Estimating structural business statistics based on administrative data: the case of the Italian small and medium enterprises

Luzi O., Rinaldi M., Seri G., Guarnera U., De Giorgi V.

Italian National Statistical Institute (Istat)

EESW, Neuchâtel, 12 – 14 September 2011

Outline

- **Use of admin data for business surveys**
- **Experimental context: SEAS - Italy, 2008**
- **Admin sources used: FS and SS**
- **Experimental study**
- **Conclusions**

Use of admin data for business surveys

Main objectives

1) Reduction of statistical burden on enterprises

2) Reduction of survey costs for NSI's

3) Increase of population coverage:
   - Integration of unit and item non responses from admin sources
   - Increase of response rate

Experimental context: SEAS (1)

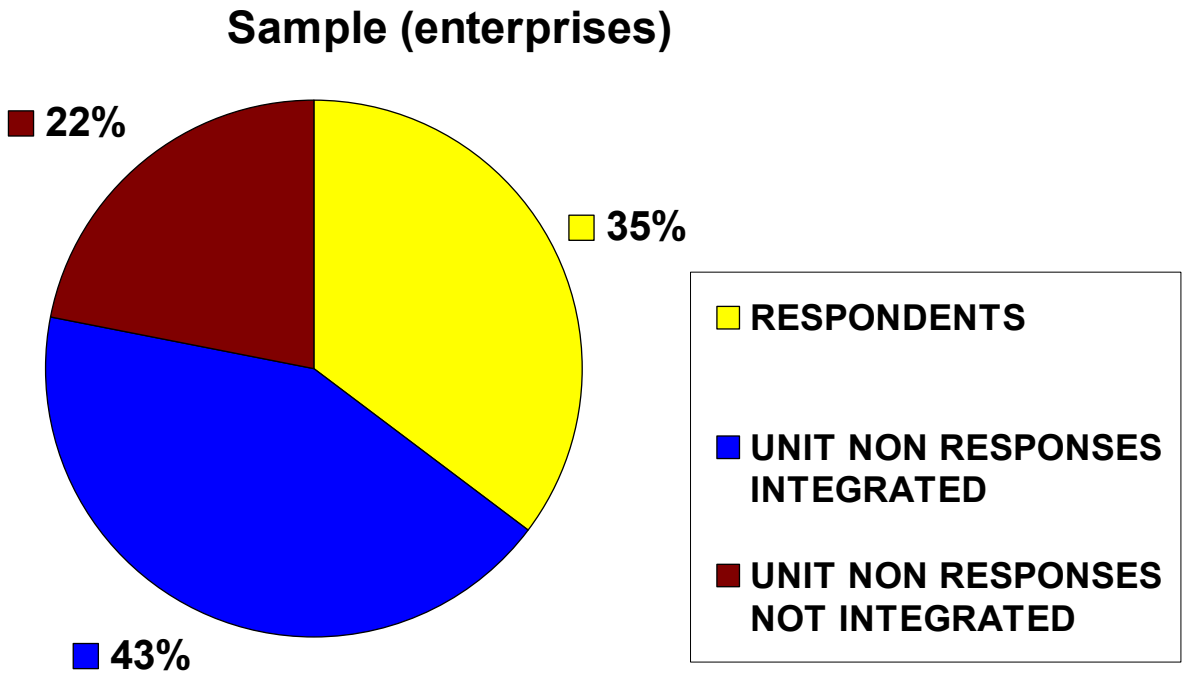
System of Enterprise Account Surveys (SEAS) [Italy, year of reference 2008]

- **Target population: enterprises operating in industry, construction, trade and services**
- **Frame: the Italian business register (Asia)**
- **2 distinct annual surveys to estimate mainly profit-and-loss accounts, employment, investment, personnel costs of Italian enterprises:**
  1. **SME, sample survey on 4,5 million enterprises with less than 100 persons employed**
  2. **SCI, census survey on 11,000 enterprises with 100 or more persons employed**

Experimental context: SEAS (2)

1) SME

- Population: 4,5 mil enterprises, 13,5 mil persons employed
- Sample: 105,000 ent, 1,250,000 emp
- Respondents: 37,000 ent, 415,000 emp

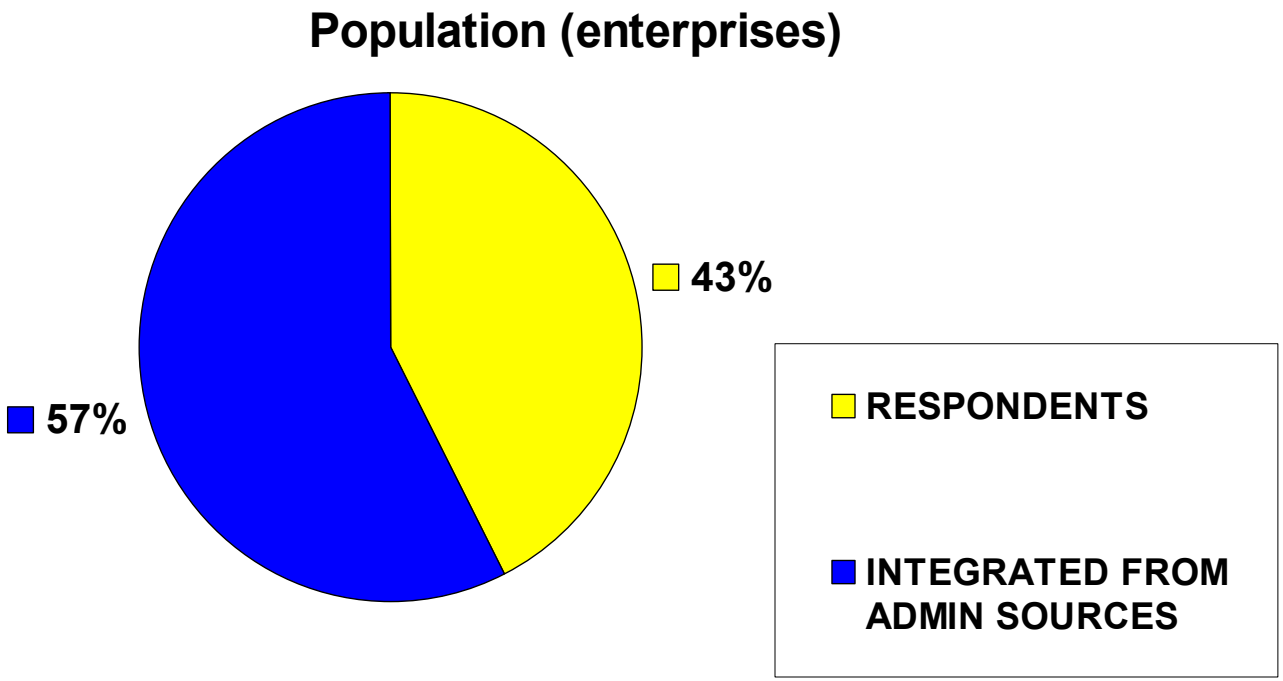


Istat

Experimental context: SEAS (3)

2) SCI

- **Population: 11,000 ent, 4,4 mil emp**
- **Respondents < 5,000 ent, 2,6 mil emp**



Admin data used in the experiment

Financial statements (FS)

- 650,000 corporate firms
- the best harmonized source with SBS definitions

Sector studies (SS)

- 3,6 million enterprises
- fiscal survey on enterprises with:
  30.000 € ≤ turnover ≤ 7,5 million €

Coverage of SEAS population and respondents

	Population		Respondents	
	Enterprises	Persons employed	Enterprises	Persons employed
FS	14,4	51,8	49,4	93,1
SS	77,2	55,4	60,5	7,6
FS or SS	80,7	89,1	87,4	95,3

Objectives of the study

Evaluating the statistical adequacy of information provided by FS and SS for estimating key SBS parameters, assuming that external data can be used in different ways:

- **as auxiliary information to improve the efficiency of the statistical survey process (in particular, error detection and non response imputation)**
- **as primary source of information for SBS parameters estimation (complemented by direct surveys to estimate either non covered sub-populations, or variables which are not directly available from external sources)**

The experimental studies: improving the efficiency of error detection (1)

Objective

Evaluating the effectiveness of selective editing for the identification of influential measurement errors on surveyed variables by exploiting related auxiliary information from FS and SS

The focus is on variables which are directly available in FS or SS, and restrict the attention on the sub-population on SMEs

Selective editing is a data editing approach which allows to identify the subset of units potentially affected by influential measurement errors on which editing efforts are to be mostly (or even exclusively) spent in order to target predefined levels of estimates' accuracy. The result is a more efficient use of survey resources and reduced respondent burden.

The experimental studies: improving the efficiency of error detection (2)

Method

- For each variable Y, estimation domain and admin source, compute a simple local score as the weighted difference between survey and admin micro data

- Identify the first ordered k units so that, once they are replaced by the corresponding admin values, the relative difference (*Diff_rep*) between the resulting estimate of the Y total differs less than 2% from the corresponding estimate computed on the original Y values

The experimental studies: improving the efficiency of error detection (3)

Overall results

In the majority of cases, few units are found as responsible of discrepancies among survey and "corrected" estimates. This means that admin data are reliable as benchmarking information for the considered variables, and that they can be efficiently used to improve the error detection stage

Economic Sector	2-dig. Nace code	Variable									
		Turnover					Personnel Costs				
		N	n influential	% Influential	Diff_ori (%)	Diff_rep (%)	N	n influential	% Influential	Diff_ori (%)	Diff_rep (%)
C-Manifacture	**26**	254	1	0.4	20.07	-0.14	254	3	1.2	-2.90	-1.81
	27	227	1	0.4	-2.05	-0.02	227	1	0.4	-2.61	-1.61
M- Professional, scientif. Techn. Activities	**72**	96	1	1.0	-2.04	1.60	96	2	2.1	-4.57	-0.19
D -Electricity, gas, steam and air conditioning supply	**35**	263	1	0.4	-5.40	0.71	263	3	1.1	-5.28	-1.60
E- Water supply, sewerage, waste management and remediation act..	**37**	59	1	1.7	2.24	0.48	59	3	5.1	-2.82	-1.05

The experimental studies: improving the efficiency of non response imputation (1)

Objective

Evaluating the effectiveness of data from FS and SS as auxiliary information for predicting (imputing) survey non responses for key SBS surveyed variables

The focus is on variables which are directly available in either FS or SS, and limit the attention on the sub-population on SMEs

<u>Imputation method</u>: within-cells regression imputation
For each response variable Y, the model's covariate is the corresponding item directly available from either FS or SS

Cells defined in terms of Economic activity & size class

The experimental studies: improving the efficiency of non response imputation (2)

Method

For each variable Y, a MonteCarlo experiment has been performed consisting in I=100 iterations of three steps:

- **Simulation of missing values at random w.r.t. economic activity on the subset of responding units (5% and 10%);**
- **Regression model estimation, "non response" imputation and estimation;**
- **Evaluation**

Evaluation indicators

- **Relative Root Mean Squared Error (RMSE)**
- **Relative estimation error due to imputation (REEI)**
- **Weighted Relative average imputation error (WRIE) (microdata level)**

The experimental studies: improving the efficiency of non response imputation (3)

Overall results

At estimation level, imputation does not significantly affect estimates for the analyzed variables and economic activity sections

Worse effects can be seen at elementary data level (WRIE): in most cases those values which are predicted unsatisfactorily correspond to original SME values with high discrepancies w.r.t. the administrative ones - this is a further confirmation of the need a more efficient use of external information at the editing stage

Use of admin data for estimation: evaluation of source effect

Objective

Evaluating the statistical impact on final estimates of using admin data instead of survey data

Evaluations methods:

I. **Compare estimates of variables' totals calculated from survey data with the same estimates calculated from admin data**

II. **Check if *true* admin totals belong to 95% confidence intervals associated to estimates of totals obtained from survey data**

Use of admin data for estimation: evaluation of source effect

- **Population: SME+SCI 2008**
- **Frame: ASIA 2008**
- **Admin sources: FS and SS, separately**
- **Restrict attention to some variables directly available from FF and SS: *turnover*, *purchases of good and services* and *personnel costs***
- **Restrict attention to sub-populations where FS and SS are –respectively– available**
- **Domain of interest: Nace Rev.2 sections**
- **Estimate totals and their 95% confidence intervals by calibration procedure on (SME+SCI) respondents**
- **Calculate *true* totals of variables from admin sources**

Evaluation of source effect: I - difference between admin and survey estimates

Method

1. For each variable Y and source (survey, FS and SS), calculate calibration estimate of totals by section
2. For each variable Y, admin source and section, compare estimates of totals calculated from survey data with the ones obtained from admin data, to evaluate if survey and admin sources can be considered coherent

Results

For *turnover* and *personnel costs:* in most sections, estimate of totals obtained from survey are similar to the ones obtained from both FS and SS

For *purchases of good and services:* they are not

Evaluation of source effect:
I - difference between admin and survey estimates

Results: % difference between estimate of totals

	Survey vs.*FS*		Survey vs.*SS*	
Nace Rev 2 - Section	Turnover	PersCosts	Turnover	PersCosts
B mining and quarrying	- 0,02	0,42	- 0,15	2,31
C Manufacturing	0,07	- 1,00	0,31	- 0,92
D Electricity, gas, steam, etc.	- 1,39	11,39	- 2,47	- 7,31
E Water supply, sewerage, etc.	0,14	- 0,56	9,61	0,25
F Construction	0,64	- 1,80	0,96	- 2,46
G Wholesale and retail trade; etc.	- 0,93	- 1,28	- 0,37	- 2,90
H Transportation and storage	- 1,60	- 0,99	0,75	- 1,51
I Accommodation and food service	- 0,19	- 0,18	- 0,37	- 1,07
J Information and communication	8,60	2,66	- 0,60	- 2,72
K Financial and insurance	0,01	- 1,65	1,03	- 3,26
L Real estate activities	- 2,42	- 2,00	- 0,41	1,09
M Professional, scientific and technical act's	21,99	- 1,68	- 1,88	- 2,80
N Administrative and support service act's	1,58	2,66	- 0,44	- 0,46
P Education	0,65	- 1,12	0,91	- 2,86
R Arts, entertainment and recreation	1,25	- 4,63	- 1,67	- 5,71
S Other service activities	0,71	- 2,81	3,21	- 2,31
Total	**0,53**	**- 0,53**	**0,02**	**- 1,83**

Evaluation of source effect: II – check if admin totals belong to 95% C.I.

Method - For each variable Y, admin source and section

1. **Estimate from survey data 95% confidence interval for totals of Y, by a calibration procedure**
2. **Calculate *true* totals of Y from admin sources**
3. **Check if *true* total from admin sources belong (or not) to corresponding confidence Interval estimated from survey data**

Results

For both FS and SS, in several cases *true* admin totals do not belong to confidence intervals, especially for industry and trade activities

Evaluation of source effect:
II – check if admin totals belong to 95% C.I.

Results

	FS			SS		
Section	PurchGS	Turnover	PersCosts	PurchGS	Turnover	PersCosts
B mining and quarrying						
C Manufacturing						
D Electricity, gas, steam, etc.						
E Water supply, sewerage, etc.						
F Construction						
G Wholesale and retail trade; etc.						
H Transportation and storage						
I Accommodation and food service						
J Information and communication						
K Financial and insurance						
L Real estate activities						
M Professional, scientific and technical act's						
N Administrative and support service act's						
P Education						
R Arts, entertainment and recreation						
S Other service activities						
Total						

true totals do not belong to C.I.

true totals belong to C.I.

Use of admin data for estimation: evaluation of source effect

Evaluation of source effect

I. Calibrated estimates of total obtained using admin data are similar to the ones obtained using survey data. It encourages us in extending the use of admin data at estimation stage.

II. *True* admin totals do not belong to estimated 95% confidence intervals in several cases. It can depend on:

- Sampling elements
  need to produce estimates for a large and complex population within detailed domains, while minimizings costs and burden
- Non sampling aspects
  measurement errors in admin data with strong effects on totals

Conclusions

- **Istat project: from a production system mainly based on survey data, to a new system based on a more extensive use of admin data**

- **Go on analyses and experimental studies to assess "usability" of administrative data and to evaluate potential biasing effects due to integrating administrative and survey data for SBS estimations**

- **Improve error detection phase, in particular by experimenting new methodologies for selective editing (e.g. the ones based on contamination models)**