

# Use of Administrative Data in the Italian quarterly OROS survey

**Fabio Massimo Rapiti**

Short-Term Statistics on Employment and Labour Incomes  
Central Directorate for Short-Term Business Statistics  
Istat

OECD STESEG MEETING  
Paris 27-28 June 2005

# Outline of presentation

1. Italian background and main characteristics of the OROS survey;
2. why administrative data;
3. the INPS data: content and timing;
4. the treatment of data (retrieving variables; check; editing);
5. estimation methodology;
6. some recent changes;
7. final remarks.

## Use of administrative data at Istat

- Compare to other NSOs Istat is a latecomer as administrative data user (no tradition, no trust).
- Recently Istat has made some steps towards use of administrative sources. Two examples:
  - Business Register ASIA (Archivio Statistico delle Imprese Attive - statistical archive of active businesses): six administrative data sources;
  - annual businesses accounting data (Financial statements register) captured by the Chambers of Commerce or other intermediaries.

# Main characteristics of the OROS Survey

- The OROS short term survey was designed to fill a crucial gap in Italian statistics and meet EU Regulations (STS, LCI- Labour Cost Index);
- OROS stands for **O**ccupazione (Employment), **R**etribuzioni (Wages), **O**neri **S**ociali (Other labour cost);
- the aim is to produce quarterly information on the evolution (and levels) of gross wage, other labour cost and employment;
- the OROS survey uses administrative data (INPS-National Social Security Institute) for Small and Medium Enterprises (SME);
- the SME estimation from administrative data is combined with the data coming from Istat Large Enterprises (LE) monthly census survey (>500 employees).
- Every quarter two new estimations are released: the “preliminary” estimate based on a “non-random” sample of INPS data, with a delay of about 75 days from the reference quarter, and a revised estimate, called “final”, based on the “total population” of INPS data, with a delay of 15 months from the reference quarter.

## Development of the OROS project

<b>Years</b>	<b>Activity</b>
1999	Start of the project
2000-01	design and development of survey method and procedures
2002	first preliminary release (100-90 days delay ) of three OROS indicators at national level: wage, other labour costs, total labour cost per FTE unit
2003	regular release of OROS indicators (90-80 days delay)
2004	EU STS Regulation delivery to Eurostat
2005	EU LCI Reg. (75-70 days delay); in autumn release OROS index of number of jobs

Only after using data for 3 years we learned how to cope with the more peculiar and subtle shortcomings of the admin. data.

In the past short term statistics for employment, wages, labour costs and hours worked were based on

monthly business survey, covering firms with more than 500 employees (accounting for 23% of total wage employment).

## EU short term statistics requirements (late '90s)

### **STS Regulation**

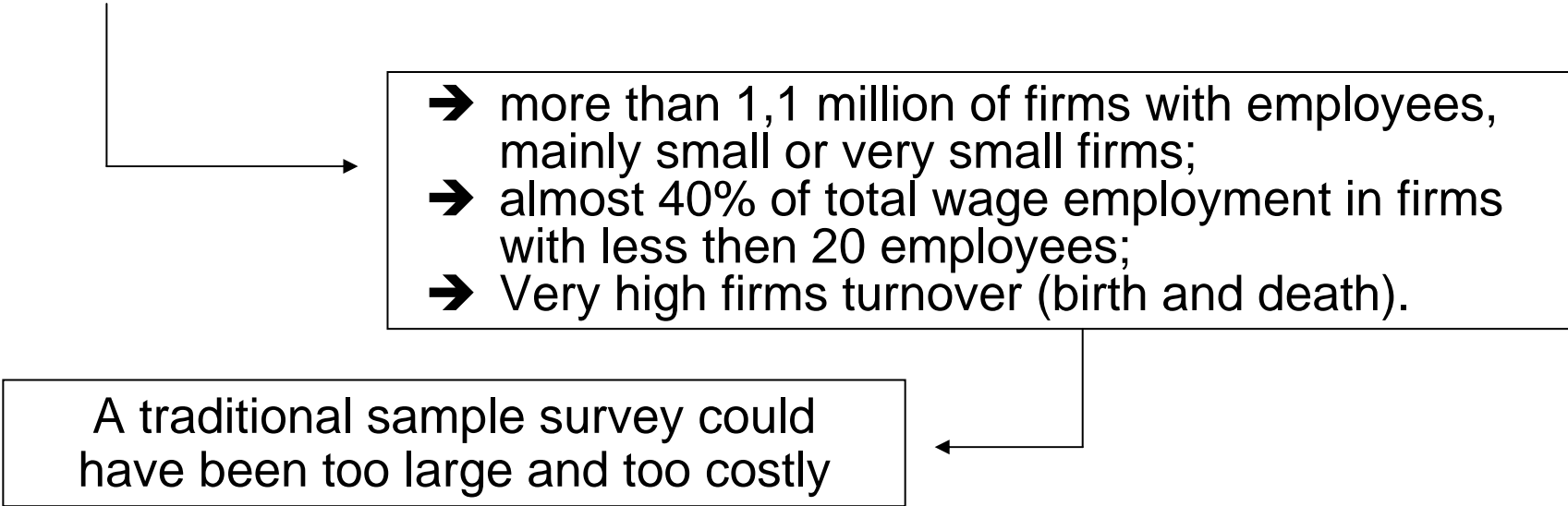
- employment;
- gross wages;
- hours worked;
  
- coverage of all enterprises with employees;
- C to I +K (two digits Nace rev.1) for employment, C to F for wages and hours worked;
- 90 days (70 or less in future).

### **LCI (Labour cost index)**

- hourly gross wages;
- hourly other labour cost;
- hourly total labour cost;
  
- coverage of all enterprises with employees;
- C to K (sections Nace rev. 1);
- 70 days.

## Why administrative data? (1)

### Given the Italian structure of firms

- 
- more than 1,1 million of firms with employees, mainly small or very small firms;
  - almost 40% of total wage employment in firms with less than 20 employees;
  - Very high firms turnover (birth and death).

A traditional sample survey could have been too large and too costly

Only using administrative data Istat could :

- meet the requirements of the UE short-term regulations (coverage, quality, timeliness);
- without increasing enormously the statistical burden on firms.

## Why administrative data? (3)

The availability in electronic form of a mass quantity of INPS data has stimulated Istat to tune strategy from

a typical “one collection-for one single survey”

to focus on “data source”  
(the wage and contributive system)

which can be used for many statistical objectives.

**short term  
economic statistics  
(STS, LCI)**

**other economic and  
social statistics**  
(Non Pension Cash  
Benefits, etc.)

**OROS  
DATABASE**

**input for annual economic  
statistics (SBS), National  
Account, etc (also for  
editing and imputation)**

**Satellite Register  
on employment  
(ASO)**

## The INPS (National Social Security Institute) two archives

All Italian non-agriculture firms in the private sector, with at least one employee (roughly 10 million employees and 1.3 million employers per year), have to pay social security contributions to INPS.

### **INPS register**

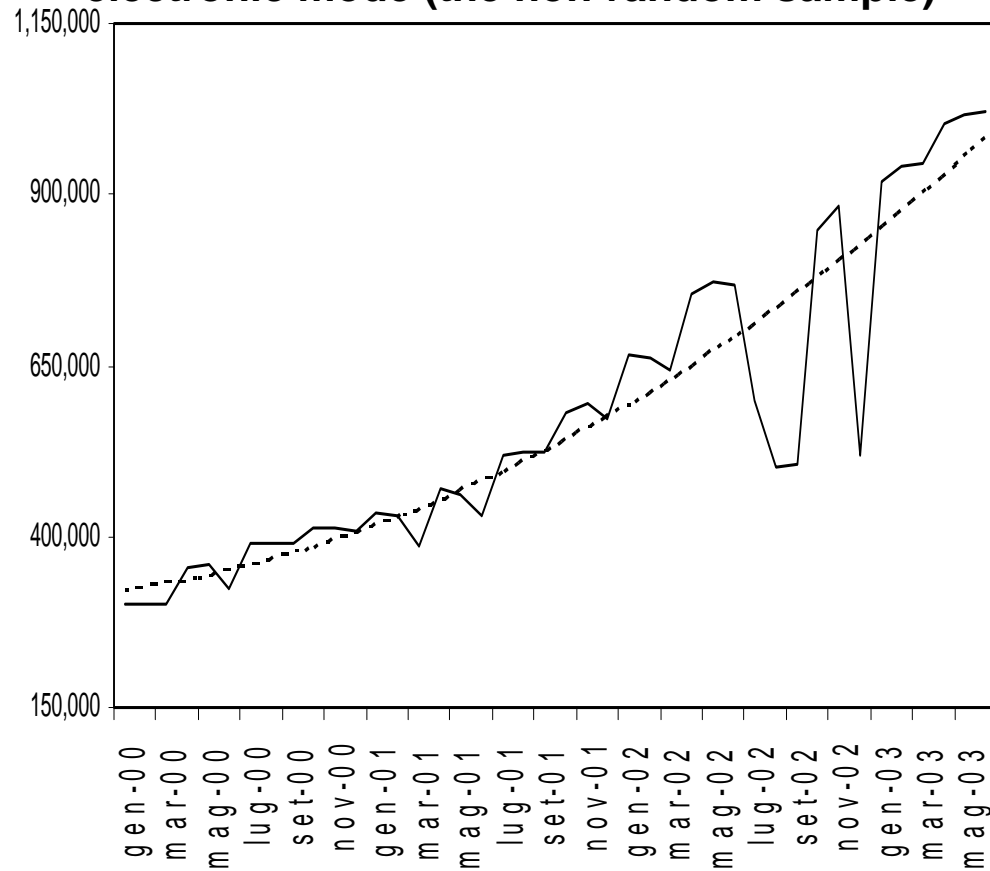
identification firm code  
(as administrative entity),  
fiscal code, name,  
address, legal form, dates  
of registration and  
cancellation, INPS  
industry code, etc.

### **Employers monthly declaration (DM10 form)**

identification number of  
the firm, total monthly  
employment and the  
associated wage-bill,  
paid days, overtime  
hours, social  
contributions, etc.

## DM10 forms in electronic mode

Time serie of the number of DM transmitted by electronic mode (the non-random sample)



DM10 forms were delivered to INPS in **different modes** (electronic and non-electronic); more and more firms used electronic communication mode (from 2001, Internet);

DM10 forms are usually available at the local INPS offices 30 days after the reference month;

INPS collects in a special file all the electronic forms and transmits them to ISTAT after about 45 days from the end of the reference quarter.

## DM10 forms in electronic mode (2): the (non random) sample

The sample is, obviously, “non random” but :

- it is extremely large (about 1 million units);
- it covers all firm sizes, economic activities and geographical areas;
- it represents new births;
- once the firms enter in the sample they normally do not exit (they do not change delivering mode).

Istat uses them:

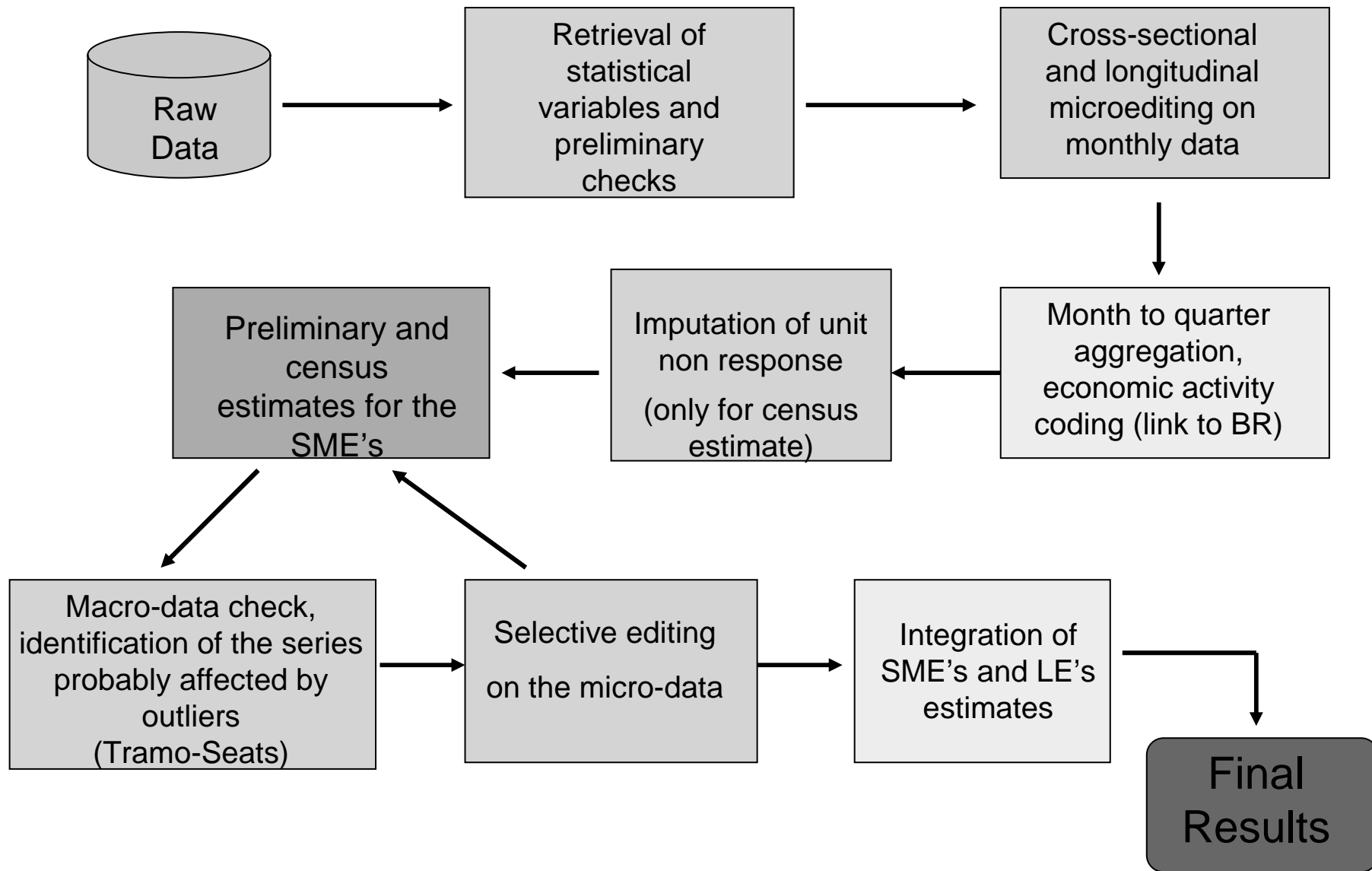
- to produce a “preliminary” estimate of current quarter  $t$ .

## Total population of DM10: the universe

- Because of delays in the delivery and registration of the non-electronic DM10s, INPS transmits 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;
- Istat uses them:
  - as auxiliary information (referred to  $t-4$ ) to improve the “preliminary” estimate of current quarter  $t$ ,
  - to produce a final (census) estimate of quarter  $t-5$ .

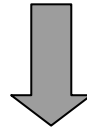
# Procedures Flow

every quarter there are two parallel processes: preliminary t and final estimate t-5



## Retrieving the statistical variables

- The translation of the administrative data into the required statistical variables imply complex computational aspects.
- There are many items in the DM10 form and more than 400 codes have to be identified (related to type of employment and associated wage bills, contributions, credit terms etc..).



We had to build a metadata database of codes and to keep it up-to-date to account for new codes and suppression of old ones.

Unfortunately italian social security rules change continuously

Survey variable	estimation
Gross wages	DM10 gross wage (as a result of <u>aggregation</u> inside the single form of wages related to different type of employment)
Number of jobs	DM10 jobs (as a result of <u>aggregation</u> inside the single form of different type of employment)
Other labour cost	DM10 total debit of the unit – (DM10wage* worker social security rate) + <u>estimation</u> of other labour costs (INAIL, TFR)

## Retrieving the statistical variables (2)

- Retrieving Other Labour Cost (OLC) has not been an easy task; we have at least two kind of problems:
  - 1) we can get from DM10 only total (employer + employee) social contributions due to INPS: we have to identify employee contributions, already included in the gross wages, according to the legal rates;
  - 2) only a part (though the largest one) of the OLC is recorded in the DM10: we need to impute the other labour cost (e.g. Employers' injuries insurance premiums - INAIL, severance payment - TFR).



We had to build a metadata database of  
the legal rates  
and to keep it up-to-date

# Check and Editing

- Check and editing Strategy
  - Microediting at monthly level data:
    - Interactive for top 100 firms;
    - Very selective and interactive over the firms with larger wage bills (we assign a score).
    - cross-sectional coherence and longitudinal (month to month) checks.
  - Macro Series check (after the estimates have been produced):
    - Comparison with other similar quarterly indicators (QNA);
    - automatic detection of outlier in the time series (Tramo ERROR).
  - Selective editing on the anomalous sectors:
    - identification of the units that have influenced most the change of the series;
    - correction, if needed.

# Imputation

- Imputation of unit non-response (only for census estimate)
  - separating the actual non-responses from “true” absences of the DM10s due to temporary inactivity of the enterprise (seasonal activity), or to the death of the unit (to avoid over-imputation);
  - using mainly the pattern of presence of the DM10s to identify the non-response of the units;

## Target estimation methodology

- Information available from INPS is:
  - INPS register (available at the end of each quarter);
  - the DM10 sample (available quarterly after 45 days);
  - the DM10 universe (available monthly after 13-14 months);
- short terms surveys have the problem to estimate levels and changes with reference to the current population without a current Business Register. In our case the BR (ASIA) has up to two years delay;
- given the huge size and timeliness of the sample data and the availability of the INPS register we developed an methodology to estimate the level and trend of the variables in the current population and not based on a fixed population in the past;
- at this stage the BR (ASIA) is used only to get the right Nace rev.1 economic classification.

## Target estimation methodology (2)

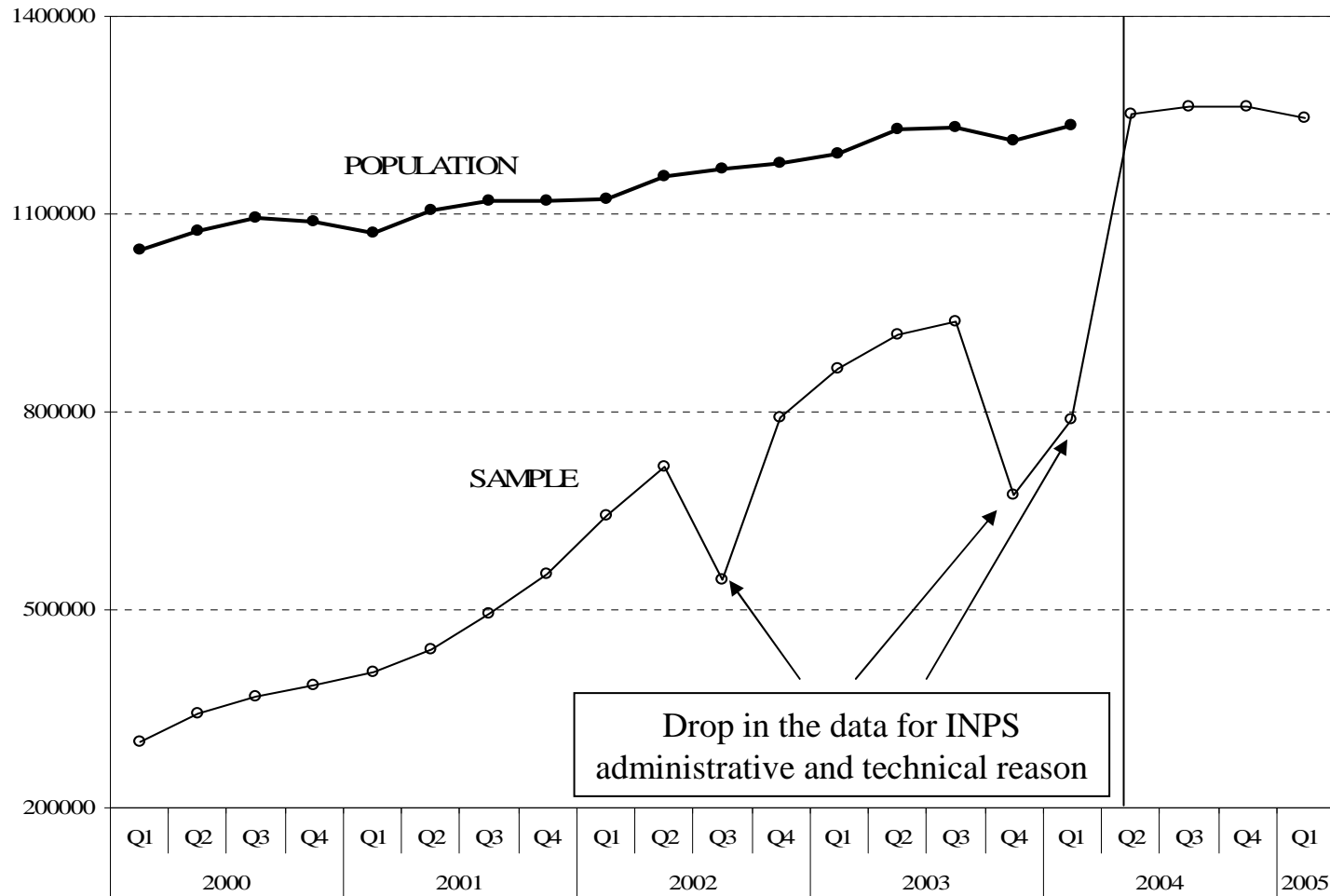
- In our estimates we assume that a relation exists between the target variables at  $t$  and the same variables at  $t-4$  (the auxiliary variables);
- the estimates are obtained applying to each unit in the sample a weight;
- weights are calculated to satisfy the condition that the auxiliary variables of the units in the sample multiplied by these expanding factors are equal to the known totals of the auxiliary variables (calibration);
- these totals are obtained by summing up over the current population the auxiliary variables available in the universe of  $t-4$ ;

## More on the preliminary estimates

- the calibration is performed at model groups level.
- The non randomness of the sample is faced with:
  - the partition of the population in model groups (50 economic activities, 4 firm size classes, 4 geographical areas, 2 age of firm classes);
  - the calibration of the weights.

# Recent change in the size of the “sample”

Since spring 2004 for administrative reason for all firm is compulsory to use “fast” delivering mode: internet; INPS do not accept anymore paper declaration.



## Implication of this change (sample = population)

- The target variables for the population can simply be calculated by summing up the data (no need of calibration methodology);
- There are potentially almost no limitation in the details (economic activity, size, etc) for tabulating and releasing the aggregate results;
- Still we need a calibration (or any other) methodology in case of a new drop in the data to expand to the population.
- A project to better the calibration methodology is ongoing.

## Summary of the INPS administrative data weaknesses and the methodology used to manage them

Weakness	Method for OROS	
	Preliminary (70 days)	Final (365+140 days)
Timeliness	Calibration, model groups	-
Missing data	Calibration, model groups	Imputation
Under coverage (just few large units)	Integration with LES	Integration with LES
Definitions differ	Not deemed necessary	Not deemed necessary
Incomplete set of variables	Estimation procedure using external information	Estimation procedure using external information
Measurement errors	Selective micro editing at monthly level Selective micro editing at quarterly level	Selective micro editing at quarterly level
Wrong or missing economic activity code	Link (through fiscal code) to the BR to get the right Nace rev.1 economic classification	Link (through fiscal code) to the BR to get the right Nace rev.1 economic classification

## Final remarks (1)

### Advantages

- ↙ very low collection cost;
- ↙ complete firm size coverage, timeliness, estimates precision
- ↙ a lot of different kind of data (can be used in different ways);
- ↙ no statistical burden on firms.

### Disadvantages

- ↙ huge handling of data (millions of records every month);
- ↙ 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.



commitment to co-operate;  
framework Istat-INPS agreement;  
high level co-ordination committee.

## Final remarks (2)

- in OROS not only big, but also small admin. change may jeopardise the quarterly release; strong link with admin. data suppliers (persons ready to respond to any different kind of questions relative to the data);
- in short term statistics we have to achieve and maintain:
  - **Timeliness and punctuality;**
  - We cannot waste time;
  - We have to prevent any unexpected kind of problem in advance: delays in delivery or quality problem (technological; administrative)
  - Be prepared for alternative solutions (rescue net).

# Thank you

[fabio.rapiti@istat.it](mailto:fabio.rapiti@istat.it)

[oros-info@istat.it](mailto:oros-info@istat.it)