Q 2016
European Conference on Quality in Official Statistics

Designing the integration of register and survey data in earning statistics

Session:18 Multi-Source Statistics

C. Baldi, C. Casciano, M. A. Ciarallo, M. C. Congia, S. De Santis, S. Pacini

National Statistical Institute, Italy

INE Instituto Nacional de Estadística eurostat
Madrid, May 31 - June 3

Contents

- Toward a system of integrated data in earning statistics
- The availability of the job Register on Earnings
- The new register-base Structural Earning Survey on 2014
  - The new role of the survey
  - The sampling and questionnaire redesign
- From sampling survey estimates to new register-based statistics: ongoing work...

Toward a system of integrated data in earnings statistics

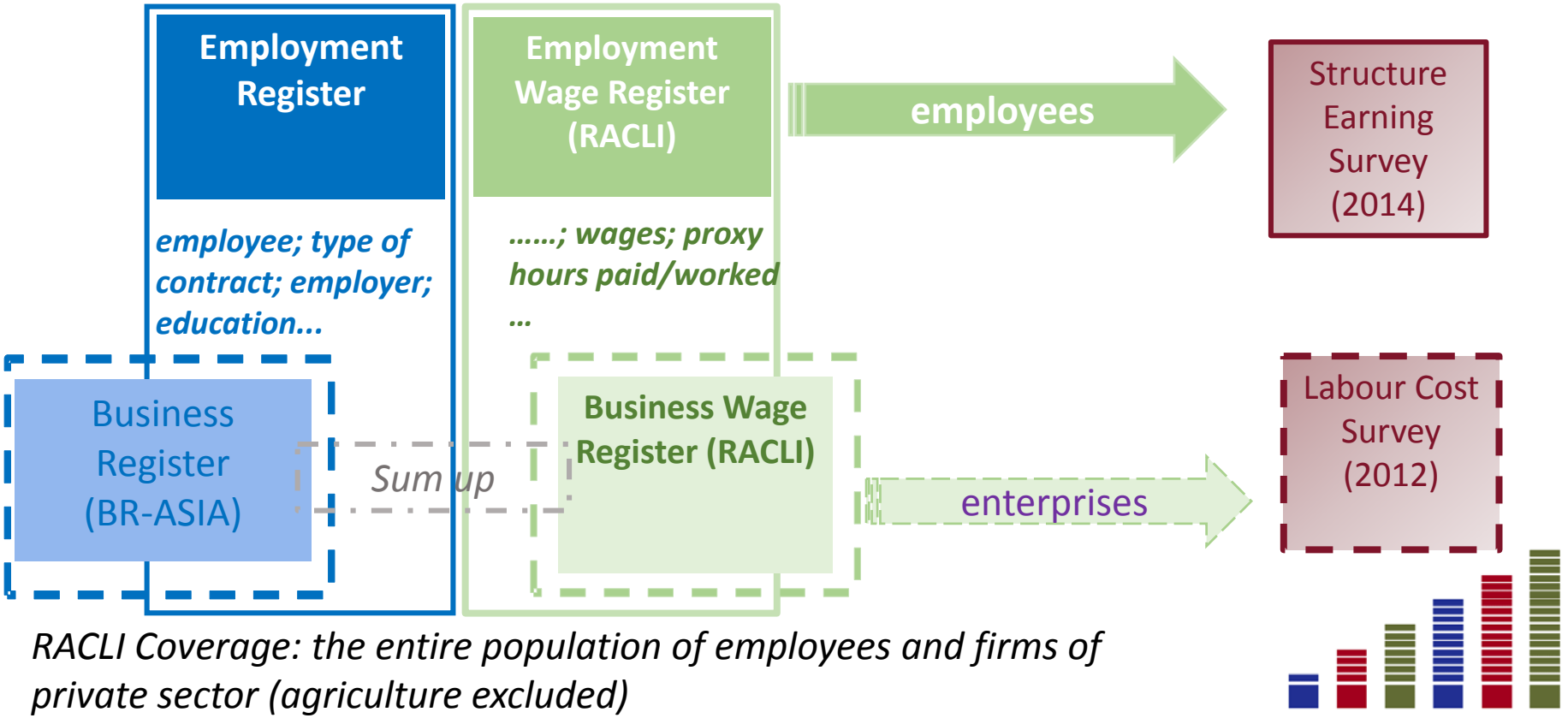
- Up to recent years earning statistics based on 4-yearly Structure of Earning Survey (SES - Regulation CE n. 530/99)
- Evolution in the statistical register available ➔ new coherent system of statistics on earnings and hours paid/worked based on integrated data

  - Annual statistics on earnings broken down for characteristics of job, enterprise and employee
  - Annual statistics of the Gender Pay Gap (GPG) on earnings per hour paid
  - 4-yearly release of microdata on employees for SES Regulation
  - 4-yearly release of Labour Cost Statistics for LCS Regulation

RACLI Register on job earnings

Up to 2011: Business Register (BR) with information on total number of employees

Since 2011: Employment Wage Register (RACLI) employees-level register with information on type of contract, wages, a proxy of paid time, other labour costs



RACLI Coverage: the entire population of employees and firms of private sector (agriculture excluded)

The RACLI variables and their status

The RACLI variables and their different status referring to their usability (for SES statistical purposes):

SV: statistical variables

➔ can be used directly *definition fulfill the statistical one*

(e.g. contractual working time, type of employment contract, wages...)

PV: proxy variables

➔ can be used after harmonization *definition not exacly fulfill the statistical one (e.g. hours paid)*

AV: auxiliary variables

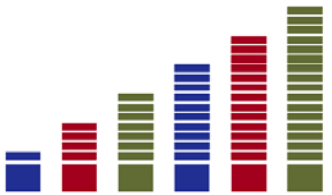
➔ *Variables that can be used as auxiliary ones*

(e.g. a proxy of non regular earnings)

NAV: not available

➔Statistical variables not available in the Register

(e.g. ISCO, overtime hours paid)



Survey unavoidable but...

Survey necessary to ask for

- **unavailable variables in the Register**
- **information to harmonize Register proxy variable to statistical concepts**

…but completely redesigned to be as much complementary as possible to the register

- the "one variable collected only once" principle inspired the new survey design
- the approach is for a "circular" integration:
  - register data assist the survey
  - survey data planned to check and support the harmonization of register variables
- Several solutions ex-ante has been experimented:
  - the pre-filling and the redesigned of the questionnaire
  - the new sampling method

Old SES sampling strategy

First stage of sampling: enterprises

- ***Sampling design***: *stratified random (StrRS)*
- ***Sampling size***: *via optimal multivariate multidomain allocation (Bethel, 1989)*
- ***Auxiliary variable:*** *monthly gross earning per employee* ***Target CV:*** *3%*
- ***Primary units' selection scheme***: *Pareto pps sampling*
- $N_E = 180K$ *enterprises in the register (>10 employees)* $n_E = 20K$ **sampled**

Second stage of sampling: employees

- ***Sampling design***: *stratified random (StrRS)*

☹ ***Sampling size***: *No optimal allocation algorithm!* Dimensions 2nd stage sample established a priori in absence of auxiliary information on employees

☹ **Selection of the employees:** Left to the enteprises according to a

- ***Secondary units' selection scheme***: *systematic sampling*
- $N_W = 8{,}2\ ML$ $n_w = 480K$ **sampled!!!**

N.B.: no certainty enteprises followed the proposed scheme

The introduction of RACLI Register: main methodological innovations

Register used to assist SES estimate production in all sampling phases

1. Sample directly the employees, instead of delegating this phase to the enteprises
2. Improve the allocation:
   a. *the use of guide variables at the employees-level that are (close to) the core variable of the survey*
   b. *the use of characteristics of employees in the stratification (e.g. FT/PT)*
3. Calculate a close approximation of the ex-post errors on subpopulations (e.g. the sample to be provided to Eurostat)
4. Measure ex ante the bias on core variables: this has been very useful in testing some methological innovation
5. Reduce dramatically the sample size

Q 2016
European Conference on Quality in Official Statistics

Redesigning the new SES 2014 – our solution: *TWO* stage *stratified random Sampling(StrRS)*

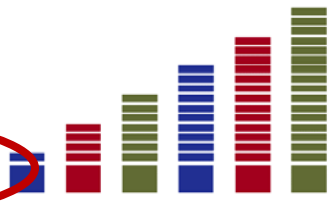
First stage of sampling: enterprises

- ***Sampling design***: *stratified random (StrRS)*
- ***Strata***: *2 digit NACE X Nuts 1 X size classes*
- ***Sampling size***: *via optimal multivariate multidomain allocation (Bethel, 1989)*
- ***Auxiliary variables:*** *average monthly hourly earning; average yearly earning per employee*
- ***Target CV:*** *3.6%*
  - $N_E = 168K$ *enterprises in the register* **$n_E = 25K$ sampled enterprises**

Second stage of sampling: employees

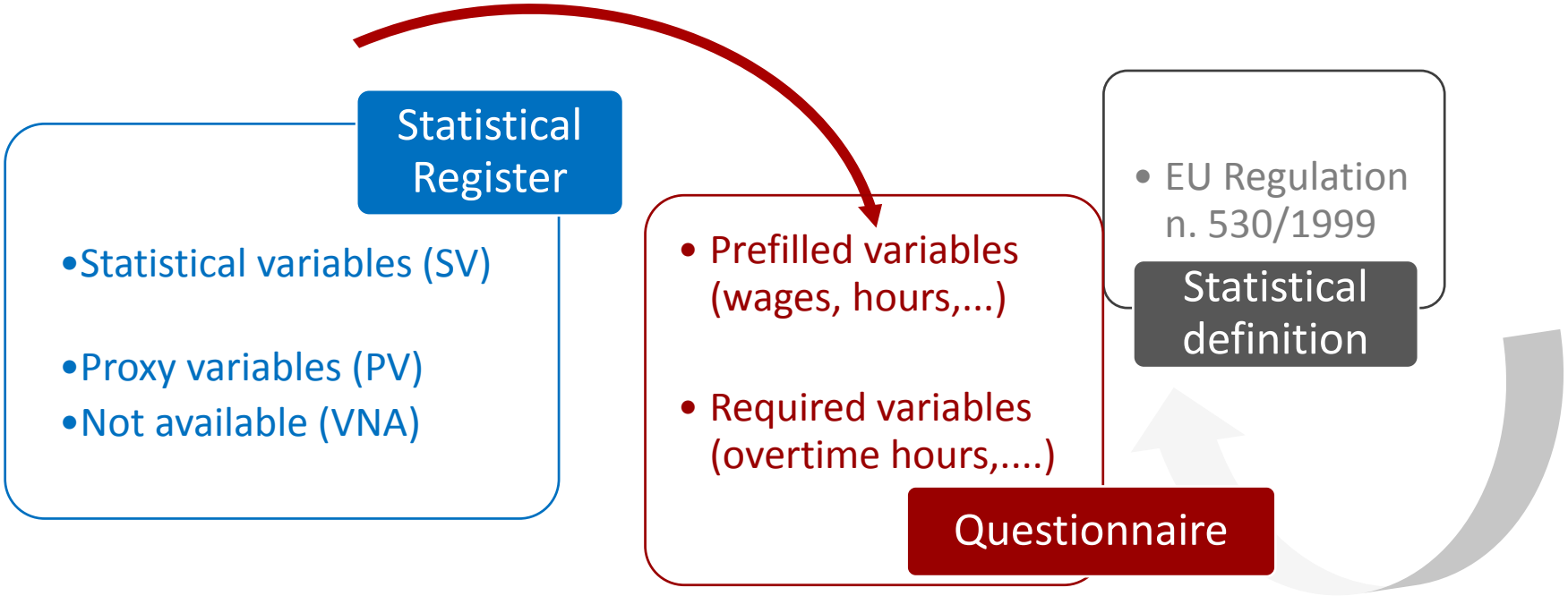
- ***Sampling design***: *stratified random (StrRS)*
- ***Strata***: *Nuts1 localization of worker X working time (FT/PT) X Broad Occupation (Blue collar/White Collar) N.B.: Each enteprise partitioned into groups corresponding to strata*
- ***Auxiliary variables:*** *monthly hourly earning; yearly earning;*
- ***Sampling size***:*via optimal multivariate multidomain allocation, with the constraints:*
  - Minimum number of sampled worker per enterprise=3
  - Maximum number of sampled worker per enterprise=500
  - $N_W = 8{,}5ML$

$n_w = 212K$ sampled



The new questionnaire as a bridge

Information related to wages and hours has been pre-filled from Register in order to reduce measurement errors and correct administrative data.

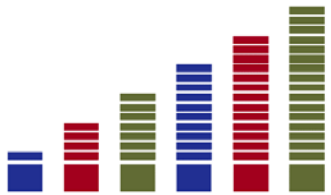


- Register used to build a simplified questionnaire
- Survey information collected also to come back to register
  - to impute the missing data or
  - to adjust the nearly statistical variables

Ongoing work

- C&C of survey data:
  - Reconciliation with register data statistical variables
  - Error localization under edit constraints
  - Other variables not in the register through minimum distance donor
- Imputation of total non responses
- Reweighting of survey data through calibration to RACLI known totals on **number of employees and earnings** → Estimates from survey data consistent with those on Register
- Harmonization of Register proxy variables: PV→SV
  - Survey data correction factors
  - Register auxiliary variables
  - Information from other surveys as known/estimated totals
  - Mass imputation

    e.g. **proxy** hours paid → **statistical** hours paid



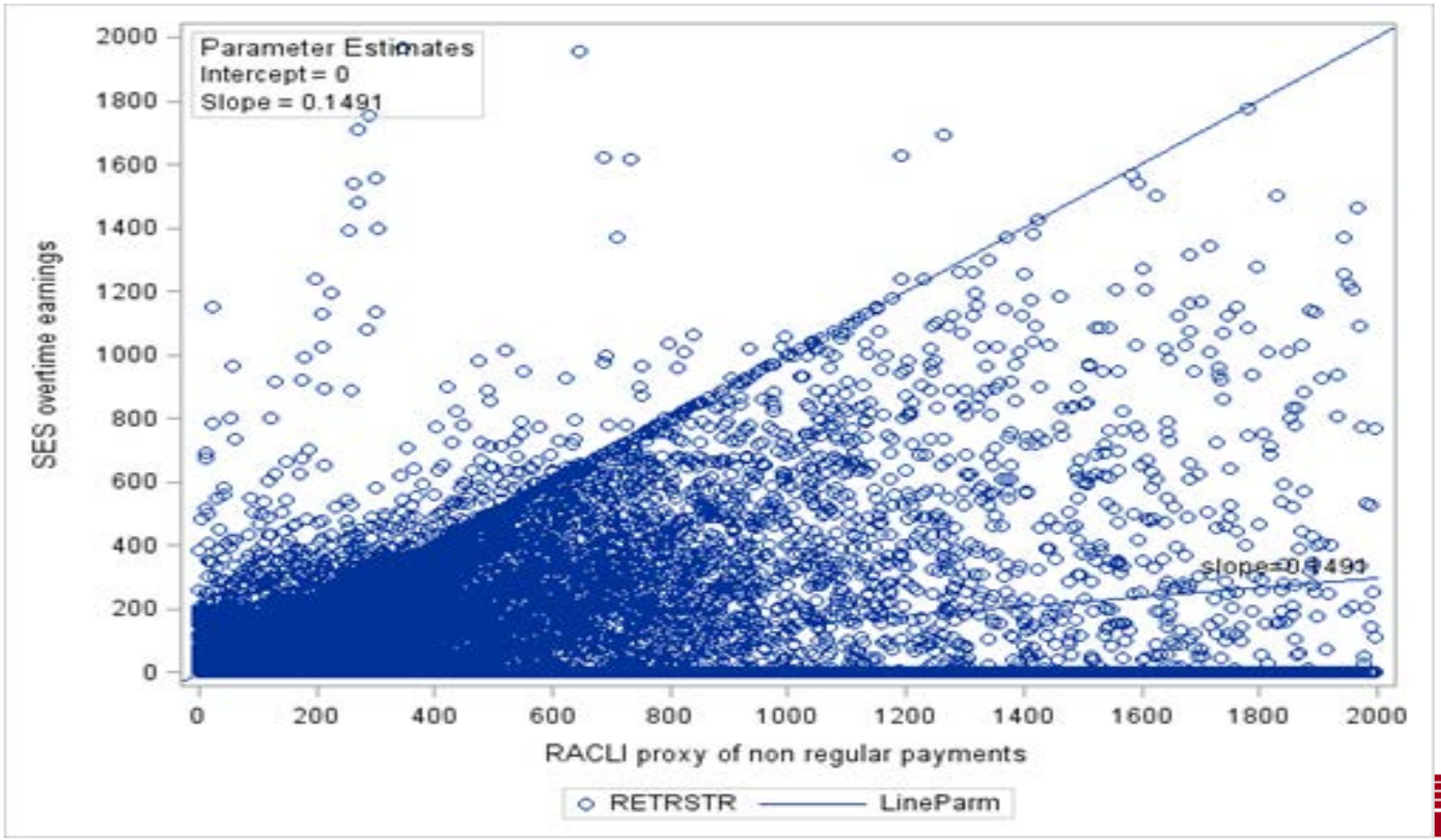
From Proxy to Statistical Variables

- RACLI provides only a proxy for the number of hours paid because overtime hours are not included
  - Share of overtime hours is a small part of the number of hours paid (2-3%) but its absence make the variable slightly downward biased
  - The Register earnings includes overtime wages

- The aim is to impute overtime hours in RACLI
- Sources on overtime hours:
  - SES survey: data on the number of overtime hours and earnings at job level
  - Monthly/quarterly enterprise surveys data➔ quarterly estimates of the number of overtime hours from VELA (SRS on enterprises between 10 and 500) and monthly information on the number of overtime hours for each enterprises with at least 500 employees (GI)
  - RACLI auxiliary variables (e.g. a proxy of non regular earnings)

RACLI register and survey data

Relationship between SES survey respondent overtime earnings and a RACLI proxy of non regular earnings (retrstr)



To conclude

Wide use of the register data in all phases of the survey process

- More efficient sample and reduction of sample size ➔ lower burden and higher response rate

- Prefilling questionnaire➔ reduction of measurement errors

- Improving the quality of register variables: use of multiple sources (register and surveys data) and different class of models (multilevel models?)

….toward a system of integrated data and coherent earning statistics.

Q 2016
European Conference on
Quality in Official Statistics

Thank you for your attention

silvia.pacini@istat.it