WP.8
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (ii): Editing administrative data and combined sources

THE EDITING PROCESS IN THE ITALIAN SHORT-TERM SURVEY ON LABOUR COST BASED ON ADMINISTRATIVE DATA

Invited Paper

Prepared by M. Carla Congia, Silvia Pacini and Donatella Tuzi,
Italian National Statistical Institute (Istat), Italy*

Abstract
The quarterly Italian Oros Survey on Wages and Labour Cost is a interesting example of an innovative use of administrative sources combined with survey data to produce short term statistics. The peculiarities of the Survey, characterized by a huge number of records, a highly detailed level of the raw data, differences among the population subgroups and timeliness in the release of the indicators, imply an extensive and complex check and editing procedure, covering the whole production process. The translation of administrative data into statistical information is the first aspect described in this paper: check and editing is necessary not only to correct administrative micro data but also to modify translation procedures and metadata errors to follow frequent changes in administrative concept and definitions. Once the statistical variables have been made available a more traditional micro level check procedure becomes necessary. Outliers editing and imputation of unit non-responses may consequently be needed, particularly in the case of influential observations. A focus is also put on the micro level integration between the administrative records and the Large Enterprises Survey data involving record linkage and the computation of harmonised variables. Finally, the paper describes the macro data validation process which implies, among other aspects, time series analysis and macro level comparisons to other statistical sources.

I. INTRODUCTION

1. This paper describes the main peculiarities of the editing and imputation (E&I) process of the quarterly Italian Oros Survey on wages and labour cost. This survey is based on an innovative and extensive use of administrative data with the purpose of producing timely short-term business indicators. Until 2002 these indicators were produced only by a monthly survey covering firms with 500 employees or more. The use of administrative data has allowed, on a quarterly basis, the extension of the target population to all enterprises with at least one employee, avoiding any further statistical burden on firms.

2. In Section II a short summary of the survey and a description of its sources are presented, while the peculiarities of the E&I process are briefly discussed in Section III. The E&I process in a survey based on administrative data differs substantially from that of traditional surveys: the preliminary checks and the retrieval of the statistical variables (Section IV) characterize the first and basic step. Afterwards, a more traditional micro level check is performed (Section V) followed by the imputation of the unit non-responses of a specific sub-population which has a relevant influence on the estimates: the temporary employment agencies (Section VI). Section VII focuses on the combination with survey data of estimates

* This paper has benefited from comments and suggestions from Leonello Tronti and Fabio Rapiti. All remaining errors are those of the authors.

for large firms. Finally, the macrodata checks with possible drill-down on microdata, are described in Section VIII.

II. THE MAIN FEATURES OF THE OROS SURVEY

3. The Oros Survey[1] produces information on quarterly changes in gross wages, other labour costs and total labour cost of the Italian firms with at least one employee in the private non-agricultural sector[2] (Baldi et al, 2004). Until 2002, Italian labour cost short-term statistics were produced only for large enterprises, with 500 employees or more, by the Monthly Survey on Labour Input variables in Large Firms (hereafter Large Enterprises Survey - LES). Considering that the Italian business population is mainly composed of small and medium enterprises, the Oros Survey was designed to extend the coverage to all business size classes, and avoid further statistical burden on enterprises through the use of administrative data collected by the Italian National Social Security Institute (INPS). The survey was also planned to satisfy two European Community requirements on short-term business statistics: the STS Regulation (n.1165/98) and the LCI-Labour Cost Index Regulation (n.450/2003).

4. Since 2003, quarterly provisional Oros indicators are released with a delay of about 70 days from the reference quarter. The estimates are revised after 15 months.

5. The INPS source is mainly used for the estimation of small and medium enterprises, while for large firms data are drawn from the LES because of their higher quality due to direct checks constantly carried out on each enterprise by the LES experts. In the Oros Survey four sub-populations are singled out and subjected to a specific editing:

(i) small and medium enterprises (SME);
(ii) large enterprises surveyed by LES (estimated with LES data);
(iii) large enterprises not surveyed by LES (estimated with INPS data);
(iv) temporary employment agencies.

6. The INPS sources used in the Oros Survey are:

- the archive of the monthly social contribution declarations (i.e. DM10 forms) that all firms with at least one employee have to transmit electronically to INPS within the 30th day after the end of the reference month. Given the Oros release deadline, Istat asked INPS to transmit the electronic monthly data as soon as they are uploaded on the central database, before going through the administrative check procedures. That set of information, transmitted to Istat after about 35 days after the end of the reference quarter, is used to produce the provisional estimates. That "provisional population" file is extremely large and covers about 95-98% of the full population which is available only with a delay of about 12 months and used to produce the final estimates (about 1.3 million employers covering 10 million employees each quarter);
- the INPS Administrative Register (AR), containing structural information on the single administrative unit, is available at the end of each quarter and regularly updated. It represents the current population but suffers over-coverage problems. In fact, while the entrance of new born units are correctly registered, the temporary suspensions and firm closures are under-recorded because firms have no administrative incentive to comply with the obligation to communicate these events[3]. In order to make the AR suitable for statistical purposes, some checks are carried out to improve the fiscal code quality, and to exclude units not belonging to the survey target population. Furthermore, AR is matched with the Italian Statistical Business Register (BR–ASIA) to acquire the official economic activity code. Although the BR is available with about a two-years delay from the reference quarter, roughly the 90% of active units get the economic classification from it while for the remaining ones it is drawn from the current AR.

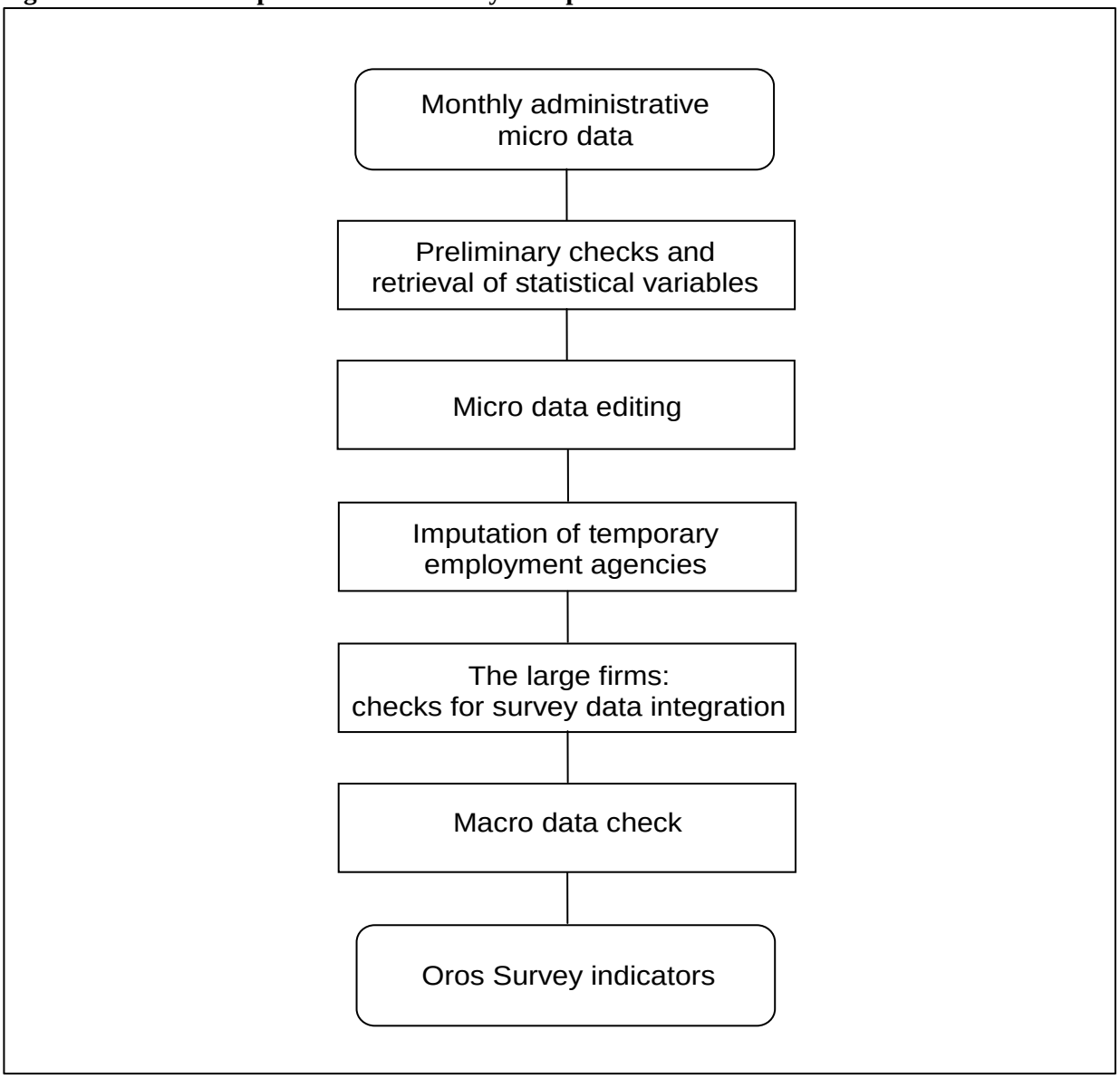
[1] The acronym Oros stands for Occupazione (Employment), Retribuzioni (Wages), Oneri Sociali (Other Labour Costs). At the moment, figures on employment are produced to calculate the per capita values but they are not released.
[2] Sections from C to K of the European Community classification (Nace Rev.1.1).
[3] The over-coverage problem interests about the 20% of the AR units.

III. THE PECULIARITIES OF THE EDITING PROCESS

7. Because of the administrative nature of the main sources and their integration with survey data, the Oros Survey's editing and imputation (E&I) process has some peculiarities.

- While in the traditional surveys many non-sampling errors may be prevented and/or reduced *ex ante* during the design phase, the general quality of administrative data is completely independent from statisticians' control. Moreover, some *ex post* traditional check and editing techniques, like the questionnaire revision and the enterprise recalling, are also not applicable to the Oros Survey.
- Capturing the whole raw DM10 declaration without any previous aggregation and check implies preliminary checks and a complex retrieval process of the statistical variables based on a "metadata database" that has to be built and continuously updated to take into account changes of laws, regulations, contribution rates.
- The high number of units included in the provisional population quarterly acquired from INPS makes necessary a selective editing.
- Because of the difficulties in the use of administrative information, a systematic sequence of checks is required along the whole process, also at macro data level and for the different sub-populations.
- A systematic documentation of the whole E&I process is carried out. Considering the frequent changes in administrative rules, the different versions of the procedures are quarterly saved (versioning). Moreover, all checks are recorded to maintain a useful time series of errors detected and corrections carried out.

Figure 1: The main steps of the Oros Survey E&I process



IV. PRELIMINARY CHECKS ON ADMINISTRATIVE DATA AND RETRIEVAL OF STATISTICAL VARIABLES

8. A peculiar procedure was implemented to exploit INPS declarations for statistical purposes. The administrative raw data needs to be preliminarily checked before retrieving the statistical variables. This step implies complex computational aspects due to the highly detailed level of the raw data, their huge quantity, and the frequent changes in the administrative and legislative rules.

9. The monthly form, used by the employers to declare compulsory contributions, appears as an extremely detailed grid partitioned in sections where information about the firm, the number of employees by type of employment, the wage bill, the paid days and the social contributions, credit terms and tax relieves are registered. Information is identified by four digits (character/number) codes[4], used to classify the employment relationships, the working time, the contributions due or rebates to be received, the wage peculiarities, etc. Some statistical information is also requested, but not always filled in by the firms because unnecessary to administrative purposes. For a proper translation of the administrative data into statistical variables a deep knowledge of these codes is required, so a list of codes has been set up and its updating is a fundamental task in the process: each quarter legislation has to be examined to take note of the introduction of new codes and the elimination of other ones.

10. Before the translation of administrative information, the DM10 forms go through a complex preliminary check procedure aimed at investigating and possibly correcting errors on codes, record duplications, incoherencies with current legislation, etc. The original dimension of raw data is about 10 million records per month because each form is split up into several records. In this step, information referred to each DM10 is summarized in a single record, for a total about 1.3 million records per month.

11. Finally, the retrieval of the target variables is carried out in two steps: the calculation of employment and wages, and the computation of social contributions. At first, the appropriate codes identifying the number of employees and the related wage bill have to be unambiguously selected, trying to avoid possible duplications. Secondly, other labour costs (OLC) have to be calculated. As DM10's codes refer to total social contributions (employer + employee), the employee social contributions have to be removed from this total through the application of appropriate legal rates, because they are already included in the gross wages.

12. Besides, other labour costs such as employer injuries insurance premiums (INAIL) and termination of employment relationship allowance (TFR) have to be added as they are not recorded in the DM10 form.

13. To make all this work feasible an input "metadata database" has been built, containing information on laws and regulations, contribution rates, codes and other technical aspects concerning Social Security (Banca Dati Normativa su retribuzione e contribuzione - BDN). To be effective, the BDN must be continuously updated and this requires a very hard and time consuming effort.

V. THE MICRO DATA EDITING

14. After administrative data have been translated in the required statistical variables, a more traditional check procedure becomes necessary in order to find out possible anomalous values and correct them at a monthly micro level.

15. The micro editing procedure, developed with a specific statistical application (SAS FSEDIT procedure), is set up on very selective criteria. The selection is based on a weight representing the probability of an error in the target variables, assigned to each of the 1.3 million of units.

[4] At the moment, each quarter more than 5,000 different codes are in force.

16. Units are checked through some functional relations among the analysed variables aimed at evaluating both cross-sectional and longitudinal consistency using the information on the previous month. 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 number of employees recorded in the current month should not significantly differ from that of the previous month;
- the gross per capita wages, or the per capita paid days, should have similar and acceptable amounts in the analysed period;
- the rate of social contributions on gross wages should fall within an expected range, etc.

17. In the past, the largest values identified by the procedure as measurement errors were automatically corrected, but in the latest years the experience has suggested to avoid these automatic corrections because of the specific nature of the administrative data. So the most anomalous values are selected according to established cut-off thresholds, then they are interactively analysed and, if necessary, corrected.

18. The number of edits performed is globally not high but sometimes even the omission of one correction can have a substantial impact on the final estimates. Just to have an idea, in the 3rd quarter of 2007 an erroneous value of a single firm's number of employees would have determined an year-on-year change of 3.0% in the section G (wholesale and retail trade; repair of motor vehicles and motorcycles) instead of a 0.8% correct change.

19. The peculiarities of the administrative data used have a considerable impact also on gross wage and other labour cost distributions making the identification of the cut-off thresholds particularly problematic.

- The distribution of per capita gross wages, for example, shows that, besides the usual right tail area of the distribution, in the INPS data there is also a significant left tail area where a high number of units with very low per capita wages are concentrated. Normally these observations should be considered erroneous, but in this case they are the right representation of economic phenomena (for example firms with very few employees and all receiving only supplementary earnings by the employer). In this left tail area of the distribution it is very hard to distinguish wrong figures and the final risk is an asymmetrical correction of errors.
- The other labour costs distribution may show negative values, because of social contribution rebates. This aspect must be taken into consideration both to calculate correct check indicators and to single out all possible wrong data.

VI. IMPUTATION OF THE UNIT NONRESPONSES IN TEMPORARY EMPLOYMENT AGENCIES

20. Forms that arrive to INPS central office with a delay longer than 35 days from the reference quarter are treated as unit non-responses. The provisional population covers the 95-98% of the entire population, and the evidence shows that the non-responses have not a significant effect on the Oros wages and other labour costs changes except for those referred to temporary employment agencies. This group of firms has a considerable weight: about one hundred enterprises employ 3% of workers in the private sector (C to K) and 20% in section K (Real estate, renting and business activities) where they are all classified by INPS. These are very large units (about 1,500 employees on average) and subjected to frequent changes. The imputation of these units is essential because if only one of them is missing, it may have a significant impact not only on levels but also on changes of the per capita indicators.

21. The most crucial aspect is the identification of unit non-responses. In fact, in traditional surveys the list of non-respondents in each reference period is available as difference between units belonging to the sample and the set of the respondents. In the Oros Survey, a list of the units which should send the DM10 form is not available considering that the AR suffers over-coverage problems (see Section II).

22. In order to find out the units to be imputed, the distinction between an absence of a monthly declaration due to (even seasonal) inactivity and a real non-response due to a delay in the declaration delivery, is needed. So the activity state of the units must be predicted through the analysis of the patterns of the DM10 in a pre-determined span of time (one year), with the help of some auxiliary information.

23. First of all, a list of reference units is built: it includes all the units which are active according to the information of the AR and that have presented at least a DM10 form in one of the months of the reference quarter or of the previous four quarters. The use of a set of quarters preceding the reference one is due to the hypothesis that the quarters close to that of estimation can be informative on the latter. Indeed, the evidence shows that the probability that a latecomer position in a quarter could be latecomer also in its near quarters is actually low.

24. Furthermore, considering the dynamism of these units, before imputing it is necessary to check possible absences due to firm changes, like mergers or split-ups. At this aim, it is opportune to follow the employment flows among all units.

25. The imputation of the employment and wages variables is mainly based on the longitudinal information available on each missing unit. Firstly, suitable values for the two variables are selected from the closest quarter when the current missing unit was respondent. Secondly, these values are fairly updated using panel information drawn from the current respondents. The reconstruction of the other labour costs is based on the multiplication between the estimated wages and a contribution rate (other labour costs on wages) calculated on the information available from the closest quarter.

26. The imputation process of unit non-responses in temporary employment agencies determines an increase in the number of employees of about 1-3% of total temporary employment. The effect on per capita wages and total labour cost is less relevant (up to 0.5% in some quarters).

VII. LARGE ENTERPRISES: CHECKS AND COMBINATION WITH SURVEY DATA

27. The INPS sources could guarantee the coverage of all firms in the private sectors, but for the estimation of large firms, the use of LES data is preferable mainly because of the specific characteristics of the large firms themselves.

- One thousand enterprises that employ about 2 millions of workers (more then about 20% of total employees in the target sectors). Each of these firms has a considerable influence on the estimates.
- They frequently change over time and only a survey can assure a direct check on them: specialized Italian NSI employees contact enterprises when some data problem occurs, assuring a higher quality of the data and a more rapid and efficient management of their changes.

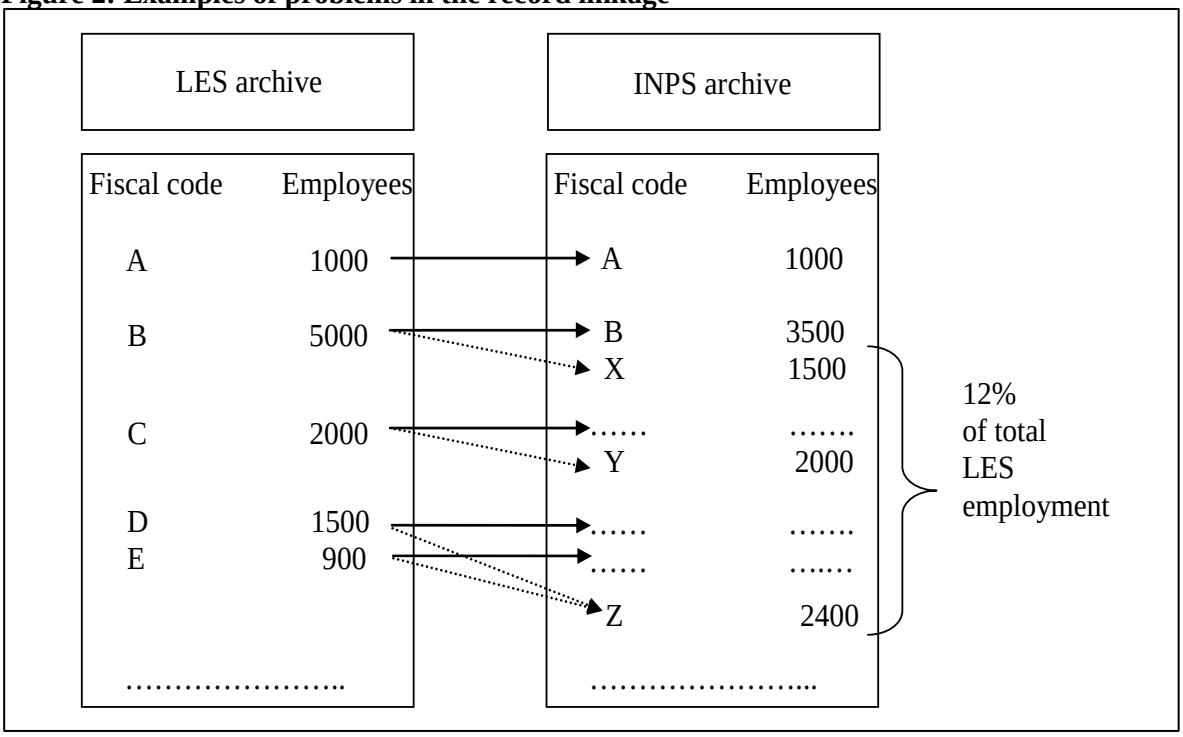
28. The micro level integration between INPS and LES data has two main aspects.

(i) First of all the production of variables harmonised with those produced using administrative data. There are specific operations to do for the selection of wage and other labour cost components to be included in the Oros target variables. Starting from monthly data, the computation of the quarterly ones is also needed.

(ii) A check and editing procedure to correctly single out the LES enterprises among the INPS data is necessary for the substitution of their economic values with those drawn from LES. Starting from the list of firms belonging to the survey, a complementary list of INPS firms must be realized avoiding omissions or duplications. There is only a matching variable between the two sources, the fiscal code, which is unfortunately not sufficient to identify the same firms in the two archives. This identification code may be affected by formal errors or updated in different times. Moreover, changes like mergers, take-overs, hive-offs, split-ups these firms are subjected to, have to be taken into consideration. These events imply that a population does not consist of exactly the same units in different reference periods and problems arise when these changes are recorded in different times and according to different rules in the two sources (Istat, 2006). Figure 2 shows some of the problems that have to be faced of when the two archives are linked. Actually, the majority of firms have the same fiscal code even if this is not always enough to

guarantee that it correctly represents the same enterprise. Meanwhile, an enterprise may have different fiscal codes in the two archives In such cases a significantly different number of employees is used as a signal of possible problems.

Figure 2: Examples of problems in the record linkage



On average, 12% of the total employees surveyed by the LES have to be manually checked and joined to the correspondent INPS firms using auxiliary information. One of them is the firm's name, but its use is not easy because it is not standardized in the two sources.
Other information to check the quality of the two lists created are drawn from external sources like the Italian Business Register that contains information on the enterprises' history concerning their activities and changes over the time.

29. Some large enterprises not included in the LES panel are estimated using INPS data[5]. Since all these firms are influential for their size, first of all it is fundamental to be sure changes occurred to them do not involve any LES enterprise. Then, it is important to evaluate the possible impact on the indicators of a change in one or more of these firms from a quarter to another like an enterprise that modifies its economic activity. These checks are mainly realised using the BR-ASIA through a record linkage procedure based on the fiscal code.

VIII. THE MACRO DATA CHECK

30. Once indicators have been produced, macro data are submitted to further quality controls to identify possible anomalous values that may significantly affect the series released. This is a key step in the E&I process because the difficulties to be faced in the use and translation of administrative data make possible residual errors, in spite of the several previous checks. Since changes in contribution legislation with an impact on macro data are frequent, irregular but acceptable trends due to economic or legal factors must be as possible distinguished from anomalies due for example to an erroneous updating of the "metadata database" (see § 4) or outliers/errors not singled out and corrected in the micro data editing step (see § 5).

31. These controls, mainly based on the *analytic inspection* of the time series at a sub-population detail, are carried out through some statistical measures which have to respect some pre-defined

[5] These firms provide work for about 1% of total employment.

acceptance thresholds. In order to extend checks to a more disaggregated level and to fully consider the time series information that can be affected by seasonal patterns, noise, or special events, an *automatic detection* of outliers is also performed. This analysis is based on TERROR, an application of the software TRAMO-SEATS (Caporello and Maravall, 2002) which detects suspected errors in the last observations comparing them with their forecasts estimated trough REG-ARIMA models. This procedure is very rapid and permits to handle a very large number of series in few seconds.

32. The final indicators are also evaluated using figures drawn from other Istat statistical sources. Given the definition differences, Oros indicators are compared to the LES and quarterly National Account estimates on wages and total labour cost. Some evaluations on wage bargaining effects are also possible by comparing Oros estimates to the Indices of wages according to collective agreements (contractual wages).

33. Furthermore, certain variable relationships are deeply examined, whose coherence has always to be guaranteed, for example the ratio of other labor costs on wages, the evolution of their trends, etc.

34. If the anomalous values emerged in the macro-data checks hide outliers, a drill-down to micro data is required despite very rigid time restrictions (2/3 days). A set of further *ad hoc* checks on micro data (implemented through new Sas programs) helps the understanding of the problem origin. Finally, if necessary the errors correction is carried out at micro level in order to guarantee the coherence between macro and micro data.

IX. CONCLUSION

35. In the Italian experience, it has been possible for the NSI to produce statistical business indicators for wages and labour cost, covering also all SMEs without adding any further burden on firms and with very low survey costs, only using social security data. The quality problems of administrative data cannot be addressed *ex ante* but only *ex post* through a very complex and pervasive E&I process.

36. This paper describes the E&I process of the quarterly Italian Oros Survey which was developed without any previous experience in the use of administrative data for the production of short-term indicators. The E&I procedure has been gradually implemented learning by the experience. The sharp deadline requirements forced Istat to capture the raw INPS data in a very detailed form, and without any previous check. This has implied the planning and implementation of a complex translation process of administrative information into statistical data including the set up of a “metadata database”. The evolution of the social security legislation has implied continuous updating of E&I procedures.

37, The use of administrative data in a survey characterized by high timeliness and frequent social security law and regulation changes, implies a greater dependence on the data supplier and a higher risk to incur in data quality problems than in a traditional survey.

38. The strategy used to reduce those risks was based on:

- a more strict relationship to and coordination with INPS data suppliers;
- a systematic sequence of checks and editing steps which should assure the interception of the errors.

39. Some other peculiarities of the E&I process are due to the integration with survey microdata on large enterprises which are used to complement administrative data. The linkage between the two sources is a non-trivial task implying a continuous and careful control of the firm’s identification variables that in some cases requires a manual check.

40. On the whole, the E&I procedures developed *ad hoc* turn out to be reliable both in terms of effectiveness (quality of the entire process) and efficiency (relatively limited time consuming and low use of human and economic resources).

References

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

Caporello G., Maravall A. (2002) A tool for quality control of time series data. Program TERROR. Bank of Spain.

Istat (2006) Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese, Metodi e Norme n.29, Roma.

Istat, CBS, SFSO, Eurostat (2007) Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, available on the web site: http://edimbus.istat.it/dokeos/document/document.php?openDir=%2FRPM_EDIMBUS