Quality improvement of individual data and statistical outputs based on combined use of administrative and survey data

Emmanuel Gros
Insee
Ljubljana, May 2011

Context

- ➢ Re-engineering of the French system for the production of structural business statistics :

  - ✓ in the previous system, two « parallel » processes : statistical survey and process using administrative data ;

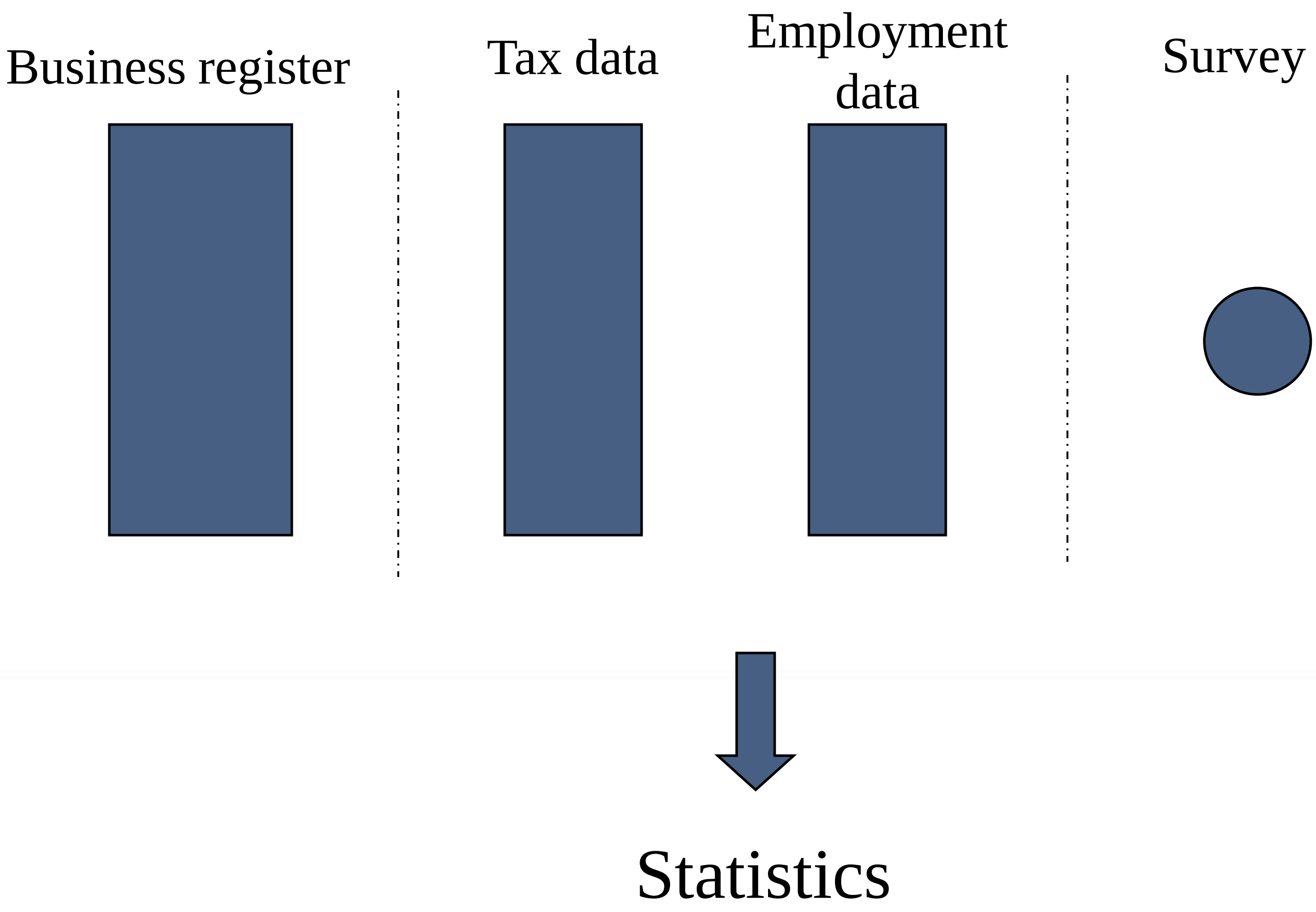
  - ✓ new system based on combined use of administrative and survey data ⇒ improve coherence of the statistical results.

➔ Complicate the statistical production, but in compensation opens up new horizons :

  - ✓ for data editing process ⇒ consistency monitoring between sources ;

  - ✓ for quality of estimates ⇒ calibration techniques, specific estimators.

Structure of the new system Esane



Business register

Tax data

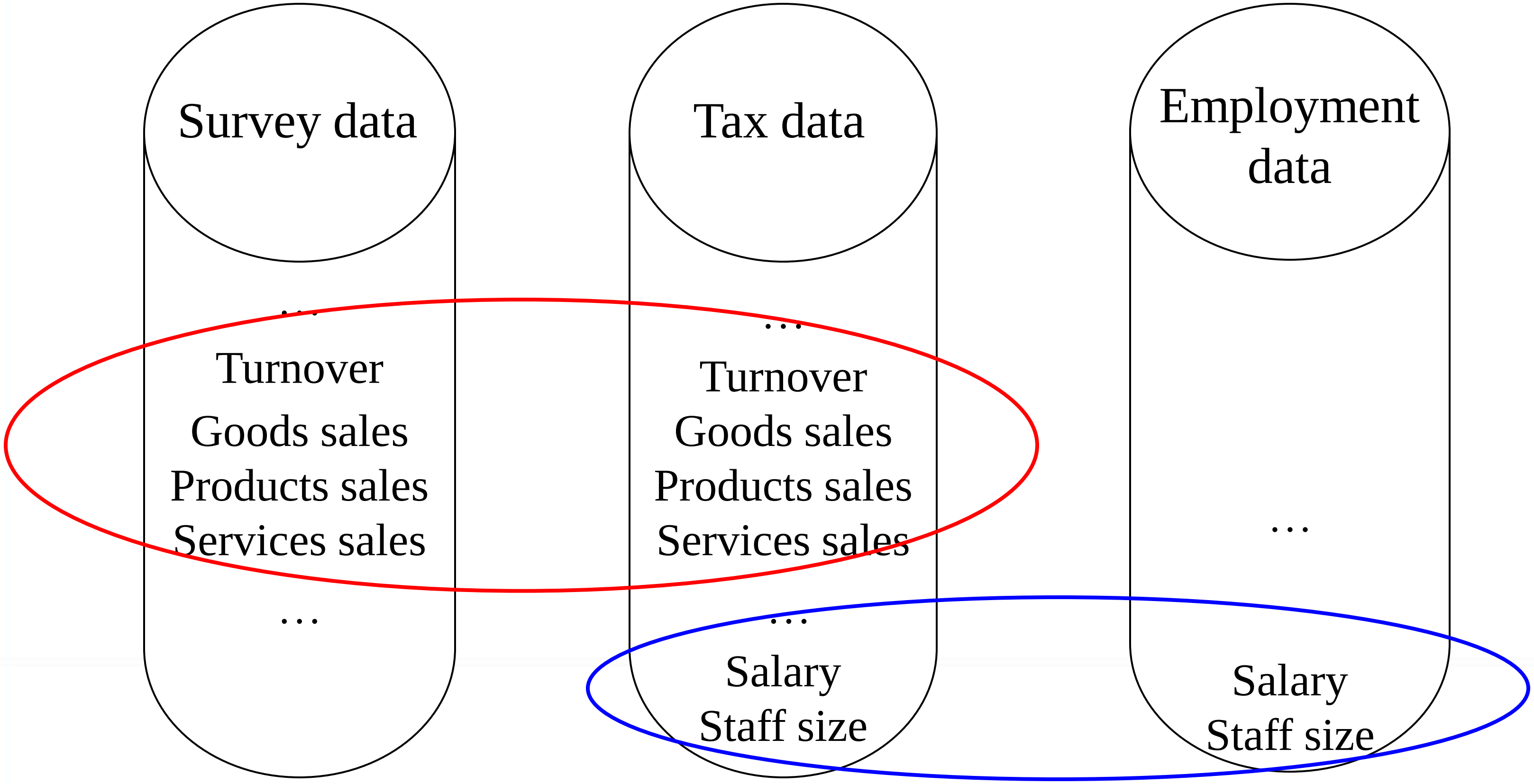
Employment data

Survey

Statistics

Data editing process

- Common characteristics exist in the different sources :



Survey data

…

Turnover

Goods sales

Products sales

Services sales

…

Tax data

…

Turnover

Goods sales

Products sales

Services sales

…

Salary

Staff size

Employment data

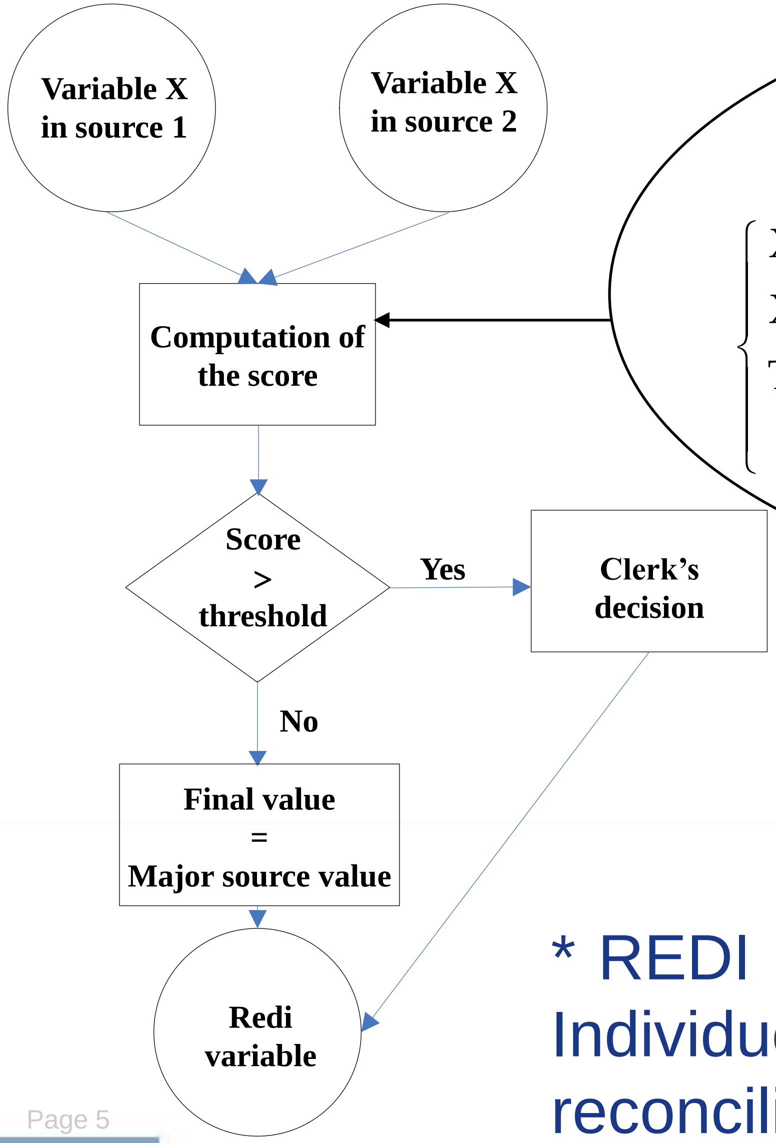
…

Salary

Staff size

→ We can use this redundancy of information to set up a consistency monitoring of individual records.

The REDI* process



Variable X in source 1

Variable X in source 2

Computation of the score

Score > threshold

Yes

Clerk's decision

No

Final value = Major source value

Redi variable

$$\text{score} = \left| \frac{X_{S1} - X_{S2}}{T(X_p)} \right|$$

$$\begin{cases} X_{S1} = \text{value of characteristic X in source 1} \\ X_{S2} = \text{value of characteristic X in source 2} \\ T(X_p) = \text{total of characteristic X in the major source} \\ \qquad \text{at the level of aggregatio n used for the control} \end{cases}$$

* REDI : « REconciliation des Données Individuelles », i.e. individual data reconciliation

Assessment of the REDI process (1)

- For the turnover and its sales' breakdown, for year 2008

- For a threshold of 1% and a NACE "3-digit" level of aggregation

- A first campaign of the new system, disrupted by many problems ⇒ in 2008, the individual data reconciliation was mainly performed in an automatic way ⇒ generally, choice by default of the major source value for the Redi variables :

  - ✓ for turnover, major source = tax data retained for 97% of the units, accounting for 99% of the total turnover ;

  - ✓ for turnover's breakdown between "commercial activities", "service activities" and "production of goods", major source = survey ⇒ structure stemmed from the survey retained for 81% of the units, accounting for 93% of the total turnover .

Assessment of the REDI process (2)

- For turnover, tax data and survey data are globally consistent : only 1,1% of difference.

- Most important discrepancies observed for the sale's repartition

Variable	Survey total	Tax total	Final Total
Turnover	3 509	3 469	3 466
Goods sales	1 436 40,9%	1 433 41,3%	1 411 40,7%
Product sales	1 232 35,1%	925 26,7%	1 180 34,1%
Service sales	840 23,9%	1 111 32,0%	875 25,2%

Amount in billion €

→ The impact of this process was not negligible.

Assessment of the REDI process (3)

- Efficiency of the REDI selective editing process :

  - ✓ less than 3% of the sample's units are detected by the process as seriously inconsistent...

  - ✓ ... but these units account for 81% of the difference in terms of turnover !

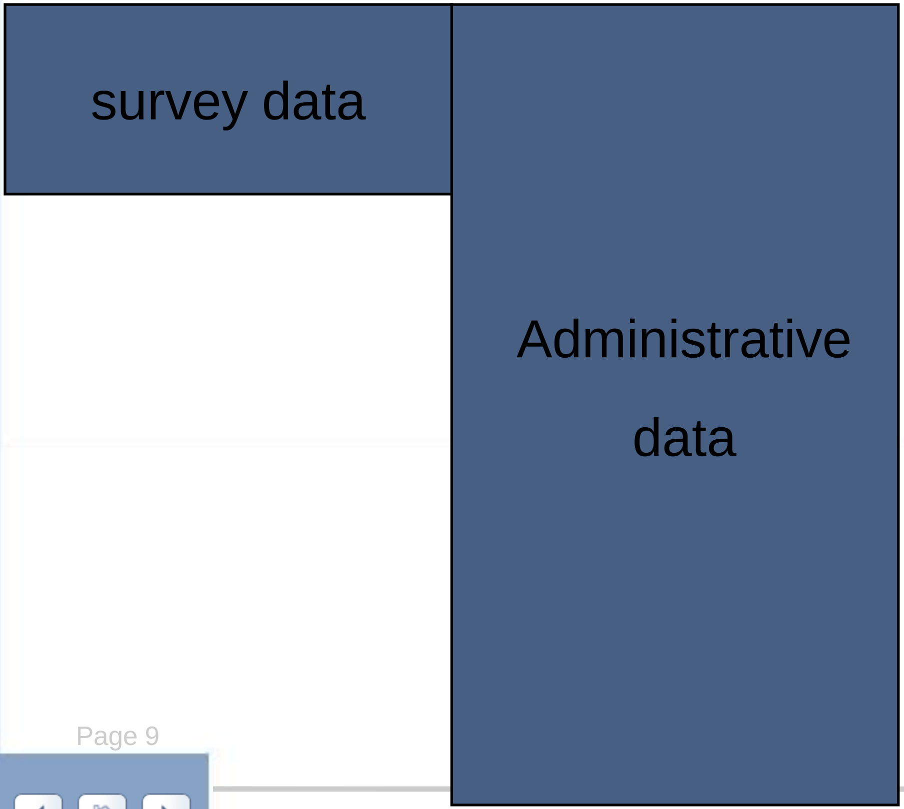
- In 2008, only 200 enterprises actually called back by a clerk.

- In 2009, the selective editing process was fully operative, so most of the serious inconsistencies were detected and checked by a clerk.

Statistical estimates (1)

- Methodological issue : produce statistical estimates jointly using, in the most efficient way, both administrative data and statistical survey.

Framework


survey data

Administrative data

Statistical device

⇒ Use of calibration techniques

⇒ Use of specific estimators, mainly difference estimator

Statistical estimates (2)

- Starting point : the standard estimator $\sum_{i \in R} d_i Y_i$
- First step : use of calibration techniques :

$\Rightarrow$ modify weights according to the following calibration equations :

$$\begin{cases} \sum_{i \in R} w_i T^{tax}(i)\, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U} T^{tax}(i)\, \mathbb{1}_{APE\_rep=X}(i) \\ \sum_{i \in R} w_i\, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U} \mathbb{1}_{APE\_rep=X}(i) \end{cases}$$

where APE_rep is the value of the APE code in the register and T(i) the value of the turnover of enterprise i in tax data.

$\Rightarrow$ implemented at the “3-digit” level for turnover and “2-digit” level for number of enterprises, to limit weights distortion.

Statistical estimates (3)

- Second step : for sector-based estimates and variable Y available for all units, use of a difference estimator, confronting the APE code of the register (APE_rep) and the APE code coming from the survey (APE_enq) :

$$\sum_{i \in R} w_i Y_i \, \mathbb{1}_{APE\_enq=X}(i) + \sum_{i \in U} Y_i \, \mathbb{1}_{APE\_rep=X}(i) - \sum_{i \in R} w_i Y_i \, \mathbb{1}_{APE\_rep=X}(i)$$

- For variable Y available only on the survey, use of the Horvitz-Thompson estimator using the final weights :

$$\sum_{i \in R} w_i Y_i \, \mathbb{1}_{APE\_enq=X}(i)$$

Impact of the new statistical estimates (1)

- Objective : assess the impact of the methodological improvements – use of calibration techniques and specific estimators – implemented in the new system.

➔ We reproduced estimators as in the previous system, and compare their CVs with CVs of the new estimators.

- Estimators « as in the previous system » : $\sum_{i \in R} d_i Y_i \, \mathbb{1}_{APE\_enq=X}(i)$

- Estimators in the new system :
  - ✓ for turnover and sales : $\sum_{i \in R} w_i Y_i^{redi} \, \mathbb{1}_{APE\_enq=X}(i)$
  - ✓ for variables “number of enterprises”, “salary” and “employer's social contributions” :

$$\sum_{i \in R} w_i Y_i \, \mathbb{1}_{APE\_enq=X}(i) + \sum_{i \in U} Y_i \, \mathbb{1}_{APE\_rep=X}(i) - \sum_{i \in R} w_i Y_i \, \mathbb{1}_{APE\_rep=X}(i)$$

Impact of the new statistical estimates (2)

- CVs computed thanks to a self-made SAS macro, which takes into account :

  - ✓ for the mimicked estimators of the previous system : sampling error of the survey, due to the stratified sample design, and unit non-response adjustment using the RHG model ;

  - ✓ for the current estimators : same things + use of calibration techniques and use of the difference estimator when applicable.

- Comparison for sector-based estimates at the “3-digit” level of the NACE

Impact of the new statistical estimates (3)

Means and quintiles of the ratio between new estimators' CVs and CVs relating to the previous system

	Turnover	Goods sales	Products sales	Services sales	Number of enterprises	Salary	Employer's social security contributions
Mean	0,67	0,94	0,88	0,86	0,74	0,63	0,64
Max	2,50	3,31	2,38	1,67	2,99	2,15	2,98
Q90	0,99	1,03	1,03	1,02	1,03	1,00	1,03
Q75	0,90	1,00	1,00	1,00	0,94	0,87	0,88
Median	0,70	0,98	0,96	0,97	0,79	0,66	0,62
Q25	0,45	0,88	0,76	0,79	0,57	0,36	0,36
Q10	0,18	0,56	0,50	0,45	0,27	0,12	0,10
Min	0,00	0,11	0,09	0,00	0,00	0,00	0,00

➔ global improvement of the estimator’s accuracy

Conclusion

- The new French system for the production of structural business statistics presents many advantages...

  - ✓ improves coherence of the statistical results ;
  - ✓ permits a consistency monitoring on key variables, which reduces the bias due to response errors ;
  - ✓ allows the use of more sophisticated estimates, involving calibration techniques and specific estimators, which improve estimates' accuracy in most of the cases.

- ... at the cost of a more complex device, which is not without presenting some practical problems.