di questo lavoro è quello di analizzare la differenza tra la stima preliminare e la stima finale e descrivere le sperimentazioni condotte per ridurre tale discrepanza. In questo paragrafo si discute dell'importanza e della natura concettuale di questa differenza e si presentano le misure sintetiche che verranno usate nella parte empirica.

L'errore di stima preliminare, derivante dalla revisione delle stime, è un parametro fondamentale nella valutazione dell'affidabilità di una metodologia di stima anticipata. La sua importanza nelle indagini congiunturali nasce dalla necessità di non esporre gli utenti dei dati a revisioni sostanziali che minerebbero l'utilità delle stime preliminari stesse.

[14] Nella metodologia di base, il numero dei register error groups è pari a circa 380.

Per comprendere meglio la natura dell'errore nell'ambito della rilevazione Oros è utile provare a confrontare le caratteristiche delle stime Oros con quelle di indagini più classiche. L'errore di cui trattiamo è per natura simile a quello cui va incontro ogni rilevazione congiunturale che presenta una stima preliminare, basata su un set informativo ridotto di rispondenti rapidi, e una stima rivista, che sfrutta tutta l'informazione disponibile. In questo contesto l'errore dipende sostanzialmente dalla quantità e dalla qualità dell'informazione usata nelle due stime. Vanno però sottolineate almeno due differenze tra le stime Oros e quelle di indagini campionarie classiche. La prima deriva dalla diversità dell'informazione disponibile per la stima definitiva che, per un'indagine classica si basa su un campione di rispondenti totali (che in assenza di mancate risposte coincide col campione teorico), mentre per la rilevazione Oros è rappresentata dall'universo dei dati effettivi. La seconda, non meno importante, è legata alla conoscenza della lista dei rispondenti totali attesi: mentre in un'indagine congiunturale classica questa è nota nel momento in cui si rilascia la stima preliminare, nella rilevazione Oros la lista dei rispondenti totali, che coincide con la popolazione corrente (lista di stima), non è nota nel trimestre di stima preliminare, in cui si dispone solo di una sua rappresentazione fornita dal registro anagrafico dell'INPS e affetta da errori di copertura. In altri termini, in Oros si deve aggiungere lo specifico errore di stima della lista di unità attive all'errore, cui va incontro anche un'indagine congiunturale classica, dovuto alla stima dei valori di interesse sulla base di un sottoinsieme di rispondenti. Come si vedrà in seguito, queste due fonti di errore (l'errore di lista e l'errore di riporto alla popolazione di riferimento) contribuiscono in maniera molto simile all'errore totale.

L'attenzione verrà focalizzata sugli errori nei livelli e nelle variazioni tendenziali. In particolare, l'errore nei livelli di uno specifico trimestre t è definito come:

$$e_t = \frac{\hat{y}_t^a - \hat{y}_t^f}{\hat{y}_t^f} * 100 \qquad (7)$$

dove $\hat{y}_t^a$ è il livello della stima anticipata (denotata dall'apice a) dell'occupazione dipendente al tempo t, mentre $\hat{y}_t^f$ è il livello della stima finale (denotata dall'apice f) al tempo t.

L'errore della variazione tendenziale è, invece, espresso come:

$$_{vt}e_t = {}_{vt}\hat{y}_t^a - {}_{vt}\hat{y}_t^f \qquad (8)$$

dove:

$$_{vt}\hat{y}_t^a = \frac{\hat{y}_t^a - \hat{y}_{t-4}^a}{\hat{y}_{t-4}^a} * 100 \qquad (9)$$

è la variazione tendenziale calcolata sulle stime preliminari e

$$_{vt}\hat{y}_t^f = \frac{\hat{y}_t^f - \hat{y}_{t-4}^f}{\hat{y}_{t-4}^f} * 100 \qquad (10)$$

è la variazione tendenziale calcolata sulle stime finali.

Naturalmente l'errore nelle variazioni e quello nei livelli sono strettamente connessi: si può mostrare che l'errore nella variazione tendenziale dipende dalla differenza dell'errore nei livelli a *t* e a *t*-4 e dall'entità della variazione tendenziale finale[15]. Ne consegue che, se gli errori nei livelli a *t* e *t*-4 fossero uguali, l'errore nella variazione risulterebbe nullo.

Come misure di sintesi dell'errore di stima preliminare su più periodi temporali si usano il MAPE (*Mean Absolute Percentage Error*), espresso da:

$$MAPE = \frac{\sum_{t=1}^{T} |e_t|}{T} \qquad (11)$$

e l'MPE (*Mean Percentage Error*), definito come[16]:

$$MPE = \frac{\sum_{t=1}^{T} e_t}{T} \qquad (12)$$

5. L'errore nelle stime preliminari: una quantificazione sul totale delle imprese

Prima di affrontare i problemi di stima delle PMI, in questo paragrafo si fornisce una quantificazione dell'errore medio, calcolato sul totale delle unità, che si commette nel fornire stime preliminari.

Le analisi che hanno condotto a tale misurazione sono state effettuate sui sette trimestri che coprono il periodo dal II trimestre 2001 al IV trimestre 2002 con riferimento al quale, nel momento in cui sono state condotte le sperimentazioni, erano disponibili sia stime anticipate che stime definitive[17].

Nel periodo considerato, l'errore totale di stima nei livelli è pari, in media, al 2,1% ed è sistematicamente positivo (tabella 1). La sua stabilità nel tempo, tuttavia, implica errori sui tassi di variazione molto più contenuti: il MAPE è pari allo 0,4%, mentre l'MPE è quasi

[15] Si ringrazia Leonello Tronti per avere derivato algebricamente questa relazione.

[16] Si ricordano alcune caratteristiche dei due indicatori, che possono facilitarne la lettura. Le due misure sono complementari in quanto, mentre MPE misura la media tout court, MAPE dà una misura al lordo delle compensazioni tra valori positivi e negativi. Considerati insieme i due indicatori mostrano la presenza di errori sistematici: MPE e MAPE uguali stanno ad indicare che si commettono errori sistematicamente positivi; uguali, ma di segno opposto, evidenziano errori sistematicamente negativi.

[17] Le sperimentazioni, di cui si presentano alcuni risultati in questo documento, rientrano in un progetto di revisione metodologica ancora in corso. L'estensione delle analisi a un intervallo temporale più lungo è un aspetto certamente importante per confermare la robustezza dei risultati ottenuti sul quale si sta lavorando.

nullo. Per definizione, l'errore è nullo per le grandi imprese della rilevazione GI e, dato che tale sottopopolazione conta per oltre il 20,3% dell'occupazione totale di Oros, l'assenza di errore contribuisce sensibilmente al contenimento dell'errore complessivo di stima. La sottopopolazione con il maggiore errore in valore assoluto è quella delle imprese interinali (INTER), relativamente alle quali, tuttavia, l'introduzione di alcune innovazioni metodologiche tra il 2001 e il 2002 ha comportato un notevole contenimento nel tempo dell'errore18. Anche sulle grandi imprese rilevate soltanto dall'INPS (GI-INPS) si riscontra un errore molto elevato; per queste, è in fase di valutazione il disegno di una metodologia di stima più specifica, al pari delle interinali. La sottopopolazione più problematica, comunque, rimane quella delle PMI, sia per l'entità dell'errore sia per la quota di occupazione che essa rappresenta, in media pari al 76% del totale. È su questa sottopopolazione, quindi, che si è focalizzata l'attenzione e ad essa è dedicato il resto dell'analisi.

Tab. 1 – Errore totale di stima per sottopopolazione in termini di livelli, variazioni congiunturali e tendenziali. Periodo II trimestre 2001 – IV trimestre 2002 (valori percentuali medi di periodo)

Sottopopolazione di stima	Occupazione sul totale	Livelli		Variazioni Tendenziali	
		MPE	MAPE	MPE	MAPE
PMI	76,0	3,1	3,1	-0,6	0,6
GI-INPS	2,2	-2,7	3,8	7,7	7,7
INTER	1,5	-17,7	17,7	50,4	50,4
GI	20,3	0,0	0,0	0,0	0,0
TOTALE	100,0	2,1	2,1	0,2	0,4

Fonte: Elaborazioni su dati della rilevazione Oros

6. Errore nelle stime preliminari delle PMI: scomposizione in errore di lista e errore di riporto

L'impianto della metodologia di stima delle PMI consente di ricondurre l'errore a due possibili fonti. La prima è legata ad alcuni aspetti critici di cui soffre il registro anagrafico utilizzato per individuare la lista di stima; la seconda fonte invece, tipica delle indagini campionarie, dipende dal metodo applicato per il riporto alla popolazione di riferimento.

Le basi dati a disposizione consentono ex-post di sovrapporre i set informativi con cui si ottengono la stima preliminare e quella finale; in riferimento allo stesso istante temporale si ha, quindi, sia la lista stimata della popolazione di riferimento, sia quella effettiva. In questo contesto di sperimentazione, l'errore totale nei livelli dell'occupazione nella [7], pertanto, può essere scomposto in due componenti:

[18] L'adozione di procedure ad hoc di imputazione nei valori economici relativi alle unità interinali consentono di presupporre che gli errori nei trimestri del 2003 e del 2004 si siano ridotti ad un livello accettabile.

$$e_t = \left(\frac{\hat{y}_t^a - \hat{y}_t^{a_{LC}}}{\hat{y}_t^f} + \frac{\hat{y}_t^{a_{LC}} - \hat{y}_t^f}{\hat{y}_t^f} \right) * 100 \qquad (13)$$

Nella [13] $\hat{y}_t^{a_{LC}}$, sottratto ed aggiunto al numeratore nella [7], rappresenta il livello della stima anticipata dell'occupazione dipendente al tempo *t* nell'ipotesi in cui il modello di riporto potesse essere applicato alla lista delle posizioni realmente attive (lista certa o LC). Il primo addendo della scomposizione rappresenta, dunque, l'errore di lista in quanto misura lo scostamento tra stima anticipata e stima finale dovuto unicamente alla mancanza a *t* di una lista certa. Il secondo addendo, invece, rappresenta l'*errore di riporto* misurato come scostamento tra stima anticipata e finale nell'ipotesi in cui non vi fosse errore nella definizione della lista.

Nei sette trimestri della sperimentazione si stima che, se fosse possibile eliminare la componente dovuta all'errore di lista, resterebbe una sovrastima media dell'occupazione totale dell'1,7% tutta imputabile al metodo di riporto alla popolazione di riferimento (tabella 2).

Tab. 2 - Errore, totale e al netto degli errori di lista anagrafica, nella stima dell'occupazione delle PMI. Periodo II trimestre 2001 – IV trimestre 2002 (valori percentuali)

Sezione di attività Economica	*Errore totale*		*Errore al netto degli errori di lista anagrafica*	
	MPE	*MAPE*	*MPE*	*MAPE*
C Estrazione di minerali	3,1	3,1	2,9	2,9
D Attività manifatturiere	2,1	2,1	1,2	1,2
E Produzione di energia elettrica, gas ed acqua	2,8	4,2	2,5	3,6
F Costruzioni	6,6	6,6	3,5	3,5
G Commercio e riparazione di beni di consumo	3,6	3,6	2,3	2,3
H Alberghi e ristoranti	2,0	2,0	0,3	0,8
I Trasporti, magazzinaggio e comunicazioni	3,3	3,3	0,7	0,7
J Intermediazione monetaria e finanziaria	5,2	5,2	4,0	4,0
K Altre attività professionali ed imprenditoriali	3,2	3,2	1,5	1,5
C-K TOTALE	3,1	3,1	1,7	1,7

Fonte: Elaborazioni su dati della rilevazione Oros

Nel paragrafo 7 si espongono i risultati di alcune proposte di miglioramento alla metodologia attualmente usata, volte a ridurre le cause di errore legate alla lista anagrafica (restante 1,4%), mentre nel paragrafo 8 si sperimentano alcune tecniche di trattamento dei dati campionari per migliorare la stima per calibrazione utilizzata.

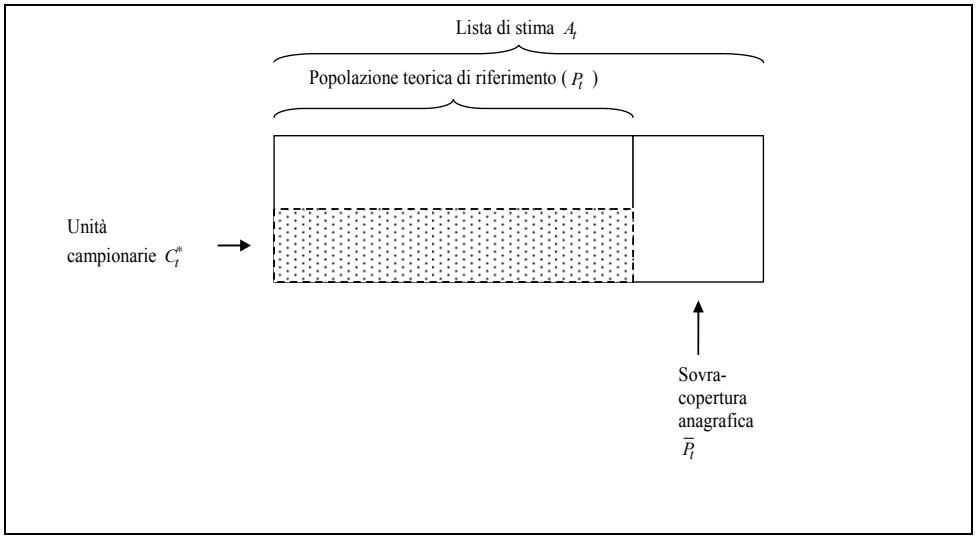
7. L'errore di lista nella stima delle PMI. Alcune proposte di miglioramento

L'analisi precedente mostra che l'errore di stima preliminare è dovuto quasi nella stessa misura ad errori di lista e a errori nel modello di riporto. Lo scopo di questo paragrafo è

quello di esporre alcune proposte di modifica alla metodologia attuale che consentano di ridurre l'effetto sulle stime dell'occupazione di problematiche legate alla definizione della lista di stima.

Se il registro anagrafico fosse aggiornato in tempo reale sui cambiamenti di stato di attività delle posizioni contributive, la lista di stima (A_t) rappresenterebbe correttamente la popolazione corrente di riferimento delle unità (P_t). Esso invece, da una parte, è caratterizzato da errori di sovracopertura dovuti al ritardo di registrazione degli eventi di sospensione e/o cessazione dell'attività (cessazioni non registrate), dall'altra è affetto da errori di sottocopertura causati dai ritardi nella comunicazione e/o registrazione delle nuove posizioni o delle riattivazioni di posizioni sospese. Nel primo caso si determina l'inclusione nella lista di stima di alcune posizioni che risultano erroneamente attive e una conseguente sovrastima dei totali di calibrazione. Nel secondo caso, al contrario, vengono considerate attive un minor numero di posizioni rispetto a quelle realmente esistenti, con conseguenze di segno opposto sui totali. Si è calcolato, però, che l'errore di sovracopertura risulta predominante rispetto a quello di sottocopertura la cui entità è assolutamente trascurabile. Ciò ha indotto a riflettere su una serie di modifiche alla metodologia di base volte a contenere soltanto la sovracopertura considerato il rilevante peso che questa ha sulla lista di stima. La figura 2 rappresenta graficamente le informazioni disponibili.

Fig. 2 – Informazioni disponibili e problematiche nella definizione della popolazione teorica[19]



L'unica informazione certa nel trimestre corrente deriva dalla presenza delle unità nel campione (C_t^*). Il problema più rilevante per la stima della popolazione è costituito, però, dalle unità non campionarie per le quali non si ha certezza sullo stato di attività non essendo pervenuto il DM10. Il loro stato di attività, infatti, potrebbe non essere stato assegnato in

[19] Il soprassegno indica il complemento all'insieme delle unità del registro anagrafico.

maniera corretta nell'anagrafe per ritardo di cancellazione/sospensione (sovracopertura $\overline{P_t}$).

Il confronto tra la lista di stima e la popolazione di riferimento dello stesso trimestre permette di misurare la sovracopertura. In media sui sette trimestri considerati essa interessa circa un quarto delle posizioni contributive considerate "formalmente attive" (tabella 3). La dimensione del fenomeno ha indotto alla messa a punto di una procedura preliminare di *data cleaning*, finalizzata all'esclusione delle unità formalmente definite attive ma che con probabilità elevatissima sono cessate. Il "taglio" sulle unità definite attive è stato effettuato su quelle unità che al dato trimestre *t* non avevano presentato neanche un DM10 nei precedenti tre anni. L'effettiva utilità della variabile "numero di periodi di assenza del DM10" nel discriminare quale unità potessero essere correttamente predette come inattive è stata testata con un modello logistico. L'utilizzo di tale metodo di *data cleaning* consente un abbattimento dell'errore di sovracopertura dal 25,4% al 16,3% (tabella 3)[20]. Sebbene l'errore sia significativamente ridotto l'entità della sovracopertura anagrafica rimane ancora consistente.

Tab. 3 - Errori medi di sovracopertura dell'anagrafe per sezione di attività economica, in termini di numero di posizioni contributive PMI, prima e dopo il data cleaning. Periodo II trimestre 2001 – IV trimestre 2002 (valori percentuali)

Sezioni di attività economica	*Senza data cleaning*	*Con data cleaning*
C Estrazione di minerali	17,6	10,0
D Attività manifatturiere	18,3	11,4
E Produzione di energia elettrica, gas ed acqua	18,0	10,5
F Costruzioni	38,5	23,6
G Commercio e riparazione di beni di consumo	23,9	15,1
H Alberghi e ristoranti	28,7	18,9
I Trasporti, magazzinaggio e comunicazioni	26,6	17,4
J Intermediazione monetaria e finanziaria	18,9	12,4
K Altre attività professionali ed imprenditoriali	24,3	17,5
C-K TOTALE	25,4	16,3

Fonte: Elaborazioni su dati della rilevazione Oros

Il secondo passo nella direzione di ridurre l'errore di lista consiste nel rivedere il metodo di correzione correntemente utilizzato. Come detto nel paragrafo 3, il metodo di stima attuale tiene in considerazione l'incertezza nello stato di attività che le informazioni anagrafiche producono, mediante l'applicazione, ad ogni unità, di una probabilità di essere attiva, calcolata per gruppi omogenei di unità (*register error groups*). Tale probabilità viene applicata nel calcolo dei totali noti (nella [4]), consentendo di ridurre il peso di ogni unità sui totali. Nel metodo attuale (metodo M0), come mostra la relazione [6], la probabilità di essere attive viene stimata in *t*-4 e, sotto ipotesi di invarianza temporale, viene applicata nel trimestre corrente *t* alle unità della lista di stima non appartenenti al campione. Alle unità

[20] Il taglio dovuto al data-cleaning ha interessato 54 mila posizioni circa, pari al 3,8% sul totale, oltre alle posizioni mai presenti negli universi INPS. Tale metodo può naturalmente implicare che alcune imprese che sono in realtà attive nel trimestre corrente siano considerate inattive e quindi eliminate dalla lista: tuttavia sui 7 trimestri su cui è possibile confrontare la lista di stima e la popolazione di riferiemento è emerso che questo numero di "falsi negativi" è irrisorio.

appartenenti al campione, invece, viene attribuita probabilità pari ad 1[21].

L'applicazione delle probabilità così calcolate consente un ulteriore e significativo abbattimento del fenomeno della sovracopertura anagrafica: l'errore, infatti, scende al 3,3% (tabella 4). Tale metodo, tuttavia, pur riducendo l'entità della sovracopertura continua a produrre un errore sistematico di sovrastima nella definizione delle unità ritenute attive.

La sperimentazione di metodi alternativi a M0 si è mossa in due direzioni: da un lato, utilizzando in modo più efficace l'informazione disponibile e, dall'altro, usando approcci di regressione, anziché stime per rapporto, nel calcolo della probabilità delle unità di essere attive.

Tab. 4 - Errori medi di sovracopertura della lista di stima sul numero di posizioni contributive delle PMI per sezione di attività economica: metodi a confronto. Periodo II trimestre 2001 – IV trimestre 2002 (valori percentuali)

Sezioni di attività economica	*Metodo M0*	*Metodo M1*	*Metodo L6*
C Estrazione di minerali	1,2	-1,2	-0,8
D Attività manifatturiere	2,6	0,0	0,0
E Produzione di energia elettrica, gas ed acqua	2,0	-1,5	-1,2
F Costruzioni	4,9	0,1	0,1
G Commercio e riparazione di beni di consumo	3,0	-0,2	-0,2
H Alberghi e ristoranti	4,0	0,0	0,2
I Trasporti, magazzinaggio e comunicazioni	4,0	0,3	0,1
J Intermediazione monetaria e finanziaria	2,3	-0,4	-0,1
K Altre attività professionali ed imprenditoriali	2,7	-0,9	-1,3
C-K TOTALE	3,3	-0,2	-0,2

Fonte: Elaborazioni su dati della rilevazione Oros

Di seguito si presentano i risultati delle due sperimentazioni apparse più soddisfacenti. Il primo metodo (M1) calcola le probabilità utilizzando unicamente l'informazione di lista di stima e popolazione di riferimento di *t*-4 così come in M0. Tale probabilità misura la riduzione che bisogna applicare alla lista di stima per derivare la numerosità di unità effettivamente attive. Nell'applicazione della probabilità a *t,* come nel metodo M0, per non rinunciare all'uso dell'informazione certa sullo stato di attività derivante dalla presenza delle posizioni nel campione totale ($C_t^* \equiv CT_t$), a queste ultime viene comunque applicata una probabilità pari ad 1. Alle restanti unità, diversamente da M0 in cui viene semplicemente applicata la probabilità stimata sulle informazioni di *t*-4, la probabilità viene proporzionalmente ridotta in modo da non influenzare la numerosità stimata delle unità. Come vedremo, questa semplice modifica pone rimedio alla distorsione implicita nel metodo M0.

In formule le probabilità calcolate con il metodo M1 possono essere espresse come:

[21] Nel metodo M0 le unità del campione a cui vengono applicate le probabilità di essere attive sono quelle del campione ridotto (C_t), mentre nei metodi alternativi sperimentati (M1 e L6 illustrati di seguito in questo paragrafo) si riferiranno al campione totale (CT_t).

$$\hat{p}_{it} = 1 \qquad i \in CT_t$$

$$\hat{p}_{it} = \frac{n\left(P_{t-4}\right)}{n\left(A_{t-4}\right)} * q_1 - q_2 \qquad i \notin CT_t \tag{14}$$

in cui le quote $q_1 = \frac{n(A_t)}{n(\overline{C}\overline{T}_t)}$ e $q_2 = \frac{n(CT_t)}{n(\overline{C}\overline{T}_t)}$ sono i fattori di proporzionamento che consentono di ricalibrare le probabilità tra le unità presenti nel campione totale e le restanti[22]. Attraverso q_1 e q_2 non si fa altro che trasferire una probabilità maggiore ad unità certamente attive perché appartenenti al campione totale e una minore a quelle restanti, senza influenzare la numerosità stimata delle unità rispetto al caso in cui fosse applicata $\hat{p}_{it} = \frac{n(P_{t-4})}{n(A_{t-4})}$ a tutte le unità attive secondo la lista di stima ($i \in A_t$).

Tale assunto implica l'invarianza tra *t*-4 e *t* nella struttura delle unità campionarie e non, rimuovendo la possibile distorsione che si commette con il metodo M0 imponendo alle unità del campione corrente una probabilità diversa da quella stimata.

L'altro metodo (definito metodo L6) prevede sia una diversa selezione delle unità eleggibili, sia un criterio differente per il calcolo della probabilità, basato su un approccio di tipo modellistico, in cui le probabilità di essere attive vengono stimate mediante un modello *logit*. Questo viene stimato sulle unità non campionarie di *t*-4 ($\overline{C}\overline{T}_{t-4}$), dalle quali vengono ulteriormente escluse quelle unità certamente attive che non sono presenti nel campione totale di *t*-4 solo perché non avevano adottato la modalità di invio telematico del DM10 (si tratta delle unità che si trovano nell'universo P_{t-4} e contestualmente nel campione CT_t). In tal modo si riesce a tener conto anche nella fase di calcolo delle probabilità (e non solo nella fase di applicazione come avviene nel metodo M1) della crescita strutturale del campione dovuta al fenomeno del passaggio graduale alla modalità di invio telematico. Nello stesso tempo si riesce a tener conto anche di eventuali altri fattori di natura amministrativa (modifica di procedure adottate dall'INPS per scaricare i dati negli archivi centrali, ecc.) che possono aver indotto l'assenza di un numero rilevante di unità dal campione di *t*-4.

Le probabilità applicate alle varie unità possono essere espresse come:

$$\hat{p}_{it} = 1 \qquad i \in CT_t$$

$$\hat{p}_{it} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}} \qquad i \notin CT_t \tag{15}$$

[22] Cfr. nota 14.

Si osservi come, anche con il metodo L6, le unità del campione vengono considerate certamente attive. Nella (15) $\hat{\beta}$ è il vettore dei coefficienti stimati relativi al vettore delle variabili esplicative X, che vengono opportunamente selezionate a seconda delle sotto-popolazioni considerate[23]. In riferimento ad ogni sottopopolazione i *logit* vengono calcolati separatamente per gruppi di stima, individuati utilizzando informazioni strutturali sulle posizioni contributive.

Entrambi i metodi sperimentati si caratterizzano per una ridefinizione, rispetto al metodo base, dei raggruppamenti della popolazione su cui vengono stimate le probabilità di attività (*register error groups*), sulla base di nuove variabili, quali il ritardo nell'invio dei moduli DM10 e l'età della posizione contributiva[24]. La rilevanza di tali variabili nel modellizzare la probabilità di attività è stata testata con l'ausilio di una metodologia basata su uno *split* binario ricorsivo delle osservazioni (*Classification and Regression Trees*). Tale metodo ha anche consentito di individuare intervalli di queste variabili in cui la probabilità di attività è più omogenea. Tali intervalli sono stati usati nella definizione dei gruppi di stima.

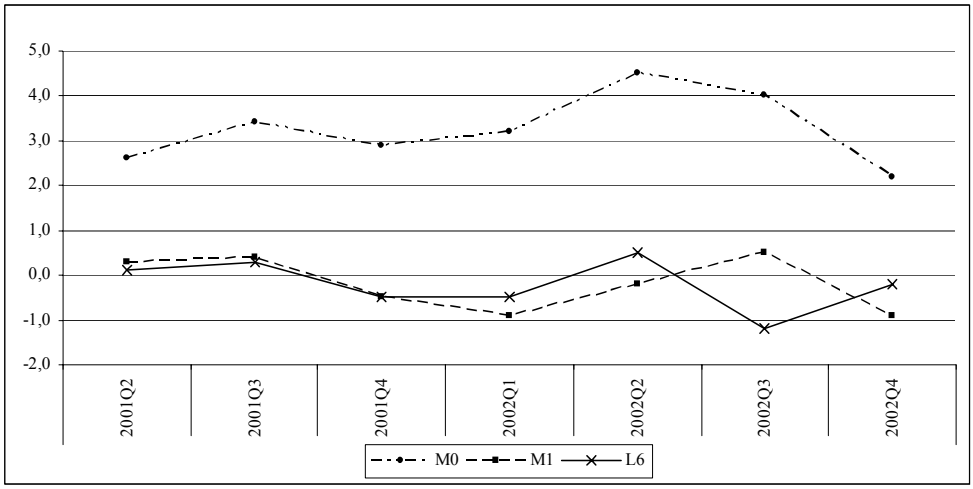
Le modifiche apportate al metodo M0 producono notevoli miglioramenti al problema della sovracopertura (tabella 4). Tra i due, il metodo L6 sembra preferibile poiché, valutato nelle singole occorrenze della serie storica di sperimentazione, genera sempre un errore minore rispetto a M1, tranne nel III trimestre 2002 (figura 3). Ciò avviene poiché L6, a differenza di M1, sfrutta anche l'informazione sul campione corrente rispetto all'ausiliario e, come tale, risente in modo maggiore di eventuali cadute della numerosità del campione. In particolare, nel III trimestre 2002 in corrispondenza di una riduzione delle unità campionarie, avvenuta per motivi amministrativi, il campione di t-4 utilizzato per il calcolo delle probabilità di essere attive non rappresenta adeguatamente la struttura informativa del campione corrente[25].

[23] In particolare, vengono individuate 4 diverse sottopopolazioni a cui applicare altrettanti modelli: le unità panel senza variabili ausiliarie, le unità panel con variabili ausiliarie, le unità nuove nate appartenenti alla sezione I dell'Ateco'02 e le unità nuove nate non appartenenti alla sezione I dell'Ateco'02. L'isolamento delle unità neo nate della sezione I si rende necessario in quanto tale raggruppamento presenta caratteristiche di forte variabilità rispetto alle unità classificate nelle altre sezioni. Per questo raggruppamento è necessario ricorrere ad un modello estremamente semplificato per la stima delle probabilità delle unità di essere attive.

[24] Nella fase attuale di sperimentazione, la nuova partizione in register error groups in M1 è stata prevista solo per le unità panel senza variabile ausiliaria, introducendo il ritardo tra le variabili di stratificazione ed affinando il livello di dettaglio della variabile età, fortemente correlata con la variabile ritardo. Nel metodo L6, invece, i register error groups sono appositamente costruiti per l'applicazione dei logit e presentano caratterizzazioni completamente differenti rispetto ai gruppi predisposti per M0 ed M1. Per ogni sotto-popolazione considerata vengono infatti definiti dei gruppi omogenei, individuati secondo le modalità di alcune delle variabili strutturali di maggior rilievo (settore di attività economica, classe dimensionale, ripartizione territoriale, classe di età etc.).

[25] Il campione di tale trimestre presenta, infatti, uno sbilanciamento a favore delle unità neo nate mentre sono assenti solo per motivi amministrativi unità che invece erano già presenti nel campione del trimestre ausiliario, su cui si calcolano le probabilità di essere attive. Ciò comporta una sottostima delle probabilità delle unità assenti nel campione corrente solo per coincidenza amministrativa (unità alle quali in una situazione ordinaria, nel metodo L6, si sarebbe attribuita una probabilità pari ad 1).

Fig. 3 - Errori medi di sovracopertura della lista di stima sul numero di posizioni contributive delle PMI secondo i diversi metodi. Periodo II trimestre 2001 – IV trimestre 2002 (valori percentuali)



Fonte: Elaborazioni su dati della rilevazione Oros

7.1 Gli effetti dell'errore di sovracopertura sulla stima dell'occupazione nelle PMI

Nel paragrafo precedente, si è mostrato l'abbattimento dell'errore sulla stima della numerosità delle unità attive dovuta ai miglioramenti introdotti. In questo paragrafo, l'analisi viene completata presentando l'impatto dei metodi M0, M1 ed L6 sugli errori di stima dei dipendenti, sia in termini di livello che di variazioni. Per produrre queste statistiche, quindi, le stime dei dipendenti sono state ricalcolate applicando di volta in volta nella procedura di calibrazione le probabilità calcolate nei tre metodi.

Prima di scendere nel dettaglio dei vari metodi si noti che l'errore totale per il metodo M0 qui riportato (tabella 5) è più basso di quello evidenziato in precedenza (tabella 2). Questo miglioramento nelle stime a parità di metodo va attribuito essenzialmente all'operazione di *data-cleaning* dell'anagrafe accennata in precedenza, che ha comportato un abbattimento dell'errore totale di 0,3 punti percentuali (da 3,1% a 2,8%). L'effetto della pulizia è particolarmente forte nei settori dell'intermediazione monetaria e finanziaria, del commercio e dell'estrazione di minerali (rispettivamente, i settori J, G e C).

L'errore di stima sui livelli, a parità di data cleaning, si riduce in termini di MAPE dal 2,8% del metodo M0 allo 0,8% del metodo M1 e all'1,1% del metodo L6.

Tab. 5 – Errore di stima nei livelli dell'occupazione delle PMI. Periodo II trimestre 2001 – IV trimestre 2002 (valori percentuali medi di periodo)

Sezione di attività economica	*Metodo M0*		*Metodo M1*		*Metodo L6*	
	MPE	*MAPE*	*MPE*	*MAPE*	*MPE*	*MAPE*
C Estrazione di minerali	2,4	2,4	0,7	0,9	1,3	1,3
D Attività manifatturiere	1,8	1,8	0,4	0,4	0,7	0,7
E Produzione di energia elettrica, gas ed acqua	2,4	3,7	0,2	2,3	0,6	2,2
F Costruzioni	6,1	6,1	2,8	2,8	3,2	3,2
G Commercio e riparazione di beni di consumo	2,9	2,9	0,8	0,8	1,6	1,6
H Alberghi e ristoranti	3,2	3,2	0,5	0,8	1,3	1,3
I Trasporti, magazzinaggio e comunicazioni	3,3	3,3	0,9	1,0	1,0	1,0
J Intermediazione monetaria e finanziaria	3,0	3,0	1,0	1,4	1,9	2,3
K Altre attività professionali ed imprenditoriali	3,1	3,1	0,6	0,8	-0,3	0,9
C-K TOTALE	2,8	2,8	0,8	0,8	1,1	1,1

Fonte: Elaborazioni su dati della rilevazione Oros

Dato l'errore teorico di riporto, quantificato pari all'1,7% e invariato in questa fase dell'analisi (tabella 2), l'errore di lista calcolato per differenza dall'errore totale diviene negativo nei due metodi sperimentati, indicando che l'originario errore di sovrastima viene corretto in eccesso, tanto da indurre una sottostima. Tali metodi alternativi consentono una riduzione generalizzata dell'errore di stima nei livelli tra tutte le sezioni di attività economica, comportando un abbattimento dell'errore di circa 2 punti percentuali, sia in termini di MAPE che di MPE. Questa riduzione, tuttavia, appare meno netta nell'errore sulle variazioni tendenziali, per cui il MAPE passa da 0,6% di M0 a 0,4% e 0,5%, rispettivamente, dei metodi M1 e L6 (tabella 6). L'errore medio (MPE), invece, registra un apparente peggioramento, passando da +0,1% nel metodo M0 a -0,4% e -0,5%, rispettivamente, nei due metodi alternativi. In realtà, il valore quasi nullo del MPE nel metodo M0 nell'aggregato totale (da C a K dell'Ateco 2002) sembra essere il frutto di effetti di composizione particolarmente favorevoli. Valutando più attentamente i risultati per sezione i tre metodi non comportano differenze rilevanti nel MPE.

Tab. 6 – Errore di stima nelle variazioni tendenziali dell'occupazione delle PMI. Periodo II trimestre 2002 – IV trimestre 2002 (punti percentuali medi di periodo)

Sezione di attività economica	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	1,6	1,6	1,2	1,6	1,2	1,5
D Attività manifatturiere	0,4	0,5	0,1	0,3	0,0	0,1
E Produzione di energia elettrica, gas ed acqua	5,9	5,9	4,8	4,8	4,3	4,3
F Costruzioni	0,8	1,5	-0,3	0,8	-0,2	0,5
G Commercio e riparazione di beni di consumo	-1,0	1,0	-1,5	1,5	-1,7	1,7
H Alberghi e ristoranti	-0,1	1,9	-0,9	1,1	-0,9	1,5
I Trasporti, magazzinaggio e comunicazioni	1,5	1,5	1,0	1,0	0,6	1,3
J Intermediazione monetaria e finanziaria	-1,9	1,9	-2,5	2,5	-2,8	2,8
K Altre attività professionali ed imprenditoriali	-0,8	0,9	-1,0	1,0	-0,9	0,9
C-K TOTALE	0,1	0,6	-0,4	0,4	-0,5	0,5

Fonte: Elaborazioni su dati della rilevazione Oros

Ulteriori evidenze emergono dalla scomposizione dell'errore per le diverse sottopopolazioni di stima delle PMI che, come detto nel paragrafo 3, si distinguono per la diversa disponibilità di informazione ausiliaria (imprese nuove nate, che pesano in termini di dipendenti il 5%, imprese panel con variabili ausiliarie, il cui peso occupazionale è il

93%, imprese panel senza variabili ausiliarie, che pesano il 2%). Il comportamento degli errori per tali sottopopolazioni, sottoposte ad un trattamento piuttosto diverso nei tre metodi (in particolare L6 rispetto ad M0 e M1, si veda la nota 19), non si differenzia molto dai risultati generali visti in precedenza. Nei livelli dell'occupazione, per tutte le tipologie, la riduzione del *bias* è netta passando da M0 agli altri due metodi, con una performance migliore di M1, rispetto a L6, nelle unità panel senza variabili ausiliarie e un risultato opposto per le nuove nate (tabella 7). Nelle variazioni tendenziali il metodo M1 consente di ottenere un miglioramento consistente rispetto a M0 soltanto nelle panel senza variabili ausiliarie. Al contrario il metodo L6 conduce, generalmente, ad errori maggiori rispetto a M0 in tutte le sottopopolazioni di stima.

In sintesi i metodi M1 e L6 consentono una drastica riduzione dell'errore nei livelli mentre lasciano quasi inalterato l'errore sulle variazioni.

Tab. 7 – Errore di stima dell'occupazione nelle PMI per tipo di sottopopolazione delle PMI. Periodo II trimestre 2002 – IV trimestre 2002 (punti percentuali medi di periodo)

Tipo di sottopopolazione delle PMI	Metodo M0		Metodo M1		Metodo L6	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
	Errore di stima nei livelli					
Unità nuove nate	9,6	9,6	4,3	4,3	3,6	3,6
Unità panel senza variabili ausiliarie	26,8	26,8	4,5	5,5	9,9	9,9
Unità panel con variabili ausiliarie	2,0	2,0	0,5	0,5	0,8	0,8
Unità totali	2,8	2,8	0,8	0,8	1,1	1,1
	Errore di stima nelle variazioni tendenziali					
Unità nuove nate	-1,0	3,4	-2,7	4,8	-3,8	3,8
Unità panel senza variabili ausiliarie	2,9	7,4	-2,3	4,5	-2,4	7,6
Unità panel con variabili ausiliarie	0,0	0,3	-0,2	0,2	-0,2	0,3
Unità totali	0,1	0,6	-0,4	0,4	-0,5	0,5

Fonte: Elaborazioni su dati della rilevazione Oros

8. L'errore di riporto nella stima delle PMI

La seconda fonte di errore nella stima preliminare di Oros risiede nel modello di riporto all'universo. Questo aspetto della stima, presenta elementi di notevole complessità, fortemente connessi alla natura delle mancate risposte che caratterizzano i *set* informativi a disposizione per le stime anticipate. In questo paragrafo, dopo aver descritto il problema della non risposta di Oros nei termini dell'analisi di Rubin (1976), si illustrano in dettaglio le motivazioni che hanno condotto alla riduzione del campione ai fini del riporto, ma anche le conseguenze che questa ulteriore selezione implica. Sono state esplorate due possibili strade verso la riduzione dell'errore di riporto da cui si traggono alcuni insegnamenti rilevanti.

8.1 Stime da campioni non casuali e natura della non risposta

Le stime anticipate di Oros si basano su un campione di convenienza, rappresentato dall'insieme dei DM10 che giungono per via telematica (campione totale). Questo insieme di unità non rispecchia alcun disegno campionario, pertanto non vi è alcuna garanzia che sia rispettato uno schema di casualità.

Come noto, lo strumento per ridurre il potenziale *bias* che deriva da un campionamento non casuale consiste nel ricorso ad una metodologia in cui vengano definite delle ipotesi sul

processo generatore dei dati mancanti.

Volendo configurare il caso della rilevazione Oros in un contesto teorico, conviene riqualificare il problema in termini di mancate risposte parziali, ovvero pensare che al tempo corrente si osservi l'intero campione teorico coincidente con l'universo dei dati: per tutte le unità si ha risposta sulle variabili ausiliarie (ad esempio le informazioni riferite a *t*-4 per le unità panel), mentre sulle variabili correnti si ha risposta solo per una parte delle unità. In riferimento a tale contesto, seguendo Rubin (1976), le mancate risposte nella rilevazione Oros possono derivare da uno schema MAR (*Missing at Random*) dove le non risposte sulle variabili correnti dipendono solo dalle variabili ausiliarie o da uno schema NMAR (*Not Missing At Random*) dove le mancate risposte dipendono, invece, dalle variabili mancanti stesse. Nel caso MAR un approccio che sfrutti opportunamente le variabili ausiliarie può consentire una riduzione della distorsione che deriva dalle mancate risposte; nel caso NMAR, invece, ciò non è garantito[26].

La metodologia di stima anticipata di Oros è fortemente basata sull'utilizzo di variabili ausiliarie che, in particolare, intervengono in due punti strategici: nella formazione dei *model groups* e nella procedura di calibrazione dei pesi. La presenza di una distorsione sistematica, nonostante l'utilizzo di variabili ausiliarie, suggerisce quindi che il processo generatore delle mancate risposte di Oros sia di natura NMAR.

8.2. Campione totale e campione ridotto

Le caratteristiche del campione sulla base del quale si produce la stima anticipata dipendono, in parte dalla autoselezione delle imprese che scelgono la compilazione telematica del DM10, in parte dall'ulteriore selezione, attuata in una fase precedente alla stima, il cui scopo è quello di escludere le unità che, pur essendo attive, non hanno risposto per tutti e tre i mesi del trimestre. Per comprendere appieno questo procedimento va ricordato che, sebbene i dati siano raccolti con riferimento ai singoli mesi del trimestre, le variabili mensili vengono successivamente aggregate per ciascuna unità in valori medi trimestrali. L'occupazione media trimestrale di ciascuna unità è semplicemente la somma dell'occupazione mensile dei DM10 pervenuti in quel trimestre diviso 3, indipendentemente dal numero di mesi per cui è presente il DM10. Questa operazione è corretta quando l'assenza del DM10 rappresenta un segnale di inattività dell'unità, mentre è errata nel caso di una mancata risposta[27]. Se fosse possibile discriminare le mancate risposte, per queste unità sarebbe utile dividere l'aggregato trimestrale per i mesi di effettiva attività inducendo una imputazione implicita del dato assente. Purtroppo, a causa dei ritardi di aggiornamento, l'informazione anagrafica non può essere utilizzata per effettuare questa discriminazione. L'operazione di riduzione del campione (che genera il

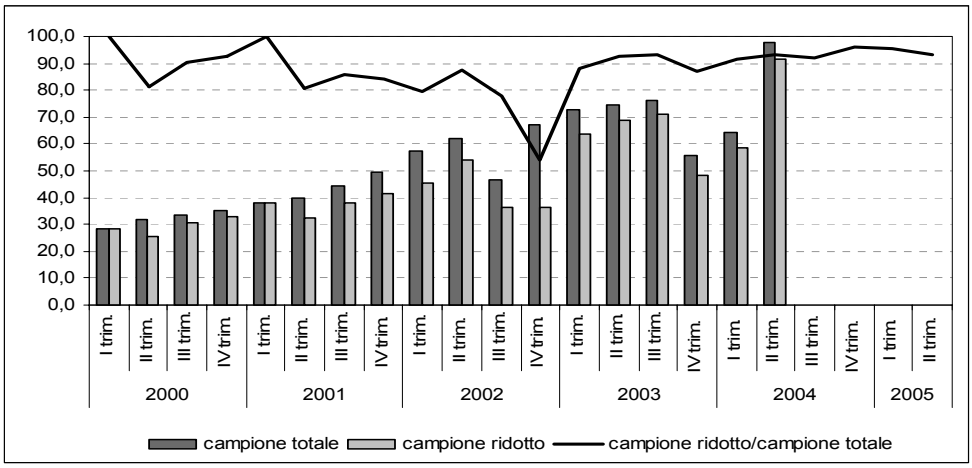
[26] Nel caso NMAR, secondo la letteratura, l'obiettivo di riduzione del bias si può raggiungere, da una parte, tentando di ridurre il fenomeno delle mancate risposte, dall'altra, accumulando informazioni circa le differenze tra rispondenti e non rispondenti sulle variabili di riferimento. In questo secondo contesto, cioè, il modello ipotizzato per la descrizione del processo generatore delle mancate risposte andrebbe adeguatamente incorporato nel modello statistico finalizzato alla stima (Little, Rubin, 1987).

[27] Due ordini di motivazioni sono alla base delle mancate risposte: i tempi con cui vengono acquisiti i dati dall'INPS, che comportano un numero minore di informazioni disponibili sull'ultimo mese di ogni trimestre; i problemi di gestione interna all'INPS (aggiornamento di software ecc.) che hanno implicato, in alcune occasioni, una ridotta disponibilità di dati indistintamente nei vari mesi del trimestre. Va osservato che, sebbene questi fenomeni siano rari sono abbastanza tipici di rilevazioni fondate su dati amministrativi e, mettono bene in evidenza la dipendenza di chi compie la raccolta secondaria di dati da chi ne compie la raccolta primaria ed i rischi ad essa connessi.

passaggio dal campione totale al campione ridotto), quindi, si rende necessaria al fine di attenuare l'effetto distorsivo che le mancate risposte mensili presenti nel campione totale dei rispondenti "rapidi" indurrebbero sulle stime. Utilizzare il campione totale senza alcun pretrattamento condurrebbe ad una sottostima dell'occupazione media mensile, non potendo con certezza stabilire il numero dei mesi, all'interno del trimestre, di effettiva attività di ciascuna unità.

Il metodo di stima correntemente utilizzato si basa, dunque, solamente sull'insieme di unità con informazione completa. In particolare, concorrono alla stima le unità attive che hanno inviato il DM10 in tutti e tre i mesi del trimestre o che sono caratterizzate da qualche assenza che sia configurabile secondo l'anagrafe come inattività (nascita, sospensione, riattivazione o cessazione o reputate stagionali sulla base dell'informazione sui pattern di presenza nel trimestre attuale e in quello ausiliario). Il campione che risulta da tale selezione (campione ridotto) ha una dimensione in termini di unità inferiore del 10-15 per cento rispetto al campione totale (figura 4), con cadute che superano il 40% nei trimestri interessati da problemi amministrativi. Si è tuttavia osservato come esso garantisca una elevata copertura in termini di alcune importanti variabili di stratificazione e, all'interno di esse, una struttura equilibrata, che solitamente non si discosta in misura significativa dalla copertura garantita dal campione totale.

Fig. 4 – Incidenza delle posizioni contributive del campione ridotto su quelle del campione totale e grado di copertura delle posizioni del campione totale e del campione ridotto rispetto a quelle dell'universo. Periodo I trimestre 2001-II trimestre 2005 (valori percentuali)



Fonte: Elaborazioni su dati della rilevazione Oros

8.3 L'errore del modello di riporto utilizzando il campione ridotto: quantificazione dell'errore di riporto

Come per la sperimentazione mostrata nel paragrafo 6, è stato possibile quantificare l'errore dovuto unicamente al modello di riporto alla popolazione di riferimento sostituendo la rappresentazione della popolazione di riferimento a *t* con la popolazione certa a *t* (Metodo a lista

certa o LC). Nei sette trimestri considerati, l'errore è sistematicamente positivo sui livelli e pari all'1,9%[28] in termini sia di MAPE, sia di MPE, mentre sulle variazioni tendenziali l'errore medio è più contenuto ma anch'esso sistematico, con un valore dello 0,7% (tabella 8).

Tab. 8 - Errori sui livelli e sulle variazioni tendenziali della stima delle PMI con il campione ridotto. Periodo II trimestre 2002 – IV trimestre 2003 (valori e punti percentuali medi di periodo)

Sezione di attività economica	*Livelli*		*Variazioni tendenziali*	
	MPE	*MAPE*	*MPE*	*MAPE*
C Estrazione di minerali	2,7	2,7	0,6	1,4
D Attività manifatturiere	1,2	1,2	-0,2	0,4
E Produzione di energia elettrica, gas ed acqua	2,6	3,0	3,1	4,3
F Costruzioni	3,9	3,9	-0,3	0,6
G Commercio e riparazione di beni di consumo	2,3	2,3	-1,8	1,8
H Alberghi e ristoranti	2,3	2,3	-1,2	1,2
I Trasporti, magazzinaggio e comunicazioni	1,6	1,6	-0,2	0,6
J Intermediazione monetaria e finanziaria	1,1	2,3	0,9	3,6
K Altre attività professionali ed imprenditoriali	2,1	2,1	-2,7	2,7
C-K TOTALE	1,9	1,9	-0,7	0,7

Fonte: Elaborazioni su dati della rilevazione Oros

L'errore è diverso nei tre gruppi di unità che contribuiscono alla stima (cfr. par. 3). Questo dipende dal numero di unità che rientrano nei sottogruppi, ma anche dalla diversa disponibilità di informazione ausiliaria connessa a ciascuno di essi: l'errore appare meno elevato per le unità per cui si possono utilizzare sia una partizione in *model groups* più dettagliata, sia un maggior numero di variabili nella procedura di calibrazione dei pesi.

Indagando sulla natura dell'errore di riporto, si deve anzitutto considerare l'effetto di distorsione sulle stime dovuto al passaggio dal campione totale a quello ridotto. L'operazione comporta una drastica riduzione dell'errore di misura altrimenti generato dalle mancate risposte, ma induce una sovrarappresentazione di unità sempre presenti nel trimestre e quindi con valore dei dipendenti mediamente più elevato.

L'effettiva distorsione sulle stime dipende dalla "capacità di aggiustamento" del modello di riporto e dall'utilizzo di variabili ausiliarie. Laddove si dispone di più variabili ausiliarie la dimensione della sovrastima è inferiore, tuttavia, anche per i gruppi in cui l'informazione ausiliaria è massima l'entità della sovrastima è troppo elevata, suggerendo che le variabili ausiliarie non contengono informazione sufficiente sul processo che genera la selezione del campione.

E' interessante riflettere su quale potrebbe essere l'effettivo meccanismo che genera la sovrastima nel processo di riduzione del campione, facendo riferimento per semplicità solo alle unità panel con variabili ausiliarie. Il tipo di selezione effettuato agisce sui pattern di presenza del campione a t, mantenendo nel campione le unità con pattern completo o assenze giustificate (non configurabili come mancate risposte). Queste unità sono quelle per cui le variabili ausiliarie saranno calibrate (si veda la [2]). Il fatto che al tempo corrente t, queste unità abbiano pattern completo di risposta, non implica che abbiano un pattern completo anche in t-4. Quello che ragionevolmente accade è che una parte di esse

[28] La discordanza con gli errori di riporto della tabella 2 (pari a 1,7% per il totale C-K) è dovuta ad una lieve differenza nella classificazione delle unità che entrano nelle due stime.

presentano in *t*-4 un'assenza parziale di DM10 dovuta al fatto che l'unità non era attiva in quei mesi. I pesi di riporto rifletteranno la struttura dei pattern di presenza tra campione ed universo in *t*-4, ma verranno applicati in *t* a unità campionarie, con pattern più completo, che non rappresentano in maniera speculare la relazione con l'universo corrente, rispetto a quella calcolata a *t*-4. In altri termini, il rapporto tra l'occupazione di *t*-4 del campione e il corrispondente totale di popolazione è inferiore al rapporto tra l'occupazione di *t* e il corrispondente totale di popolazione. Ciò implica il calcolo di pesi troppo alti per le unità del campione con la conseguente sovrastima che si osserva sull'occupazione totale. Ciò avviene in conseguenza del fatto che, la riduzione del campione incidendo solo sulle informazioni correnti non riguarda le variabili ausiliarie. Nel paragrafo 8.3.1 si propone una possibile soluzione a questo problema, che consiste nel ridisegnare i pesi rispetto alla struttura del campione corrente.

E' infine interessante notare che l'errore sulle variazioni tendenziali è molto inferiore a quello sui livelli. Una ragionevole spiegazione è che, essendo basato prevalentemente su unità con informazione completa, il campione ridotto tende ad accentuare le caratteristiche panel tra i trimestri, conferendo una maggiore stabilità nelle stime che si riflette in variazioni tendenziali di migliore qualità. Il dettaglio sui singoli trimestri, non riportato nel documento, evidenzia che una stima derivata prevalentemente da unità panel, e rafforzata dalla crescita delle dimensioni del campione, comporta anche una lieve riduzione dell'errore: questo effetto giustifica il segno negativo dell'MPE sulle variazioni tendenziali, rispetto al segno positivo del MAPE (tabella 8).

8.3.1. Bilanciamento del campione rispetto ai pattern di presenza

Le analisi esposte nel paragrafo precedente hanno evidenziato che la riduzione del campione provoca uno sbilanciamento tra variabili correnti e variabili ausiliarie. Se l'analisi è corretta, dunque, una diversa selezione del campione che comporti un migliore bilanciamento delle variabili ausiliarie e correnti, dovrebbe implicare una riduzione del bias. Per verificare questa ipotesi, si è svolto un esercizio in cui il campione è stato ridotto selezionando le posizioni con pattern di presenza completo sia per quanto riguarda il trimestre corrente, sia per quanto riguarda il trimestre ausiliario (selezione LCb).

I risultati di stima sui dipendenti sono riportati nella tabella 9, in cui gli errori si riferiscono ai livelli. L'effetto è positivo; l'errore di stima si riduce dall'1,9% allo 0,5% (MPE) e dall'1,9% allo 0,8% (MAPE). È interessante notare come i risultati di stima siano migliorati sebbene la numerosità del campione si sia drasticamente ridotta.

Tab. 9 – Errori per sezione nella stima dei livelli delle PMI da campione ridotto, nel metodo a lista certa (LC) e con selezione sul pattern di presenza in t-4 (LCb). Periodo II trimestre 2002 – IV trimestre 2003 (valori in percentuale)

Sezione di attività economica	*LC*		*LCb*	
	MPE	*MAPE*	*MPE*	*MAPE*
C Estrazione di minerali	2,7	2,7	0,6	0,6
D Attività manifatturiere	1,2	1,2	0,3	0,6
E Produzione di energia elettrica, gas ed acqua	2,6	3	0,8	2,2
F Costruzioni	3,9	3,9	1,5	1,5
G Commercio e riparazione di beni di consumo	2,3	2,3	0,7	1,1
H Alberghi e ristoranti	2,3	2,3	0,9	1,0
I Trasporti, magazzinaggio e comunicazioni	1,6	1,6	-0,1	0,7
J Intermediazione monetaria e finanziaria	1,2	2,6	1,5	2,1
K Altre attività professionali ed imprenditoriali	2,1	2,1	0,2	1,6
C-K TOTALE	*1,9*	*1,9*	*0,5*	*0,8*

Fonte: Elaborazioni su dati della rilevazione Oros

Purtroppo, gli incoraggianti risultati sui livelli non trovano conferma negli errori riferiti alle variazioni, non presentate in tabella. Sebbene gli errori dei singoli trimestri siano minori rispetto a quelli ottenuti con il metodo base, infatti, essi presentano una minore stabilità nel tempo e, conseguentemente, conducono ad un aumento degli errori delle variazioni tendenziali. Il probabile motivo di ciò risiede nel fatto che si è ridotto l'errore sulle unità più stabili (con presenza completa nel trimestre corrente e in quello ausiliario), mentre si è indotta una minore rappresentatività e, quindi, un peggioramento della stima, sulle unità più instabili che presentano un pattern incompleto in uno o in entrambi i trimestri. Sebbene la prima popolazione sia la più consistente, la seconda contribuisce in maniera rilevante a determinare la dinamica dell'occupazione.

8.4 Un possibile utilizzo del campione totale: l'imputazione delle mancate risposte mensili

Nel paragrafo 8.2 sono state descritte le motivazioni della non fattibilità nel ricorso al campione totale finché sussistono posizioni contributive con dati incompleti nel trimestre: mediamente, le posizioni del campione totale che presentano meno di tre DM10 nel trimestre sono il 10-15%.

In generale, l'uso dei dati del campione totale senza alcuna correzione per i DM10 incompleti comporterebbe una sottostima del livello dell'occupazione che in alcuni casi può essere drammatica, come ad esempio nel IV trimestre 2002 in cui, a causa di fenomeni amministrativi, la quota di posizioni contributive con dati incompleti ha superato il 40%.

Se si volesse procedere all'imputazione delle mancate risposte del campione totale, la ricostruzione delle informazioni sui trimestri con DM10 mancante dovrebbe necessariamente seguire due passi successivi:

- l'identificazione dell'effettivo stato di attività della posizione contributiva per i mesi con DM10 mancante (cioè discriminare tra effettiva inattività della posizione e ritardo del DM10);
- l'imputazione delle variabili mancanti (l'occupazione, in primo luogo, ma anche le retribuzioni e gli oneri sociali) per i mesi identificati come ritardi.

Un primo tentativo di identificazione dello stato di attività, che fa uso delle informazioni anagrafiche e del pattern di presenza nel trimestre *t*-4, è già eseguito nell'attuale selezione del campione ridotto a partire dal campione totale. Come evidenziato nei paragrafi precedenti, però, l'operazione conduce ad una sostanziale sottorappresentazione delle unità con mesi di inattività, con conseguente eccessiva esclusione di posizioni contributive dal campione, in gran parte per effetto del problema delle cessazioni non registrate. Questo approccio, che per la natura della selezione dello stato di attività può essere definito deterministico, sebbene sfrutti pienamente le informazioni disponibili, ha chiaramente dei limiti imposti dai ritardi di aggiornamento dei dati anagrafici.

Una via alternativa per l'individuazione dello stato di attività potrebbe essere quella di ricorrere ad un approccio probabilistico, in cui sia possibile attribuire a ciascuna posizione con pattern incompleto una misura di probabilità per ogni pattern reale che tale unità può assumere. Tra le modalità che possono condurre ad una stima di tale probabilità si potrebbe, ad esempio, fare ricorso all'informazione rilevata a *t*-4, in cui si dispone sia del campione, sia dell'universo. La stima della probabilità potrebbe essere effettuata per gruppi omogenei di unità, individuati secondo qualche variabile discriminante, mettendo a confronto il pattern del campione con quello dell'universo corrispondente.

La misura di ciascuna delle probabilità individuate potrebbe essere inserita nella

relazione individuata per l'imputazione del dato mancante, come fattore di peso per i diversi pattern realizzabili. Per l'imputazione dell'occupazione, può essere opportuno utilizzare come base di riferimento per la ricostruzione del dato mancante, l'informazione in *t*. La proposta appena fatta rappresenta solo un possibile metodo per imputare il reale pattern di attività e la conseguente occupazione. Naturalmente la sua efficacia andrebbe testata e possibilmente confrontata con metodi alternativi, ma tali analisi vanno oltre i limiti del presente documento. Qui di seguito però si riporta una misura del possibile guadagno che si avrebbe, in termini di riduzione dell'errore, se si effettuasse un'imputazione dei dati mancanti. Per far ciò si simula l'effetto sull'errore di stima dell'uso di un campione totale imputato, attribuendo alla lista del campione totale i dati economici dell'universo corrispondente. Poiché l'universo fornisce, per definizione, un'informazione definitiva, corretta e completa, è come se il campione totale fosse stato sottoposto ad una operazione di imputazione delle mancate risposte mensili, con errore nullo sulla determinazione dello stato e con conseguente corretta attribuzione del dato medio mensile. In altri termini la misura di errore che si riporta di seguito rappresenta il minimo errore possibile che si commetterebbe utilizzando un campione totale imputato, ovvero nel caso in cui l'imputazione fosse perfetta. L'effetto sugli errori di livello è decisamente favorevole, comportando un passaggio di MPE e MAPE da 1,9% della stima con campione ridotto a 0,5% nel caso di uso del campione totale. Tuttavia, il passaggio al campione totale non implica alcun miglioramento sugli errori nelle variazioni tendenziali (MPE e MAPE sono pari allo 0,7% con l'uso del campione totale, mentre erano pari a -0,7% e 0,7% con il campione ridotto). Come già detto, si può supporre che la riduzione del campione, accentuando le caratteristiche panel dei dati utilizzati per le stime, tende a stabilizzare l'errore sui livelli nel tempo; al contrario l'uso del campione totale, seppure con dati corretti, risente del cambiamento della struttura del campione nel tempo, comportando errori crescenti sui livelli e quindi errori sostanziali sulle variazioni

Tab. 10 - Errori sui livelli e sulle variazioni tendenziali della stima delle PMI con il campione totale. Periodo II trimestre 2002 – IV trimestre 2003 (valori e punti percentuali medi di periodo)

Sezione di attività economica	Livelli		Variazioni tendenziali	
	MPE	MAPE	MPE	MAPE
C Estrazione di minerali	0,5	1,7	3,7	3,7
D Attività manifatturiere	0,1	0,4	0,8	0,8
E Produzione di energia elettrica, gas ed acqua	-0,4	2,5	6,3	6,3
F Costruzioni	0,8	1,4	2,9	2,9
G Commercio e riparazione di beni di consumo	0,9	0,9	0,1	0,1
H Alberghi e ristoranti	0,7	0,7	0,4	0,4
I Trasporti, magazzinaggio e comunicazioni	-0,1	0,8	1,9	1,9
J Intermediazione monetaria e finanziaria	4,3	4,3	-5,2	5,6
K Altre attività professionali ed imprenditoriali	0,4	0,7	-0,6	0,8
C-K TOTALE	0,5	0,6	0,7	0,7

Fonte: Elaborazioni su dati della rilevazione Oros

Dall'analisi della stima basata sul campione totale "perfettamente imputato" si può concludere che tale campione deriva da un processo di selezione non ignorabile (le risposte mancanti sono NMAR) e il processo di riduzione del campione, effettuato per eliminare l'errore di misura delle variabili trimestrali, accentua tali caratteristiche NMAR.

Una buona procedura di imputazione delle unità con pattern incompleto, tale da

ricostruire un campione totale con errori di misura poco rilevanti, ridurrebbe l'errore di stima, ma probabilmente non riuscirebbe ad eliminarne la sistematicità. La strada alternativa di tentare di bilanciare il campione rispetto ai pattern di presenza, pur riducendo la dimensione del campione appare più promettente sia nei termini di abbassamento dell'errore medio, sia nei termini di rendere l'errore meno sistematico. Le varie strade seguite, tuttavia, mostrano la difficoltà di tradurre i miglioramenti negli errori sui livelli in miglioramenti negli errori sulle variazioni tendenziali.

9. Conclusioni e prospettive future

Questo lavoro documenta le sperimentazioni svolte nell'ambito degli studi per il rilascio di indicatori sull'occupazione dalla rilevazione Oros tese a ridurre l'entità dell'errore di stima preliminare. Questa esperienza si riferisce ad una situazione informativa poi superata dal nuovo scenario di disponibilità dei dati. Nonostante ciò, presentare queste analisi rimane enormemente rilevante sia per la stessa rilevazione Oros, sia più in generale per rilevazioni che rilasciano stime preliminari basate su rispondenti rapidi, sia per il disegno di rilevazioni future basate su archivi amministrativi. Per quanto riguarda Oros, il nuovo scenario informativo, che vede un notevole incremento nella dimensione del campione in seguito ai recenti cambiamenti normativi, implica una necessaria revisione della metodologia di stima. Tuttavia, almeno due aspetti rendono ancora attuale la metodologia finora utilizzata. Da una parte, l'obbligo di abbreviare i tempi di rilascio dei dati, come previsto dai Regolamenti europei, può essere soddisfatto solo anticipando i tempi di acquisizione dei dati INPS, con inevitabile riduzione della quantità di informazioni disponibili, soprattutto relativamente all'ultimo mese del trimestre di riferimento. Dall'altra, non si possono escludere ne prevedere fenomeni di riduzione della numerosità delle dichiarazioni che pervengono in tempi rapidi, per motivi di natura amministrativa, con ovvie ricadute sul grado di copertura del campione. Entrambe queste circostanze potrebbero rendere necessaria una metodologia più simile a quella utilizzata sotto la vecchia situazione informativa.

L'importanza del presente lavoro per le rilevazioni congiunturali che rilasciano stime preliminari è duplice. Da un lato si mostrano i problemi e si discutono possibili soluzioni quando le stime possono essere basate solo su un sottoinsieme non casuale di rispondenti rapidi. La possibilità che questo insieme derivi da una selezione non ignorabile delle unità deve fare riflettere attentamente sulla sufficienza di metodi basati su variabili ausiliarie. Dall'altro, come si è visto, la riduzione dell'errore medio sui livelli non comporta necessariamente una riduzione dell'errore sulle variazioni, configurando un possibile trade off nella qualità tra questi due parametri. Metodi differenti possono collocarsi diversamente su rispetto a questo trade off e la scelta del metodo migliore dipende dall'importanza relativa che si attribuisce ai due parametri.

Infine l'utilità di questo lavoro per il disegno di future indagini basate su dati amministrativi è molteplice. La disponibilità di un registro anagrafico corrente apre la strada alla possibilità di stimare parametri relativi ad una popolazione non più riferita ad un tempo passato. D'altra parte la possibilità che il registro corrente non riesca a tenere conto degli eventi di cessazione fa si che metodi appropriati debbano essere utilizzati per non produrre stime distorte. Nella rilevazione Oros è stata colta la sfida derivante dall'informazioni disponibile studiando metodi e varianti per correggere la sovracopertura del registro. In questo lavoro sono state illustrate diverse metodologie per calcolare ed

applicare una probabilità di attività, usando al meglio l'informazione disponibile.

L'altra sfida metodologica consiste nel gestire un campione non casuale e molto variabile nel tempo. Diversi accorgimenti possono essere usati per migliorare le stime prodotte a partire da questi dati: in questo lavoro è stato mostrato che tecniche di bilanciamento del campione e/o metodi di imputazione delle mancate risposte possono andare in questa direzione.

Infine è importante citare come la disponibilità di dati di un campione preliminare e di dati finali su tutta la popolazione di riferimento consenta una notevole ricchezza nell'analisi dell'errore. Da questa possibilità è derivato uno dei contributi più rilevanti di questo lavoro, quello di scomporre l'errore di stima preliminare tra un errore dovuto alla sovracopertura della lista anagrafica e un errore dovuto al modello di riporto alla popolazione di riferimento. Questa scomposizione ha permesso di identificare le aree di intervento. Le metodologie proposte consentono tutte un notevole abbattimento dell'errore sulla stima dei livelli. Ulteriore lavoro di analisi sarebbe comunque necessario per modulare varianti che consentano anche una riduzione significativa dell'errore nelle variazioni.

Riferimenti bibliografici

Abbate C., Garofalo G. (1997) "Use of Integrated Administrative Sources in Order to Improve the Quality of Business Register Statistics", *Proceedings of the International Workshop "Use of Administrative Sources for Statistical Purposes"*, EUROSTAT, Luxembourg.

Baldi C., Ceccato F., Pacini S. e Tuzi D. (2005) La stima anticipata Oros sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento. Collana *Contributi Istat* n.13, Istat, Roma.

Baldi C., Ceccato F., Congia M.C., Cimino E., Pacini S., Rapiti F., Tuzi D. (2004) "Use of Administrative Data for Short Term Statistics on Employment, Wages and Labour Cost". *Essays* n. 15, Istat, Roma.

Baldi C., Falorsi P. D., Pallara A., Succi R., e Russo A. (2001) "A method for short-term estimation of labour input using current preliminary data from administrative sources having coverage errors", *Proceedings of Statistics Canada Symposium* 2001.

Deville J.C. e Särndal C.E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, 376-382.

Falorsi P. D., Pallara A., Russo A. e Succi R. (2003) "Utilizzo dei dati di fonte amministrativa per la stima congiunturale di occupazione e retribuzioni". *Temi di ricerca ed esperienze sull'utilizzo a fini statistici di dati di fonte amministrativa*. Franco Angeli, Milano.

Istat (2005) Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzione nelle grandi imprese. Collana *Metodi e norme*, Istat, Roma; in corso di pubblicazione.

Little R.J.A, Rubin D.B. (1987) *Statistical Analysis with missing data*. Wiley and Sons.

Rubin, D.B. (1976) Inference and missing data. *Biometrika* 63, 581-592.

Investimenti pubblici e sostenibilità: decidere meglio con la contabilità ambientale

Raffaello Cervigni[1], Cesare Costantino[2], Federico Falcitelli[3], Aldo Maria Femia[4], Aline Pennisi[5], Angelica Tudini[6]

Sommario

Qual è lo scenario di variazione delle emissioni di inquinanti che si può prefigurare a fronte della crescita di determinati settori economici? E in che misura i settori produttivi più rilevanti e più dinamici dell'economia regionale dipendono dalla disponibilità di risorse naturali? I territori con un maggiore inquinamento e degrado sono anche quelli che spendono di più per la protezione dell'ambiente?

Queste sono soltanto alcune delle domande chiave utili nella definizione di strategie di sviluppo di un territorio. Informazioni derivate dalla contabilità ambientale, specialmente se disaggregate a livello regionale, consentono non solo di fornire risposte a tali domande ma anche di quantificare le sinergie e i trade-off tra variabili economiche e ambientali, di cui le politiche di investimento ed incentivazione devono tenere conto. Tramite l'integrazione di informazioni economiche ed ambientali in un quadro standardizzato rispondente ai criteri di contabilità nazionale, i Conti Ambientali consentono infatti di condurre analisi sistematiche delle interazioni tra economia e ambiente con un notevole valore aggiunto rispetto ad altre fonti statistiche. In questo lavoro, frutto della collaborazione tra il Ministero dell'Economia e delle Finanze (Dipartimento per le Politiche di Sviluppo e coesione - Unità di valutazione degli investimenti pubblici) e l'Istat (Direzione Centrale della Contabilità Nazionale), si presentano possibili usi della contabilità ambientale per aiutare i decisori a scegliere quali territori, quali settori economici e quali comparti ambientali privilegiare nelle politiche di sviluppo e in che misura, utilizzando esempi desunti dai dati disponibili e vagliando le priorità per la produzione di Conti Ambientali a livello regionale.

Abstract

What is the possible future scenario in terms of variation in the emissions of pollutants that can arise due to a given level of growth in certain economic sectors? To what extent do the economy's key economic sectors depend on the availability of the various natural resources? Are the territories with the greatest pollution and degradation the same ones that spend more on environmental protection?

These are only some of the key questions to ask in the process of defining the development strategies of a given territory. The statistical information derived from environmental accounting, especially if developed at regional level, would allow not only to find answers

[1] Senior Economist, (TheWorld Bank), email: rcervigni@worldbank.org; in precedenza, Unità di Valutazione degli Investimenti Pubblici (DPS-MEF)

[2] Dirigente di ricerca (Istat), e-mail: cecostan@istat.it

[3] Ricercatore (Istat), e-mail: falcitel@istat.it

[4] Ricercatore (Istat), e-mail: femia@istat.it

[5] Componente dell'Unità di Valutazione degli Investimenti Pubblici (DPS-MEF) e-mail: aline.pennisi@tesoro.it

[6] Ricercatore (Istat), e-mail: tudini@istat.it

to the questions above, but also to quantify the positive interactions as well as trade-offs between economic and environmental objectives, that development policies need to take into account. Environmental accounting, by comprehensively integrating information on economics and the environment in a standardised accounting framework consistent with national accounts, provide a significant value added compared to other statistical sources in the thorough analysis of the interaction between the economy and the environment. This paper, which is the result of joint work carried out by the Department of Development Policies (Public Investment Evaluation Unit) of the Ministry of Economy and Finance and the National Accounts Directorate of the Italian National Statistics Institute, illustrates how environmental accounting can help decision-makers choose which territories, economic activities and sectors of the environment should be supported and to what extent, by using examples from available data and setting priorities for the production of environmental accounts at the regional level.

Parole chiave: contabilità ambientale, politiche di sviluppo, Conti Ambientali regionali, sostenibilità

1. Introduzione

L'integrazione della dimensione ambientale nella programmazione e pianificazione dello sviluppo è un'esigenza sempre più sentita nell'ambito comunitario e nazionale. Questo deriva dalla necessità, da una parte, di tener conto di impegni specifici di tutela ambientale a livello internazionale (come la riduzione delle emissioni di gas ad effetto serra prevista dal Protocollo di Kyoto) e, dall'altra, di garantire in modo adeguato la disponibilità delle risorse naturali ed un'elevata qualità dell'ambiente - fattori indispensabili per lo svolgimento delle attività economiche, per la competitività del territorio e per la salute e il benessere dei cittadini.

Nell'ottica di una maggiore attenzione alla sostenibilità ambientale degli interventi pubblici è stato previsto, ad esempio, l'inserimento di criteri ambientali nei regimi di sostegno alle attività produttive finanziati dai Fondi Strutturali Comunitari nel periodo 2000-2006. Tuttavia, non è sempre facile definire nella pratica criteri ambientali stringenti. Da una indagine della Rete Nazionale delle Autorità Ambientali e delle Autorità della Programmazione dei Fondi Strutturali Comunitari[7] è emerso che su novantanove regimi di aiuto a diretta o indiretta finalità ambientale, che risultavano alla fine del 2003 co-finanziati con Fondi Strutturali , il grado di specificità degli incentivi è stato piuttosto limitato: circa l'80 per cento delle risorse sono state messe a bando tramite regimi di aiuto non particolarmente focalizzati né dal punto di vista del tema ambientale affrontato, né dal punto di vista del soggetto destinatario (tipologia settoriale e dimensionale dell'impresa). Il rischio è dunque quello di non riuscire ad agire efficacemente sui segmenti dell'economia che esplicano una maggiore pressione sull'ambiente e di non rispondere ai bisogni di protezione ambientale più rilevanti per un dato territorio.

Un requisito essenziale per contribuire a rendere più mirati gli strumenti di sviluppo territoriale è la disponibilità di un supporto informativo adeguato ai fini della valutazione

[7] Cfr. Rete Nazionale delle Autorità Ambientali e delle Autorità della Programmazione dei Fondi Strutturali Comunitari 2000-2006 (2004).

delle implicazioni ambientali delle politiche di sviluppo. I Conti Ambientali, tramite la standardizzazione delle informazioni ambientali e di quelle economiche secondo i criteri della contabilità nazionale, consentono già oggi di condurre analisi sistematiche delle interazioni tra economia e ambiente a livello nazionale, e di confrontare le *performance* di diversi paesi. Tuttavia, per fornire indicazioni utili per le politiche di sviluppo in un paese in cui la struttura economica, l'avanzamento tecnologico e il patrimonio naturale sono molto diversificati e disomogenei, è importante identificare le differenze territoriali nei fenomeni di interazione tra economia e ambiente, almeno a livello regionale. La programmazione regionale degli interventi di sviluppo rappresenta, d'altronde, una parte consistente di tutta la programmazione per lo sviluppo, motivando ulteriormente l'urgenza di strumenti dettagliati su questa scala.

In questo lavoro si individuano le principali potenzialità di utilizzo della contabilità ambientale per le politiche di sviluppo, evidenziando il valore aggiunto di queste informazioni rispetto ad altre tipologie di statistiche. Le informazioni fornite dai Conti Ambientali possono infatti aiutare i decisori a scegliere quali territori, settori economici e comparti ambientali privilegiare e in che misura, fornendo una quantificazione delle sinergie e dei *trade-off* esistenti tra obiettivi ambientali ed economici.

Il lavoro è espressione di una collaborazione tra il Ministero dell'Economia e delle Finanze (Dipartimento per le Politiche di Sviluppo e coesione - Unità di valutazione degli investimenti pubblici) e l'Istat (Direzione Centrale della Contabilità Nazionale), che hanno realizzato nel corso del 2004 e del 2005 un progetto congiunto finalizzato all'ampliamento della base informativa utilizzata ai fini del disegno e valutazione delle politiche di sviluppo con dati che consentano di tenere conto in modo appropriato dell'interazione tra fenomeni ambientali ed economici. Oltre ad alcuni aggregati di contabilità ambientale a scala regionale (relativi in particolare al Lazio), il progetto ha prodotto una riflessione metodologica – ripresa in questa sede – circa il potenziale di utilizzo di Conti Ambientali regionali per il disegno e la valutazione delle politiche di sviluppo[8].

Una discussione ad ampio spettro sugli usi possibili della contabilità ambientale è stata avviata dalla comunità internazionale con il manuale "*Integrated Environmental and Economic Accounting 2003 (SEEA 2003)*", adottato dalla Commissione Statistica dell'ONU (sul quale si tornerà in seguito). Il presente lavoro inquadra in modo specifico il potenziale di utilizzo dei Conti Ambientali in relazione all'esigenza di meglio informare le politiche di sviluppo.

Per meglio mettere in luce tale potenziale vengono preliminarmente proposti una schematizzazione funzionale delle politiche di sviluppo (§ 2) e un quadro delle principali tipologie di "domande" che il *policy maker* può porsi in relazione ai diversi tipi di scelte da effettuare nella fase di disegno delle politiche stesse (§ 3). Attraverso una serie di esempi basati prevalentemente sui dati disponibili a livello nazionale viene illustrato il contributo dei Conti Ambientali in termini di possibili "risposte" alle domande del *policy maker* (§ 4). Successivamente, ampliando il quadro degli aspetti trattati attraverso gli esempi, viene offerta una panoramica più generale e sistematica delle principali "domande" del *policy maker* che possono trovare risposta nei Conti Ambientali; in tal modo viene proposta una

[8] Ministero dell'Economia e delle Finanze – Istat (2005), *Ambiente e politiche di sviluppo: le potenzialità della contabilità ambientale per decidere meglio* (Materiali UVAL Numero 5 – Anno 2005, Roma, http://www.dps.tesoro.it/materialiuval/ml.asp).

visione più ampia dei potenziali usi dei Conti Ambientali per le politiche qui considerate, di particolare ausilio, in questa fase, per l'individuazione di priorità nello sviluppo di Conti Ambientali (§ 5). In tale ottica vengono quindi effettuate alcune brevi riflessioni conclusive, con particolare riferimento all'opportunità di valorizzare i dati oggi disponibili e di ampliare l'offerta di Conti Ambientali, estendendo le stime alla scala regionale e a temi ambientali non ancora sufficientemente coperti (§ 6).

2. Una schematizzazione funzionale delle politiche di sviluppo e del ruolo dell'ambiente

Con il termine politiche di sviluppo ci si può riferire, in senso lato, all'insieme di decisioni delle autorità di governo volte a influenzare, in modo diretto o indiretto, la conservazione e aumento dello *stock* di capitale produttivo (pubblico e privato) di una data collettività. Implicita in questa accezione vi è la nozione che lo sviluppo - inteso come aumento nel tempo del benessere economico della collettività - non possa verificarsi in assenza di un'adeguata allocazione di risorse alla conservazione e all'aumento della capacità del sistema produttivo di generare reddito, la quale è a sua volta collegata alla quantità e qualità di beni capitali (materiali e immateriali) utilizzabili dagli agenti economici.

Alcune precisazioni sono opportune, in vista della considerazione della variabile "ambiente", ossia delle implicazioni ambientali delle politiche di sviluppo. Questa infatti comporta ai fini del presente lavoro il riferimento ad un'accezione più ampia del concetto di benessere, comprendente non solo aspetti strettamente economici, ma anche ad esempio l'esistenza di opportunità di fruizione di risorse naturali o culturali cui non corrispondono transazioni di mercato. Ciò a sua volta corrisponde, ai fini del presente esercizio, all'adozione di un'accezione ampia di capitale, tale da comprendere capitale manufatto, capitale umano, capitale naturale, capitale di conoscenza e capitale sociale[9]. Il capitale naturale, in particolare, viene definito essenzialmente sulla base di tre funzioni, cruciali per la sostenibilità ambientale dello sviluppo, che ad esso si possono ascrivere:

- fornire materie prime e risorse per i processi di produzione e di consumo (*resource functions*);
- assorbire i residui dei processi di produzione e di consumo (*sink functions*);
- fornire l'habitat a tutte le specie viventi inclusa l'umanità (*service functions*)[10].

Nella misura in cui tali funzioni sono riconducibili alla nozione di "capacità produttiva", la salvaguardia e l'espansione del capitale naturale costituiscono elementi

[9] Gli organismi internazionali hanno adottato concetti, definizioni e classificazioni a proposito del capitale, che costituiscono *standard* della statistica ufficiale; nel presente documento si fa puntualmente riferimento a tali *standard* per quanto concerne in particolare quelle componenti del capitale che sono più direttamente connesse con la sostenibilità ecologica dello sviluppo. Ad esempio, per capitale manufatto si intende ciò che nel Sistema europeo dei conti (SEC 1995) viene denominato 'attività non finanziarie prodotte' definite come '…le attività non finanziarie che sono state ottenute quale prodotto dei processi di produzione' (cfr. SEC 1995, paragrafo 7.14 e Tavola 7.1). Sono compresi: il capitale fisso, le scorte e gli oggetti di valore. Il capitale fisso, in particolare, comprende ad esempio abitazioni, fabbricati non residenziali, il software, ecc..

[10] Per la definizione di capitale naturale il riferimento metodologico è il manuale SEEA2003 ("Integrated Environmental and Economic Accounts 2003"); cfr. United Nations et al. in via di pubblicazione; cfr. in particolare paragrafo 1.23.

di conservazione e aumento della capacità del sistema produttivo di generare reddito. Alla considerazione di tale aspetto si aggiunge inoltre quella delle capacità dell'ambiente di supportare la vita e di generare benessere direttamente e indipendentemente dalla capacità di generare reddito.

Presupposto delle politiche di sviluppo è che, in loro assenza, si verificherebbero condizioni tali da far diminuire la dotazione del capitale di interesse o da impedirne la accumulazione a ritmi adeguati. Laddove le politiche di sviluppo siano orientate in particolare alla riduzione dei divari di reddito (è il caso delle politiche finanziate da risorse "aggiuntive"), si riconosce che la sottodotazione di capitale, nelle sue varie forme, riguarda alcuni territori più di altri e che questo è alla base delle disparità di reddito e benessere tra territori.

In Italia, il complesso della politica per lo sviluppo si articola in due distinte componenti: politica ordinaria e politica regionale. Presentano entrambe un'articolazione territoriale, ma si distinguono per la finalità specifica e l'origine delle risorse finanziarie che le alimentano. La politica regionale, "aggiuntiva" rispetto agli interventi ordinari condotti dalle Amministrazioni centrali e regionali, è diretta in particolare a garantire che gli obiettivi di competitività siano raggiunti da tutti i territori regionali, anche dalle aree che presentano significativi squilibri economico-sociali, tramite una strategia di offerta fondata sul miglioramento dei servizi collettivi e accompagnata da interventi per la promozione diretta degli investimenti privati. È alimentata dalle risorse nazionali destinate alle aree sottoutilizzate determinate annualmente attraverso il Fondo per le aree sottoutilizzate (FAS) – stanziate dalla Legge Finanziaria e ripartite con Delibere Cipe – e dai Fondi Strutturali Comunitari dell'Unione Europea. Per il periodo 2000-2006 quest'ultima componente ammonta complessivamente a circa 18 miliardi di euro medi annui di assegnazioni per l'Italia, di cui oltre 14 miliardi destinati al Mezzogiorno. Il complesso della spesa in conto capitale, incluse anche le risorse ordinarie, è stimato pari a circa 390 miliardi di euro per lo stesso periodo (di cui circa 153 miliardi di euro per il Mezzogiorno)[11].

Nella Tabella 1 sono messi in evidenza alcuni elementi caratteristici tramite cui si possono declinare le politiche di sviluppo[12]: gli obiettivi delle politiche in termini territoriali e in termini settoriali (a seconda della forma di capitale che si vuole ampliare, rafforzare o conservare); i soggetti destinatari ovvero gli agenti su cui le politiche si prefiggono di agire; e, infine, gli strumenti o meccanismi tramite cui le stesse possono essere implementate. Una politica di sviluppo sarà, infatti, indirizzata ad alcuni territori piuttosto che ad altri; diretta ad ampliare e/o mantenere specifiche componenti dello stock di capitale (quello manufatto, umano o naturale) piuttosto che altre e/o a migliorare la composizione interna della forma di capitale prescelta; inoltre coinvolgerà alcuni soggetti piuttosto che altri. Gli strumenti in cui si può esplicitare l'intervento pubblico sono infine genericamente classificabili in strumenti di spesa, fiscali, di regolazione e di *capacity building*.

[11] Cfr. Ministero dell'Economia e delle Finanze, Dipartimento per le Politiche di Sviluppo e di Coesione, Rapporto Annuale 2004 e 2005 (capitolo III e capitolo IV).

[12] La Tabella 1 non vuole proporre una classificazione rigorosa delle politiche, ma un tentativo di astrazione per individuare i caratteri comuni dei processi di decisione delle politiche di sviluppo.

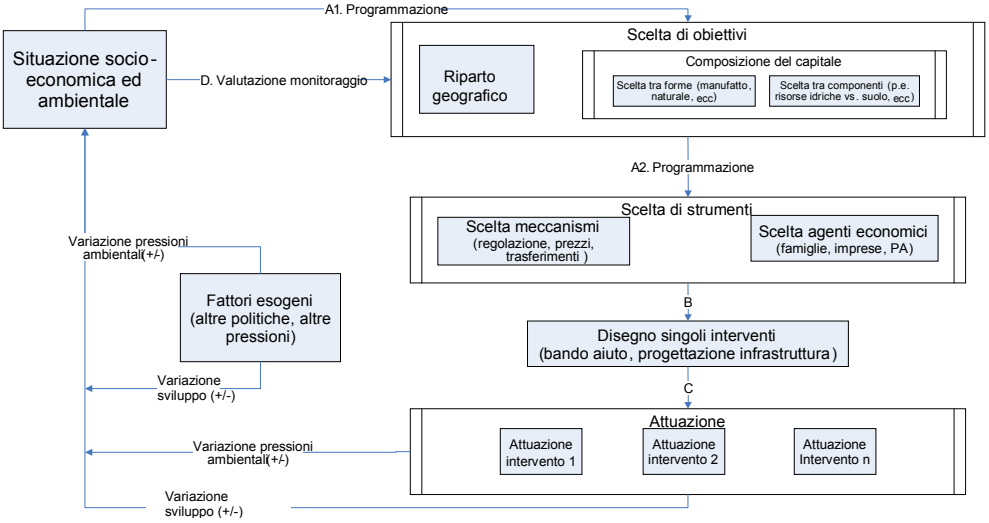
Tabella 1 – Caratteristiche delle politiche di sviluppo

gli obiettivi della politica	Incrementare la disponibilità della forma di capitale maggiormente utile o necessaria – decisione su quale componente dello stock di capitale ampliare e/o conservare: per esempio infrastrutture per i trasporti o impianti industriali (capitale manufatto), versus istruzione (capitale umano), versus qualità dei corpi idrici (capitale) Migliorare la composizione interna della forma di capitale prescelta – selezione di priorità all'interno di ogni forma di capitale: per esempio all'interno del capitale naturale: qualità dell'aria versus qualità dell'acqua; all'interno del capitale manufatto pubblico: strade versus ferrovie; all'interno del capitale umano: istruzione versus formazione professionale Ridurre i divari territoriali – decisione su dove spendere: in quali regioni, province, etc. in base alle priorità territoriali individuate (i territori con maggiore ritardo economico, con le risorse naturali più degradate, sottoposte a pressioni ambientali relativamente maggiori, con un'economia maggiormente dipendente dalle risorse naturali, caratterizzate da livelli di spesa per la protezione dell'ambiente relativamente minori)
i soggetti destinatari	le famiglie, imprese, amministrazioni pubbliche destinatarie dell'intervento pubblico o i soggetti dei quali si vuole cambiare il comportamento; e, all'interno di ciascuna tipologia, scelta dei sottoinsiemi di interesse (imprese manifatturiere versus imprese estrattive; giovani versus adulti, ecc.)
il meccanismo / strumento della politica	strumenti di spesa: allocare risorse pubbliche destinate direttamente all'ampliamento e/o alla conservazione dello stock di capitale selezionato, nei territori individuati come prioritari, dando preferenza a particolari componenti della forma di capitale prescelta strumenti fiscali: strutturare il sistema fiscale (basi imponibili, aliquote impositive, regime di esenzioni, ecc.), per influenzare le scelte private di mantenimento e/o formazione dello stock di capitale in modo da incoraggiare (o scoraggiare) investimenti privati in determinate forme di capitale e loro componenti, in determinati territori strumenti di regolazione: ricorrere a strumenti normativi di regolamentazione del mercato per indurre le variazioni dei comportamenti privati necessarie ad ampliare (o ridurre) investimenti in determinate forme di capitale e loro componenti, in determinati territori rafforzamento della capacità tecnica e amministrativa delle pubbliche amministrazioni: nella misura in cui ciò influisce sulla quantità e soprattutto sulla qualità della spesa pubblica e privata per investimenti

Fonte: Ministero dell'Economia e delle Finanze – Istat (2005).

Nella Figura 1 è riportata una possibile generalizzazione di come il processo decisionale delle politiche di sviluppo si può articolare nelle diverse tipologie di scelte allocative.

Figura 1 – Schema delle decisioni allocative e del ciclo delle politiche di spesa per lo sviluppo



Fonte: Ministero dell'Economia e delle Finanze – Istat (2005).

3. Le domande del *policy-maker*

In questo lavoro l'attenzione è rivolta prevalentemente alle politiche di spesa. Sotto questo profilo, in coerenza con l'accezione di politiche di sviluppo introdotta in precedenza (§ 2), l'implementazione di una strategia di sviluppo richiede delle scelte su come allocare le risorse finanziarie disponibili in modo da determinare o favorire quelle forme di accumulazione di capitale utili a perseguire - nel modo più efficace - obiettivi di reddito così come di benessere. Il compito del *policy maker* è pertanto quello di stabilire quante risorse assegnare:

(a) ai diversi territori (regioni, province o altro),

(b) alle diverse forme di capitale (materiale, immateriale, umano, naturale, ecc..) ed eventualmente alle loro componenti più specifiche,

(c) ai diversi soggetti (famiglie, imprese, amministrazioni pubbliche, settori produttivi).

Come ogni altra forma di attività antropica, le politiche di sviluppo generano, tra le altre cose, pressioni sull'ambiente (o riducono quelle esistenti), e modificano la disponibilità di risorse naturali in termini sia quantitativi sia qualitativi. Dato che nel medio e lungo periodo la consistenza quantitativa e qualitativa del capitale naturale è un elemento indispensabile per la produzione sia di reddito che di benessere in senso ampio, diventa essenziale assicurare che le decisioni allocative siano il più possibile informate anche sulle loro potenziali implicazioni ambientali. Tipicamente si è interessati a quantificare le sinergie e i *trade-off* delle possibili ripartizioni delle risorse e, in particolare, tra i risultati attesi dalle politiche in termini economici, sociali ed ambientali.

Ad esempio, se viene indotta la crescita di un determinato settore produttivo, sotto il profilo economico ci si chiederà qual è la variazione attesa in termini di occupazione e di reddito. D'altra parte ci si domanderà anche in che misura lo sviluppo di tale settore dipenda dalla disponibilità di risorse naturali (risorse energetiche, minerali, ecc.) e in che misura per soddisfare tale fabbisogno il sistema dipenda da altri territori. Inoltre, sarà importante sapere, a fronte dello sviluppo economico indotto nel settore, qual è la corrispondente variazione attesa in termini di emissioni di inquinanti, di produzione di rifiuti, ecc., nonché quali sono i maggiori utilizzatori delle risorse naturali a rischio di deterioramento, e in che misura possono vedersi ridotta la disponibilità quali-quantitativa di tali risorse. Ancora, in relazione a un dato obiettivo minimo di qualità dell'ambiente, è utile domandarsi quanto è necessario intervenire con un'azione di protezione dell'ambiente a fronte di un insufficiente sforzo in tal senso da parte delle attività produttive (miglioramenti tecnologici e altre spese) e delle famiglie.

Tenendo conto delle caratteristiche delle politiche di sviluppo sopra individuate si possono distinguere altrettante tipologie di domande utili per contribuire a determinare la quantità di risorse da allocare ai diversi territori, alle diverse forme di capitale e ai diversi soggetti economici nell'intento di raggiungere gli obiettivi assunti in relazione all'economia, all'ambiente e al sociale. Un'esemplificazione di tali domande è riportata in Tabella 2.

Tabella 2 – Tipologie di ripartizioni e di domande

Tipologie di ripartizioni	Esempi di domande del *policy maker*
Ripartizione territoriale	(1.1) Quali e quante risorse naturali utilizza l'economia dei diversi territori, e in che misura questi dipendono da altri territori per l'approvvigionamento? (1.2) Ci sono tra i vari territori differenze significative nella disponibilità delle varie risorse naturali e nel loro stato qualitativo? (1.3) Tra le attività produttive, quali sono quelle che più contribuiscono all'emissione di determinati inquinanti nei vari territori? (1.4) I territori con un maggiore inquinamento e degrado sono anche quelli che spendono di più per la protezione dell'ambiente?
Ripartizione tra forme di capitale e componenti	(2.1) Quanta parte dei materiali utilizzati nei processi di produzione e consumo si trasforma in residui dannosi per l'ambiente e quanta in capitale manufatto? (2.2) In quale misura il prelievo delle risorse naturali serve a soddisfare il fabbisogno dei settori economici cruciali dell'economia? (2.3) Quanto spendono le imprese, le famiglie e le Amministrazioni pubbliche per la protezione dell'ambiente e quanto incide tale spesa sul totale della spesa di ciascuna di queste tipologie di operatori? Su quali comparti ambientali si concentra la spesa?
Ripartizione tra destinatari	(3.1) Quanta parte dell'inquinamento è generato dai consumi delle famiglie e quanta invece dalle attività produttive? (3.2) In che relazione sono la *performance* economica e quella ambientale delle varie attività produttive? (3.3) Le attività produttive più inquinanti sono anche quelle che spendono di più per la protezione dell'ambiente?

Fonte: Ministero dell'Economia e delle Finanze – Istat (2005).

Si noti che le domande rimandano alle scelte allocative (tra territori, tra forme di capitale/settori e loro componenti, tra destinatari) effettuate essenzialmente nella fase di programmazione delle politiche di sviluppo. Ma le eventuali risposte possono essere altrettanto utili nella fase di attuazione per definire *benchmark* sulla base dei quali stabilire criteri di assegnazione delle risorse ai diversi soggetti dell'economia nei vari territori o per selezionare interventi che assicurino una maggiore sostenibilità ambientale (tramite criteri di eligibilità o di premiazione nei bandi di gara).

4. Risposte dalla contabilità ambientale

Un adeguato supporto informativo alle scelte allocative della spesa per lo sviluppo richiede numerosi dati sul sistema economico e sull'ambiente, ma soprattutto sulla loro interazione. La contabilità ambientale è la branca dell'informazione statistica ufficiale che descrive in modo sistematico e comprensivo le interrelazioni tra economia e ambiente attraverso una pluralità di conti, standardizzati in ambito internazionale[13]. La Tabella 3 sintetizza le caratteristiche dei principali tipi di Conti Ambientali maggiormente definiti nel contesto internazionale, mettendo in evidenza come ciascuno sia focalizzato su aspetti specifici del rapporto tra sistema naturale e sistema antropico[14].

[13] Un importante punto di riferimento a livello internazionale per comprendere i contorni e i contenuti della contabilità ambientale è il già citato manuale "Integrated Environmental and Economic Accounting 2003 (SEEA 2003)", adottato dalla Commissione Statistica dell'ONU e predisposto congiuntamente da Nazioni Unite, Unione Europea, Fondo Monetario Internazionale, Banca Mondiale e Organizzazione per la Cooperazione e lo Sviluppo Economico (cfr. United Nations et al., op. cit., http://unstats.un.org/unsd/envAccounting/seea2003.pdf).

[14] Per una panoramica dettagliata dei vari conti definiti in ambito internazionale e per la relativa manualistica di riferimento si rinvia a Ministero dell'Economia e delle Finanze – Istat (2005), op. cit., Appendice 2. Oltre ai conti già adottati e in qualche modo standardizzati nella statistica ufficiale, esiste una varietà di strumenti contabili e analitici, molti dei quali già ampiamente in uso nella ricerca economico-ecologica (ad esempio, nell'ambito della MFA, la Substance Flow Analysis e i Life Cycle Inventories).

Tabella 3 – Principali tipi di Conti Ambientali e rispettive finalità

Tipo di conto	Principale finalità
Conti dei flussi di materia dell'intera economia (EW-MFA: Economy-wide - Material Flow Accounts)	Costruzione di un bilancio complessivo, a livello di intera economia, degli scambi di materia tra il sistema antropico e il sistema naturale, ai fini dell'analisi dell'utilizzo delle risorse naturali in relazione all'andamento dell'economia
Conti dei flussi di tipo NAMEA (National Accounts Matrix including Environmental Accounts)	Registrazione dei flussi fisici intercorrenti tra economia e ambiente (emissioni atmosferiche, uso e inquinamento dell'acqua, uso dell'energia, ecc.) e associazione degli stessi alle attività economiche che li determinano, in corrispondenza con le rispettive grandezze economiche (produzione, val. aggiunto, occupazione, ecc.)
Conti economici dell'ambiente (SERIEE)	Registrazione delle transazioni economiche connesse all'ambiente (spese per la protezione dell'ambiente – EPEA/ Environmental Protection Expenditure Account - e per l'uso e la gestione delle risorse naturali – RUMEA/ Resource Use and Management Expenditure Account -, tasse ambientali, ecc.) e descrizione delle attività economiche che producono beni e servizi per l'ambiente (anche dette "eco-industrie")
Conti patrimoniali fisici delle risorse naturali	Costruzione di bilanci patrimoniali in termini fisici per le diverse risorse naturali (stock ad inizio e a fine periodo, variazioni intercorrenti nel periodo dovute a cause naturali o antropiche; si tiene conto anche della qualità delle risorse con opportuni indicatori e/o articolando i bilanci per classi di qualità)

Fonte: Ministero dell'Economia e delle Finanze – Istat (2005).

A livello italiano l'Istat, che partecipa attivamente alla definizione del quadro di riferimento metodologico internazionale, produce regolarmente aggregati a scala nazionale relativi ad alcuni dei principali Conti Ambientali definiti appunto nel contesto internazionale. Si tratta dei Conti dei flussi di materia dell'intera economia, dei conti di tipo NAMEA delle emissioni atmosferiche e del prelievo di risorse naturali vergini, e del conto satellite EPEA delle spese per la protezione dell'ambiente[15]. Per la scala regionale l'Istat ha realizzato la serie storica 1995-2001 delle spese per la protezione dell'ambiente sostenute dall'Amministrazione regionale del Lazio e i conti NAMEA delle emissioni atmosferiche per la regione Lazio e per la regione Toscana, relativi all'anno 2000[16].

La caratteristica comune di questi conti è la possibilità di confrontare i fatti economici e i fatti ambientali correlati, consentita dalla stretta connessione stabilita con i conti economici nazionali. Tale confrontabilità è realizzata attraverso l'uso di un sistema comune di principi, definizioni e classificazioni (coerente con quello di contabilità nazionale), grazie al quale le informazioni relative a differenti dimensioni (economica, ambientale e in taluni casi anche sociale) sono riferite ad entità univocamente identificate. Ciò rende l'uso che può essere fatto dei Conti Ambientali sostanzialmente simile a quello che può essere fatto in generale dei conti economici: da un parte è possibile derivarne indicatori significativi, dall'altra è possibile approfondire i meccanismi di interrelazione tra le variabili e tra soggetti e, infine, è possibile utilizzare gli aggregati per costruire modelli di simulazione e scenario.

Grazie a questa loro peculiarità gli strumenti di contabilità ambientale possono contribuire in modo significativo alla rappresentazione della realtà sulla quale il decisore vuole agire:

[15] I dati sono disponibili sul sito dell'Istat (http://www.istat.it/conti/ambientali/).

[16] I dati 2001 della spesa per la protezione dell'ambiente e i dati NAMEA per il Lazio sono stati prodotti nell'ambito del progetto congiunto dell'Istat (Direzione centrale della contabilità nazionale) e del Ministero dell'Economia e delle Finanze (Dipartimento per le Politiche di Sviluppo e coesione – Unità di Valutazione degli Investimenti Pubblici), denominato "Contabilità ambientale e sviluppo". I dati NAMEA per la regione Toscana sono stati prodotti nell'ambito di una Convenzione Istat-IRPET (Istituto Regionale per la Programmazione Economica della Toscana). I dati NAMEA per la Toscana e per il Lazio non sono confrontabili per l'eterogeneità dei dati di base relativi alle emissioni in atmosfera utilizzati come input.

- *direttamente*, fornendo il quadro delle evidenze disponibili sulle interazioni tra economia e ambiente;
- *indirettamente*, fornendo input informativi per la costruzione/verifica di ipotesi sulle relazioni di causa-effetto, e per la stima degli effetti delle *policy* sui sistemi economico e ambientale.

I paragrafi seguenti presentano alcuni esempi di come i dati dei Conti Ambientali maggiormente sviluppati in Italia forniscano interessanti indicazioni per rispondere ad alcune tipiche domande del *policy maker*. Vengono affrontate soltanto alcune delle domande della Tabella 1, rimandando al § 5 per una trattazione più ampia[17].

4.1 Conti dei Flussi di Materia dell'Intera Economia

Domanda (1.1) Quali e quante risorse naturali utilizza l'economia e in che misura questa dipende da altri territori per l'approvvigionamento?

Il principale prodotto dei Conti dei flussi di materia dell'intera economia (EW-MFA) consiste nel bilancio complessivo degli scambi di materia tra il sistema economico e il sistema naturale[18]. Tipici indicatori derivati dalla EW-MFA (il cui calcolo non necessita la costruzione dell'intero bilancio) sono *l'Input materiale diretto* (DMI – *Direct Material Input*) e il *Consumo materiale diretto* (DMC – *Direct Material Consumption*). Il primo comprende tutti i materiali estratti nel paese e destinati all'utilizzo, unitamente ai materiali contenuti nelle importazioni. Il secondo esclude i materiali esportati e rappresenta, in sostanza, la quantità di materia che dopo le trasformazioni subite nel sistema economico rimane incorporata in beni d'investimento e durevoli o viene restituita all'ambiente naturale in forma degradata.

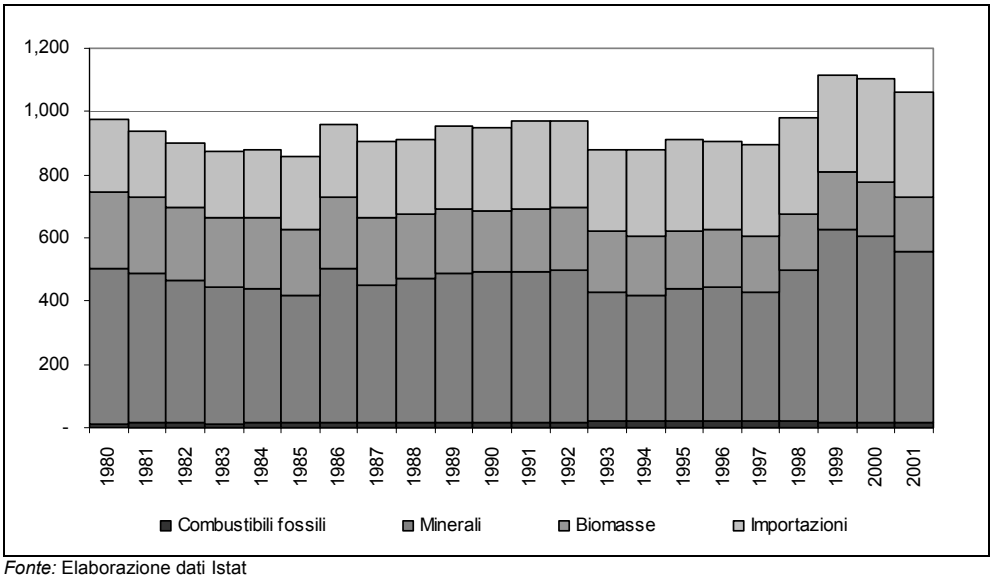
Questi indicatori forniscono una idea della dimensione fisica complessiva dell'economia con particolare riferimento al fabbisogno di risorse naturali, e possono essere analizzati a diversi livelli di aggregazione/disaggregazione delle componenti. La Figura 2, che mostra l'andamento dell'input materiale diretto dell'economia italiana dal 1980 al 2001, illustra un primo livello di disaggregazione. Si può osservare come l'economia italiana nell'intero periodo considerato abbia complessivamente estratto dal proprio ambiente naturale circa 10 miliardi di tonnellate di risorse minerali[19] (dato che di per sé fornisce anche una prima misura, sebbene molto aggregata ed indicativa, dell'intensità dei processi di modificazione del territorio e del paesaggio attuati); come il prelievo di biomasse sul territorio interno si vada riducendo nel tempo; come la quantità di minerali energetici fornita dal territorio interno abbia un ruolo del tutto marginale; e infine come l'andamento di queste due ultime voci sia più che compensato dal ricorso crescente alle importazioni.

[17] Cfr. anche Ministero dell'Economia e delle Finanze – Istat (2005).

[18] Il set completo dei dati e degli indicatori derivati dai conti dei flussi di materia è periodicamente diffuso sul sito web dell'Istat (http://www.istat.it/conti/ambientali/).

[19] Il dato è calcolato cumulando il valore dell'input materiale diretto registrato per i vari anni dell'intervallo 1980-2001.

Figura 2 – Input materiale diretto dell'Italia. Anni 1980-2001 *(milioni di tonnellate)*



Fonte: Elaborazione dati Istat

Tutto ciò può essere messo in relazione con l'andamento dell'economia sia calcolando l'intensità dell'utilizzo di materia per unità di PIL – in netta diminuzione nel tempo, data la sostanziale stabilità del DMI a fronte della crescita economica verificatasi – sia facendo ricorso ad un ulteriore livello di disaggregazione, che mostrerebbe ad esempio come le importazioni sono cresciute per far fronte ad un fabbisogno crescente di minerali metalliferi, di materiali di origine organica per l'industria e di energia, sia nella forma di biomasse per l'alimentazione sia nella forma di combustibili fossili.

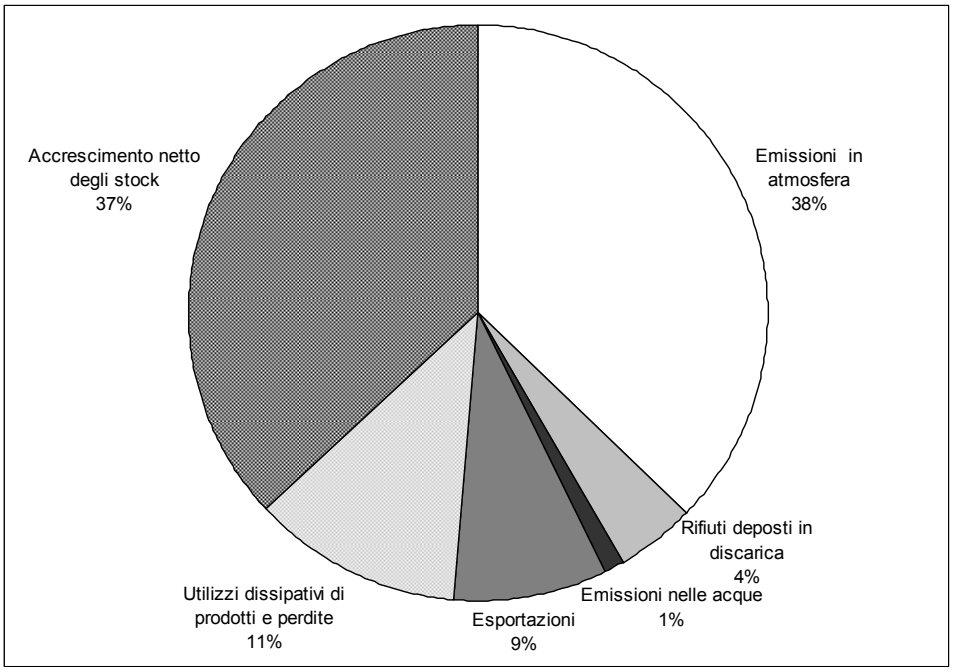
Domanda (2.1) Quanta parte dei materiali utilizzati si trasforma in residui dannosi per l'ambiente e quanta in capitale manufatto?

Il bilancio completo dei flussi di materia evidenzia oltre alla provenienza e alla composizione delle risorse utilizzate, la destinazione e la composizione del risultato fisico del loro utilizzo. Ad esempio, la Figura 3 presenta la composizione della materia utilizzata nell'anno 1997 al termine del ciclo annuale di trasformazione nelle attività di produzione e consumo. L'atmosfera si rivela essere la principale "pattumiera" delle attività antropiche, superando di poco il secondo destinatario della materia trasformata, rappresentato dal capitale fisico: l'accumulo netto[20] di stock, ossia la quantità di materia incorporata a lungo termine in edifici, infrastrutture, beni durevoli e scorte ha dimensioni paragonabili a quelle dei prelievi di minerali non energetici che sono stati evidenziati commentando la Figura 2. Si può apprezzare in tal modo la pressione esercitata sul territorio con l'utilizzo di tali risorse, aggiuntiva rispetto a quella che si genera nel momento del loro prelievo. Le

[20] Si intende lo stock fisico di materia incorporata nel sistema antropico in edifici, infrastrutture, beni durevoli e scorte, al netto dei decrementi subiti dallo stock stesso. Ad esempio demolizioni, beni durevoli e macchinari dismessi, animali macellati o comunque morti.

perdite e gli usi dissipativi deliberati (consistenti soprattutto nello spargimento di concimi sui suoli agricoli, con le note conseguenze per le acque) costituiscono la terza voce del conto, e sono superiori sia alle esportazioni di prodotti che alle deposizioni in discarica (cui pure corrispondono ulteriori pressioni sul territorio).

Figura 3 – Composizione dei materiali risultanti dall'impiego di risorse naturali da parte dell'economia italiana. Anno 1997 *(valori percentuali)*



Fonte: Elaborazione dati Istat

In sintesi, questi conti e indicatori illustrano quanto l'attività economica è dipendente dall'utilizzo diretto di risorse naturali e quanto da risorse importate da altri territori, permettendo di misurare l'efficienza con cui vengono impiegate tali risorse. Le considerazioni che derivano dalla loro analisi possono orientare le politiche di sviluppo ad agire in termini di maggiore o minore tutela delle risorse naturali interne, di riconversione dei sistemi produttivi, di supporto all'uso di tecnologie a minore consumo di materia, ad intervenire sulle perdite e sulla dissipazione, ecc.

Inoltre, tramite indicatori più completi non discussi in questa sede, anch'essi forniti dai conti dei flussi di materia, è possibile comprendere meglio i fenomeni di induzione dell'utilizzo di risorse e generazione di residui a livello inter-territoriale; tali indicatori, infatti, contabilizzano oltre alle pressioni (prelievi ed emissioni) generate sul territorio di riferimento anche tutte quelle associate ai flussi di materiali importati, che ricadono su altri territori.

4.2 Conti di tipo NAMEA

Domanda (3.1) Quanta parte dell'inquinamento è generato dai consumi delle famiglie e quanta invece dalle attività produttive?

Nei conti di tipo NAMEA le principali pressioni ambientali generate dalle varie attività produttive e dai consumi delle famiglie – misurate in unità fisiche – sono messe a confronto con i corrispondenti aggregati economici di contabilità nazionale. Per le attività produttive vengono dunque confrontati due differenti risultati congiunti della attività esercitata: da un lato i valori economici creati (produzione, valore aggiunto, occupazione) e dall'altro le pressioni sull'ambiente generate per creare tali valori (emissioni atmosferiche, rifiuti, prelievi diretti di risorse naturali vergini, ecc); in particolare, ad ogni attività economica vengono associate sia le pressioni direttamente causate dai processi produttivi tipici del settore, sia quelle generate dalle attività di supporto alla produzione (per esempio il trasporto in conto proprio e il riscaldamento degli ambienti di lavoro). Per le famiglie, le pressioni ambientali generate dai diversi consumi (per esempio le emissioni atmosferiche generate nel trasporto privato e per il riscaldamento delle abitazioni) vengono associate alle spese sostenute dalle famiglie stesse per acquistare i prodotti il cui uso è all'origine delle pressioni in questione (per esempio il combustibile).

Attraverso i dati della NAMEA è dunque possibile apprezzare il contributo delle attività produttive e delle famiglie alla generazione delle emissioni di un particolare inquinante. La Figura 4 mostra, ad esempio, nel contesto di una diminuzione complessiva delle emissioni di piombo in Italia tra il 1992 e il 2000, uno scambio di ruoli tra i responsabili delle emissioni stesse: mentre all'inizio del periodo l'apporto delle attività produttive alle emissioni complessive era superiore (seppur di pochi punti percentuali) a quello delle famiglie, nel tempo le famiglie sono diventate la fonte principale dell'inquinante in questione. Ciò è il risultato della dinamica differenziata delle emissioni generate da famiglie e imprese nel corso del periodo considerato: mentre le emissioni delle attività produttive hanno fatto registrare una riduzione molto significativa, per le famiglie la riduzione è risultata più contenuta[21].

Domanda (3.2) In che relazione sono la performance economica e quella ambientale delle varie attività produttive?

Grazie alla sua disaggregazione settoriale, la NAMEA fornisce informazioni dettagliate sull'interrelazione tra il ruolo svolto nell'economia dalle diverse attività produttive e il ruolo dalle stesse svolto nella generazione di pressioni sull'ambiente. Un esempio di analisi possibile a partire dai dati della NAMEA italiana è fornito dalla Figura 5, che mostra il contributo percentuale di ciascuna attività produttiva del settore manifatturiero alle emissioni totali di gas ad effetto serra[22] e al valore

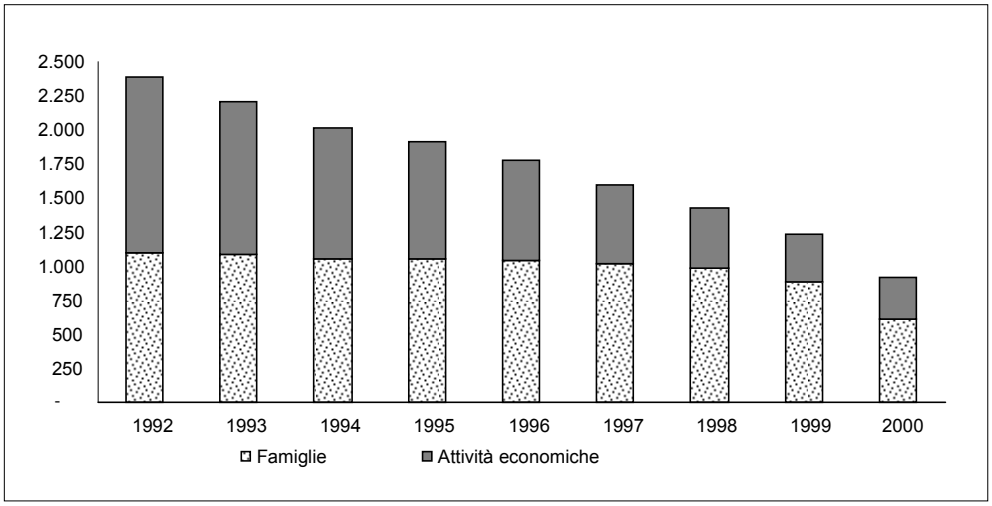
[21] Analisi analoghe a quelle della Figura 4 sono possibili per tutti gli inquinanti atmosferici per i quali sono disponibili i dati NAMEA: anidride carbonica (CO_2), protossido di azoto (N_2O), metano (CH_4), ossidi di azoto (NO_x), ossidi di zolfo (SO_x), ammoniaca (NH_3), composti organici volatili non metanici (NMVOC), monossido di carbonio (CO), piombo (Pb) e particolato (PM10).

[22] Per il calcolo delle emissioni ad "effetto serra" vengono sintetizzate con pesi adottati a livello internazionale le emissioni di CO_2, N_2O e CH_4, calcolando il cosiddetto G*lobal Warming Potential* (GWP), espresso in termini di tonnellate di CO_2 equivalente. Analisi analoghe a quelle della Figura 5 sono possibili con riferimento ai singoli inquinanti atmosferici per i quali sono disponibili i dati NAMEA (cfr. precedente nota 21) come pure con riferimento al tema ambientale dell'"acidificazione", per il quale vengono sintetizzate le emissioni di SO_x, NO_x e NH_3 con pesi adottati a livello internazionale, calcolando il cosiddetto *Potential Acid Equivalent* index (PAE) sulla base del numero di ioni idrogeno (H^+) per tonnellata di gas emesso.

aggiunto complessivo del settore[23].

Al di sopra della diagonale si trovano le attività economiche il cui contributo alle emissioni è maggiore rispetto al contributo al valore aggiunto. Considerando inoltre l'andamento nel periodo 1990 – 2000 del peso di ciascuna attività in termini di valore aggiunto totale e di emissioni complessive (non esplicitato nella figura), dai dati NAMEA si evince come alcune attività abbiano peggiorato la propria *performance* complessiva, poiché nel corso del tempo il loro peso economico all'interno del settore si è ridotto mentre è aumentata la quota di emissioni di cui sono responsabili (si tratta delle attività DA; DB e DF, contraddistinte da un motivo a quadretti nella Figura 5); al contrario, alcune attività hanno migliorato la propria *performance* in quanto a fronte di un aumento del proprio peso economico hanno registrato nel tempo una quota decrescente di emissioni (attività DI e DJ, contrassegnate con un motivo a righe nella Figura 5). Queste considerazioni, se non esauriscono un'analisi tesa a valutare l'evoluzione delle diverse attività produttive in termini di eco-efficienza, consentono quantomeno di individuare alcuni casi su cui è utile un approfondimento[24].

Figura 4 - Emissioni di piombo delle famiglie e delle attività economiche. Italia – Anni 1992-2000
(tonnellate)

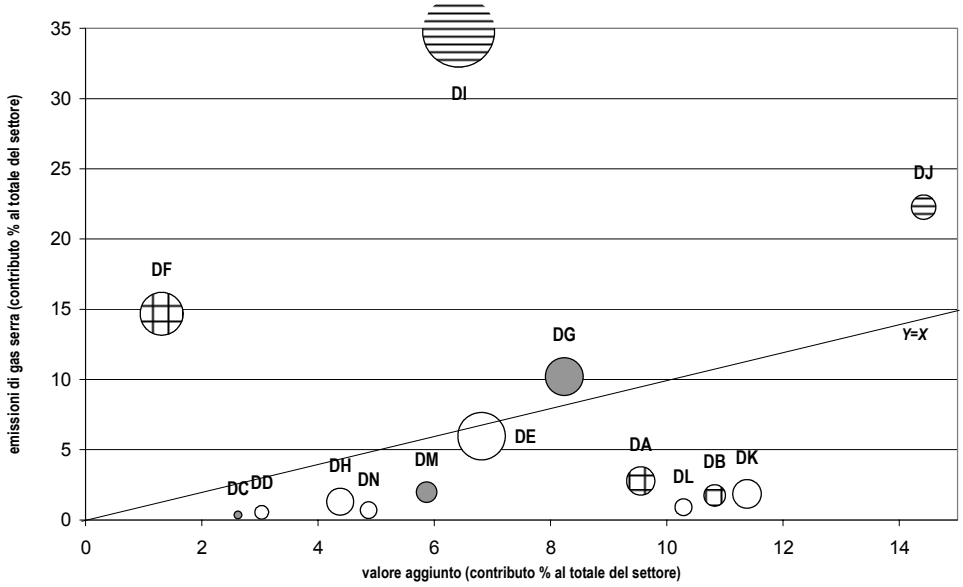


Fonte: Elaborazione dati Istat

[23] La quota del settore manifatturiero nel suo complesso sul valore aggiunto totale è del 21% circa e il contributo del settore alla emissione di gas serra da parte del totale delle attività economiche è del 25% circa.

[24] Una rappresentazione analoga può essere effettuata considerando altre variabili economiche, come la produzione o l'occupazione, e altri tipi di pressione ambientale (produzione di rifiuti, acque reflue, ecc.).

Figura 5 - Emissioni ad effetto serra e valore aggiunto nel settore manifatturiero per attività economica. Italia, Anno 2000 *(contributi percentuali al totale del settore)*



Legenda:

DA. Industrie alimentari, delle bevande e del tabacco; DB. Industrie tessili e dell'abbigliamento; DC. Industrie conciarie, fabbricazione di prodotti in cuoio, pelle e similari; DD. Industria del legno e dei prodotti in legno; DE. Fabbricazione della pasta-carta, della carta e dei prodotti di carta; stampa e editoria; DF. Fabbricazione di coke, raffinerie di petrolio, trattamento dei combustibili nucleari; DG. Fabbricazione di prodotti chimici e di prodotti chimici artificiali; DH. Fabbricazione di articoli in gomma e materie plastiche; DI. Fabbricazione di prodotti della lavorazione di minerali non metalliferi; DJ. Produzione di metallo e fabbricazione di prodotti in metallo; DK. Fabbricazione di macchine e apparecchi meccanici, compresi l'installazione, il montaggio, la riparazione e la manutenzione; DL. Fabbricazione di macchine elettriche e di apparecchiature elettriche e ottiche; DM. Fabbricazione di mezzi di trasporto; DN. Altre industrie manifatturiere.

Lungo la diagonale si registra l'uguaglianza tra i contributi in termini di valore aggiunto e di emissioni.

La simbologia utilizzata per il riempimento delle bolle si riferisce al risultato del confronto tra i valori delle due variabili nel 2000 e nel 1990:

quadretti - nel 2000 risulta diminuito il contributo al v.a. e aumentato il contributo all'emissione di gas serra
righe - nel 2000 risulta aumentato il contributo al v.a. e diminuito il contributo all'emissione di gas serra
pieno - nel 2000 risulta diminuito sia il contributo al v.a. sia il contributo all'emissione di gas serra
vuoto - nel 2000 risulta aumentato sia il contributo al v.a. sia il contributo all'emissione di gas serra.

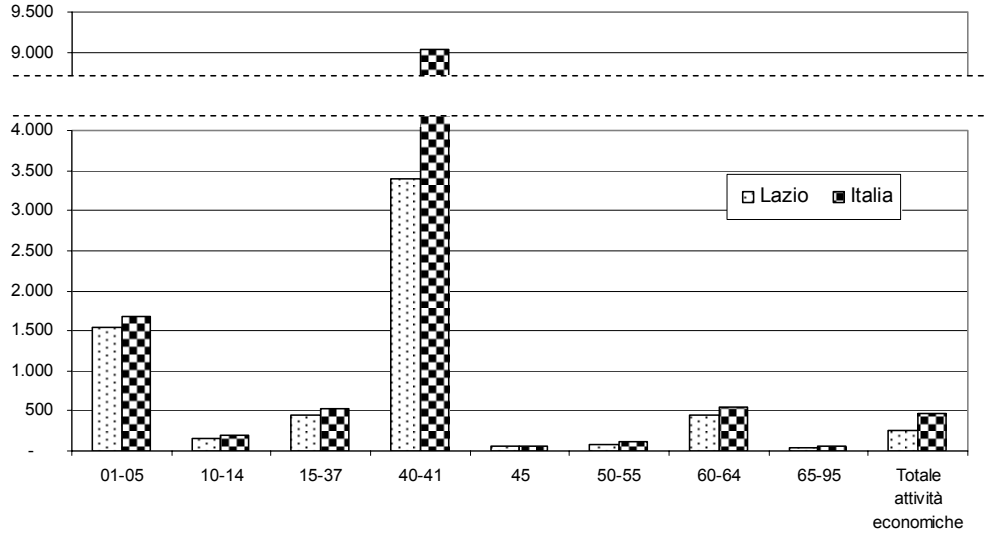
Fonte: Elaborazione dati Istat

Un altro tipo di utilizzo dei dati NAMEA che può contribuire a rispondere alla stessa domanda si ottiene attraverso il calcolo dell'"intensità di emissione", un indicatore di sintesi dato dal rapporto – per un dato inquinante e per una data attività economica – tra emissioni e produzione o valore aggiunto. Questo rapporto può essere utilizzato come un indicatore della efficienza di una data attività e servire come base per confronti nel tempo o nello spazio. Ad esempio dalla Figura 6, in cui viene confrontata l'efficienza delle attività produttive nel Lazio e in Italia[25], emerge come in Italia per ogni Meuro di valore aggiunto sono state generate nel 2000 in media circa 450 tonnellate di gas ad effetto

[25] Alternativamente è possibile effettuare confronti intertemporali relativi ad una stessa attività economica (una riduzione del rapporto nel tempo indica un aumento di efficienza e viceversa) o confronti tra attività diverse in uno stesso paese.

serra mentre nel Lazio l'intensità di emissione è stata significativamente inferiore (pari a circa 260 tonnellate per Meuro). La migliore efficienza dell'economia laziale in termini di intensità di emissione vale per tutti i principali settori di attività produttiva e, in particolare, per la "produzione di energia elettrica, gas e acqua", la cui intensità di emissione è pari a meno del 40% della corrispondente media nazionale. Queste differenze possono essere dovute alla diversa composizione interna dei settori produttivi (prevalenza di attività produttive più o meno inquinanti) ma anche alla diversa eco-efficienza delle tecnologie utilizzate. Informazioni di questo tipo sono utili per individuare le realtà territoriali che più contribuiscono al raggiungimento di obiettivi stabiliti in ambito internazionale per l'intero paese (come è il caso del Protocollo di Kyoto), e quelle per le quali la situazione attuale di intensità di emissione presenta maggiori margini di miglioramento.

Figura 6 - Intensità delle emissioni di gas ad effetto serra per raggruppamento di attività economica. Lazio e Italia – Anno 2000 *(tonnellate di CO_2 equivalente/milioni di euro di valore aggiunto)*



Legenda: 01-05 Agricoltura, silvicoltura e pesca; 10-14 Estrazione di minerali; 15-37 Attività manifatturiere; 40-41 Energia elettrica, gas e acqua; 45 Costruzioni; 50-55 Commercio, alberghi e ristoranti; 60-64 Trasporti, magazzinaggio e comunicazioni; 65-95 Altri servizi

Fonte: Elaborazione dati Istat

I dati presentati consentono di calcolare, a parità di tecnologie utilizzate, le quantità aggiuntive di emissioni (o di altri tipi di pressioni) direttamente generate dalla crescita di un dato settore. Inoltre, applicando ai dati della NAMEA l'analisi Input-Output si possono calcolare le pressioni ambientali delle attività produttive direttamente ed indirettamente connesse al soddisfacimento della domanda finale, rispondendo così a domande sulle cause ultime del degrado ambientale di notevole importanza per la politica ambientale[26].

[26] Alcune possibili applicazioni analitiche degli aggregati di tipo NAMEA, con esempi basati sui dati nazionali, sono illustrate in Femia A. – Panfili P. (2005).

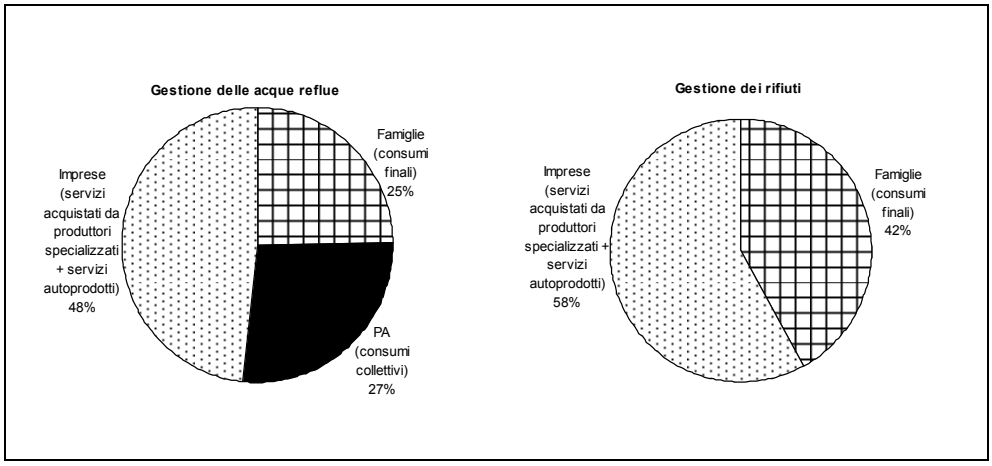
4.3 Conto satellite EPEA delle spese per la protezione dell'ambiente

Domanda (2.3) Quanto spendono le imprese, le famiglie e le Amministrazioni pubbliche per la protezione dell'ambiente?

Accanto alle pressioni ambientali, un altro aspetto dell'interazione economia-ambiente oggetto della contabilità ambientale è rappresentato dalle risposte del sistema socio-economico ai problemi ambientali, colte in particolare attraverso l'analisi delle spese per la protezione dell'ambiente, cui è dedicato il conto EPEA. Questo conto è finalizzato principalmente all'analisi della domanda e dell'offerta di servizi di protezione dell'ambiente (come la riduzione e l'abbattimento delle emissioni atmosferiche, la gestione delle acque reflue, la gestione dei rifiuti, il disinquinamento del suolo, ecc.), nonché a stabilire su chi grava (famiglie, imprese, amministrazioni pubbliche) il carico finanziario per la protezione dell'ambiente e in che misura.

Ad esempio, nel 1997 la spesa per il consumo intermedio e finale di servizi di gestione delle acque reflue e di gestione dei rifiuti (pari a poco più di 10 miliardi di euro, di cui il 77% per la gestione dei rifiuti) ha gravato sui diversi utilizzatori, al netto dei trasferimenti intercorrenti tra i diversi soggetti (tasse, tariffe e sussidi), come illustrato nella Figura 7[27]. Il settore delle acque reflue, diversamente da quello della gestione dei rifiuti, appare caratterizzato da un forte ruolo delle amministrazioni pubbliche che coprono il fabbisogno della collettività per più di un quarto. Anche la maggiore accumulazione di capitale produttivo da parte della PA si registra nel campo della gestione delle acque reflue, ferma restando, comunque, la preponderanza del capitale privato a testimonianza del notevole livello di *outsourcing* dei servizi considerati (Figura 8).

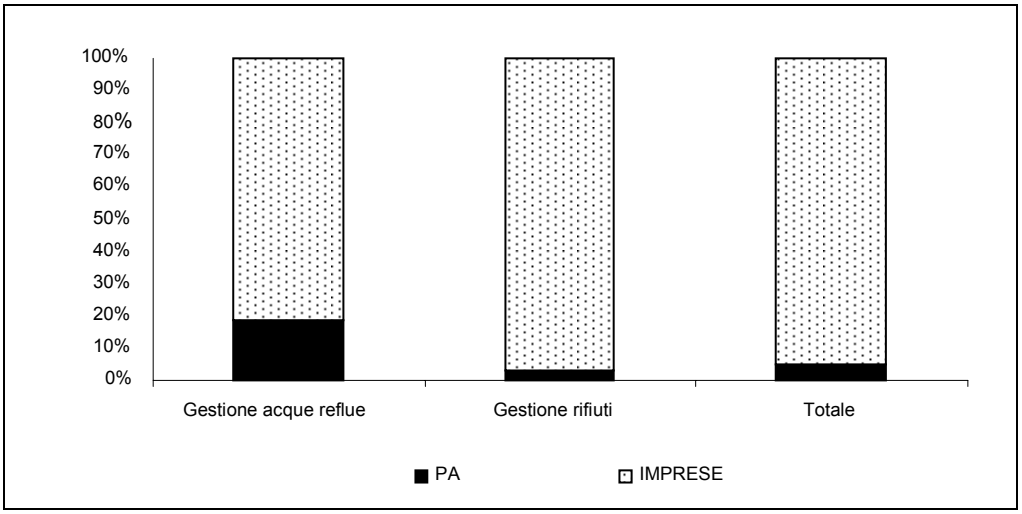
Figura 7 – Spesa per il consumo intermedio e finale di servizi di gestione delle acque reflue e di gestione dei rifiuti. Italia – Anno 1997 *(valori percentuali)*



Fonte: Elaborazione dati Istat

[27] In generale la spesa per la gestione delle acque reflue e dei rifiuti rappresenta a livello europeo, mediamente, circa il 70% del complesso della spesa per la "protezione dell'ambiente" come definita ai fini del conto EPEA (cfr. Istat, 2003a).

Figura 8 – Investimenti dei produttori specializzati nella fornitura per conto terzi di servizi di gestione delle acque reflue e di gestione dei rifiuti. Italia – Anno 1997 *(valori percentuali)*



Fonte: Elaborazione dati Istat

Informazioni di questo tipo, specie se disponibili per tutti i settori ambientali rilevanti e articolate anche per settore di attività economica e per regione, consentono di mirare meglio le politiche di spesa pubblica, tenendo conto degli interventi già in atto e avendo cura di evitare duplicazione di sforzi.

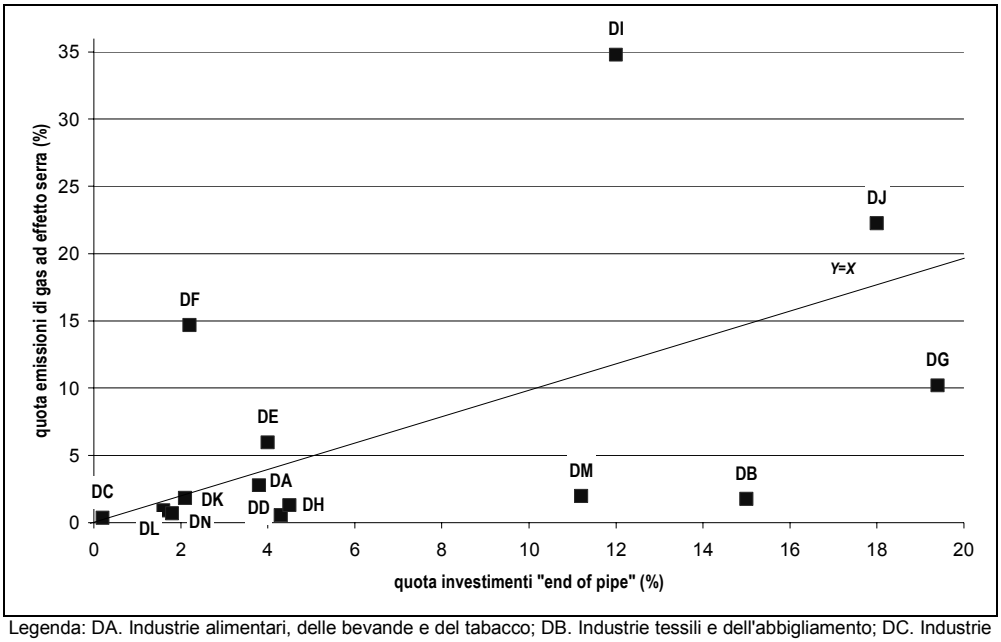
Domanda (3.3) Le attività più inquinanti sono anche quelle che spendono di più per la protezione dell'ambiente?

Generalmente i dati sulle spese per la protezione dell'ambiente, considerati singolarmente, non sono sufficienti per definire politiche di intervento pubblico tese, ad esempio, a riequilibrare situazioni di divario territoriale o settoriale. Occorre infatti tenere conto di altri fattori quali ad esempio il livello delle pressioni ambientali cui si intende far fronte con le "risposte" in atto da parte del sistema socio-economico, nonché il grado di efficacia delle "risposte" stesse in relazione ai problemi ambientali cui sono indirizzate.

La Figura 9 fornisce un esempio di elaborazione che mette in relazione dati di spesa per la protezione dell'ambiente con dati relativi alle pressioni ambientali, questi ultimi desunti dai conti di tipo NAMEA. In particolare viene confrontato, per ciascuna industria del settore manifatturiero, il contributo percentuale al totale degli investimenti di tipo "end-of-pipe" per la prevenzione e l'abbattimento delle emissioni atmosferiche nocive per il clima e la qualità dell'aria e il contributo percentuale al totale delle emissioni atmosferiche generate. I dati rappresentati possono essere utili per valutare l'attuazione del principio "chi inquina paga", individuando ad esempio le attività economiche il cui contributo agli investimenti per la protezione dell'ambiente è minore del corrispondente contributo al problema ambientale in questione (i punti al di sopra della diagonale nel grafico). In una situazione di bassa diffusione di tecnologie a ridotto impatto ambientale, si potrebbe valutare se avviare, proprio a partire da queste attività, l'incremento di investimenti in tecnologie più favorevoli per l'ambiente attraverso politiche di incentivazione mirate[28].

[28] Nel caso particolare delle emissioni di gas serra, occorre tenere presente che il piano nazionale di riduzione delle emissioni (piano attuazione di Kyoto) prevede un ampio ricorso alle misure internazionali di riduzione dei gas serra (come

Figura 9 – Investimenti "end of pipe" per la protezione dell'aria e clima e emissioni di gas ad effetto serra per le attività economiche del settore manifatturiero. Italia – Anno 2001
(contributi percentuali al totale del settore)



Legenda: DA. Industrie alimentari, delle bevande e del tabacco; DB. Industrie tessili e dell'abbigliamento; DC. Industrie conciarie, fabbricazione di prodotti in cuoio, pelle e similari; DD. Industria del legno e dei prodotti in legno; DE. Fabbricazione della pasta-carta, della carta e dei prodotti di carta; stampa e editoria; DF. Fabbricazione di coke, raffinerie di petrolio, trattamento dei combustibili nucleari; DG. Fabbricazione di prodotti chimici e di prodotti chimici artificiali; DH. Fabbricazione di articoli in gomma e materie plastiche; DI. Fabbricazione di prodotti della lavorazione di minerali non metalliferi; DJ. Produzione di metallo e fabbricazione di prodotti in metallo; DK. Fabbricazione di macchine e apparecchi meccanici, compresi l'installazione, il montaggio, la riparazione e la manutenzione; **DL**. Fabbricazione di macchine elettriche e di apparecchiature elettriche e ottiche; **DM**. Fabbricazione di mezzi di trasporto; **DN**. Altre industrie manifatturiere.
Lungo la diagonale si registra l'uguaglianza tra i contributi in termini di investimenti e di emissioni.
Fonte: Elaborazione dati Istat

In generale, il conto EPEA – e lo stesso vale per il conto RUMEA – è concepito per fornire informazioni disaggregate rispetto a molteplici profili: ad esempio per settore ambientale, per settore istituzionale, per settore di attività economica, per tipologia di transazione (investimenti, costi di produzione, acquisti, tariffe, tasse ambientali, trasferimenti, ecc.). Questo tipo di articolazione, combinato con i dati sulle pressioni generate dai vari operatori economici, presenta grandi potenzialità di utilizzo ai fini della definizione di politiche pubbliche, anche, ad esempio, in materia di tariffazione e imposizione fiscale.

quelle istituite dalla direttiva EU "Emission Trading" - 2003/87/EC- e dal cosiddetto "Clean Development Mechanism", articolo 12 del Protocollo di Kyoto). Quindi se un dato settore investe relativamente poco in misure "domestiche" di riduzione delle emissioni, puo' benissimo darsi che spenda molto in acquisto di *allowances* sul mercato europeo o di certificati di riduzione delle emissioni sul mercato internazionale, contribuendo così per altre vie al raggiungimento degli impegni di Kyoto.

5. Potenziali usi dei Conti Ambientali per le politiche di sviluppo

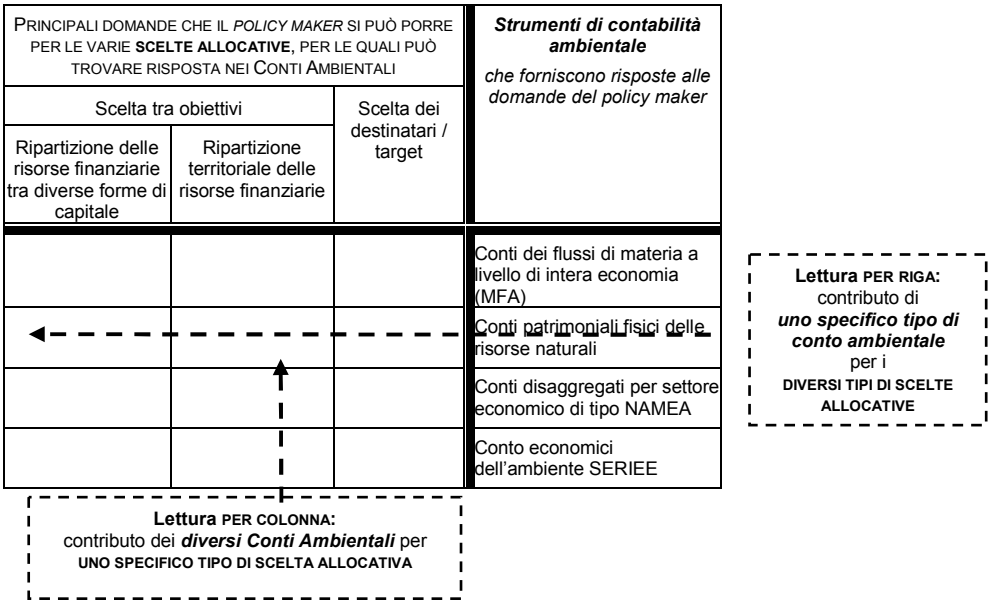
In aggiunta alle esemplificazioni riportate nei paragrafi precedenti, e al fine di abbozzare un quadro delle principali possibilità di uso dei Conti Ambientali per le politiche di sviluppo, il presente paragrafo intende offrire una panoramica più ampia delle domande conoscitive alle quali i Conti Ambientali possono fornire una risposta per loro vocazione, date le peculiarità specifiche che distinguono questo tipo di strumenti rispetto ad altre tipologie di informazioni.

Tale panoramica è fornita attraverso la successiva Tabella 4, nella quale sono raccolte, in particolare, le domande che possono essere considerate di maggiore rilievo ai fini del disegno e della valutazione delle politiche di sviluppo. Rispetto alla schematizzazione di tali politiche introdotta in precedenza (cfr. § 2, Figura 1), le domande qui raccolte riguardano le fasi più "a monte" del processo decisionale, ossia quelle in cui vengono compiute le diverse decisioni allocative della spesa (tra forme di capitale, territori e destinatari delle politiche), per le quali il contributo dei Conti Ambientali appare particolarmente strategico.

Le diverse domande sono raccolte e organizzate in Tabella 4 secondo i criteri illustrati in Figura 10: consultando la tabella per riga si individuano le domande che il *policy maker* può porsi in funzione dei diversi tipi di scelte allocative e a cui può trovare risposta in uno specifico strumento di contabilità ambientale indicato nell'ultima colonna; consultando lo schema per colonna si individuano le domande che il *policy maker* può porsi in funzione di uno specifico tipo di scelta allocativa e a cui può trovare risposta nei diversi strumenti offerti dalla contabilità ambientale.

Va sottolineato che, poiché allo stato attuale non vengono ancora prodotti regolarmente Conti Ambientali regionali, la Tabella 4 offre soprattutto una visione delle potenzialità d'uso dei dati di contabilità ambientale a supporto delle politiche di sviluppo. L'analisi di tali potenzialità è quindi di particolare ausilio – in questa fase – per l'individuazione di priorità per lo sviluppo di Conti Ambientali a scala regionale.

Figura 10 – Quali Conti Ambientali per quali esigenze conoscitive: principali chiavi di lettura



Fonte: Ministero dell'Economia e delle Finanze – Istat (2005)

Tabella 4 - Scelte allocative e contabilità ambientale: principali "domande" del *policy maker* e "risposte" dei Conti Ambientali

PRINCIPALI DOMANDE CHE IL *POLICY MAKER* SI PUÒ PORRE PER LE VARIE SCELTE ALLOCATIVE, PER LE QUALI PUÒ TROVARE RISPOSTA NEI CONTI AMBIENTALI			***Strumenti di contabilità ambientale che forniscono risposte alle domande del policy maker***
Scelta tra obiettivi		**Scelta dei destinatari / target**	
Ripartizione delle risorse finanziarie tra diverse forme di capitale	**Ripartizione territoriale delle risorse finanziarie**		
Il funzionamento del sistema economico richiede un fabbisogno di risorse materiali molto elevato?	In quali territori il funzionamento del sistema economico richiede il fabbisogno di risorse materiali più elevato?		Conti dei flussi di materia a livello di intera economia (MFA)
In particolare il funzionamento del sistema economico richiede un elevato fabbisogno di risorse materiali importate, determinando così pressioni ambientali localizzate altrove?	In particolare in quali territori il funzionamento del sistema economico richiede un maggiore fabbisogno di risorse materiali importate, determinando così pressioni ambientali localizzate altrove?		
Qual è la disponibilità delle varie risorse naturali e il loro stato qualitativo?	Ci sono tra i vari territori differenze significative nella disponibilità delle varie risorse naturali e nel loro stato qualitativo?		Conti patrimoniali fisici delle risorse naturali:
Qual è il livello delle pressioni antropiche sulle varie risorse naturali?	Tali differenze dipendono da un diverso livello, nei vari territori, delle pressioni antropiche sulle varie risorse naturali?		*foreste, acque, risorse del sottosuolo, uso e copertura del suolo, altre risorse naturali*
In quale misura le varie emissioni di inquinanti sono attribuibili a settori economici cruciali per l'economia?	In quali territori le varie emissioni di inquinanti sono attribuibili a settori economici cruciali per l'economia e in che misura?		Conti disaggregati per settore economico di tipo NAMEA:
	Quante tonnellate di inquinanti sono causate nei vari territori dai consumi delle famiglie e quante dalle attività produttive?	Quante tonnellate di inquinanti sono causate dai consumi delle famiglie e quante dalle attività produttive?	*pressioni in termini di flussi di sostanze inquinanti (emissioni atmosferiche, rifiuti, reflui, ecc.)*
	Tra le attività produttive, quali sono quelle che più contribuiscono all'emissione di determinati inquinanti nei vari territori?	Tra le attività produttive, quali sono quelle che più contribuiscono all'emissione di determinati inquinanti?	
	Ci sono tra i vari territori differenze significative nella relazione tra la *performance* economica e quella ambientale delle varie attività produttive (per esempio in termini di rapporto emissioni/valore aggiunto, emissioni/occupati, ecc.)?	In che relazione sono la *performance* economica e quella ambientale delle varie attività produttive (ad es. in termini di rapporto emissioni/valore aggiunto, emissioni/occupati, ecc.)?	

PRINCIPALI DOMANDE CHE IL *POLICY MAKER* SI PUÒ PORRE PER LE VARIE SCELTE ALLOCATIVE, PER LE QUALI PUÒ TROVARE RISPOSTA NEI CONTI AMBIENTALI			***Strumenti di contabilità ambientale che forniscono risposte alle domande del policy maker***
Scelta tra obiettivi		**Scelta dei destinatari / target**	
Ripartizione delle risorse finanziarie tra diverse forme di capitale	**Ripartizione territoriale delle risorse finanziarie**		
	Quale scenario si prefigura nei vari territori in termini di variazione delle emissioni di inquinanti a fronte di un dato livello di crescita di determinati settori economici (ad es. quante tonnellate di inquinanti sono causate dalle diverse attività produttive per un dato incremento della domanda finale, dell'occupazione, etc. di determinati settori economici)?	Quale scenario si prefigura in termini di variazione delle emissioni di inquinanti a fronte di un dato livello di crescita di determinati settori economici (ad es. quante tonnellate di inquinanti sono causate dalle diverse attività produttive per un dato incremento della domanda finale, dell'occupazione, ecc. di determinati settori economici)?	Conti disaggregati per settore economico di tipo NAMEA: *pressioni in termini di flussi di sostanze inquinanti (emissioni atmosferiche, rifiuti, reflui, ecc.)*
In quale misura il prelievo delle varie risorse naturali serve a soddisfare il fabbisogno dei settori economici cruciali per l'economia?	In quali territori il prelievo delle varie risorse naturali serve a soddisfare il fabbisogno dei settori economici cruciali per l'economia e in che misura? Quante tonnellate di risorse naturali sono prelevate nei vari territori per soddisfare i consumi finali delle famiglie e quante per soddisfare i consumi intermedi delle attività produttive? Tra le attività produttive, quali sono nei vari territori quelle con un maggior fabbisogno di risorse naturali? Ci sono tra i vari territori differenze significative nella relazione tra la *performance* economica e quella ambientale delle varie attività produttive (ad es. in termini di rapporto fabbisogno di risorse naturali/valore aggiunto, fabbisogno di risorse naturali/occupati, ecc.)? Quale scenario si prefigura nei vari territori in termini di variazione del fabbisogno di risorse naturali a fronte di un dato livello di crescita di determinati settori economici (per esempio quante tonnellate di risorse naturali sono necessarie per soddisfare i consumi intermedi delle varie attività produttive per un dato incremento della domanda finale, dell'occupazione, etc. di determinati settori economici)?	Quante tonnellate di risorse naturali sono prelevate per soddisfare i consumi finali delle famiglie e quante per soddisfare i consumi intermedi delle attività produttive? Tra le attività produttive, quali sono quelle con un maggior fabbisogno di risorse naturali? In che relazione sono la *performance* economica e quella ambientale delle varie attività produttive (ad es. in termini di rapporto fabbisogno di risorse naturali/valore aggiunto, fabbisogno di risorse naturali/occupati, ecc.)? Quale scenario si prefigura in termini di variazione del fabbisogno di risorse naturali a fronte di un dato livello di crescita di determinati settori economici (ad es. quante tonnellate di risorse naturali sono necessarie per soddisfare i consumi intermedi delle varie attività produttive per un dato incremento della domanda finale, dell'occupazione, etc. di determinati settori economici)?	Conti disaggregati per settore economico di tipo NAMEA: *pressioni in termini di flussi di prelievo di risorse naturali (Vapore endogeno, Combustibili fossili, Minerali, Biomasse)*

PRINCIPALI DOMANDE CHE IL *POLICY MAKER* SI PUÒ PORRE PER LE VARIE SCELTE ALLOCATIVE, PER LE QUALI PUÒ TROVARE RISPOSTA NEI CONTI AMBIENTALI			***Strumenti di contabilità ambientale che forniscono risposte alle domande del policy maker***
Scelta tra obiettivi		**Scelta dei destinatari / target**	
Ripartizione delle risorse finanziarie tra diverse forme di capitale	**Ripartizione territoriale delle risorse finanziarie**		
Quanto incide la spesa per la protezione dell'ambiente sul totale della spesa dell'economia?	Quanto incide nei vari territori la spesa per la protezione dell'ambiente sul totale della spesa dell'economia?	Quanto spendono per la protezione dell'ambiente le imprese delle diverse attività produttive e per quali settori ambientali di intervento?	Conto satellite delle spese per la protezione dell'ambiente EPEA: *tutela della qualità dell'ambiente da fenomeni di inquinamento e degrado*
In quali settori di intervento ambientale si concentra la spesa?	In quali settori di intervento ambientale si concentra la spesa nei vari territori?	Considerando le pressioni generate dalle diverse attività produttive quali risultano dai conti di tipo NAMEA, le attività più inquinanti sono anche quelle che spendono di più per la protezione dell'ambiente?	
Quanto spendono le imprese, le famiglie e le Amministrazioni pubbliche per la protezione dell'ambiente e quanto incide tale spesa sul totale della spesa di ciascuna di queste tipologie di operatori?	Quanto spendono nei vari territori le imprese, le famiglie e le Amministrazioni pubbliche per la protezione dell'ambiente e quanto incide tale spesa sul totale della spesa di ciascuna di queste tipologie di operatori?	Con riferimento alle pressioni più forti generate dalle diverse attività produttive e dalle famiglie quali risultano dai conti di tipo NAMEA, il carico finanziario per la protezione dell'ambiente è sostenuto per la maggior parte dalle stesse imprese e famiglie o prevale l'intervento pubblico?	
In quali settori ambientali di intervento si concentra la spesa delle varie tipologie di operatori?	In quali settori ambientali di intervento si concentra la spesa delle varie tipologie di operatori nei vari territori?	Quanto incidono le tasse ambientali sul carico finanziario totale per la protezione dell'ambiente gravante su imprese e famiglie?	
	Considerando l'evoluzione delle pressioni sulla qualità dell'ambiente quale risulta dai conti del patrimonio naturale e dai conti di tipo NAMEA i territori con un maggiore inquinamento e degrado sono anche quelli che spendono di più per la protezione dell'ambiente?	Quanto incidono le tariffe ambientali sul carico finanziario totale per la protezione dell'ambiente gravante su imprese e famiglie?	
		Qual è l'importanza economica dell'industria della protezione dell'ambiente (ad es. in termini di fatturato, redditi da lavoro dipendente, occupati, investimenti, ecc.)?	

PRINCIPALI DOMANDE CHE IL *POLICY MAKER* SI PUÒ PORRE PER LE VARIE SCELTE ALLOCATIVE, PER LE QUALI PUÒ TROVARE RISPOSTA NEI CONTI AMBIENTALI			***Strumenti di contabilità ambientale che forniscono risposte alle domande del policy maker***
Scelta tra obiettivi		**Scelta dei destinatari / target**	
Ripartizione delle risorse finanziarie tra diverse forme di capitale	**Ripartizione territoriale delle risorse finanziarie**		
Quanto incide la spesa per l'uso e la gestione delle risorse naturali sul totale della spesa dell'economia? Su quali risorse naturali si concentra la spesa? Quanto spendono le imprese, le famiglie e le Amministrazioni pubbliche per l'uso e la gestione delle risorse naturali e quanto incide tale spesa sul totale della spesa di ciascuna di queste tipologie di operatori? Su quali risorse naturali si concentra la spesa delle varie tipologie di operatori?	Quanto incide nei vari territori la spesa per l'uso e la gestione delle risorse naturali sul totale della spesa dell'economia? Su quali risorse naturali si concentra la spesa nei vari territori? Quanto spendono nei vari territori le imprese, le famiglie e le Amministrazioni pubbliche per l'uso e la gestione delle risorse naturali e quanto incide tale spesa sul totale della spesa di ciascuna di queste tipologie di operatori? Su quali risorse naturali si concentra la spesa delle varie tipologie di operatori nei vari territori? Considerando l'evoluzione delle pressioni sullo *stock* delle risorse naturali quale risulta dai conti del patrimonio naturale e dai conti di tipo NAMEA i territori con un maggiore prelievo di risorse naturali sono anche quelli che spendono di più per l'uso e la gestione delle risorse naturali?	Quanto spendono per l'uso e la gestione delle risorse naturali le imprese delle diverse attività produttive e per quali risorse naturali? Considerando le pressioni generate dalle diverse attività produttive quali risultano dai conti di tipo NAMEA, le attività che denotano il maggior fabbisogno di risorse naturali sono anche quelle che spendono di più per l'uso e la gestione delle risorse naturali? Con riferimento alle pressioni più forti generate dalle diverse attività produttive e dalle famiglie quali risultano dai conti di tipo NAMEA, il carico finanziario per l'uso e la gestione delle risorse naturali è sostenuto per la maggior parte dalle stesse imprese e famiglie o prevale l'intervento pubblico? Quanto incidono le tasse ambientali sul carico finanziario totale per l'uso e la gestione delle risorse naturali gravante su imprese e famiglie? Quanto incidono le tariffe ambientali sul carico finanziario totale per la protezione dell'ambiente gravante su imprese e famiglie? Qual è l'importanza economica dell'industria dell'uso e gestione delle risorse naturali (ad es. in termini di fatturato, redditi da lavoro dipendente, occupati, investimenti, ecc.)?	Conto satellite delle spese per l'uso e la gestione delle risorse naturali RUMEA: *tutela e gestione dello stock di risorse naturali da fenomeni di esaurimento*

Gli esempi della Tabella 4 suggeriscono che il valore aggiunto informativo della contabilità ambientale può essere valutato secondo due diverse prospettive tra loro complementari: a) dal punto di vista del contributo che ogni snodo del processo decisionale può trarre dai vari tipi di Conti Ambientali (lettura per colonna della tabella), e b) dal punto di vista del contributo che ciascun tipo di Conti Ambientali può fornire ai diversi snodi decisionali (lettura per riga).

a) Il valore aggiunto della contabilità ambientale ai singoli snodi decisionali

Le decisioni di riparto territoriale delle risorse finanziarie per lo sviluppo sono quelle in cui vi è sicuramente un uso più ampio e analitico dei Conti Ambientali. Queste informazioni possono consentire di tener conto delle differenze nelle rispettive situazioni ambientali ritenute in grado di influire sui divari di sviluppo. Ad esempio, possono essere determinati alcuni criteri per assegnare maggiori finanziamenti a Regioni con risorse naturali più degradate, sottoposte a pressioni ambientali relativamente maggiori, o attualmente caratterizzate da livelli di spesa per la protezione dell'ambiente relativamente più bassi.

Nelle decisioni di riparto tra forme di capitale, le informazioni desumibili dalla contabilità ambientale possono suggerire di allocare risorse ad alcune forme di capitale naturale in presenza di una diminuzione quantitativa e/o qualitativa della risorsa, oppure laddove i settori più rilevanti e/o dinamici dell'economia dell'area sono altamente dipendenti da alcune risorse naturali e/o hanno un forte impatto su di esse. In tali casi, tra l'altro, il degrado delle risorse naturali in questione può nel medio-lungo termine mettere in pericolo le prospettive di crescita di settori economici chiave e più in generale lo sviluppo della collettività interessata.

Le informazioni di contabilità ambientale appaiono di particolare rilievo per la scelta degli operatori e dei soggetti cui orientare gli strumenti delle politiche; possono anche fornire indicazioni per calibrare i parametri degli strumenti di intervento (per esempio determinazione delle variazioni di prezzo necessarie a indurre cambiamenti di comportamento). In particolare, l'utilizzo di dati di contabilità ambientale consente di mettere in luce le sinergie e i trade-off tra diminuzione delle pressioni ambientali e possibili ricadute su reddito, occupazione, ecc..

b) Il valore aggiunto dei singoli strumenti di contabilità ambientale

Vi sono strumenti di contabilità ambientale che per loro natura possono fornire un utile supporto per alcuni tipi di scelte e non per altre. È il caso dei Conti dei flussi di materia e dei Conti patrimoniali delle risorse naturali: dal momento che tali conti producono – quale che sia la scala territoriale di analisi – un'informazione aggregata a livello di intera economia, essi possono contribuire alle scelte di allocazione delle risorse tra territori, ma non trovano un utilizzo specifico ai fini della scelta degli strumenti di *policy*.

D'altra parte, vi sono strumenti di contabilità ambientale che possono fornire un utile supporto per tutti i tipi di scelte allocative, sebbene di volta in volta in modo diverso, ossia

privilegiando talvolta la lettura di certe informazioni, talvolta la lettura di altre. È il caso dei conti NAMEA e EPEA/RUMEA, le cui informazioni, direttamente riconducibili a quelle dei conti economici (in virtù del fatto, ad esempio, che si adotta le stessa articolazione in settori istituzionali e settori di attività economica), sono suscettibili di essere lette a vari livelli e per vari obiettivi. Per esempio, in relazione ad una decisione di ripartizione tra diverse forme di capitale, si può prefigurare una prima lettura parziale dei dati di tipo NAMEA, limitata a verificare se i settori economici più inquinanti siano anche quelli "trainanti" dell'economia, informazione che può incidere sulla determinazione dei pesi da assegnare nella funzione obiettivo alle finalità di natura economica e a quelle di natura ambientale. Una lettura più analitica e completa può essere fatta invece in fase di ripartizione territoriale e/o di scelta degli strumenti, in cui è rilevante confrontare in modo sistematico la *performance* economica e quella ambientale di tutti i settori dell'economia in tutti i territori.

I conti patrimoniali delle risorse naturali e i conti EPEA/RUMEA forniscono un quadro, rispettivamente, dello stato dell'ambiente di un dato territorio e dell'intensità di alcune azioni di risposta da parte degli operatori pubblici e privati alle pressioni ambientali e alle perdite di capitale naturale. Questa informazione, in particolare quando esaminata in serie storica, può suggerire al *policy maker* dove sia più urgente orientare gli sforzi sul territorio, privilegiando un'allocazione di risorse ai territori in cui lo stato dell'ambiente è più degradato (qualità) oppure quelli in cui il capitale naturale è diminuito maggiormente (quantità). Può inoltre servire a giustificare una scelta settoriale verso le problematiche a cui meno corrispondono risposte dirette da parte degli operatori sia pubblici che privati (rifiuti, qualità dell'aria, foreste, riserve di fauna, ecc.), valutando la propensione alla spesa per la tutela ambientale da parte di famiglie, imprese e enti pubblici, e avendo cura di evitare duplicazione di sforzi.

Il contributo della Contabilità ambientale appare alquanto diverso se considerato in fase di programmazione (ossia in relazione ai diversi tipi di scelte allocative sopra considerate) e di attuazione delle politiche, o in fase di monitoraggio e valutazione (ex ante, in itinere e ex post). Nel primo caso i Conti Ambientali, quale che sia la scala territoriale di riferimento, forniscono dati di "contesto", che possono essere utilizzati tali e quali:

- in fase di programmazione orientando le scelte allocative del *policy maker* come illustrato negli esempi riportati nel paragrafo precedente (§ 4) e in Tabella 4;
- in fase di attuazione fornendo, per esempio, parametri di *benchmark* sulla base dei quali stabilire criteri di assegnazione delle risorse ai diversi soggetti dell'economia nei vari territori, al fine di selezionare interventi che assicurino una maggiore sostenibilità ambientale (tramite criteri di eligibilità o premiazione nei bandi di gara).

Per condurre invece una valutazione degli impatti di un programma, è necessario separare gli effetti attribuibili agli interventi finanziati, da effetti di altra origine, tipicamente attraverso le diverse tecniche impiegate in letteratura per la costruzione e l'analisi di ipotesi di situazione "controfattuale". La costruzione di tali schemi di valutazione è senz'altro facilitata quando i sistemi di monitoraggio adottano gli stessi standard (definizioni, classificazioni, schemi, ecc.) della statistica ufficiale. Nel caso particolare in cui interessi valutare gli effetti ambientali dei programmi, può tornare utile classificare gli interventi di tutela ambientale secondo le classificazioni standard del SERIEE, o definire indicatori che misurano le pressioni ambientali degli interventi secondo l'articolazione della NAMEA. In questo modo, è possibile per esempio mettere in relazione il comportamento delle imprese beneficiarie degli interventi con quello medio delle imprese a esse confrontabili.

6. Conclusioni

Lo schema generale descritto in questo lavoro consente di individuare le potenziali risposte fornite dai Conti Ambientali a quesiti che tipicamente il *policy maker* può porsi in sede di decisioni allocative nell'ambito delle politiche di sviluppo. Attraverso alcuni esempi basati su dati recenti, si mostra che un'analisi su scala regionale di dati di contabilità ambientale può aiutare ad orientare l'allocazione delle risorse per lo sviluppo in modo da tener conto di criteri connessi alla situazione ambientale ed economica specifica dei territori e da quantificare gli eventuali *trade-off* tra sviluppo economico e disponibilità quali-quantitativa del patrimonio naturale, indotti direttamente o indirettamente dalle politiche di investimento ed incentivazione. Ciascun tipo di conto, specialmente se disaggregato territorialmente, consente di rispondere a domande sulle dinamiche che legano le politiche di sviluppo e l'ambiente (effetti negativi sull'ambiente, retroazione negativa sull'economia, interventi di tutela ambientale) tipiche della fase di programmazione degli interventi (e in qualche modo importanti anche nella fase di valutazione).

Per l'Italia si dispone di dati di contabilità ambientale, aggiornati annualmente, per un sottoinsieme di Conti Ambientali. Questi dati, allo stato attuale, restituiscono un'immagine delle pressioni generate e delle risposte dei diversi soggetti economici per l'intero territorio nazionale. Tuttavia, per fornire indicazioni utili per le politiche di sviluppo in un paese come l'Italia è importante evidenziare le differenze territoriali nei fenomeni di interazione tra economia e ambiente. Se sviluppati a scala regionale, gli strumenti di contabilità ambientale consentirebbero di confrontare le diverse realtà territoriali e di evidenziare divari in termini non solo di struttura economica e di patrimonio naturale, ma anche di eco-efficienza delle attività produttive e di consumo.

È in quest'ambito che si colloca il progetto congiunto del Ministero dell'Economia e delle Finanze e dell'Istat, che ha contributo ad avviare attività per lo sviluppo di Conti Ambientali su scala regionale (Ministero dell'Economia e delle Finanze – Istat, 2005). Con riferimento in particolare alla comunità di decisori e tecnici che operano nel contesto delle politiche di sviluppo il progetto ha considerato prioritario lo sviluppo su scala regionale dei conti di tipo NAMEA e dei conti delle spese per la tutela dell'ambiente (EPEA/RUMEA). La sperimentazione effettuata nell'ambito del progetto ha contribuito a rendere oggi disponibili per la regione Lazio i conti NAMEA delle emissioni atmosferiche per l'anno 2000 e l'aggiornamento al 2001 della serie storica della spesa per la protezione dell'ambiente dell'amministrazione regionale (EPEA). Una produzione a regime di tali conti per tutte le regioni italiane – e in prospettiva di un set completo di Conti Ambientali – costituirebbe un contributo essenziale per estendere l'analisi degli squilibri territoriali alle variabili ambientali e per disegnare politiche in cui economia e ambiente siano ambiti integrati piuttosto che paralleli.

Riferimenti bibliografici

CIPE (2002), *Strategia d'azione ambientale per lo sviluppo sostenibile in Italia, Delibera 2 agosto 2002*, Roma.

Commissione delle Comunità Europee (1994), *Orientamenti per l'UE in materia di indicatori ambientali e di contabilità verde nazionale – Integrazione di sistemi di informazione ambientale e economica*, Comunicazione della Commissione delle Comunità Europee al Consiglio e al Parlamento Europeo, (COM (94) 670) def., 21.12.1994, Bruxelles.

Eurostat (1994), *SERIEE – 1994 Version*, Theme Environment, Series Methods, Luxembourg.

Eurostat (1996), *Sistema europeo dei conti SEC 1995*, Lussemburgo.

Eurostat (1999a) *Towards Environmental Pressure Indicators for the EU: indicator definition*, Working document, Theme 8 Environment and Energy, Luxembourg.

Eurostat (1999b), *Pilot Studies on NAMEAs for air emissions with a comparison at European level*, Office for Official Publications of the European Communities, Theme 2: Economy and Finance, Collection: Studies and research (catalogue number: CA-23-99-338-EN-C), Luxembourg.

Eurostat (2000), *Economy-wide material flow accounts and derived indicators. A methodological guide*, Luxembourg.

Eurostat (2001), *NAMEAs for air emissions – Results of Pilot Studies*, Office for Official Publications of the European Communities, Theme 2: Economy and Finance, Collection: Studies and research (catalogue number: CA-23-99-338-EN-C), Luxembourg.

Eurostat (2002a), *SERIEE Environmental Protection Expenditure Accounts – Compilation Guide*, Luxembourg.

Eurostat (2002b), *The European Framework for Integrated Environmental and Economic Accounting for Forests – IEEAF*, Luxembourg.

Eurostat (2002c), *NAMEAs for air emissions – Results of pilot studies*, Numero di catalogo KS-39-01-093-EN-N, Luxembourg.

Eurostat (2002d), *The European Strategy for Environmental Accounting*, Luxembourg.

Eurostat (2003), *Decomposition analysis of carbon dioxide-emission changes in Germany – conceptual framework*, Office for Official Publications of the European Communities, Theme 2: Economy and Finance, Collection: Working Papers and Studies, Luxembourg.

Eurostat (2005), *Environmental Expenditure Statistics: Industry Data Collection Handbook*, Luxembourg.

Falcitelli F. – Femia A. – Tudini A. – Vetrella G. (2005), "Focus 1 – Contabilità ambientale: 'pressioni' e 'risposte' dell'economia nel contesto della Contabilità nazionale", in Istat, *Statistiche Ambientali*, Collana Annuari, Annuario n.8 – 2005, Roma.

Femia A. – Panfili P. (2005), "Analytical Applications of the NAMEA", in Società Italiana di Statistica - SIS (2005), *Statistics and Environment – Contributed Papers*, Atti del Convegno intermedio "Statistics and Environment", Università di Messina, 21-23 settembre 2005, CLEUP, Padova.

INSEE (1986), *Les comptes du Patrimoine Naturel*, N. 535-536 des Collections de l'INSEE, série D n° 137-138, Paris.

Istat (1996), *Contabilità ambientale*, Annali di Statistica, anno 125, serie X – vol. 13, 1996, Roma.

Istat (1999), *Indicatori e Conti Ambientali: verso un sistema informativo integrato economico e ambientale*, Annali di Statistica, Anno 128, Serie X – vol. 18, Roma.

Istat (2002), *Spese delle imprese per la protezione dell'ambiente – anno 1997*, http://www.istat.it/conti/ambientali/

Istat (2003a), *Prima applicazione dell'EPEA per l'Italia. Conto satellite delle spese per la protezione dell'ambiente per i settori della gestione delle acque reflue e della gestione dei rifiuti. Anno 1997*, http://www.istat.it/conti/ambientali/

Istat (2003b), *Indicatori e conti dei flussi di materia dell'economia italiana, 1980-1998*, http://samoa.istat.it/Economia/Conti-nazi/index.htm

Istat (2003c), *Classificazione delle attività economiche. ATECO 2002. Derivata dalla NACE Rev. 1.1*, Istat, Metodi e Norme n. 18, Roma.

Istat (2003d), *Contabilità ambientale e risposte del sistema socio-economico: dagli schemi alle realizzazioni*, Annali di Statistica, Anno 132, Serie XI, Vol. 1, Roma.

Istat (2005), *La spesa per la protezione dell'ambiente delle Amministrazioni dello Stato – anni 1995-2002*, http://www.istat.it/conti/ambientali/

Istat (2006), *La NAMEA: Conti economici nazionali integrati con Conti Ambientali – anni 1990-2002*, http://www.istat.it/conti/ambientali/

Istat-Ministero dell'ambiente e della tutela del territorio e del mare (2007), *Il calcolo della spesa pubblica per la protezione dell'ambiente. Linee guida per riclassificare i rendiconti delle amministrazioni pubbliche*. Istat, Metodi e Norme n. 33/2006, Roma, disponibile all'indirizzo: http://www.istat.it/dati/catalogo/20070212_00/

Istat (in via di pubblicazione), *Contabilità ambientale e pressioni sull'ambiente naturale: dagli schemi alle realizzazioni*, Annali di Statistica, Roma.

Ministero dell'ambiente (2001), *Relazione sullo stato dell'ambiente*, capitolo "Gli strumenti economici, la spesa pubblica e la contabilità ambientale", Roma.

Ministero dell'Economia e delle Finanze (2000), *Guida metodologica per la costruzione di conti consolidati della finanza pubblica a livello regionale*, disponibile all'indirizzo: http://www.dps.mef.gov.it/cpt/cpt_guidametodologica.asp

Ministero dell'Economia e delle Finanze, Dipartimento per le Politiche di Sviluppo e Coesione (2005), *Rapporto Annuale 2004 e Appendice*, disponibile all'indirizzo: http://www.dps.tesoro.it/documenti_elenco.asp#specificiDPS

Ministero dell'Economia e delle Finanze, Dipartimento per le Politiche di Sviluppo e Coesione (2006), *Rapporto Annuale 2005 e Appendice*, disponibile all'indirizzo: http://www.dps.tesoro.it/documenti_elenco.asp#specificiDPS

Ministero dell'Economia e delle Finanze – Istat (2005), *Ambiente e politiche di sviluppo: le potenzialità della contabilità ambientale per decidere meglio*, Materiali UVAL N. 5, Ministero dell'Economia e delle Finanze, Dipartimento per le Politiche di Sviluppo, Unità di Valutazione degli Investimenti Pubblici, Roma, disponibile all'indirizzo: http://www.dps.tesoro.it/documentazione/uval/materiali_uval/Muval5_Contabilita_Ambientale.pdf

OCDE (1993), *Corps central d'indicateurs de l'OCDE pour l'examens des performances environnementales*, Rapport de synthèse du Groupe sur l'États de l'Environnement, Monographies sur l'environnement, Paris.

OECD/Eurostat (1999), *The Environmental Goods & Services Industry. Manual for data collection and analysis*, Paris.

Regione Lazio (2005), *Rapporto sullo stato dell'ambiente della Regione Lazio*, Roma.

Rete Nazionale delle Autorità Ambientali e delle Autorità della Programmazione dei Fondi Strutturali Comunitari 2000-2006 (2004), *'RAITA': Regimi di Aiuto alle Imprese che prevedono esclusivamente o meno Interventi di Tutela Ambientale. Indagine statistica sul comportamento delle imprese*, relazione presentata alla riunione della Rete tenutasi a Roma il 15 Dicembre 2004, disponibile all'indirizzo: http://www.minambiente.it/SVS/fondi/fondi.htm, link 'documenti di indirizzo'.

Siesto V. (1996), *La contabilità nazionale italiana*, Il Mulino, Bologna.

United Nations (1993a), *Integrated Environmental and Economic Accounting*, Studies in Methods, Series F, No. 61, New York.

United Nations (1993b), *System of National Accounts*, Series F/2/Rev. 4, New York, ch. XXI.

United Nations (1999), *Classification Of the Functions Of Government*, disponibile all'indirizzo: http://unstats.un.org/unsd/cr/registry/

United Nations et alii (in via di pubblicazione), *Integrated Environmental and Economic Accounting 2003 (SEEA 2003)*, disponibile all'indirizzo: http://unstats.un.org/unsd/envAccounting/seea2003.pdf

Norme redazionali

La Rivista di statistica ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche Istat corredati, a parte, da una nota informativa dell'Autore contenente: appartenenza ad istituzioni, attività prevalente, qualifica, indirizzo, casella di posta elettronica, recapito telefonico e l'autorizzazione alla pubblicazione firmata dagli Autori. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di un referente scelto tra gli esperti dei diversi temi affrontati. Gli originali, anche se non pubblicati, non si restituiscono.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file Template.doc disponibile on line o su richiesta. In base a tali standard la lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 30–35 pagine.

I lavori devono essere corredati di un sommario in inglese e, se l'articolo è in italiano, anche in italiano della lunghezza massima di 12 righe ciascuno. La bibliografia, in ordine alfabetico per autore, deve essere riportata in elenco a parte alla fine dell'articolo. Quando nel testo si fa riferimento ad una pubblicazione citata nell'elenco, si metta in parentesi tonda il nome dell'autore, l'anno di pubblicazione ed eventualmente la pagina citata. Ad esempio (Bianchi, 1987, Rossi, 1988, p. 55). Quando l'autore compare più volte nello stesso anno l'ordine verrà dato dall'aggiunta di una lettera minuscola accanto all'anno di pubblicazione. Ad esempio (Bianchi, 1987a, 1987b).

Nella bibliografia le citazioni di libri e articoli vanno indicate nel seguente modo. Per i libri: cognome dell'autore seguito dall'iniziale in maiuscolo del nome, il titolo in corsivo dell'opera, l'editore, il luogo di edizione e l'anno di pubblicazione. Per gli articoli: dopo l'indicazione dell'autore si riporta il titolo tra virgolette, il titolo completo in corsivo della rivista, il numero del fascicolo e l'anno di pubblicazione. Nei riferimenti bibliografici non si devono usare abbreviazioni.

Nel testo dovrà essere di norma utilizzato il corsivo per le parole in lingua straniera e il corsivo o grassetto per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare il Comitato di redazione delle pubblicazioni scientifiche Istat e per inviare lavori: rivista@istat.it. Oppure scrivere a:

Comitato di redazione delle pubblicazioni scientifiche
C/O Carlo Deli (cadeli@istat.it)
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012007003000000

ISSN 1828-1982

€ 10,00