

Proceedings of the “17<sup>th</sup> Roundtable on Business Survey frames”  
Rome, 26-31 October 2003

## **Use of Administrative Data for Short Term Statistics on Employment, Wages and Labour Cost\***

*Ciro Baldi, Francesca Ceccato, Maria Carla Congia, Eleonora Cimino,  
Silvia Pacini, Fabio Rapiti, Donatella Tuzi*

### **1. Introduction**

In the last decade the Italian National Statistical Institute (Istat) started successfully to fill the gap in the use of administrative data with the most advanced countries. After the seminal experience of ASIA archive (Italian official statistical business register), big steps ahead have been done in the use of administrative data for short term and structural statistics. The OROS Survey is a good example of this extensive and innovative use of administrative data for current short term statistics. With the aim of meeting the requirements of the UE regulations on statistics, satisfying the growing national demand in this field and reducing the statistical burden on firms, Istat in 1999 started to work to exploit the rich administrative database of Italy's major Social Security institution, namely INPS (Istituto Nazionale di Previdenza Sociale). The statistical potential of INPS data was well known since the first nineties but only at the end of the last decade Istat succeeded in exploiting it for current statistics.

In the Italian statistical context the OROS Survey experience is unique for many reasons:

1. for the first time administrative data has been used for current quarterly economic statistics; this means that for the first time an incredible bulk of data has been processed in a very short time schedule;
2. an original methodology to use a non-random sample of administrative data has been developed;
3. a complex approach to the quality of the Survey has been used to keep control of all different factors of bias and non-sample errors .

In this paper we discuss the main methodological aspects of the OROS Survey and try to illustrate the main innovative aspects. This paper is divided in nine parts. In the next section (§ 2) there is a summary of the main characteristics of the survey and the requirement of the UE Regulations on wage, labour cost and employment short term statistics. The third paragraph deals with the reasons why the main source used is an administrative one and the consequences that this choice impressed on Istat strategies. The fourth section deeply examines the OROS data sources i.e. the different INPS archives and the Monthly Survey on Labour Input variables in Large Firms. The fifth paragraph summaries the whole survey production process: from the data capturing to the final estimation, through all the check and editing procedures. The sixth paragraph describes the retrieving and transformation of the administrative data in statistical variables. The seventh

---

\* Notwithstanding the very strong cooperation among the authors the single paragraph can be attributed to:

p. 1, 3 and 9 Fabio Rapiti;

p. 2 Carla Congia and Silvia Pacini;

p. 4.3 and 8.2 Silvia Pacini;

p. 4.1, 4.2, 5 and 8.3 Donatella Tuzi;

p. 6 Eleonora Cimino, Maria Carla Congia, Francesca Ceccato;

p. 7 Eleonora Cimino and Francesca Ceccato.

p. 8.1 Ciro Baldi and Francesca Ceccato;

p. 8.4 Ciro Baldi.

paragraph addresses the preliminary and final estimation methodology and describes the imputation of unit non-responses. The macro checks are treated in the eighth section. The ninth paragraph contains some general conclusions.

## **2. Main characteristics of the OROS Survey**

The acronym OROS stands for Occupazione (Employment), Retribuzioni (Wages), Oneri Sociali (Other Labour Costs). The main aim of this survey is to produce short term information on the quarterly changes and levels of employment, gross wage, other labour cost and total labour cost for Italian firms in the private sector (sections C to K of the Nace Rev.1 classification) with at least one employee.

In the past this information was collected only for those enterprises in industry and services classified as “large”, that is with more than 500 employees, through the Monthly Survey on Labour Input variables in Large Firms (hereinafter Large Enterprises Survey - LES). The new OROS survey based on the administrative data collected by INPS is aimed to cover all firms classes, included enterprises with less than 500 employees (Small and Medium Enterprises - SME), without increasing the statistical burden on firms. The survey has been designed also to satisfy the Community requirements on short term statistics (STS and LCI-Labour Cost Index Regulations) .

At the moment, OROS releases indexes on gross wage, other labour cost and total labour cost per full time equivalent (FTE). In the very next future the survey will implement estimates of employment.

Each quarter two new estimations are released: the “preliminary” estimate based on a “non-random” sample of INPS data, with a delay of about 90 days from the reference quarter, and a “final” estimate based on the “total population” of INPS data, with a delay of 15 months from the reference quarter.

The OROS survey has to fulfill the requirements of two recent European Community regulations concerning short term statistics. These Regulations require that each Member State must make available a large volume of information on, at least, a quarterly basis, for monitoring the economic situation. These Regulations do not indicate rigid survey criteria, but explicitly refer to various estimation methods and the possibility of using administrative data. In fact, the Regulations are inspired by criteria of “harmonisation of the output and not the input”.

The first European Community Regulation (n.1165/98) is aimed at the establishment of a common framework for the production of short term statistics on the business cycle (hereafter STS-R). The variables required are the number of persons employed, the hours worked and the gross wages and salaries and they should represent the entire population of firms of all size classes.

The STS-R’s coverage in terms of economic activities is based on the NACE Rev.1 classification and concerns, broken down at a high level of detail, industry, construction, retail trade and repair and other services. The Regulation was approved in May 1998, but the implementation times are rather long. In fact, it was foreseen that all countries might have derogations up to 5 years on single variables, thus until spring 2003. For the variables “number of persons employed”, “hours worked”, and “gross wage” Istat satisfied the Regulation using LES data, obviously, covering only large firms. Since winter 2003 the process of substituting LES source with OROS is gradually occurring.

Secondly, the most recent Regulation (n.450/2003) concerns the labour cost index (LCI). In order to produce comparable short-term labour cost statistics in the Community, member States have to estimate quarterly labour cost index for the economic activities in sections C to O of NACE Rev.1<sup>1</sup>. Labour cost indices are to be produced separately for three labour cost categories: total labour cost, gross wages and salaries, employers social contributions plus taxes paid by the employer less subsidies received. Data is to be transmitted within 70 days from the end of the reference period. A

---

<sup>1</sup>The inclusion of sections L, M, N and O shall be determined after the evaluation of the results of feasibility studies.

transition period relating to the implementation of the LCI Regulation is allowed to the member States. In particular, Italy shall be able to produce the indices per hour worked (at the moment the OROS survey estimates the labour costs per FTE) and reduce the delay of transmission, within a year from the date of entry into force of the Regulation.

### 3. Why Administrative Data?

In recent years National Statistical Institutes (NSIs), for all business statistics including wages and employment, faced to opposite needs :

- to increase the flow of information supplied at national and international level (this often means covering all firm size classes);
- reduce the statistical burden on businesses.

With those two conditions in countries like Italy, with a very large share of SME (small and medium size enterprise), the set of possible choices was constrained. In practice, in business statistics in some cases there is no alternative but to use administrative sources to fill the data gaps, although it must be done “cum grano salis”. In this case what we should be talking of is *integration* and *complementarities* of sources rather than *substituting* data obtained by means of the traditional surveys with data drawn from administrative sources. However, we should also define the tasks assigned to the various sources: the administrative source yields the mass of data, i.e. the universe of the population, while direct survey on small samples of the population provides a basis for appropriate processing of administrative data guaranteeing quality in terms of accuracy and comparability. In some cases, therefore, the complementarity and integration we are dealing with appear markedly skewed, statistical survey serving largely as a qualitative support to the administrative source<sup>2</sup>.

In the case of the OROS Survey the use of social security data (INPS) as main source depends on the following simple and well known evidence: the population of firms in Italy is largely composed of SME’s.

**Table 1: Number of firms and employees for firm size classes in the private sector, years 2001**

	firms		employment	
	abs. val.	%	abs. val.	%
<b>1-9</b>	998,245	87,2	2,450,254	25,5
<b>10-19</b>	84,285	7,4	1,152,965	12,0
<b>20-99</b>	52,744	4,6	2,041,333	21,2
<b>100-499</b>	7,935	0,7	1,543,296	16,0
<b>500 and more</b>	1,266	0,1	2,439,195	25,3
<b>Tot</b>	1,144,475	100,0	9,627,043	100,0

Source: Istat, Business Register (Asia)

Looking at the table 1, it appears that the *micro* firms, those with less than 20 employees, have an incredible large share of employment (37,5%). Thus, estimating wages, labour cost and employment without covering the very small firms, could produce very biased results. Moreover, especially for the estimation of employment variable, it is essential to include very small firms

<sup>2</sup> Generally speaking, administrative data can be used generally directly or in model based estimation, for example as auxiliary variable in regression models. In both cases, reversing the traditional relation between administrative data and survey data, the latter can be used to evaluate and support administrative data production process.

because of the role of births and deaths in the creation and destruction of jobs<sup>3</sup>. As we will see later, the complex methodology developed to estimate quarterly indexes referred to the current population tends to take into account all those problems.

Once the statistical institute turned to the strategic option to capture and use a large quantity of INPS data it was worth value to exploit all the positive externalities of the data itself. The availability in electronic form of a mass quantity of INPS data has stimulated Istat to tune the strategy from a typical “one collection-for one single survey” to focus on the “data source” i.e. the wage and social security system which can be used for different statistical objectives. Thus the data capturing wasn’t limited only to the little set of aggregated variables needed for the short term statistics but covered the whole, very detailed, information contained in the INPS Social Security monthly declaration form. This choice from one side complicates the retrieving and translation process<sup>4</sup>, while from the other side, allows the exploitation of the very rich statistical potential of INPS data.

Although this process of exploitation is still in its first phase, at the moment the OROS Survey databases, apart the quarterly indexes, already satisfy other statistical objectives:

- it is an important support to the preliminary estimation of wages and employment annual SBS (structural Business Statistics);
- it is the main source for wage employment variables in the new Employment Satellite Archive;
- it is used in many different ways in national account;
- it is an source which contribute to some annual Social Security statistics;
- it is frequently used for ad hoc analysis on business and labour market variables.

#### **4. The OROS data sources**

OROS Survey main source is the archive of the monthly contribution declarations (i.e. DM10 forms) that all firms in the private business sectors, with at least one employee, have to transmit to INPS to pay compulsory contributions for Social Security. The companion source is the INPS register which collects individual information on the administrative units. The entire private sector is covered (roughly 10 million employees and 1.2 million employers per year); agriculture and public administration are excluded. Although the use of this data guarantees the coverage of all firms in the private sectors, at the moment the INPS sources are mainly used for the SME estimation, while for large firms OROS integrate also data coming from LES.

##### **4.1. INPS register**

When a firm is going to start an activity with wage employment it became subject to compulsory social security contributions and has to register to INPS through an appropriate form, in which it declares individual information. These information are collected in a register. We will refer to it as the INPS Register or the Administrative Register (AR).

The administrative nature of INPS register implies that some information are almost always present and have a good quality because relevant for administrative purposes. These refer to the identification number of the administrative unit, its fiscal code, the date of registration, the address, etc.

AR is composed of administrative units that may or may not correspond to the enterprises. In fact an enterprise can open more than one administrative unit for reasons that are sometimes related to the fact that the firm is multi-localized sometimes to the fact that the for the firm is more convenient

---

<sup>3</sup> Births and deaths are largely concentrated in very small firms. Besides, different sources indicate that almost one third of job creations and job destructions is caused by demographic events.

<sup>4</sup> In chapter 7 the very complex operation of computing and aggregating of raw micro data and their translation into statistical variables is described.

declare in different units different types of job arrangement or different portion of the firm with different economic activity and so on. In other words, the administrative unit does not correspond to any of the official statistical units.

INPS update regularly the AR. Changes are essentially related to the registration of new units and the change of information over units which already exist. This latter case can be promoted on the side of the firm or when INPS indirectly acquires information on it. This especially happens about the activity status of the units which do not give communication for a long period (e.g. a possible death after a period of economic inactivity).

The communication about the activity status (or the lack of it) introduces another characteristic of the AR. While all births are covered because it is obligatory to register to INPS in order to set up an enterprise with employees, the deaths are not frequently removed from the AR because, although there is an obligation to communicate the cessation of the enterprise, there is no administrative penalty to enforce it.

The communication about a change has the effect to replace previous information. That is, in each moment INPS register gives a picture of the situation according to the information available until that moment. No past history is preserved. Given the longitudinal nature of OROS Survey, it is instead necessary to give to the information context an historical configuration, where a trace about the changes occurred in the past would be maintained. That is accomplished by periodically downloading the register. In fact INPS register is made available to ISTAT at the end of each reference quarter and contains information for about 2 millions of administrative units.

To make it suitable for the statistical purposes it undergoes some phases of check and it is matched with the Business Register, ASIA, mainly to acquire the economic activity code. In fact the classification rule consist in drawing the NACE code from ASIA where the two register matches and in using the NACE code of the AR or the translation of the Contributive Statistical Code for the residual units. About the 70% of the units in the register get the economic classification from the Business Register. The reason for the non-correspondence between the two archives is mainly due to the temporary release gap: the Business Register is available with up to two years delay from the reference quarter of INPS register. Regarding the problem of the units of analysis, while the transformation of the data from a list of administrative units to a list of firms has some advantages as the possibility of comparisons at micro level with other enterprise surveys, a feasibility analysis conducted to study the criteria and the problems of this transformation has shown that it has a number of drawbacks. The study has pointed out several methodological difficulties mainly due to the individuation of unambiguous rules for the definition of the new units and the attribution of some characters. Those difficulties in turn make it difficult following a unit over time, whereas the availability of INPS identification number of the administrative units let this task be quite simple and qualitatively good. In conclusion the problem of the units of analysis has been solved in a quite heuristic way: the administrative unit has been preserved as the building block of OROS archive while some enterprise level has been linked to them.

#### **4.2. Employers monthly declaration archive**

Each month enterprises with at least one employee must submit to the competent local INPS offices the forms (DM10) in which Social Security compulsory contributions are declared. The deadline for this presentation is the 20<sup>th</sup> day of the month following the one the declarations refers to.

Several delivering modes have been made available to firms, that can decide to send the forms through electronic or non-electronic supports: the most common ways are paper, floppy disc or other magnetic means, internet, etc.. Although the paper support is the traditional mean, it has been observed that more and more firms, during the years, have been using the electronic communication mode (figure 1). Particularly, since 2001, the use of Internet is becoming very usual. At the moment about 80% of the total forms are sent electronically, while the remaining must be entered by hand or passed through an optical reader by INPS operators.

The DM10 forms delivered by diskettes or paper are usually available at the local INPS offices after 30-40 days from the end of the reference month. Once arrived, they are registered into a local archive and then gradually uploaded to the central database. Data sent in electronic mode are available very quickly, while those on paper arrive much more slowly.

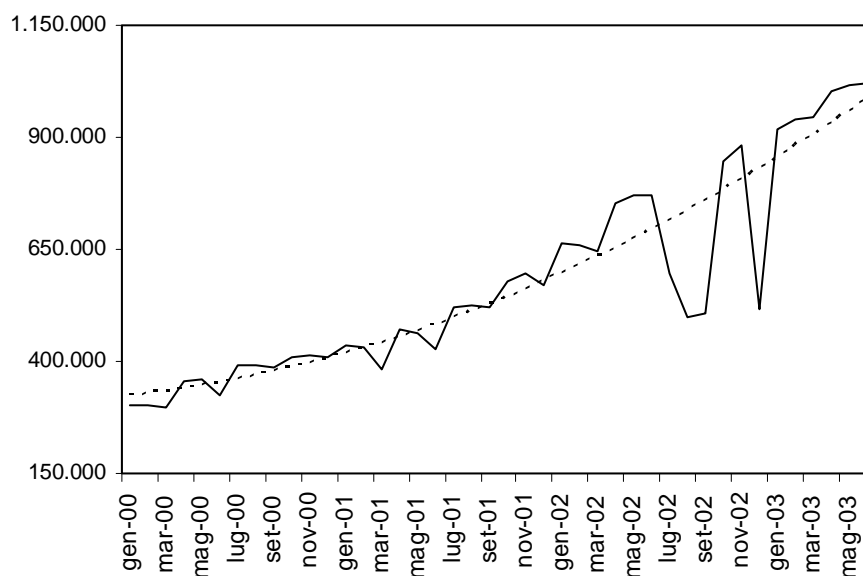
Since OROS delay in the production of the indicators may not be greater than 90 days from the reference quarter, Istat asked to INPS to withdraw the electronic monthly data as soon as they are uploaded on the central database, i.e. before they go through the check procedure that INPS normally carries out on the DM10. INPS collects these data in a special file and transmits it to ISTAT after about 45 days from the end of the reference quarter.

This sub set of data is a sample of the entire population of the DM10. Of course, it is “non random” because of the way of selection of the units (forms filled out by the enterprises on magnetic media). Nevertheless it is extremely large (now about 1 million of units) and it covers all firm sizes, economic activities and geographical areas. It represents new births and it has been observed that once the firms enter in the sample they normally do not exit (they do not change delivering mode).

OROS makes use of this set of data to produce “preliminary” estimates referring to current quarter  $t^5$  (see par. 7.1). Because of delays in the delivery and registration of the non-electronic DM10, INPS can transmit to Istat the complete information about the whole firm population referring to month  $t$  (1,3 million of units per month) only after 13-14 months from the end of the reference period.

OROS uses this data to produce the final (census) estimates referred to quarter  $t-5$ . Furthermore, given the methodological approach for the production of the provisional estimates, census data are also used as auxiliary information (referred to  $t-4$ ) to improve the preliminary estimate of current quarter  $t$  (see par.7.2).

**Figure 1 - Number of DM10 transmitted by electronic mode (January 2000 – June 2003)**



#### 4.3. The Monthly Survey on Labour Input variables in Large Firms

The Monthly Survey on Labour Input variables in Large Firms (LES) refers to enterprises employing 500 persons or more, in the sector from C to K of the NACE Rev. 1<sup>6</sup>. The survey covers approximately the 21,9 per cent of the total amount of employed persons according to the ASIA

<sup>5</sup> It is important to note that when estimates are performed, not all the electronic DM10 can be used. For statistical purposes it is necessary to make a selection of the forms that can enter into the estimation process.

<sup>6</sup> The construction sector has been included in the survey since the revision of the base realized in the year 2000.

register referred to the year 2000 (17,3 per cent in industry and 27,0 per cent in services). The LES survey was established in the early seventies to produce monthly indexes, with a delay of about 70 days from the end of the reference month, measuring the change of working hours, gross wages, labour cost<sup>7</sup> and of the number of employed persons, broken down for the following three categories: manual workers; non-manual workers; total employees (not including executives). The indexes base changes every five years. In the actual base year (2000) the number of kind of activity units (KAU)<sup>8</sup> in the survey is equal to 1337. According to the fiscal codes these KAU correspond to 1001 enterprises.

Different reasons explain why it is necessary to integrate OROS data with information coming from the LES. First, two extremely large firms (Public Railways and Posts and Telegraphs) are absent in the INPS register: because of their public nature they pay social contributions to a different Social Security Institute (INPDAP). Secondly, large firms are not covered completely in the non-random INPS sample, causing problems in the estimation of the preliminary indexes. On the contrary it is necessary to have a census coverage of the large firms (the traditional “take all strata”). Only 920 enterprises originally presents in the LES, are in the INPS population. Furthermore, in the sample we found 171 large firms but only 102 of them are completely covered. In fact in INPS Register firms, especially the largest ones, may have more than one contributive unit according to different administrative needs (see par. 4.2). This implies that if not all the contributive units are present, the firm is not completely covered. In brief, in 2000 the sample covers about the 10% of LES. As a consequence it becomes necessary to integrate the two data sources, INPS and LES, to produce adequate estimation of the target variables for all firm sizes.

Thus, since 2000 OROS uses INPS data to produce SME estimations, while it mainly uses data drawn by the LES for large firms<sup>9</sup> for both final and the preliminary estimates.

The most important thing to do in order to realize the integration of the two sources is to find out the firms of the survey from the INPS Register and substitute them with the economic information coming from the LES. The main problem is that there is only one matching variable, i.e. the fiscal code, which is not sufficient to identify the same firms in the two different archives, because it can be affected by formal errors or be updated in different time. In order to correctly identify LES enterprises in INPS a complex work based on the use of auxiliary information has become necessary.

## 5. Procedures Flow

Data transmitted by INPS must be submitted to a complex treatment procedure, of which a summary is schematically given (figure 2):

1. the first stage consists in the retrieval of the statistical variables that consists in trans-coding and aggregation of the elementary variables present on the raw data. In this phase, it is necessary to re-aggregate several “employment” type and “contribution” type variables associated to codes present on the DM10 on the basis of the contribution homogeneity;
2. once raw data has been reorganised in a much more handy format, a check and correction procedure is performed: this step is carried out at the level of monthly data through cross-sectional and longitudinal edits;
3. since estimations are provided at quarterly figures, aggregation of monthly data to quarterly information is necessary. At this step some correct individual characters are attributed to each unit, mainly on the basis of the BR;
4. before going on to the estimates, the individuation and correction of unit non responses is performed. At the moment imputation of missing variables does not concern the preliminary

---

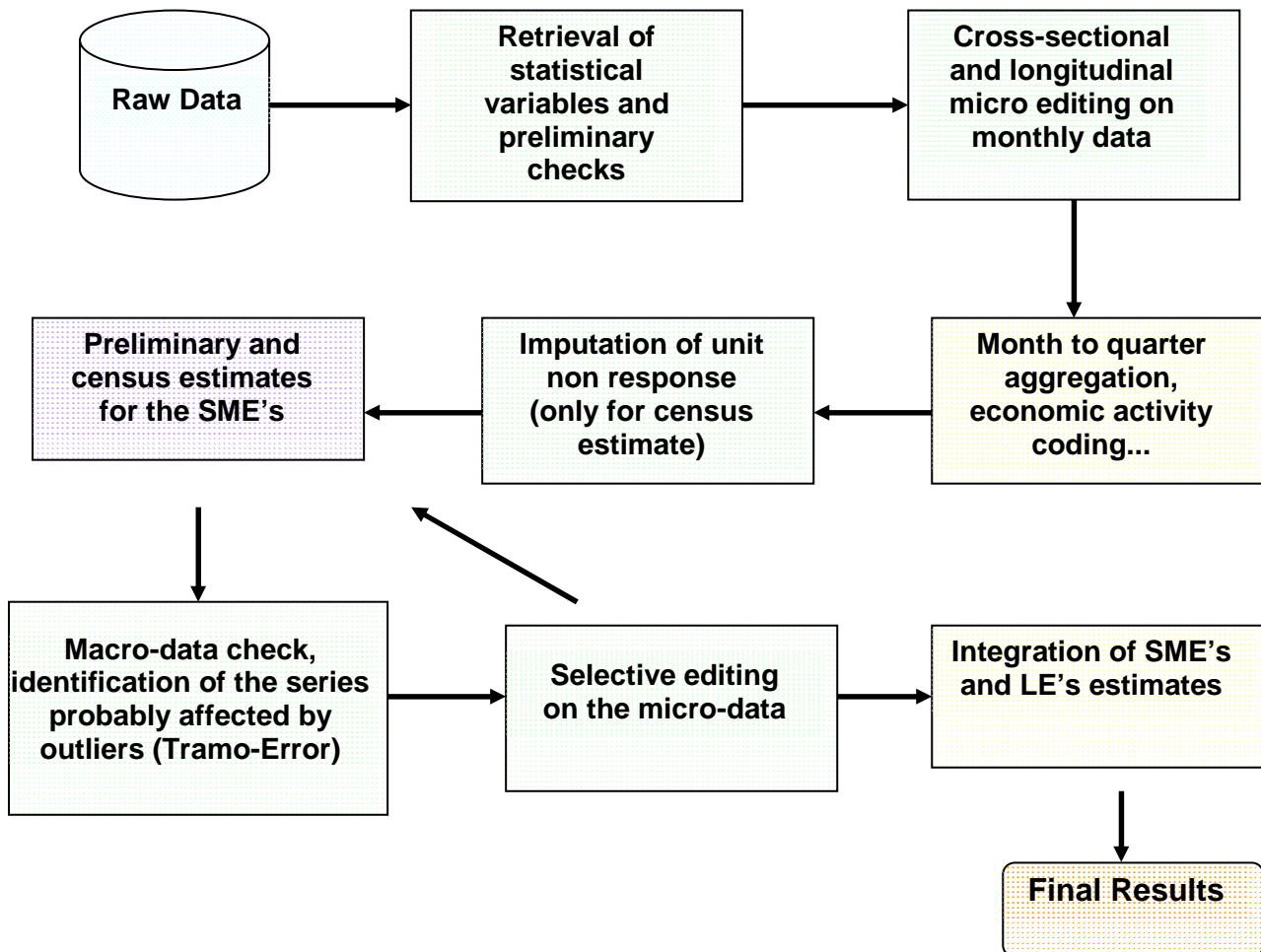
<sup>8</sup> KAU is the statistical unit on which the survey is based.

<sup>9</sup> LES covers the majority of the biggest units but not all (for example the very new ones) . The missing unit are covered with INPS data.

estimates;

5. preliminary estimates on quarter  $t$  and final estimates on quarter  $t-5$  are realized using all the information shown above: at this step estimates are calculated using only the INPS sources. Preliminary estimates are calculated on the basis of a calibration estimator that makes use of the sample and, as auxiliary information, of the variables of the universe of  $t-4$
6. once estimates have been produced, a macro-data check is carried out for the identification of the series probably affected by outliers by looking at the evolution of the series and through the use of a software for the analysis of time series (TRAMO for ERROR);
7. in some cases, the individuation of anomalous data in the time series brings back to checks on micro data, through a selective editing;
8. an eventual correction of micro-data implies a return to step 5;
9. estimates on SME's are integrated with the data of the Monthly Survey on Large firms.

**Figure 2 – OROS Procedure Flow**



## **6. Retrieving and translating administrative information in statistical variables**

The conversion of the administrative data into the required statistical variables implies complex computational aspects. INPS-DM10 form appears as a detailed grid partitioned in four sections. Individual data about firms fills partition A; the number of employees classified by type of employment, associated wage bills, paid days and social contributions constitute partition B/C; finally the partition D concerns credit terms and the tax relieves. Each information is identified by an id-code: more than 800 different codes can be found.

Each code consist of four digits (character/number). Many of them, the “employment codes”, result as the combination of three blocks:

- the first block, that is the first digit, identifies both the type of employment (worker, clerk, executive, manager, etc.) and the working time (full-time, part-time);
- the second block, that are the second and the third digits, defines the Social Security contributions;
- the third block, that is the fourth digit, identifies some wage peculiarities.

Besides, some other four digit codes, the “contribution codes”, are associated to particular contribution rebate defined by law. Finally, others, that we call “statistic codes”, are related to statistical purposes. These latter, because not necessary to compute the contributions to be paid, are not always filled in by the firms.

Retrieving the statistical variables strictly depends on a deep knowledge of the meaning of the DM10’s codes. For this purpose the building and the up-to-dating of a database of codes is a fundamental task. Almost twice a year legislation should be examined to notice the introduction of new codes and suppression of old ones.

Before drawing out the statistical variables, the DM10 form passes through a complex preliminary check procedure to verify the validity of each code and of the relative information recorded. In particular, errors in the i.d. of the units, duplications and formal coherence with the current legislation are checked and, where possible, corrected.

The process for the retrieving of the interest variables consists of two steps: the calculation of employment and wages, and the computation of social contributions. The first procedure aims to select the appropriate codes that identify unambiguously the number of the employees and the associated wage bills, avoiding possible duplications. The second procedure provides the calculation of the other labour cost (OLC). As the employee contributions are already included in the gross wages, and DM10’s codes refer only to the total (employer + employee) social contributions due to INPS, from this total it is necessary to remove the employee social contributions, through the application of the rates known by law. Besides, to obtain a complete estimate of the OLC due to the employers, other labour costs such as employers’ injuries insurance premiums (INAIL) and termination of employment relationship allowance (TFR) are to be added because they are not recorded in DM10 form.

Thus to perform all the work described in this paragraph Istat needed to build and up date continuously a sort of input meta-database on 1) law and regulation, 2) contributive rate, 3) codes and other technical aspects of social security (“Banca dati normativa su retribuzione e contribuzione”) for the definition of the translation scheme. Although the help of INPS the fulfilment of the translation of social contributions administrative data into statistical information has been a very new, hard and time consuming work.

## **7. Microdata check and editing**

After data has been reorganized in the required statistical format, a check procedure becomes necessary to evaluate their quality at monthly micro level.

A first preliminary check and automatic editing is carried out for a large number of variables: this operation following the social security rules try to edit cross sectional incoherence among the number of employees, the wage bills and the social contribution bill.

Then a more complex and proper process of editing for measurement errors is carried out only on data coming from the sample. In fact, as INPS downloads for the OROS Survey the entire DM10 population with a delay of fifteen months as regards to the analysed period, this data is already partially checked for measurement errors by INPS itself. This enforces the choice to skip the microediting procedure on the population data because it should have too high costs in terms of human and time resources with a little advantage in terms of quality (see par. 8.3 for the check and selective editing of the population data). On the contrary, as INPS downloads the DM10 sample with a delay of only one and half month, this set of data is not submitted to any controls.

The microediting procedure, developed with a specific sas program, is very selective in nature and combine automatic and manual control. First of all, it must select among about five hundred thousands of units those to be checked assigning a weight in term of probability to contain an error in the main variables. The units are checked through some functional relations among the analysed variables (based both on cross-sectional and longitudinal consistency) and then sorted. The largest values, that the procedure identifies with certainty as large measurement errors, are automatically corrected, while the others are selected, interactively analysed and if necessary corrected. The main rules the editing procedure is based on are:

- a positive amount of wage bills must correspond to a positive amount of employment, and often to a particular rate of social contributions;
- the amount of employees recorded in the current month should not be much different to the amount of the previous month;
- the gross wage per FTE, or the paid days per FTE, should have similar and acceptable amounts during the analysed period;
- the share (rate) of social contributions on gross wages should be included in an expected range, etc..

The number of edit performed is globally very small and it is enough for giving an high quality to the micro data.

## **8. Target estimation methodology**

### **8.1. Preliminary estimates**

The methodology developed for yielding quarterly estimates of employees, wages and other labour costs uses the administrative data integrated with the Large Enterprises Survey (LES) data. In fact, for reasons related to the quality of largest firms data of the sample, the methodological choice has been to use the data from the INPS sample to cover the population of the SME and the data of the LES (integrated with the INPS data related to some of large enterprises excluded by the LES) to cover the population of LE. In what follows the focus is on the estimate for the SME.

The methodology has been developed taking into account the information available from INPS above described, that is the three main data sources: the INPS register, available at the end of each quarter, the DM10 sample, available quarterly and the DM10 universe, available with the delay of four quarters.

The main characteristics of these sources are:

- i.* although the sample is non-random, its size is huge and growing over time as firms adopt the electronic way to send the DM10 forms. Besides, it has a good degree of coverage of the target population as regard the breakdown by economic activity sectors and the age of firms: specifically, the sample includes a large number of births;
- ii.* the Administrative Register (AR) is the most updated available representation of the current population: whereas the Istat BR has up to two years delay, the AR is potentially updated

currently. More precisely, as previously said, all births are covered to satisfy the obligation of registration to set up an enterprise with employees; conversely, the deaths are not frequently removed because of the lack of communication of the end of activity.

The following tables illustrate some of the points addressed above. Table 2 illustrates the coverage in terms of economic activity in the first and second quarter 2002. The sample contains 508,876 units in the first quarter and grows up to 621,489 units in the second quarter. As can be seen it covers all the NACE sections with coverage rate greater than 36% (units) and 33% (employees) in the first quarter and greater than 39% (units) and 38% (employees) in the second quarter. Another point that can be noted is that the rate of employees coverage is almost always less than the rate of units coverage: it means that the sample selection is slightly biased toward the smaller firms.

**Table 2. Coverage of the sample in terms of economic activity: units and employees. Year 2002, I and II quarter.**

Economic Activity	2002 Q1				2002 Q2			
	Units		Employees		Units		Employees	
	Number	% Universe	Number	% Universe	Number	% Universe	Number	% Universe
Mining and quarrying	1,282	43.8	10,890	42.4	1,573	52.8	13,450	50.5
Manufacturing	141,409	49.0	1,434,615	44.5	168,862	58.3	1,725,526	53.1
Electricity, gas and water supply	545	36.2	11,321	33.6	610	39.7	13,073	38.2
Construction	83,244	45.4	370,689	44.1	103,290	54.5	468,395	53.5
Trade and repair	135,733	45.4	589,640	46.0	164,102	53.8	714,006	54.3
Hotels and restaurants	41,649	43.5	164,993	42.8	56,271	52.3	238,570	51.2
Transport and communication	22,395	43.8	167,158	36.3	27,098	52.6	203,230	44.3
Financial intermediation	9,586	44.1	53,044	33.7	11,506	52.5	63,503	38.9
Real estate, renting and business activities	73,033	44.7	375,820	40.2	88,177	53.3	457,361	48.2
Total	508,876	45.9	3,178,169	43.3	621,489	54.7	3,897,114	51.7

Table 3 compares the AR and the sample as regarding the number of births. Here a birth is defined as a unit whose registration date is within a year back from the end of the reference quarter. It's interesting to note that the percentage of births in the sample is very similar to that in the population. For example, in the first quarter the percentage of births in Industry in the sample is 9.6% while in the AR is 10.0.

**Table 3. Number of births in the AR and in the sample 2002 I and II quarters**

Economic activity	2002Q1		2002Q2	
	Number	%	Number	%
<b>Administrative Register</b>				
Industry	60,014	10.0	60,315	9.9
Services	96,546	12.1	96,497	11.9
Total	156,560	11.2	156,812	11.0
<b>Sample</b>				
Industry	21,760	9.6	26,625	9.7
Services	33,361	11.8	40,651	11.7
<b>Total</b>	<b>55,121</b>	<b>10.8</b>	<b>67,276</b>	<b>10.8</b>

The coverage of births, both in the sample and in the Register, has made possible to develop an ambitious methodology to estimate the level and the trend of the target variable of the current population and not referred to a fixed population in the past. By the way, the methodology has to solve the two problems related to the data sources that are the **non-randomness** of the sample and the **over-coverage** of the AR.

The methodology that was first built to cope with the non-random sample was based on a prediction model assuming that a relation exists between a target variable and the same auxiliary variables (Royall, 1988). The model specified, estimated on the sample units, was used to predict the target variables on the non-sample part. In order to control the self-selection bias the model was fitted within homogeneous subgroups of the population (named *model groups*) (Hidioglou et al., 1995). The model groups were obtained by partitioning the population according to four variables: the economic activities (at two digit level), the size classes (four modes), the geographical areas (four modes), the age of firm classes (two modes). They amounted to over five hundred groups, totally. Then, this methodology has been modified to produce coherent estimates of the target variables: in fact the main parameters consist of ratios of target variables (e.g. wage per full-time equivalent employees) so the consistency of the numerator and the denominator has to be granted. The current methodology is based on giving a weight to each unit of the sample. The weights are calculated to satisfy the condition that the sum over the units of the sample of an auxiliary variable multiplied by the weight of the unit, is equal to the known total of the auxiliary variable (calibration). These must be true for each auxiliary variables. The selected variables are the number of firms available from the AR and the employees at  $t-4$ , the wage bills at  $t-4$ , the social contributions at  $t-4$  available from the total population of DM10 referred to one year before. Certainly, the amount of the last three auxiliary variables is null for the units born within the last year, so in that case the calibration procedure is based on the number of firms and the employees declared at the registration of the firm both coming from the AR.

The calibration is performed at model groups level above mentioned.

To present the methodology in a more formal way, let:

$$E_t = \sum_{i \in P_t} e_{ii} \quad \text{the total number of employees (expressed in FTE) referred to the units that}$$

$$\quad \quad \quad \text{belongs to the population } P_t \text{ at time } t \text{ period;}$$

$$W_t = \sum_{i \in P_t} W_{ii} \quad \text{the wage bill referred to the same population;}$$

$$w_t = \frac{W_t}{E_t} \quad \text{the wage per full-time equivalent employees at time } t .$$

The estimate of the variable  $w_t$  can be written as

$$\tilde{w}_t = \frac{\sum_{i \in S_t} w_{ii} K_{ii}}{\sum_{i \in S_t} e_{ii} K_{ii}} \quad \text{where } S_t \text{ is the sample at time } t \text{ period and } K_{ii} \text{ represents the weight given to}$$

the firm  $i$  from the calibration procedure.

The weight  $K_{ii}$  is obtained by solving the following problem of constrained minimum:

$$\left\{ \begin{array}{l} \underset{\{K_{ii}\}}{\text{Min}} \left[ \sum_{i \in S_t} c_{ii} (K_{ii} - 1)^2 \right] \\ \sum_{i \in S_t} K_{ii} \mathbf{x}_{ii} = \mathbf{X}_t \end{array} \right.$$

In the objective function,  $c_{ii}$  is a constant that depends on the size classes of the firm  $i$  and 1 is the sample weight given to the unit  $i$ , because of the lack of a sample design. The constraint is a system of equations where  $\mathbf{x}_{ii}$  is the column vector of the auxiliary variables  $\mathbf{x}$  of the firm  $i$  at  $t$  period, and

$\mathbf{X}_t$  is the vector of the known totals of the auxiliary variables. These known totals are obtained by summing up over the current population the auxiliary variables available in the DM10 universe referred to one year before as regards the active units at that time and the variables available in the AR for the units born within the last year. In formulas:

$$\mathbf{X}_t = \sum_{i \in AR_t} \mathbf{x}_{ti}$$

The calibration procedure and the use of the model groups are the tools that should eliminate or reduce the possible bias involved in using a non random sample.

A solution to overcome the over-coverage of the AR has been found in reducing the known totals of the calibration procedure by applying to each unit in the AR a probability of being active.

$$\tilde{\mathbf{X}}_t = \sum_{i \in AR_t} \tilde{p}_{ii} \mathbf{x}_{ti}$$

So that the calibration procedure becomes:

$$\left\{ \begin{array}{l} \underset{\{K_{ii}\}}{\text{Min}} \left[ \sum_{i \in S_t} c_{ii} (K_{ii} - 1)^2 \right] \\ \sum_{i \in S_t} K_{ii} \mathbf{x}_{ti} = \tilde{\mathbf{X}}_t \end{array} \right.$$

The probability is estimated at  $t$  on the data referred to  $t-4$ : in fact, at  $t$  there are available the AR and the DM10 universe referred to that period. The probability is calculated dividing the number of firms for which the DM10 form is arrived by the number of active firms according to the AR. This estimate is performed within homogeneous groups (named *register error groups*)

In formula the probability we are interested in is:

$$p_{ii} = \frac{n(U_t)}{n(AR_t)}$$

where  $n(U_t)$  is the number of units in the Universe at time  $t$  and  $n(AR_t)$  is the number of active units according to the AR at time  $t$ . Since at time  $t$  the Universe of  $t$  is not available, this probability can be estimated using the ratio of time  $t-4$ , that is:

$$\tilde{p}_{ii} = \frac{n(U_{t-4})}{n(AR_{t-4})}$$

The use of the data of time  $t-4$ , to estimate the probability at time  $t$  implies the hypothesis that the phenomenon of overcoverage is roughly constant. This is confirmed by the data analysed so far.

## 8.2 Final estimates and imputation of unit non-responses

The main data sources used for the final estimates are: the AR of the reference quarter for the individual information on administrative units and the total population of DM10 arrived at INPS with a delay of 14 months after the end of the reference quarter. As for the preliminary estimates, the data related to the majority of firms with more than 500 employees is drawn by the LES.

The population of DM10 used in the final estimates do not represent the totality of forms effectively referred to the reference months because models normally continue to arrive also after several years. The reasons standing for these delays are:

- enterprises may present the DM10 to the INPS local offices later than the official date;
- data entry of local offices to INPS central database may take time.

The following example can give an idea about this problem. The temporary sequence of the DM10 arrivals for one reference month (January 1997) to the INPS central office is more or less the following: 82.5% of the forms related to January 1997 arrived within the end of the first quarter 1997. After a year of delay, their weight was not less than 97%. The residuals tend to arrive much more slowly. It was observed that DM10 arrived with a delay of more than one year concern small size units: they cover about 1% of total employment recorded in the archive and represent a stable percentage of 2-4% of the total DM10 referred to the reference quarter.

The forms which arrive to INPS central office with a delay longer than 14 months represent unit non-responses. Although they cover a small proportion of total population (both in term of units and employees), imputation is necessary for at least two reasons:

- delays could interest a big unit, with a relevant impact on the estimates;
- unit non-responses can be un-influent on the estimation of the labor costs but they can considerably affect the estimate of employment.

The most critical aspect to afford in the reconstruction of unit non-responses is their individuation. In fact, in traditional survey the list of expected non-respondents in each reference period is available because it is simply derived as a difference between the list of the units belonging to the sample and the set of the respondents. In the case of OROS, a list of the units which should sent the DM10 form is not available, because it corresponds to the population of active units in the quarter. Considering that the Administrative Register suffers of over-coverage problems, this population is clearly known only when all the DM10 are available, and this can require several years.

In order to find out the units to be imputed, a prediction of their state of activity must be done. This is made possible through the analysis of the patterns of presence of the DM10 in a pre-determined span of time, with the help of some auxiliary information.

First, a list of reference units is built: here are included all the units which have presented the DM10 in at least one of the months of:

- the reference quarter;
- the previous four quarters;
- the following quarter;

for a total of 18 months. The use of a set of quarters around the reference one is due to the hypothesis that the quarters close to that of estimation can be informative on the latter. This is true if it is low the probability that a latecomer position in a quarter is latecomer also in its quarter neighbours. If it would not be so and the latecomers are always the same units, these ones will never be in the reference list and could not be imputed. It was observed as the majority of the delays over 14 months have a low persistence. This means that the patterns of presence can effectively be used for the individuation of the delays.

The distinction between a normal absence of a DM10 and a unit non-response, intended as the omission in the sending of a DM10, is not a easy task. In fact a unit could be simply inactive in a period and its imputation could mean an over imputation of the interest variables. Furthermore there could be some units whose activity is seasonal. In order to make a correct individuation of the states, auxiliary information on the units are used.

The procedure built for the correction of units non-responses shows that on the first quarter 2001 the 80% of DM10 not present in the quarter are likely to be inactive. The majority of them are dead or still not in life. About 1% of the remaining can be attributed to seasonal activity, while for the 9% of the absences it is not possible to make clear hypothesis on the status of activity: they are

classified as “uncertain”. Finally about 12% of the DM10 absent can be submitted to a procedure of imputation.

In order to avoid over imputation due to the not absolute strictness of the procedure for the individuation of unit non-responses it was decided to exclude from imputation the units which belong to small size classes (less than 20 employees). Vice versa units with more than 500 employees are always imputed when missing.

Different calculation rules are applied depending on the variables to be reconstructed. While for employment the value fairly updated of the closest quarter is selected (under a constraint of growth), for the reconstruction of gross wages it is necessary to take into consideration the strong seasonality of this variable. Furthermore auxiliary information are selected from the correspondent quarters of the previous years or are represented by average values calculated on homogeneous cells of units. As concerns the estimation of the other labor costs, the reconstruction is done through the estimation of an average contribution rate which, given the relative short term invariance of legislation, should be selected as close as possible from the month to be reconstructed. When possible this is done using auxiliary information coming from the same unit, otherwise from cells of units which are homogeneous with the one that has to be estimated.

The imputation carried out the first quarter 2001 shows that the 90% of the selected unit non-responses is eligible for imputation, given the available information, but only the 23% of the units which are eligible can be imputed. These rates are almost stable over the time. In terms of total employment, the imputation implies an increase on the number of employees amounting to about 2% (on a share of units to be imputed which is about 1%).

After the imputation of units non-responses, estimates of the target variables are performed at level of division of economic activity, eliminating large firms that are included in the LES panel. Estimates over these firms are realized through the integration with data of the LES Survey (see par. 4.3).

### **8.3 Macro data validation, checks and selective editing**

When the preliminary and final estimates have been produced, macro series are submitted to an “manual” quality control in order to identify anomalous trends in the period of interest.

Quality controls based on manual inspection are generally conducted with the help of some simple measures, such as comparing the last two period data on the variables, checking if the new observation is an overall maximum or minimum etc.. However these measures do not take into consideration the full information contained in the series that can be affected by seasonality, noise, or special events. Moreover, non-automatic checks affect strongly the number of series that have to be analyzed, limiting controls at a relatively high level of aggregation.

When allowed processing time is short but of course data quality is required, it would be convenient to dispose of automatic methods, reliable and efficient in large scale applications, that would consider the full information on time series and permit controls at a more disaggregate level.

The detection of errors by mean of simple time-series models is the object of a vast quantity of studies. Gómez and Maravall developed a methodology for automatic identification of ARIMA models when observations may be missing and the time series may be contaminated by outliers and special effects (TRAMO program – Time series Regression with ARIMA noise, Missing values and Outliers) (Gómez and Maravall, 1996). A particular application of TRAMO gives the possibility to execute automatically the problem of quality control in time series (TERROR program, that is TRAMO for errors) (Caporello and Maravall, 2002). The TERROR judging criterion for the evaluation of a new observation to be a suspected error is to verify if it is very far from what could have been expected looking at its past history. The program identifies a REG-ARIMA model for each series and obtains the standardized forecast error for the period associated with the new data, which is not considered in the process of model identification and forecasting. When the forecast error is, in absolute value, larger than some a priori specified limits, the new observation is

identified as a possible error. Depending on how sensitive the detection of errors should be (it can be fixed by the user), the new observation can be classified as “likely” or “possible” error. TRAMO is a very rapid program, which permits to handle a large number of series in short time. The program requires quarterly series to have a minimum of 16 observations.

Each quarter, the normal procedure for the production of the interest indexes implies the detection of 47 series per each variable. TERROR performs the control of all series in a very brief time.

After the identification of the anomalous sectors through TERROR algorithm, the process turns back to micro data to perform a selective editing procedure (SE), in order to investigate among the units that have most influenced the change of the series.

This common procedure has a little different purpose if controls are made on the results of the preliminary or the final estimates. As before mentioned (see par. 7) sample units are previously submitted to the microediting procedure, so aimed at the correction of the results of the preliminary estimates, the SE is processed to check the estimation weights (see par. 8.1): the weights of the anomalous units should be reduced although they are not clearly errors. For this purpose, Eurostat advises to consider these units oneself representative. Instead, as the universe units are not controlled by the microediting procedure, the aim of the SE is to select and to correct the remaining errors after INPS checks.

Despite the different aims, in the SE the procedure to identify the units that have a relevant influence on the series, is the same for data contributing to the preliminary and the final estimates. It is based on the attribution of a weight for each unit, describing its importance to determine the analysed change on the series. The SE method takes also into account for both births and deaths that in some cases can have relevant effects on the patterns of series. The micro-data analysis is limited to those units that have a considerable impact on the estimates.

#### **8.4 The revision error**

As the final estimate is published, the data first released from the preliminary estimate are revised. The nature of this revision is comprehensive of a number of factors. The most important refers to the type of the estimation: while the preliminary estimate is based on a sample, albeit of big size, the final estimate, based on the universe of the data, can be defined a census. For this reason this revision is different in nature from all the revisions implied by those surveys that issue a fast, preliminary estimate based on the subset of respondents that have already been contacted and later release figures based on larger sample. Even if this looks like similar to OROS revision it is quite different. The number of the units on which the final OROS estimate is based is far larger than those of the sample. Consider as an example that the universe, whose size is stable at 1.3 million of units, is 2.5 times the sample used in the estimation of the first quarter of 2002 and 2.1 times the sample of the second quarter.

The revision error is used in two main ways: analysis and benchmarking. The analysis consist of the study of it to understand the causes with the aim to reduce it. They include the break down by sectors and size classes and the time series analysis. The time path of the revision error has shown that it is roughly constant and, as regards to the labour cost variables, probably due to a residual selection bias of the sample: until now the units with lower wages and labour costs seem to be over-represented. This characteristic has been found to be useful to correct the preliminary estimates currently produced. Given that the length of series of the revision errors are still too little<sup>10</sup> to use a time series model (ARIMA, State-Space Modelling), they are currently being used to implement a simple benchmarking procedure. Briefly, the preliminary estimates of time  $t$  are adjusted for the revision error of the estimate  $t-4$ .

---

<sup>10</sup> At the moment of writing they are available since the second quarter of 2002 and they can be produced back since the first quarter of 2000.

## 9. Final remarks

In the paper we summarised the main methodological and procedural aspects of the OROS Survey. On the base of the OROS experience some remarks can be made about advantages and disadvantages in the use of the administrative data for short term statistics.

The advantages are:

- very low collection cost;
- complete firm size coverage, timeliness, estimates precision;
- no statistical burden on firms.

The disadvantages are:

- huge handling of data;
- very complex process of production in a very short schedule;
- complete dependence from INPS; relative risk of inconsistency and discontinuity of the information over time if the social security declarations are modified or cancelled.

Also taking in consideration all possible future cost of the use of social security data the benefits are clearly overwhelming. To reduce the risk highlighted in the last point Istat made an effort to create very strict and cooperative relation with the Social Security institute: Istat and INPS signed a framework agreement with a strong commitment to co-operate; an high level co-ordination committee supervise all the bilateral relations; besides Istat has the right to influence the process of modification/cancellation of administrative information collected by INPS.

There are some other important issues which arise from the OROS experience.

- The importance of the conceptual/definition and translation problem which must be correctly addressed and continuously monitored. First concerning the identification and understanding of all administrative and technical concepts, definitions, names and codes, and their comparison with official statistics concepts and variable definitions; secondly referring to the translation or reconciliation of the administrative data into the required statistical variables, measuring, if possible, the effects of the residual gaps and bias; thirdly the computational aspect, which occurs when the statistical variables has to be calculated from a huge number of wage and contribution items and sub-items, by addition and or subtraction of more than 800 different employment and contribution codes.
- The capture and process of mass of data increase the necessity of a very selective check and editing procedures; it must combine selective and interactive editing.
- The possibility to use non-random (convenience) big sample in combination with appropriate statistical methods to produce an unbiased estimate of the target parameter. The tools used by OROS are an appropriate partition of the population and the calibration procedure. While overall the method works fine in estimating the change of the labour cost variables a number of issues have to be addressed, and need further analysis. The estimation of the level of the variables seems to be still biased in some cells when we compare the preliminary estimate with the final one. A second methodological problem arise when a large upswing of the sample size occur. This mainly happens for reasons related to administrative procedures. An example can illustrate the point: the sample size experienced a large fall in few quarter in 2002 and 2003, because the transmission of the data from the INPS local offices to the central one was delayed due to the software being fixed.

## References

- Baldi C. and Rapiti F. (1999) "Wages and employment official statistics using INPS data: a preliminary proposal and some methodological and quality problems". *Contributi ISTAT*, n.16/1999.
- Baldi C., Falorsi P. D., Pallara A., Succi R., and Russo A. (2000), "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.
- Baldi C., E. Cimino, F. Rapiti, P. Minicucci, R. Succi, D. Tuzi (2001), "L'utilizzo dei dati INPS per la stima trimestrale del numero dei dipendenti, le retribuzioni, il costo del lavoro e le ore lavorate", Documenti Istat, n.14.
- Baldi C., Pacini S. (2003) "L'integrazione delle grandi imprese nelle stime OROS per la stima congiunturale di occupazione e retribuzioni", Relazione tecnica, giugno 2003,
- Caporello G and Maravall A. (2002), "A tool for quality control on time series data. Program TERROR". Bank of Spain. December 2002.
- Falorsi P. D., Pallara A., Succi R. and Baldi C. (2001), "La stima delle variabili occupazione, retribuzioni, e costo del lavoro basata sui dati INPS-DM10", , documento interno Istat.
- Falorsi P. D., Pallara A., Succi R. and Russo A. (2000), "Estimating indicators of labour input from administrative records having coverage and measurement errors", ICES II – Proceedings of the Second International Conference on Establishment Surveys, (available at web site <http://www.eia.doe.gov/ices2/>).
- Gómez V. and Maravall A. (1996), "Program TRAMO (Time series Regression with Arima noise, Missing observations, and Outliers) and SEATS (Signal Extraction in Arima Time Series). Instructions for the User", Working Paper 9628, Servicio de Estudios, Banco de España.
- Hidiroglou, M.A., Latouche, M., Armstrong, B., and Gossen, M. (1995), "Improving Survey Information Using Administrative Records: the Case of the Canadian Employment Survey," *Proceedings of the Annual Research Conference, U.S Bureau of the Census*, pp. 171-197.
- Rapiti F. (2000), "Use of INPS administrative data to estimate the number of employees, wage, and hours worked" Report to Eurostat, STS Regulation Contract no. 9.442.022/1.
- Succi R., C. Baldi, E. Cimino, D. Tuzi, Pallara A. (2000), "Una sperimentazione dell'utilizzo delle dichiarazioni mensili e dell'anagrafe delle posizioni contributive dell'INPS per la stima trimestrale dell'occupazione dipendente", documento di lavoro Istat.
- Royall, R.M. (1988), "The Prediction Approach to Sampling Theory", in: Krishnaiah, P.R., Rao, C.R. (eds.), *Handbook of Statistics*, vol. 6, Elsevier Science Publishers, pp. 399-413.