WP.6
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

Quality Improvement of Individual Data and Statistical Outputs Based on Combined Use of Administrative and Survey Data

Key Invited Paper

Prepared by Emmanuel Gros, Insee, France

I. Introduction

1. The new production system for the French structural business statistics (see [1], [2], [3] and [4] for further details), which is operational since 2009, relies on a major founding principle: the use of administrative data in an intensive way in conjunction with survey data. This combined use of administrative data – which are supposed to be complete – and survey data – which are obtained on a sample of enterprises – makes the statistical production more complex, but in compensation opens up new horizons, on one hand for data editing process – through a consistency monitoring of individual data between administrative source and survey data – and on the other hand for quality improvement of statistics – thanks to the use of statistical estimates by combining survey and administrative data.

2. The first part of this paper details the mechanism of consistency monitoring between the different sources, as well as the choice of the estimators. The second part presents a first evaluation of the impact of this methodological improvement on produced statistics.

II. Main outlines of the new system

A. Structure of the data

3. The system Esane (in French "Élaboration des Statistiques Annuelles d'Entreprises") relies on a combined use of different administrative sources and a statistical survey (figure 1).

Two administrative sources are used:

- Annual income returns of enterprises to tax authorities, containing accounting variables;
- Annual social security returns, containing information about employment and wages.

All these data are theoretically complete[1] and the record linkage is made easy with the id-number of the French business register SIRENE. The statistical unit that is used is the legal unit[2] as defined in this register.

[1] In pratice, we however have to deal with some missing data.
[2] Except for specific units that will be defined for some large groups, for which profiling techniques are used.

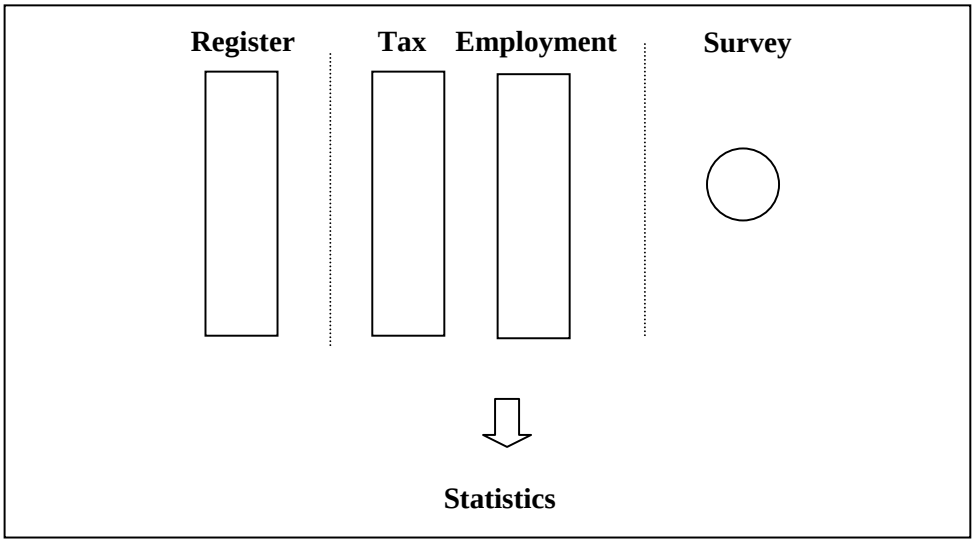
4. However, the use of administrative data alone is not sufficient to produce the required statistics, on account of the lack of some variables. Especially, one essential information, which is the basis for all sector-based statistics, is not available in the administrative sources: the breakdown of enterprise turnover. This information has two main uses. First, the national accounts need information about the "pure" economic branches turnover, which is obtained through this table. Secondly, the breakdown of the turnover is used to compute the value of the principal activity code (in French, "APE" code), referring to the French classification of activities NAF (which is derived from the European NACE). This value of the APE code is obtained through an algorithm, based on the relative share of each component of the turnover. This value is used to produce the aggregates for the economic sectors, and may differ from the value available in the register, which may have not been updated since some years.

5. In order to make up for the incompleteness of administrative sources, a statistical survey, called ESA (in French "Enquête Sectorielle Annuelle", i.e. Insee Annual Sectoral Survey on Businesses, see [4] for further details), is conducted on a sample of enterprises. This survey is the most important survey of businesses carried out in six of the main production sectors[3]. It comprises two parts for each sector:

- A core section, which includes in particular enterprise turnover and its breakdown by different activities at a very detailed level, questions on employment and legal restructuring;

- A sectoral part of the questionnaire relating to characteristics that are specific to a given sector: e.g. sales area for enterprises in the trade sector, spending of fuel for enterprises in the transport sectors, etc.

Figure 1 : The different components of the new system of structural business statistics



B. Consequences on data editing strategy: a consistency monitoring on individual data

6. Such an organisation opens up new horizons for data editing process. Indeed, in the three main sources of information for an enterprise (survey data, tax data and employment data), there are common characteristics, at least between two of these sources:

- The "turnover", "goods sales", "products sales" and "services sales" variables are present in both survey[4] and tax data;

- The "salary" and "staff size" variables are present in both tax data and employment data.

[3] Industry excluding food-processing industry, food-processing industry, transport, construction, trade and services excluding banks and insurance companies.

[4] The sales (goods, services and products) are estimated in the survey by aggregating the breakdown of turnover based on relevant activities.

When merging[5] all the data collected for a statistical unit – from the three sources for the units which responded to the survey, from tax and employment data only for the others –, we can use this redundancy of information to set up a consistency monitoring of individual records.

7. The principles of this data editing process, called REDI (for "REconciliation des Données Individuelles", i.e. individual data reconciliation), are as follows:

- For each common characteristic, a major source is defined;

Table 1: Definition of the major source for each common characteristic

Common characteristic	Condition	Major source
Turnover	We have tax data for the enterprise	Tax
	We have no tax data for the enterprise	Survey[6]
Sales	We have an answer in the survey for the year of reference	Survey
	We have no answer in the survey for the year of reference	Tax
Salary	We have tax data for the enterprise	Tax
	We have no tax data for the enterprise	Employment
Staff Size	We have employment data for the enterprise	Employment
	We have no employment data for the enterprise	Tax

- For each common characteristic, we compute the difference between the two sources. Then we calculate a score which will allow us to determine the units that the clerks are going to check on:

$$\text{score} = \left| \frac{X_{S1} - X_{S2}}{T(X_p)} \right|, \quad \text{where} \begin{cases} X_{S1} = \text{value of the characteristic X in the source 1} \\ X_{S2} = \text{value of the characteristic X in the source 2} \\ T(X_p) = \text{total of the characteristic X in the major source at the level of aggregation used for the control} \end{cases}$$

- If the score is below the threshold, the final value is the one from the major source; otherwise, the unit is manually checked[7] by a clerk, who has to determine the "good" values and also the linked characteristics. For example, if the clerk changes the goods sales, he has to verify that the goods purchases are consistent with the new values of the sales. Otherwise, he has to adjust the purchases.

8. As a result of this consistency monitoring, a new occurrence, called "Redi", is then created for each concerned[8] characteristic. For salary and staff size (and each linked characteristic), Redi variables are available for all units. For turnover and sales (and each linked characteristic), Redi variables are available only for units belonging to the sample of the survey.

C. Statistical estimates dedicated to the device

9. Thus, we may consider that we have an incomplete rectangular data base: a nearly complete data base for the administrative data, and a part available only from the survey data. Now some of these variables – mainly the principal activity code, the breakdown of turnover and the Redi variables – are cornerstone information. In order to take into account at best these information available only on the survey while using all information available in the administrative sources, we use both calibration techniques and specific estimators.

[5] Thanks to the id-number of the register SIRENE
[6] In this case, the turnover of the survey will be used to impute the tax data for the enterprise.
[7] This can lead to call back some enterprises.
[8] That is, the common characteristics, which have been checked, and linked characteristics.

10. First, having the administrative data available allows us to use calibration techniques ([5]) that lead to modifying the weights according to some calibration equations[9]. More precisely, the equations used here are:

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

where APE_rep is the value of the APE code within the register, and $T^{tax}(i)$ is the value of the turnover of the enterprise i in administrative data. The calibration on the turnover (first equation) uses a "3-digit" level for the sectoral classification, whereas the calibration on the number of enterprises (second equation) uses a "2-digit" level, in order to limit the range of changes of the weights.

11. Moreover, for sector-based estimates, the existence of two APE codes – the one of the register (APE_rep) and the one coming from the survey (APE_enq) – leads us to consider the following difference estimator:

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

This kind of estimators can be computed for any variable Y available for all units: tax and employment variables, Redi variables for salary, staff size and linked characteristics, etc. As the variables $Y_i\ \mathbb{1}_{APE\_enq=X}(i)$ and $Y_i\ \mathbb{1}_{APE\_rep=X}(i)$ are usually well correlated and even often almost identical, this difference estimator usually improves[10] the quality of sector-based estimates.

12. For the variables of the statistical survey, as well as the Redi variables relating to turnover, sales and linked characteristics, we use the Horvitz-Thompson estimator using the final weights:

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

III. First assessment of these methodological improvements

A. Impact of the REDI process

13. We present here the results of the REDI process, relating to the variables "turnover", "goods sales", "products sales" and "services sales", for the 2008 ESANE campaign.

14. First, let us note that, as this campaign was the first of the new system (see [6] for a first assessment on a more practical point of view), we have encountered many problems, which led to delaying different stages upstream of the REDI process. Consequently, telephone follow-up was carried out only a small number of enterprises and the individual data reconciliation was mainly performed in an automatic way. It results in the choice by default of the major source value for the Redi variables in the majority of cases:

- For the turnover, the value coming from tax data was retained for 97% of the units, accounting for 99% of the total turnover. The remaining 3%, for which the turnover data coming from the survey was preferred, generally corresponds to units without tax data available;

[9] This calibration step takes place after the total non-response correction in the survey, which calls on weighting methods.
[10] Compared with the Horvitz-Thompson estimator using the final weights.

- For the breakdown of the turnover between "commercial activities", "service activities" and "production of goods", the structure derived from the survey, whenever available, is chosen, unless it is less detailed than the one observed in the administrative data. Thus, the structure stemmed from the survey was retained for 81% of the units, accounting for 93% of the total turnover.

15. With regard to aggregates, for the variable "turnover", the difference between Horvitz-Thompson estimator using survey data and Horvitz-Thompson estimator using administrative data amounts to 39 billions of euros. Although important, this difference accounts for only 1.1% of the total turnover. So, for the variable "turnover", the two sources of information are globally consistent. As for the aggregated breakdown of the turnover, we observe more discrepancies between survey and tax data, as shown in Table 2. For example, the services sales represent 16.7% of the turnover of the industry with the administrative data, but only 3% with the survey data.

Table 2: Estimators for total of turnover and its breakdown between "goods sales", "products sales" and "services sales" (in million of euros) and structure of this turnover breakdown (in %)

Sector	Horvitz Thompson estimator using final weights and survey data			
	Turnover	Goods sales	Products sales	Services sales
Food-processing industry	152 776	13 765 9,0%	137 398 89,9%	1 613 1,1%
Construction	274 192	2 527 0,9%	269 340 98,2%	2 325 0,8%
Trade	1 383 321	1 291 304 93,3%	69 293 5,0%	22 698 1,6%
Industry	880 062	105 636 12,0%	747 948 85,0%	26 477 3,0%
Services	643 488	20 786 3,2%	5 801 0,9%	616 746 95,8%
Transport	174 690	2 134 1,2%	2 341 1,3%	170 215 97,4%
Total	3 508 530	1 436 153 40,9%	1 232 121 35,1%	840 074 23,9%

Sector	Horvitz Thompson estimator using final weights and tax data			
	Turnover	Goods sales	Products sales	Services sales
Food-processing industry	152 658	23 818 15,6%	121 538 79,6%	7 302 4,8%
Construction	265 675	9 453 3,6%	72 815 27,4%	183 407 69,0%
Trade	1 362 573	1 194 890 87,7%	76 388 5,6%	91 295 6,7%
Industry	883 444	131 939 14,9%	603 989 68,4%	147 515 16,7%
Services	631 147	69 724 11,0%	49 144 7,8%	512 255 81,2%
Transport	173 844	3 296 1,9%	839 0,5%	169 709 97,6%
Total	3 469 341	1 433 121 41,3%	924 713 26,7%	1 111 483 32,0%

Secteur	Horvitz Thompson estimator using final weights and Redi data			
	Turnover	Goods sales	Products sales	Services sales
Food-processing industry	153 007	15 358 10,0%	136 057 88,9%	1 592 1,0%
Construction	266 626	2 160 0,8%	261 812 98,2%	2 654 1,0%
Trade	1 360 850	1 256 966 92,4%	32 672 2,4%	71 212 5,2%
Industry	882 959	110 611 12,5%	743 771 84,2%	28 578 3,2%
Services	629 214	24 023 3,8%	3 758 0,6%	601 433 95,6%
Transport	173 590	1 893 1,1%	2 202 1,3%	169 494 97,6%
Total	3 466 247	1 411 011 40,7%	1 180 273 34,1%	874 962 25,2%

The impact of the REDI process on the aggregates is in accordance with the choice of the major source and with the data editing process detailed in § 14: for each sector, the estimator of the total for turnover using Redi data is close to the estimator using tax data, whereas the structure of its aggregated breakdown is similar to the survey's one.

16. Let us finally conclude by noting the efficiency of the REDI selective editing process detailed in § 7. Indeed, units presenting serious inconsistency between administrative data and survey data for the turnover or its aggregated breakdown, and which consequently have to be checked manually in a detailed way, account for only 2.7% of the units belonging to the sample – around 4200 enterprises –, but explain 81% of the difference in terms of turnover. For the 2008 Esane campaign, due to lack of time, only 1000 of these units were actually checked by a clerk, with only 200 enterprises called back. But starting from the 2009 ESANE campaign, this selective editing process should be fully operative and would permit to detect and correct most of the serious inconsistencies.

B. Impact of the new statistical estimates

17. In this section, we try to assess the impact of the methodological improvements implemented in the new system, namely the use of calibration techniques and specific estimators. For this purpose, we reproduced estimators as in the previous system[11], by computing Horvitz-Thompson estimators with final weights taking into account the unit non-response adjustment but not the calibration step. We then compute CVs for these estimators and the current ones. We focus on two groups of variables:

- The turnover and its aggregated breakdown – goods sales, products sales and services sales –, which are affected by the Redi process, and for which estimators used in the new system are the Horvitz-Thompson estimators described in § 12;

- The variables “number of enterprises”, “salary” and “employer's social contributions”, that are available for all units and for which estimators used in the new system are the difference estimators described in § 11;

18. The computed CVs for the mimicked estimators of the previous system take into account the sampling error of the survey, due to the stratified sample design and unit non-response adjustment in the take-some part of the survey using the RHG model. For the current estimator, we also take into account the use of calibration techniques as well as the use of the difference estimator when necessary.

19. Taking account of the first two points, we use a self-made SAS macro, which analytically computes variance under the hypothesis of stratified sample design and RHG model for unit non-response adjustment. The use of calibration techniques is taken into account by computing, thanks to our self-made SAS macro, the variance of the Horvitz-Thompson estimator for the total of the residuals derived from the weighted least squares regression of the variable of interest on calibration variables.

20. Table 3 gives the result of this comparison for sector-based estimates at the “3-digit” level of the French nomenclature:

Table 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
Q99	2,27	2,78	1,83	1,38	1,78	1,98	2,36
Q95	1,03	1,10	1,13	1,07	1,13	1,24	1,17
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
Q5	0,08	0,48	0,37	0,25	0,07	0,00	0,00
Q10	0,00	0,16	0,17	0,01	0,00	0,00	0,00
Min	0,00	0,11	0,09	0,00	0,00	0,00	0,00

So the new statistical estimators systematically lead to an average reduction of the CVs, between 6% to 37%. Moreover, we observe a depreciation of accuracy above 3% in only 10% of cases.

[11] More precisely, in the previous device coexisted two parallel systems: one using fiscal data and the other resting only on the statistical survey – which included then many questions about information existing in other sources –. We focus here on the system based on the survey, for which no administrative data was mobilized to produce the statistical estimates.

IV. Conclusion

21. After two years of functioning, the first assessment of the new production system for the French structural business statistics tends to validate the methodological choices. The comparison of the different sources of information allows a consistency monitoring on key variables which improves the quality of estimates by reducing the bias due to response errors. As for the new statistical estimators, they lead to an improvement of the accuracy in most of the cases.

References

[1] Depoutot R., 2010 : *Reengineering French structural business statistics : an overview*, paper presented at the Q2010 conference, Helsinki

[2] Brion Ph., Gros E., 2009 : *Methodological issues related to the reengineering of the of French structural business statistics*, EESW09, Stockholm

[3] Chami S., 2010 : *Reengineering French structural business statistics - an extended use of administrative data*, paper presented at the Q2010 conference, Helsinki

[4] Haag O., 2010 : *Reengineering French structural business statistics : redesign of the annual survey*, paper presented at the Q2010 conference, Helsinki

[5] Deville J.-C,. Särndal C.-E., 1992 : *Calibration estimators in survey sampling*, Journal of the American Statistical Association, 87, pp. 376-382

[6] Brion Ph., 2011 : *First elements relative to the data editing strategy used for the new system of french structural business statistics*, UN/ECE work session on statistical data editing, Ljubljana