What have we learned in almost 10-years experience in dealing with administrative data for short term employment and wages indicators?

Fabio Rapiti, Francesca Ceccato, M. Carla Congia, Silvia Pacini, Donatella Tuzi

Preliminary version 03-03-2010

1. Introduction

It's almost ten years since Istat has started the project to use administrative data to compile short-term indicators coherent with the requirements of the STS Regulation. It is time to summarise what we have learned in dealing with the Italian social security information, focusing on the specific problems and difficulties arising when administrative data must be used in a context where timeliness is a key issue.
The Oros[1] survey was designed to fill a crucial gap in Italian statistics and meet the requirements of EU Regulations (STS, LCI-Labour Cost Index). It has been the first Italian short-term survey based on a massive quantity of administrative data. The aim of the survey is to produce quarterly information on the evolution of gross wage, other labour costs and employment. The survey uses administrative data for Small and Medium size Enterprises (SME) which are combined with the data of the Istat Large Enterprises (>500 employees) monthly census survey (LES).
Working with the National Social Security Institute (INPS) data Istat faced the typical problems described in the literature on the utilisation of admin data but also other challenges deriving from the specificity of the data structure and from the degree of timeliness required. This paper describes the main peculiarities of the Oros survey and highlights the main problems arising in practice when dealing with short-term indicators. The paper is devoted specifically to illustrating non-sampling problems which often are very country-specific, depending heavily on national laws and regulations or institutional sets. But this kind of problems, indeed more than ones related to sampling, are those that often limit the use of administrative data for short term purposes or jeopardize the quality of the results.

2. The peculiarity of the Istat experience

2.1 The difficult "choice" to use admin data for short term purposes

Compared to other NSIs Istat was a latecomer in the exploitation of administrative data for statistical purposes. In the past there was no very much trust in admin data and, at the same time, an over-optimistic view of the quality of surveys data often based on the assumption their results could always comply with statistical theory. Only in the second half of the the nineties, the successful development of the ASIA Business Register, largely based on admin data, increased the interest in the statistical use of admin data also for current surveys.
In the business statistics domain, in the same years Istat had to increase the flow of information supplied at national and international level covering all firm size classes. New and very demanding Regulations were approved at Eu level (STS, SBS). At the same time all NSIs became aware on the opportunity of reducing the statistical burden on businesses. With those two conditions Italy, with a very large share of SME (small and medium size enterprises), in some cases had no alternative but the use of administrative sources to fill the new data gaps. Otherwise, the huge number of small-size enterprises (in 2001 21,4% of employees worked in micro firms, those with less than 10 employees) and the extremely dynamic nature of the Italian firms would have implied the design of too onerous traditional sample survey, with a considerable impact on the statistical burden on enterprises. On the other hand, estimating employment and wages without covering the very small firms, could have produced very biased results.

[1] Oros stands for Occupazione (Employment), Retribuzioni (Wages), Oneri Sociali (Other labour cost).

Until 2002, the Italian National Institute of Statistics had collected information on wages, labour cost and employment at a monthly frequency with a traditional census survey limited to firms with 500 or more employees (hereafter LES-Large Enterprise Survey). The new Oros survey was planned to extend the coverage to all business size classes and the INPS employers' social contribution declarations (DM10 form) was the best source available because it was rich of information on wages and employment. After preliminary studies DM10 declarations have been considered to be suitable for Oros purposes although information were extremely detailed/disaggregated and the administrative metadata were rather fragmented. Selecting and aggregating in a very strict time scheduled the DM10 data in the format required for statistical purposes was for INPS an inappropriate activity, resulting in extra costs compared to the supply of comprehensive files. Thus Istat decided the acquisition of the whole data set, i.e. all records available in the central administrative database at a certain moment in time. In other words, Istat has been obliged to capture the extremely disaggregated raw micro data, in the original format they are transmitted by firms and without any check by INPS. This implied that Istat had to:

- implement a complex pre-processing administrative data phase of checks, computation, translation and aggregation of "administrative micro codes" to get to the statistical basic variables at micro level (pre-editing);
- capture every month much more data compared to the target units and aggregated variables: 1,3 millions declarations per month split on 8 records each on average.

This constraint could also be seen as an opportunity, considering that it allows a more direct control on the aggregation/translation process and makes available a lot of detailed information useful for other different statistical purposes.

2.2. Two models of exploiting the INPS data

DM10 declarations represent a rich mine of detailed information at firm level about the composition of the workforce (qualification, type of contract, etc.), their wages and the various components of the labour cost. Considering the constrain of resources and time, Istat decided at first to design and implement the new survey exclusively to fulfill objectives related to the compilation of short-term indicators. The mandate was limited to this domain, exploiting the data potential in term of coverage and timeliness, aggregating all the information to estimate four main variables: number of jobs, full-time equivalent, gross wages and other labour costs. Thus between 2000 and 2003 the Oros survey has been set up as a single production line specialized in compiling three different kind of short term indicators: quarterly STS indicators for employment and wages, national wages and labour cost index per Fte (Full-time equivalent) and also the hourly LCI (Labour Cost Index). All the IT and statistical work has been devoted to guarantee quality results for the very aggregate variable by economic activity at quarterly level. The internal Istat architecture with two separate Divisions, one dealing only with short term business indicators and surveys and another dealing with structural business surveys had a role in this initial choice not to exploit the whole information set. Subsequently the availability in electronic form of a mass quantity of data has *de facto* stimulated Istat to tune strategy from a typical "one collection-for one single survey" (or "stove-pipe") model to an integrated system focusing on the "data source"- the "wage and social security system[2]"- to be used as datawarehouse for many statistical objectives (Figure 1). This change is along a more general shift toward a integrated systems designed to produce statistics in which more and more indicators for specific domains are no longer produced independently from each other but through the integration of data sets and by combining data from different sources, administrative data and surveys. In this sense Oros tends to became a common statistical infrastructure, specialized in employment and wages, integrated in a comprehensive production system useful:

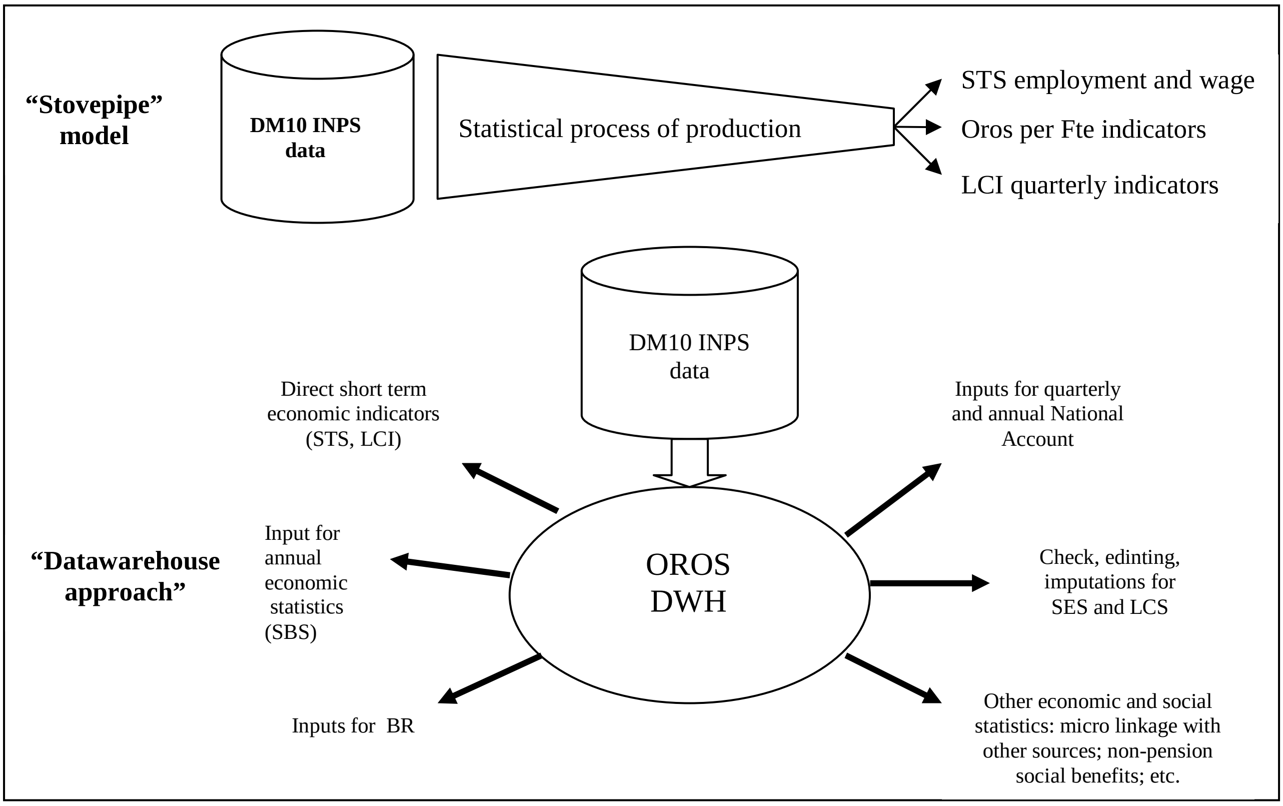
- to compile direct indicators for short-term objectives;
- as input in many National Account quarterly and annual estimation processes;

[2] In other countries sometimes is called PAYE (Pay as you earn).

- as source of variables for some structural surveys or for the BR;
- as auxiliary information for check and editing at micro level in structural wage and labour costs surveys (LCS, SES);
- as source of micro data to be linked to many other business sources, etc.

In this way the Oros survey appears quite different from other short term surveys based on administrative data (SEPH at StatCan, VAT based surveys in some countries, etc) and more similar to ERDR[3] (Elettronic Raw Data Reporting) based statistics – which automatically retrieves data collection from mainly financial bookkeeping systems of enterprises – or register-based structural business statistics,; however it must be considered that Oros operates at quarterly frequency and in a extremely changing environment.

Figure 1. The two different uses of the social security source.


"Stovepipe" model
DM10 INPS data
Statistical process of production
STS employment and wage
Oros per Fte indicators
LCI quarterly indicators
DM10 INPS data
Direct short term economic indicators (STS, LCI)
Inputs for quarterly and annual National Account
"Datawarehouse approach"
Input for annual economic statistics (SBS)
OROS DWH
Check, edinting, imputations for SES and LCS
Inputs for BR
Other economic and social statistics: micro linkage with other sources; non-pension social benefits; etc.

3. The Oros survey quarterly challenges

Since the beginning of the Oros project it was clear that there were extremely ambitious targets. Succeeding in combining coverage, quality and timeliness resulted particularly difficult for three main reasons,

a. the most difficult challenge was to succeed in keeping constant every quarter the "suitability" for the statistical purposes of the DM10 INPS data, because it couldn't be defined once and forever but obtained and verified continuously, taking into account the evolution of the informative contents of the source, due to law, administrative and technical changes;
b. for the first time in a business short term Istat survey, it was projected to cover each quarter the whole current firms population, exploiting the INPS administrative register (AR) instead of drawing the frame from the Istat Business Register (BR). The BR is updated once a year[4] and has a varying delay of 15-22 months with respect to the Oros quarterly estimations. On the other

[3] In the past it was called more frequently EDI (Elettronic Data Interchange).
[4] At the end of march of year Y is ready the annual BR for year Y-2.

hand, the AR covers timely new births but suffers of overcoverage problems due to the lack of inactive units cancellation, requiring particular solutions in the estimation step;

c. last, but not least, the sharp timeliness at 90 days after the reference quarter in the first period and then reduced at 60 days for the employment STS variable.

To design and implement the survey Istat faced old and new methodological issues. Many unusual, mainly non sample, quality problems have been overcome planning ad hoc instruments and specific check phases. Only after few years and a long trial and error process, Istat managed to cope with the more peculiar and subtle shortcomings of the social security admin. data.

4. Change is the rule not the exception

Reliance in any particular administrative source carries always a certain degree of risk of discontinuity due to legislation, administrative and technical changes which can affect coverage, variables, definitions, or even the source itself. The specificity described in the previous paragraphs explain why in the Oros survey the change is not a chance, a possibility, but the rule. First of all, it is important to distinguish between current quarterly changes, mainly affecting administrative metadata changes and other changes which are quite frequent but less regular.

The correct exploitation of the huge quantity of administrative data entails coping with very frequent changes in the basic INPS metadata which have an impact on the correct translation rules of admin data into the target variables. Those continuous changes depend on the fact that in Italy a large part of labour market, occupational and industrial policies take the form of rebates in social contributions - for example measures to stimulate the employment of specific target groups of the labour force can rely on the differentiation of social security contribution rates - and enterprises have to use the DM10 declaration to take advantage of them. This implies that the process of retrieval of statistical target variables is heavily affected because continuously new small components (administrative micro codes) of labour cost have to be included, while other information have to be excluded because they are not relevant for statistical purposes.

A prerequisite to the correct inclusion of changes in the calculation of the target variables is the availability of complete and clear input metadata. Istat could succeed in identifyng correctly and systematically all those changes and then modifyng the relative software/code to retrieve/calculate/aggregate/ the target variables only implementing and continuously updating an in-house ad hoc metadata database collecting laws, regulations and other technical aspects concerning social security contribution (Input Metadata Base - IMB).

Other changes regard the way the original INPS data are collected, stored, transmitted depending on legislative or administrative changes. Different administrative changes have taken place during the last ten years but without any relevant effect on the data[5]. Only one very big change in legislation had a very visible and positive effect on the number of units belonging to the non-random sample of DM10 electronically delivered by firms and used by Oros for the provisional estimates of the quarter *t* (figure 2). Since spring 2004 the submission via internet of DM10 declarations has became compulsory. As a result, the sample provided by INPS become a “provisional population” nearly reaching the dimension of the “final population” available after one year and used to produce the final estimate referring to quarter *t*-4. This change has allowed a significant simplification in the estimation procedures as it is described in the next paragraph.

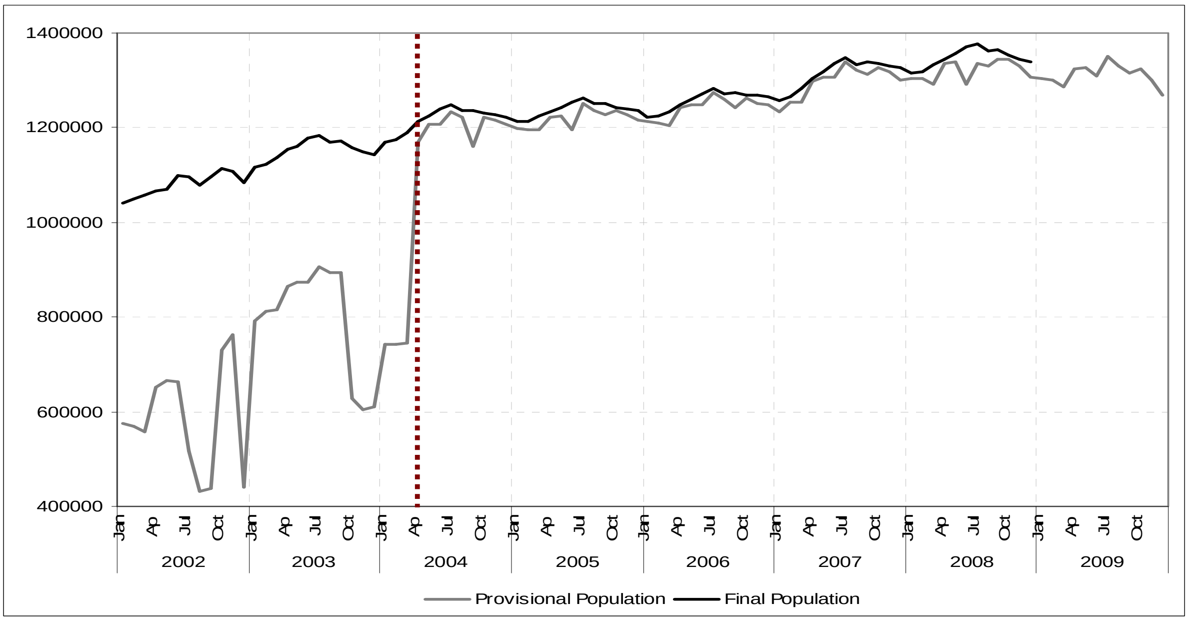
Another indirect big change was related to the transition in 2009 to the new economic activity classification Nace Rev.2. In the Oros survey the economic activity code is assigned through the official Istat Business Register because it is of better quality compared to economic activity code assigned by INPS that at the moment is still based on the old classification (Nace Rev.1.1). Since the BR has a maximum of 8 quarters of delay compared to the Oros estimates reference period, the

[5] Large swings of the sample size occurred. This mainly happened 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.

INPS activity code is assigned to all new born units. Between 2008 and 2009 Istat had to move to the new Nace classification but INPS couldn't implement it[6], therefore for the provisional estimates Istat has decided to use an alternative timely administrative source coming from the Tax Authority. Since spring 2009 this Tax Register is delivered regularly to Istat allowing to currently assign the new classification to the new born firms. Obviously this change had a significant impact on the whole process of production of the quarterly indicators.

A new big change is just happening now. It is a real administrative revolution and has huge consequences on statistics: since January 2010 the old DM10 form is cancelled and it is unified with another monthly individual wage declaration (Emens) in a new electronic declaration called Uniemens, containing a huge and interesting quantity of information on each single workers. The information contained refers to individuals and is organized in a completely different way compared to the DM10. This change can really put at risk, at least temporarily, the release of the quarterly indicators as there is not enough time to redesign the whole process of production but at the same time it represents a new challenge which in perspective could really improve the quantity and quality of short-term and structural statistics.

Figure 2. Provisional and final population of DM10 forms. January 2002 – December 2009



Source: Oros Survey.

5. What we have learned

This paragraph discusses some lessons learned dealing empirically with the survey, facing everyday administrative, statistical, IT and organizational problems. Mainly sources of non-sample errors are considered because they are indeed, since 2004, the most relevant error sources in the survey and also because they are not often covered in previous studies. The Oros survey suffers obviously also of sample errors whose properties, given the continuously changing information set used for the preliminary estimates, have never been easily generalized.

Preliminary data and estimation methods

Depending on the dimension of the preliminary data, three estimation phases can be distinguished in the Oros Survey.

Phase 1: before April 2004 the provisional population was considered as a non-random sample consisting of a set of self-selected respondents (those adopting the electronic mode to submit the

[6] The Nace code is not a relevant variable for administrative purposes.

DM10 declarations, thus collected faster by Inps). Beside its large and increasing size, this sample showed a good coverage in terms of main target population characters and included a fair number of new birth . In this information context, in the absence of a design-based framework, the preliminary estimation was set up using a calibration weighting procedure (Baldi et. al., 2004) based on a model-assisted approach. In order to cope with the bias due to the sample units self-selection, the weights were calculated within homogeneous subgroups (model groups), identified by partitioning the population according to some characters (e.g. economic activities, size classes, geographical areas, age of firms). The calibration weighting procedure has provided good results in the estimation of the ratio variables, while more serious problems have been faced in the estimation of the employment levels: the non-ignorable mechanism distinguishing the later respondents, the continuously changing sample representativeness, together with the over-coverage problems in the administrative register have implied less accurate estimates for this variable, requiring supplementary ad hoc actions based on "expert judgment" interventions.

Phase 2: since spring 2004, when over the 95% of the responders have began to be available for the short term deadlines (about 70 days from the end of the quarter), the preliminary estimation method has been radically simplified. The provisional estimates are in fact calculated, as in the final estimates, by simply summing up all the available data. This methodology guarantees non-biased estimates for the ratio variables but keeps unsolved the underestimation of the employment levels due to the later respondents (unit non responses[7]). Various approaches have been experimented in order to cope with this problem. The "expert judgment" adjustments strategy, already undertaken in the previous information context, has been deeply developed exploiting systematically information concerning the past revision errors (for some aggregates revisions show a systematic and predictable component), and studying the relationship between the employment dynamic in the preliminary and in the final data. An alternative approach based on the imputation of missing data for the non-responders at micro level has also been experimented. Given the availability of a wide set of preliminary (and final) micro data in the new informative situation, the problem has been delineated as a *wave non-response* matter, where the past behaviour of each current non respondent was considered informative on its outcome in the final population. In the reconstruction of the micro data, the information on the respondents behaviour has also been exploited.

Phase 3: in the next months data resulting from the recently changed contribution declaration mode will be made available by INPS. At the moment almost nothing is known about the coverage/representativeness of the new data framework. The radical changes are likely to imply a transition period during which firms and INPS itself will gradually adjust to the new situation, implying an instable informative perspective. To understand the consequences on the Oros estimates new analysis and, perhaps, radical changes in the estimation methods will be needed. This topic will be further discussed in the next future in the context of the WP4 ESSnet.

Stable data capturing

The survey implies to handle a huge quantity of data in a very limited time assuring fast acquisiton and efficient solutions to any unexpected possible technical problems. The data transmission between data provider (INPS) and data user (Istat) has followed different ways until now: mainly disks and, only recently, on-line direct connection. The flow ensures fast delivery and guarantee the full respect of confidentiality of statistical data but problems can arise at any stages/point of the chain and can regard: data structure, formats, software used to compress and/or encrypt files. Sometimes technical problems derivs from an unannounced little change in the INPS source database (Host). Istat experienced many technological difficulties in the first period (inadequate client/host hardware) also because it was not completely aware of the work-load necessary to manage this large amount of data. But dealing with frequent data transmission of a lot of gigabyte

[7] Inps data used for the Oros Survey are not characterized by item non responses, in the sense that the interest variables are quite always coherently available in the DM10 forms.

always arise problems. It is important not to underestimate IT and technical problems, and have personnel with suitable skills and proactive attitude, able to anticipate any possible risk of delay.

Completeness and consistency of metadata

The availability of fragmented, insufficient and not in user-friendly format INPS metadata has been overcome with the implementation of an in-house and *ad hoc* metadata database collecting information on laws, regulations and other technical aspects regarding social security contribution (Input Metadata Base - IMB). To assure a correct retrieval of the target statistical variables the IMB has to be quarterly updated. This input metadata database is indispensable for the statistician in keeping track of changes in administrative data definitions and setting the rules for the translation of input administrative data into statistical variables.

Stable and correct translation/retrieval of target statistical variables

The quarterly updated information deriving from the metadata database must be examined in depth and correctly interpreted to take into account the introduction of new administrative micro codes that identify labour cost components and to exclude new codes without statistical relevance. These changes has to be included into the procedures, requiring quarterly appropriate modification of the code that should be as automated as possible to avoid the introduction of coding errors. Unidentified metadata changes or modifications incorrectly reported on the preliminary "administrative micro codes" aggregation procedures can have big impact on the estimates. Conceptual/definitions and translation problems must be correctly addressed and continuously monitored.

Necessity to keep the direct collection (LES) and proper integration with admin data

The INPS sources could guarantee the coverage of all firms in the private sectors, except in the first period of the Oros survey, when large firms were not well represented in the "sample" used for the preliminary estimations. However for the estimation of firms with more than 500 employees, the use of direct survey data is preferable mainly because of their specific characteristics. They are one thousand enterprises that employ about 2 millions of workers, 20% of total employees in the target sectors. Each of them has a considerable influence on the estimates and is frequently involved in changes, like merger, split-up and acquisitions. The survey assures that specialized Istat staff follow these firms each month and can contact them when some data problem occurs, assuring a higher quality of the information and a more rapid and efficient management of changes. Moreover cross-checks among administrative and survey data are very informative on the different characteristics of the several sources highlighting pros and cons of them. Of course the integration of INPS data with LES, realized at micro level, implies specific procedure to avoid units (and employment) double counting and to harmonize the statistical variables.

Pervasive check

The quality problems of administrative data cannot be addressed ex ante but only ex post through a complex and pervasive check and editing process. This is particularly true in the Oros short-term survey where Istat has no ex ante control on INPS data. Therefore the data must be processed in an extensive and complex way that ensure all statistical quality problems are identified and addressed correctly and systematically each quarter. Thus the survey is characterized by pervasive checks in all phases. To assure the quality of the aggregation and translation procedure to obtain the target variables, the monthly declarations go through complex preliminary checks aimed at investigating and possibly correcting errors on administrative micro codes, record duplications, incoherencies with current legislation, etc. After administrative data have been translated into the required statistical variables, a more traditional check procedure becomes necessary in order to identify possible anomalous values and correct them at a micro level. Final key checks on macrodata are carried out to be sure that no residual influential errors, or legislation changes not correctly reported in the previous steps are still present in the data (Congia, Pacini, Tuzi 2008). It is useful to remind that the most subtle errors to look for are not the traditional outliers due to accidental, mistype, duplication but systemic changes (legislation, INPS regulations or registration and practices, INPS information systems) (Congia et al, 2008).

Specific process quality indicators

All checks are recorded to maintain useful time series of errors detected and corrections. Every quarter quality indicators are measured at each stage of the production process. They are monitored to keep under control data quality and to provide information to improve the process. Few of them are rather traditional indicators, useful mainly as documentation to assess the quality of process and output. Others are more oriented to carefully monitor every step of the quarterly process. The latter can signal decisive problems or detect sources of error helping survey managers to react quickly, solve problems and, if necessary, to run again the procedures (Congia, Rapiti 2008). Those indicator, often very survey specific are useful for monitoring the Oros process of production but only few can be standardised and compared with similar indicators in other surveys.

Very selective editing via interactive mode

Given the huge quantity of uncorrected raw declarations (1,3 millions per month), for the preliminary estimate the process of editing must necessarily be extremely selective. On the other hand the goal of the preliminary estimate is to obtain “good and fit for use” data as opposed to “perfect” data that could be obtained by editing a large amount of declarations. 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. Those rules detect and identify a very limited group of unit with possible errors. In the very first period the largest values identified as outlier were considered measurement errors and automatically corrected by the procedure. More recently 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, manually corrected.

Unusual editing rules for INPS data

The peculiarities of the INPS administrative data 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 average 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 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. Moreover the other labour costs monthly distribution may show correct negative values, because of social security contribution rebates which can be concentrated in some months even if they refer to previous periods. This aspect must be taken into consideration both to calculate correct check indicators and to single out all possible wrong data.

Key macro checks

Once the estimates 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 editing 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 far as possible distinguished from anomalies due for example to an erroneous updating of the “Input Metadata Base” or outliers/errors not singled out and corrected in the micro data editing step. If the anomalous values emerged in the macro data checks hides? outliers, a drill-down to micro data is required despite very rigid time schedule (2/3 days to complete the process). A set of further ad hoc checks on micro data (often implemented through new *ad hoc* procedures) 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.

Overcoverage

The social security declarations are compulsory also for a number of units that are out of the scope for the Oros survey. This overcoverage problem has to be faced identifying and excluding the units belonging to the public sector or running economic activities out of B-N sections of the Nace Rev.2 classification. The main source used to single out those units is the statistical BR-ASIA but it is not sufficient due to its availability with a delay of about two years from the reference period. So information drawn from the administrative register is used but also other external statistical sources are necessary, for example the Istat annual list of public institution (S13).

Relation with data provider and inter agency cooperation

Since 1996 Istat has had the right by law to access to all public administrative data and register and has the theoretical right to influence the process of modification of administrative information collected by public institutions. INPS is one of the most important public agency participating in the Italian National Statistical System (SISTAN). Moreover in 1998 Istat and INPS signed a general framewok agreement to facilitate cooperation and data flows. For a long period there was a strong commitment to co-operate and a high level co-ordination committee supervising the bilateral relations. In 2000 a sort of specific “service level agreement” for the Oros project data flows has been set up. INPS data flows have always been accurate and on time thanks to not only those formal agreements but also for the development and manteinance of a strong unformal cooperation between provider and Istat officers practically involved in the flow. To succeed in preventing any unexpected kind of problem in advance, delays in delivery or quality problem (technological, administrative) very strong link and cooperation with suppliers are necessary at all level. From a general point of view the Oros survey imply a complete dependence from INPS and there is always a risk of inconsistency or discontinuity of the information. The recent change from the DM10 to the new Uniemens elettronic declaration has shown that the role of Istat in the process is absolutely insufficient.

Staff organization and skills

Differently to traditional short-term surveys in which normally all the IT procedures are changed and improved only or mainly every five years, in the occasion of the change of base year, in the Oros survey a lot of software modules in the first part of the process have to be modified continuously and in a very short time. This imply that, differently from traditional other short term surveys in which there is a clear distinction of roles between people that work on software and people working on a wide range of statistical issues, in the Oros experience the same little group of people must focus at the same time on wages and labour cost legislation and regulations, statistical methods and software programming. About programming it is important to highligh that because the basic editing rules at “administrative micro codes” level change very frequently and often have an impact on check and editing rules in other part of the process, to keep track of changes and make them easily accessible, it has been necessary to save and catalogue in a standard way (versioning) the different quarterly version of the code (Baldi et al., 2008).

Organizational structure

The organisation of activities at Istat is based on internal differentiations with respect to the statistics domain and output to be published. This organisational structure can be characterised mainly as “stove-pipes”; there is limited coordination and integration between different departments, and sometimes nor complete statistical coordination of common concepts, definitions and variables even if the subject matter is the same. To take advantage of all the potentiality of the DM10 (or the next Uniemens) DWH sources, in a context in which administrative data become the core of statistical production in some domains, not only the single process of production but also important part of the organization of the NSI must move towards a more integrated model. The Oros survey unit is still based on a stove-pipe model organization (inside the short term statistics Divion) while part of the process could be moved outside the survey lines in a cross-departmental “input and integration” division where there aren’t the traditional boundaries of surveys because the focus in on the source system.

6. Final remarks

In the paper we summarised what we have learned dealing with social security data to produce STS indicators. Reviewing the Oros survey experience some final remarks can be made about potentiality and limitation of using large and complex administrative sources for compiling short term statistics.

1. In the domain of short-term statistics timeliness and punctuality are a key component of quality that can be jeopardised by relatively small change in the administrative rules governing the utilisation of raw and complex sources, like those embodied in the Oros process. Therefore, more than on contingency plans we have to rely in being very proactive, anticipating any sudden changes or possible risk of lag or delay. We learned and developed a general attitude to prevent any unexpected kind of problem in advance: delays in delivery or quality problem - technical (format, software update) or administrative.
2. Sometimes we had to go through alternative solutions and frequently we had to change part of the process and input sources. New problems have been faced using other registers or administrative sources: S13, Tax register, etc. More than a survey Oros is now a multi-sources process combining the following data sets: two INPS archieves (DM10 and INPS Register), the Istat BR, the Tax register, the Istat S13 register, LES survey data, other less relevant sources. This growing combination and integration of sources further increased the complexity of the whole process itself: more sources means more data flows to manage and to capture on time, more linkages, more variables, more definitions and concept to reconciliate, in other words more workload finalized to the scope to maintain and possibly increase quality.
3. Nevertheless, also taking into consideration all those growing difficulties, the present and future benefits of the use of social security data are always overwhelming. But in order to move toward a greater role of combined administrative sources for short term or structural statistics it is unavoidable to improve and adapt the staff organization and skills, and also adjust the NSI internal organization to succeed in the management of the input and integration phases.
4. In this situation as in other cases of statistical use of administrative data, Istat experienced a good cooperation with the institutions managing data collections. Anyway, given the actual institutional architecture and inter-agency relationships, when Istat comes to deal with big changes and has to influence reorganization/simplification of administrative data collections ex ante, it has scarce power. The actual legislative and cooperation framework has to be strengthened, in order to create a more cooperative environment to fulfil the EU Statistical Regulations: these are national tasks and must be accomplished by all institutional players cooperating together, sharing advantages and disadvantages. Just like in the Nordic countries (UNECE 2007) regulations should guarantee not only to the National Statistical Institute the right to access admin files but also the obligation to use them in alternative to new direct surveys, in order to avoid to duplicate data collections. At the same time all Institutions should be obliged to provide Istat admin data for statistical purposes.

References

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

Baldi C., F. Ceccato, E. Cimino, M.C. Congia, S. Pacini, F. Rapiti, D. Tuzi (2008), “Il controllo e la correzione in una indagine congiunturale basata su dati amministrativi. Il caso della rilevazione Oros”, Contributi Istat, n.13.

Congia M.C., Pacini S., Tuzi D. (2008), “Quality Challenges in Processing Administrative Data to Produce Short-Term Labour Cost Statistics”, Proceedings of Q2008 European Conference on Quality in Official Statistics, Rome.

Congia M.C., Pacini S., Tuzi D. (2008), “The Editing Process in the Italian Short-Term Survey on Labour Cost based on Administrative Data”, paper presented to the UNECE - Conference of European Statisticians,Work Session on Statistical Data Editing, 21 – 23 April, Wien.

Congia M. C., Rapiti, F. (2008) “Quality assessment and reporting in a short-term business survey based on administrative data”, Proceedings of Q2008 European Conference on Quality in Official Statistics, Rome.

WJ Keller, J Bethlehem (2000) “The impact of EDI on statistical data processing”, Computational Statistics, Citeseer

V. Koskinen (2009) “Preparing for Changes in Administrative Data for Short Term Tsatistics” OECD STESEG Meeting, September.

Wallgren A., and B. Wallgren (2007), Register-based Statistics. Administrative Data for Statistical Purposes, West Sussex: Wiley.

UNECE-United Nations Economic Commission For Europe (2007), “Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics”, United Nations, New York and Geneva.