DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 589–609, http://dx.doi.org/10.1515/JOS-2015-0036

Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics

Philippe Brion[1] *and Emmanuel Gros*[2]

Using as much administrative data as possible is a general trend among most national statistical institutes. Different kinds of administrative sources, from tax authorities or other administrative bodies, are very helpful material in the production of business statistics. However, these sources often have to be completed by information collected through statistical surveys. This article describes the way Insee has implemented such a strategy in order to produce French structural business statistics. The originality of the French procedure is that administrative and survey variables are used jointly for the same enterprises, unlike the majority of multisource systems, in which the two kinds of sources generally complement each other for different categories of units. The idea is to use, as much as possible, the richness of the administrative sources combined with the timeliness of a survey, even if the latter is conducted only on a sample of enterprises. One main issue is the classification of enterprises within the NACE nomenclature, which is a cornerstone variable in producing the breakdown of the results by industry. At a given date, two values of the corresponding code may coexist: the value of the register, not necessarily up to date, and the value resulting from the data collected via the survey, but only from a sample of enterprises. Using all this information together requires the implementation of specific statistical estimators combining some properties of the difference estimators with calibration techniques. This article presents these estimators, as well as their statistical properties, and compares them with those of other methods.

Key words: Structural business statistics; administrative data; multisources device.

1. Introduction

Using administrative data to produce official statistics is a big challenge for National Statistical Institutes (NSIs). Concerning business statistics, a lot of administrative sources are often available, and NSIs are using them more and more in an intensive way. A European ESSnet has been working on finding common ways for their use. However, an information collection carried out in 2009–2010 about existing practices among NSIs shows that various contexts do exist, especially concerning the legal basis underlying the use of administrative data, and the cooperation with administrative data holders (Costanzo 2011).

If we now consider the case of structural business statistics, the strategies of the different NSIs vary greatly, from the simple use of statistical surveys (without any use of

[1] INSEE Boulevard Adolphe Pinard F-75675 Paris Cedex 14 75675, France. Email: philippe.brion@insee.fr
[2] INSEE Department of Statistical Methodology, 18, Boulevard Adolphe Pinard, F-75675 Paris Cedex 14 75675, France. Email: emmanuel.gros@insee.fr

administrative sources) to the complete replacement of survey data with administrative sources. In between, a lot of NSIs use intermediate systems, combining administrative and survey data.

This article describes the French strategy adopted by Insee (National Institute of Statistics and Economic Studies) in order to build a new process of producing structural business statistics. As mentioned in Costanzo (2011), concerning the use of administrative data for business statistics, France is considered to have a specific model, highly centralized, due to a business register (SIRENE) that serves both administrative and statistical purposes. This model makes the use of administrative data concerning enterprises easier than in other countries, particularly due to the fact that each administration uses the same unit and the same ID number for the enterprises, which is the SIRENE ID number.

France has been using tax files to produce structural business statistics for a long time (Grandjean 1997). The richness of these files, composed of annual income statements sent by enterprises to the tax authorities, is very interesting, since the files provide detailed information about the accounting characteristics of all French businesses. However, for a long time, these files were available too late to answer certain needs, such as the supplying of preliminary results before the end of October of year $(n + 1)$ for the European Structural Business Statistics (SBS) regulation. Furthermore, they did not provide information for all kinds of needs. Thus a statistical survey, limited to a sample of enterprises, was conducted at the same time: this statistical survey was the basis for the preliminary results sent to Eurostat, as the administrative data were used for the definitive results sent later.

This double system had a significant drawback, however: the two sources sometimes told different stories, even at a highly aggregated level. Using two different sources led, obviously, to the possibility of conflicting results. Here, one of the most important reasons identified related to the classification of enterprises within the NACE nomenclature. The two sources do not obtain the same quality of information for this variable (see Subsections 2.1 and 2.2 below), tax files being mainly based on the value of the code within the register, which cannot be updated in a continuous way for all enterprises. Since the results by industry are very important for structural business statistics, the divergences of the two systems were particularly problematic.

Hence a new system of production of French Structural Business Statistics, named ESANE (as *Elaboration des Statistiques ANnuelles d'Entreprises*), has been implemented to unite the two previous systems in just one, taking advantage of each of their characteristics (Brion 2011).

The originality of this device is that within it, variables obtained in the two sources (administrative files, statistical survey) are used jointly for the same enterprises, especially for classifying them within the NACE. By contrast, in many other systems, at least in European countries (ESSnet on administrative data 2011), the two sources generally complement one another for different categories of enterprises (for example, the statistical survey being limited to large enterprises, and the administrative data used for small and medium units).

This article is mainly dedicated to the questions of statistical estimators used in the device. The next section of the article provides a quick overview of the system. The following section is dedicated to the characteristics of the estimators that have been implemented. In Section 4, some other aspects of the system are mentioned briefly.

2. The French System of Structural Business Statistics

2.1. An Intensive Use of Administrative Data Combined With a Survey

The French system is mainly based on two administrative sources, completed by a survey. It is based upon a central administrative source: the annual statements of benefits sent by enterprises to the tax authorities (Chami 2010), containing accounting variables (between 500 and 1,000 according to the size of the enterprise). It should be noted that French statistical law makes Insee's access to these files possible. This material is very rich, since it concerns every unit of the three millions of enterprises under the scope of business statistics. Of course it cannot be used directly, mainly for two reasons:

- it has to be checked, because of missing data, or of multiple declarations: hence work is done by Insee to impute missing data (Deroyon 2013) and to deal with multiple declarations,
- not all information needed to produce the structural business statistics is available in these files, and additional information has to be obtained elsewhere.

A second interesting source is composed of the annual social security returns of the enterprises to the administration, giving information about employees and wages.

Using these two sources helps lessen the statistical burden on enterprises, but some additional information has to be collected to answer some of the users' needs. This is done through a statistical survey, because the required information is not available in administrative files. One cornerstone variable in particular is obtained thanks to the survey: the detailed breakdown of the enterprise's turnover according to its different activities. This information, among others, is needed at a very detailed level for the national accounts. Since only a "rough breakdown" – between production, sales and services – of the total turnover of the enterprise is available in the tax files, one main part of the statistical survey questionnaire is dedicated to this question: enterprises are asked to fill out a table giving the value of the turnover of each industry they are performing.

Other variables are collected through the survey, concerning restructuring of enterprises, data about nonsalaries, and other specific topics related to the economic sectors (relative to professional expenses, or to other specific aspects such as, for example, the number of trucks for road transportation). This survey is limited to a sample of enterprises (Haag 2010).

2.2. The Business Register and the Classifying of Each Enterprise Using the Nomenclature of Activities

As mentioned above, the French business register, SIRENE, serves both administrative and statistical purposes. The use of its ID number is mandatory for each French administration, and this makes the use of administrative files for statistical purposes very easy. In this way, there is no problem of undercoverage of the register.

Every French enterprise has, within SIRENE, a "principal activity code" named APE (in French *Activité Principale de l'Entreprise*), classifying it within the French NAF nomenclature of activities, which is derived from the European NACE. In this article, this

value is named APE_{reg}. At the time of the creation of the enterprise, this value is coded by SIRENE clerks, according to the firm declaration.

However, this value is not necessarily updated in a continuous way for all French enterprises, especially for the numerous small ones. Some enterprises send information to modify the value of this code, but this is not the case for all of them. So directly using the value available in the register for producing statistics may raise quality-related questions: economic sectors are changing, for example, during the last years some enterprises have been moving from industry to the trade sector. The statistics that could be produced directly using the values of the code within the business register would not properly represent these changes.

Through the statistical survey, we obtain updated and rather objective information on the different activities conducted by the surveyed enterprises: each enterprise fills out a table giving a breakdown of its turnover according to the different activities it is performing, and an algorithm is then used to calculate an updated value of the APE code (using the breakdown of the turnover by activities as a proxy for the breakdown of value added of these activities, which should be, from a theoretical point of view, the basic information to classify the enterprise). This updated value, referred to in this article as APE_{survey}, may differ from the initial value of the register, and is only available for some of the enterprises, namely those that are surveyed. In the end, it is introduced into the business register, and may be used for the next drawing of samples; however, it cannot be fed back into the register and then used directly in the current survey as an auxiliary variable for statistical purposes (for example for calibration), since the partial updating of the register would lead to some bias.

2.3. An Original Kind of Database

Using administrative and survey data jointly leads, in a simplified presentation, to an incomplete rectangular data base (Figure 1). In this figure, rows represent enterprises and columns variables. The right part contains variables obtained through the administrative sources (mainly accounting variables), and the left part variables obtained through the statistical survey. This survey uses a sample stratified according to the activity and the size of the enterprises. The stratification variable used for the activity is obviously APE_{reg}, and the size is based on the number of employees. The sampling rates are different according to the size of the enterprise, and the take-all stratum (generally defined as more than 20 employees) contains the largest enterprises. The white area dominating the left part represents unobserved data (since in the sampled stratum only 85,000 enterprises are surveyed from the population of almost three million units).

This data base, where sampling weights exist for the left part only, is not easy to use, compared to an administrative data base (without sampling weights) or to a survey data base (with data limited to the sampled units, with sampling weights).

It should be noted concerning the classifying of the enterprise within the nomenclature of activities that two values may coexist in the database: the value of the register APE_{reg}, available for all three million enterprises, and the value of the updated APE_{survey}, which exists only for the units of the survey.

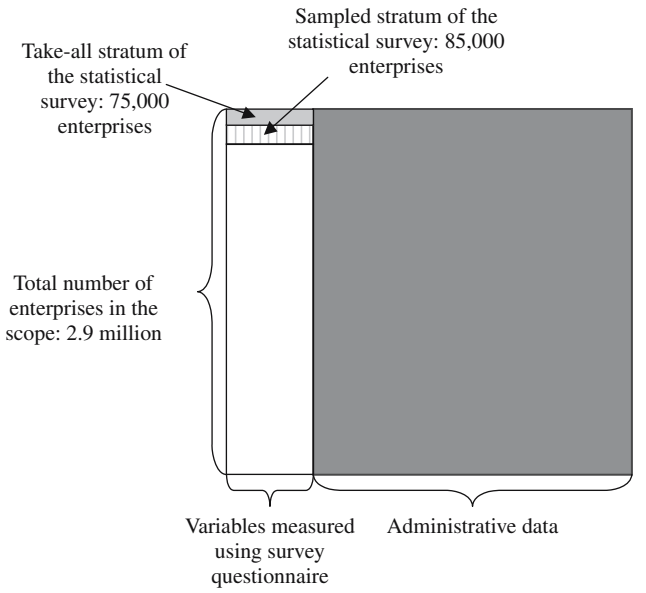

Fig. 1. ESANE, the multisources device for the French business statistics

3. The Statistical Estimators Used to Produce the Structural Business Statistics

3.1. What Kinds of Statistics Do We Want to Produce?

Structural business statistics have to give an appropriate picture of the population of enterprises, mainly concerning accounting variables (such as the turnover, the value added, the investments, etc.), but also characterizing enterprises by the industry to which they belong.

In this way, many of the produced statistics do not result from one variable only, but from a combination of two (or more) variables: a quantitative variable combined with a qualitative variable.

For example, if we consider the total turnover of an economic sector A, the quantity to estimate is:

$$\sum_{i \in U} \text{Turnover(i)} 1\text{I}_{\text{APE=A}}\text{(i)},$$

where $1\text{I}_{\text{APE=A}}\text{(i)}$ is the indicator variable relative to the classifying of enterprise *i* in industry *A* (or sector: in this article we use sometimes the wording sector, understood as economic sector, not institutional sector referring to the system of national accounts), and *U* is the global population of enterprises.

The variable "turnover" is available in the tax files, while for the activity code the survey provides fresher and richer information than the register (even if a value does exist in the register).

Other kind of statistics are produced, for example statistics based only on survey variables but the following sections of the article focus mainly on the multisource statistics presented above, since sector-based statistics are one of the main results of the device.

3.2. Different Possible Methods

The objective is to rely, as much as possible, on the exhaustiveness of administrative sources, which concern hundreds of variables. This material has to be used jointly with the information available in the statistical survey, conducted on the sample of enterprises, particularly the up-to-date activity code.

Two "families" of methods may be considered:

- mass imputation (Kovar and Withbridge 1995), taking into account the observations of the sample to generate values for the white part of the rectangle of Figure 1; in particular, it is necessary to generate an updated APE code for each enterprise of the population (that means approximately three million enterprises),
- inference using specific statistical estimates.

Methodological studies have been conducted to compare the two kinds of methods (Brion 2007, performed on past data in NACErev1. It should be noted that Kroese and Renssen (2000) present some elements on the mass imputation method that are similar to those of (Brion 2007)). More precisely, the imputation method that has been evaluated consisted of imputing an updated value of the APE code for the nonsampled units by using probabilities of moving from the economic sector in which the enterprise is classified within the register to another sector, these probabilities being estimated on the sample for categories belonging to the same "four-digit" level within the register.

The first thing to note is that the mass imputation method leads to some potential bias in the way it is proposed here. The methodological studies have quantified the value of this bias as far from negligible: more precisely, for the trade sector, composed of 119 different values of the code APE, 15 have a potential bias with the proposed imputation method that is more, in absolute value, than ten percent of the total to estimate.

Then, in order to compare the mass imputation method with other estimates, the mean square error of every method needs to be computed: for the mass imputation method, its variance needs to be evaluated and to be added to the square of the bias that has been evaluated previously. The mass imputation method is compared to a difference estimator, which is unbiased, and close to the final estimators used in ESANE that are presented in next section (Brion 2007). Results show that, for the global trade sector, the root mean square error of the difference estimator is approximately half of the root mean square error of the mass imputation estimator. A comparison at a lower level of the nomenclature (four digits of the NACE) has been found that for 13 classes mass imputation was better, as for 100 classes the difference estimator was better. For this reason, it was decided to abandon the idea of using the mass imputation method. However, the question of the different kinds of methods to use to produce official statistics remains open (see for example Little 2012 and Brion 2012a).

Then, concerning "classic" statistical estimates, two usual strategies may be considered:

1. Only using the data coming from the units in the sample (and taking into account both survey and administrative variables for these units). This is a minimum approach, because it does not exploit the exhaustiveness of the administrative sources, which is consequently unsatisfactory.

2. Using calibration techniques (Deville and Särndal 1992) to improve the efficiency of the estimators. Here, the exhaustiveness of the administrative sources is used to modify the sampling weights according to calibration equations involving some of the administrative variables. This approach, which is an extension of the general regression estimator, will lead to a better precision of estimates for variables linked to the calibration variables.

In theory, this strategy allows us to take into account all information available in the administrative sources by computing a calibration estimator which takes into account all administrative variables. However, the huge number of fiscal variables (over 500) makes this approach totally impracticable: the calibration procedure (if it converges, which is clearly not guaranteed: indeed, with so many variables, even the regression estimator, which is a special case of calibration estimator, can be incomputable, due to colinearity problems for example) will lead to some negative and/or overly extreme weights, which will induce unrealistic sector-based estimates for some economic sectors, especially at a detailed level.

To avoid these problems, another strategy could be to take into account the information available in the administrative sources "variable by variable" and "sector by sector", by computing a simple regression estimator for each fiscal variable and each sector.

For a fiscal variable Z and a sector A, such a sector-based estimator would be:

$$\hat{Z}_{reg}^{A} = \sum_{s} \frac{Z_i 1I_{APEsurvey=A}(i)}{\pi_i} + \hat{\beta}_{Z,A} \left[\sum_{U} Z_i 1I_{APEreg=A}(i) - \sum_{s} \frac{Z_i 1I_{APEreg=A}(i)}{\pi_i} \right]$$

$$= \hat{Y}_{\pi} + \hat{\beta}_{Z,A} \left[X - \hat{X}_{\pi} \right] \quad (1)$$

with $Y_i = Z_i 1I_{APEsurvey=A}(i)$ and $X_i = Z_i 1I_{APEreg=A}(i)$;

π_i is the inclusion probability of unit i;

$1I_{APEreg=A}(i)$ is the indicator variable using the value of the APE code within the register;

$1I_{APEsurvey=A}(i)$ is the indicator variable using the value of the APE code obtained through the statistical survey;

$\hat{\beta}_{Z,A}$ is the coefficient of the simple regression of Y on X (the subscript Z,A reminds us that this coefficient depends at the same time on the fiscal variable Z and on the sector A).

It should also be noted that this regression estimator can also be formulated as a weighting estimator with weights $w_i^{Z,A}$ that are different for each fiscal variable and each sector. This differs from the "classical" calibration estimator that leads to a single weight for each sampling unit regardless of the fiscal variable or sector.

Such an estimator allows us to produce sector-based estimates for all fiscal variables and all "sectoral" levels by systematically taking into account the exhaustiveness of the administrative sources. Unfortunately, such an approach is not appropriate to the context

of ESANE. And consequently, these regression estimators are not used in the final system. Indeed, statistics produced in the ESANE device are subject to many consistency constraints, both "vertical" – consistency between estimations concerning different levels of hierarchically nested nomenclature – and "horizontal" – consistency between estimations relating to variables linked by accounting relationships – that the estimation method has to respect. However, the approach detailed above is not linear, because the $\hat{\beta}_{Z,A}$ coefficients – or the weights $w_i^{Z,A}$ if the estimator is formulated as a weighting estimator – change with the fiscal variable Z and the sector A, and consequently this approach does not ensure that the consistency constraints in the ESANE method would be respected.

Let us take, for example, three fiscal variables U,V and W linked by the accounting relationship $W = U + V$ and a group G of the NACE Rev.2 divided into two classes G1 and G2. We can compute the three sector-based estimators according to Formula (1):

$$\hat{U}_{reg}^{G} = \sum_{s} \frac{U_i 1I_{Group\_survey=G}(i)}{\pi_i} + \hat{\beta}_{U,G}\left[\sum_{U} U_i 1I_{group\_reg=G}(i) - \sum_{s} \frac{U_i 1I_{group\_reg=G}(i)}{\pi_i}\right]$$

$$\hat{V}_{reg}^{G} = \sum_{s} \frac{V_i 1I_{Group\_survey=G}(i)}{\pi_i} + \hat{\beta}_{V,G}\left[\sum_{U} V_i 1I_{group\_reg=G}(i) - \sum_{s} \frac{V_i 1I_{group\_reg=G}(i)}{\pi_i}\right]$$

$$\hat{W}_{reg}^{G} = \sum_{s} \frac{W_i 1I_{Group\_survey=G}(i)}{\pi_i} + \hat{\beta}_{W,G}\left[\sum_{U} W_i 1I_{group\_reg=G}(i) - \sum_{s} \frac{W_i 1I_{group\_reg=G}(i)}{\pi_i}\right]$$

But as $\hat{\beta}_{U,G} \neq \hat{\beta}_{V,G} \neq \hat{\beta}_{W,G}$, $\hat{W}_{reg}^{G}$ is not equal to $\hat{U}_{reg}^{G} + \hat{V}_{reg}^{G}$, even if we have $W_i = U_i + V_i$ for each unit i.

And in the same way, we can compute for variable U the sector-based estimator for sectors G, G1 and G2:

$$\hat{U}_{reg}^{G} = \sum_{s} \frac{U_i 1I_{Group\_survey=G}(i)}{\pi_i} + \hat{\beta}_{U,G}\left[\sum_{U} U_i 1I_{group\_reg=G}(i) - \sum_{s} \frac{U_i 1I_{group\_reg=G}(i)}{\pi_i}\right]$$

$$\hat{U}_{reg}^{G1} = \sum_{s} \frac{U_i 1I_{Class\_survey=G1}(i)}{\pi_i} + \hat{\beta}_{U,G1}\left[\sum_{U} U_i 1I_{Class\_reg=G1}(i) - \sum_{s} \frac{U_i 1I_{Class\_reg1=G2}(i)}{\pi_i}\right]$$

$$\hat{U}_{reg}^{G2} = \sum_{s} \frac{U_i 1I_{Class\_survey=G2}(i)}{\pi_i} + \hat{\beta}_{U,G2}\left[\sum_{U} U_i 1I_{Class\_reg=G2}(i) - \sum_{s} \frac{U_i 1I_{Class\_reg=G2}(i)}{\pi_i}\right]$$

The same causes produce the same effects, as $\hat{\beta}_{U,G} \neq \hat{\beta}_{U,G1} \neq \hat{\beta}_{U,G2}$, $\hat{U}_{reg}^{G}$ is not equal to $\hat{U}_{reg}^{G1} + \hat{U}_{reg}^{G2}$.

Another strategy is hence proposed: using combined statistical estimates mixing the principles of the difference estimators (Särndal et al. 1992) and the calibration techniques. This third option is detailed in the next section of the article.

3.3. The Statistical Estimators for Sector-Based Estimates at the Group (and Upper) Level

The idea is to start from the standard Horvitz-Thompson estimator and to use the exhaustiveness of the administrative sources to improve its efficiency as much as possible while keeping to all the consistency constraints of the ESANE device. In practice, as we have to deal with unit nonresponse, the "starting point" is in fact not the Horvitz-Thompson estimator but the reweighted-expansion estimator, with weights adjusted for unit nonresponse thanks to the response homogeneity groups method RHG.

First, as the turnover is a core variable – highly correlated with both turnover breakdown and the main accounting variables of the device such as value added –, we can use calibration techniques to modify the RHG-adjusted weights according to calibration equations involving turnover by sector. More precisely, the equations used here are:

$$\begin{cases} \sum_{i\in R} w_i T(i) 1I_{APEreg=A}(i) = \sum_{i\in U} T(i) 1I_{APEreg=A}(i) \\ \sum_{i\in R} w_i 1I_{APEreg=A}(i) = \sum_{i\in U} 1I_{APEreg=A}(i) \end{cases}$$

where:

- w_i is the calibrated weight of each enterprise i of the sample of respondents *R*,
- $1I_{APEreg=A}(i)$ is the indicator variable using the value of the APE code within the register,
- $T(i)$ is the value of the turnover of enterprise i in the tax files.

That is, we perform calibration on the total turnover and the number of enterprises by sector for each sector A of the ESANE device. In practice, this calibration is generally performed at the "3-digits" level of the sectoral classification, in order to limit the range of changes of the weights.

The calibration on the sectoral total of turnover permits us to improve the accuracy of sector-based estimates for all variables correlated with the turnover, while the calibration on the number of enterprises by sector aims to avoid too much distortion concerning the estimation of numbers of enterprises by sectors.

This calibration estimator thus incorporates all information available in the tax sources for the turnover variable, but, as previously stated, it does not allow the exhaustiveness of the administrative sources to be taken into account for other variables. In order to compensate for this drawback, we can use the principle of difference estimation and consider the following "combined estimator" for sector-based estimates relating to any administrative variable *Z*, such as turnover, value added, investments and so on:

$$\hat{Z}^A_{diff} = \sum_{i\in R} w_i Z_i 1I_{APEsurvey=A}(i) + \sum_{i\in U} Z_i 1I_{APEreg=A}(i) - \sum_{i\in R} w_i Z_i 1I_{APEreg=A}(i) \qquad (2)$$

This estimator, based on the existence of two APE codes – the one of the register (APE_{reg}), available for all units, and the one derived from the survey (APE_{survey}), known only for the sample –, allows us to use all information available in the administrative

sources for the variable Z while keeping to all the linear consistency constraints of the ESANE device because of its linearity.

Indeed, if we consider again the example of three fiscal variables U,V and W linked by the accounting relationship $W = U + V$ and a group G of the NACE Rev.2 divided into two classes G1 and G2, we can compute the three sector-based estimators according to Formula (2):

$$\hat{U}^{G}_{diff} = \sum_{i\in R} w_i U_i 1I_{Group\_survey=G}(i) + \sum_{i\in U} U_i 1I_{Group\_reg=G}(i) - \sum_{i\in R} w_i U_i 1I_{Group\_reg=G}(i)$$

$$\hat{V}^{G}_{diff} = \sum_{i\in R} w_i V_i 1I_{Group\_survey=G}(i) + \sum_{i\in U} V_i 1I_{Group\_reg=G}(i) - \sum_{i\in R} w_i V_i 1I_{Group\_reg=G}(i)$$

$$\hat{W}^{G}_{diff} = \sum_{i\in R} w_i W_i 1I_{Group\_survey=G}(i) + \sum_{i\in U} W_i 1I_{Group\_reg=G}(i) - \sum_{i\in R} w_i W_i 1I_{Group\_reg=G}(i)$$

and we have:

$$\begin{aligned}
\hat{U}^{G}_{diff} + \hat{V}^{G}_{diff} &= \sum_{i\in R} w_i U_i 1I_{Group_survey=G}(i) + \sum_{i\in R} w_i V_i 1I_{Group_survey=G}(i) \\
&\quad + \sum_{i\in U} U_i 1I_{Group_reg=G}(i) + \sum_{i\in U} V_i 1I_{Group_reg=G}(i) \\
&\quad - \left[\sum_{i\in R} w_i U_i 1I_{Group_reg=G}(i) + \sum_{i\in R} w_i V_i 1I_{Group_reg=G}(i) \right] \\
&= \sum_{i\in R} w_i \underbrace{(U_i + V_i)}_{W_i} 1I_{Group_survey=G}(i) + \sum_{i\in U} \underbrace{(U_i + V_i)}_{W_i} 1I_{Group_reg=G}(i) \\
&\quad - \sum_{i\in R} w_i \underbrace{(U_i + V_i)}_{W_i} 1I_{Group_reg=G}(i) \\
&= \hat{W}^{G}_{diff}
\end{aligned}$$

In the same way, for variable U we can compute the sector-based estimator for Sector G, G1 and G2:

$$\hat{U}^{G}_{diff} = \sum_{i\in R} w_i U_i 1I_{Group\_survey=G}(i) + \sum_{i\in U} U_i 1I_{Group\_reg=G}(i) - \sum_{i\in R} w_i U_i 1I_{Group\_reg=G}(i)$$

$$\hat{U}^{G1}_{diff} = \sum_{i\in R} w_i U_i 1I_{Class\_survey=G1}(i) + \sum_{i\in U} U_i 1I_{Class\_reg=G1}(i) - \sum_{i\in R} w_i U_i 1I_{Class\_reg=G1}(i)$$

$$\hat{U}^{G2}_{diff} = \sum_{i\in R} w_i U_i 1I_{Class\_survey=G2}(i) + \sum_{i\in U} U_i 1I_{Class\_reg=G2}(i) - \sum_{i\in R} w_i U_i 1I_{Class\_reg=G2}(i)$$

and we have:

$$\hat{U}_{\text{diff}}^{G1} + \hat{U}_{\text{diff}}^{G2} = \sum_{i \in R} w_i U_i \underbrace{\left(1I_{\text{Class\_survey=G1}}(i) + 1I_{\text{Class\_survey=G2}}(i)\right)}_{1I_{\text{Group\_survey=G}}(i)} + \sum_{i \in U} U_i \underbrace{\left(1I_{\text{Class\_reg=G1}}(i) + 1I_{\text{Class\_reg=G2}}(i)\right)}_{1I_{\text{Group\_reg=G}}(i)} - \sum_{i \in R} w_i U_i \underbrace{\left(1I_{\text{Class\_reg=G1}}(i) + 1I_{\text{Class\_reg=G2}}(i)\right)}_{1I_{\text{Group\_reg=G}}(i)} = \hat{U}_{\text{diff}}^{G}$$

Moreover, as the variables $Z_i 1I_{\text{APEsurvey=A}}(i)$ and $Z_i 1I_{\text{APEreg=A}}(i)$ are usually well correlated and indeed often almost identical, this difference estimator is particularly appropriate to the ESANE device, and generally permits us to improve the quality of sector-based estimates.

It should be noted that the principle of difference estimation is used here in an unconventional way: indeed, in the conventional difference estimator (Särndal et al. 1992), the same set of auxiliary variables is used to perform estimation for all variables; conversely, in our combined estimator, the auxiliary variable $Z_i 1I_{\text{APEreg=A}}(i)$ depends at the same time on the administrative variable Z and on the sector A and is consequently suited to the considered sector-based estimation.

Let us finally conclude with two comments on the relevance and the impact of calibration in our combined estimator. First, with calibrated weights, the combined estimators coincide with the calibrated estimators at the level of the nomenclature used for the calibration equations for the sector-based estimates relating to variables "turnover" and "number of enterprises". This gives coherence between statistics based on the administrative variables and estimates based on variables available only in the survey – obtained with the calibrated estimator. Finally, the use of calibrated weights in the combined estimator leads to improvements in the accuracy of sector-based estimates when $Z_i 1I_{\text{APEsurvey=A}}(i) - Z_i 1I_{\text{APEreg=A}}(i)$ is correlated with $T_i 1I_{\text{APEreg=A}}(i)$ or $1I_{\text{APEreg=A}}(i)$.

3.4. A Quantitative Comparison of the Different Methods

In this section, we assess the impact of the methodological improvements implemented in the new system, namely the combined use of calibration techniques and difference estimators, for sector-based estimates at the "three-digit" (and above) level of the NACE Rev.2 classification. For this purpose, we consider the three following estimators:

- the reweighted-expansion estimator, with weights adjusted for unit nonresponse thanks to the response homogeneity groups method (named RHG),
- the calibrated estimator stemming from the calibration step performed in the ESANE device, which is equivalent to the GREG estimator using as auxiliary information the total turnover and the number of enterprises by sector, for each sector at the three-digit level of the sectoral classification (named GREG),
- and the combined estimator described in the previous section (named Esane).

Let us first note that, under the RHG model, these three estimators are unbiased – the reweighted-expansion estimator – or asymptotically unbiased – the GREG estimator and the Esane estimator. Consequently, we focus here on comparing the accuracy of these three estimators, measured by their coefficient of variation (CV).

To compute the coefficients of variation relating to the reweighted-expansion estimator, we use a self-made SAS macro which analytically computes variance, taking into account the stratified sampling design of the survey and the unit nonresponse adjustment using the RHG model.

The coefficients of variation relating to the GREG estimator are obtained by computing the variance of the reweighted-expansion estimator for the total of the residuals derived from the weighted least squares regression of the variable of interest $Y_i 1I_{APEsurvey=A}(i)$ on calibration variables.

Finally, the coefficients of variation relating to the Esane estimator are obtained by computing the variance of the reweighted-expansion estimator for the total of the residuals derived from the weighted least squares regression of the variable of interest $Y_i 1I_{APEsurvey=A}(i) - Y_i 1I_{APEreg=A}(i)$ on calibration variables.

We focus on a small group of core variables of the ESANE device: number of enterprises, turnover, salary, value added, gross operating profit, total assets, total liabilities and gross investments in tangible goods. Table 1 gives the result of this comparison for the six main production sectors covered by the ESANE device.

These results show that, at a global level, the Esane estimator gives better results than the two other estimators. At a more detailed level, the Esane estimator improvement performs better, as shown in Figures 2 and 3. These figures compare the different possible strategies for all variables and main sectors. They show that GREG performs better than RHG, and ESANE generally better than GREG.

But the improvement differs, obviously, depending on the relationship between the studied variable and the variables involved in the calibration procedure, especially with turnover, as Figures 4 and 5 show: for the variable "value added", the calibration step leads to an improvement of the estimators' accuracy, since the value added is positively correlated with the turnover; the "difference estimation" step leads to an another improvement of the estimators' accuracy, of the same order of magnitude as that of the calibration step.

Conversely, for the variable "gross investments in tangible goods", the improvement of the combined estimator is much more important. This is due to the richness of the tax file, which is used in the combined estimator, thanks to the principle of difference estimation, but not in the other methods – since, in the ESANE device, only the turnover and the number of enterprises by sector is used in the calibration equations. The link between the turnover and the investments is relatively weak, compared to the link between the value added and the turnover.

This first global assessment of an improvement of estimates' accuracy due to the combined use of calibration techniques and difference estimators to produce the sector-based estimates in the ESANE device is confirmed by the comparison of sector-based estimates' CV at the three-digit level of the French nomenclature, presented in Figure 6 (Table 2 in the Appendix gives the means and quintiles corresponding to these box plots). Indeed, the new statistical estimators generally lead to an average reduction of the CV, and

Table 1. Comparison between RHG, GREG and Esane estimators' coefficients of variation (CV) (using Esane's 2010 data)

	Estimators's CVs relating to RHG estimators							
Sector	Number of enterprises	Turnover	Salary	Value added	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Food-processing industry	3.9%	0.5%	1.1%	0.8%	0.7%	0.5%	0.6%	2.3%
Construction	0.9%	1.1%	0.9%	1.4%	5.5%	3.0%	3.2%	10.4%
Trade	1.1%	0.4%	0.5%	0.6%	1.8%	0.5%	0.5%	1.9%
Industry	1.3%	0.1%	0.2%	0.1%	0.4%	0.1%	0.1%	0.9%
Services	0.5%	0.4%	0.4%	0.5%	1.3%	0.9%	1.0%	4.0%
Transport	2.1%	0.4%	0.5%	0.5%	0.9%	1.2%	1.6%	8.6%
Total	0.40%	0.20%	0.21%	0.26%	0.77%	0.45%	0.53%	2.14%
	Estimators's CVs relating to GREG estimators							
Sector	Number of enterprises	Turnover	Salary	Value added	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Food-processing industry	3.2%	0.3%	0.6%	0.5%	0.6%	0.3%	0.4%	2.2%
Construction	0.3%	0.5%	0.7%	1.2%	5.6%	3.0%	3.2%	10.7%
Trade	0.5%	0.1%	0.3%	0.4%	1.7%	0.4%	0.4%	1.9%
Industry	1.3%	0.1%	0.1%	0.1%	0.3%	0.1%	0.1%	0.8%
Services	0.3%	0.2%	0.4%	0.4%	1.2%	0.8%	1.0%	4.1%
Transport	0.6%	0.2%	0.3%	0.3%	0.7%	0.5%	0.6%	5.0%
Total	0.09%	0.05%	0.16%	0.20%	0.72%	0.42%	0.49%	1.95%
	Estimators's CVs in the ESANE device							
Sector	Number of enterprises	Turnover	Salary	Value added	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Food-processing industry	3.2%	0.3%	0.4%	0.4%	0.4%	0.3%	0.4%	1.0%
Construction	0.3%	0.5%	0.2%	1.2%	6.2%	1.8%	1.8%	3.4%
Trade	0.5%	0.1%	0.2%	0.4%	1.6%	0.6%	0.5%	1.3%
Industry	1.3%	0.1%	0.1%	0.1%	0.3%	0.1%	0.1%	0.1%
Services	0.3%	0.2%	0.1%	0.3%	0.9%	0.4%	0.5%	0.5%
Transport	0.6%	0.2%	0.2%	0.1%	0.3%	0.2%	0.3%	0.3%
Total	0.09%	0.05%	0.01%	0.15%	0.56%	0.17%	0.22%	0.07%

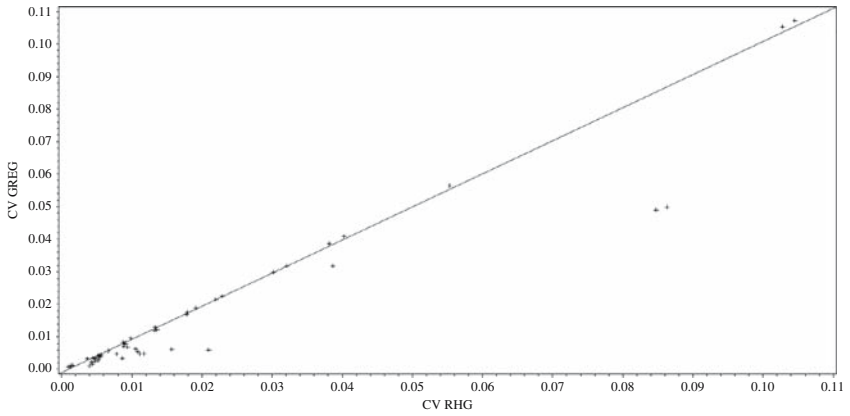

Fig. 2. Comparison of RHG and GREG coefficients of variation, estimations relating to the six main production sectors and the eight variables presented in Table 1

improve the accuracy of estimators in more than 80% of cases. Conversely, for the remaining 20%, the RHG estimator performs better than the ESANE one.

3.5. The Statistical Estimators for Sector-Based Estimates at Finer Levels

As indicated in the previous section, the implemented methods use the richness of the whole administrative data, and correct the problems of misclassifying some units within the registers.

However, these combined estimators have also some limits: particularly, they do not guarantee to always produce positive values, and can consequently lead to negative estimates even if all individual data for the variable of interest are positive. This proves

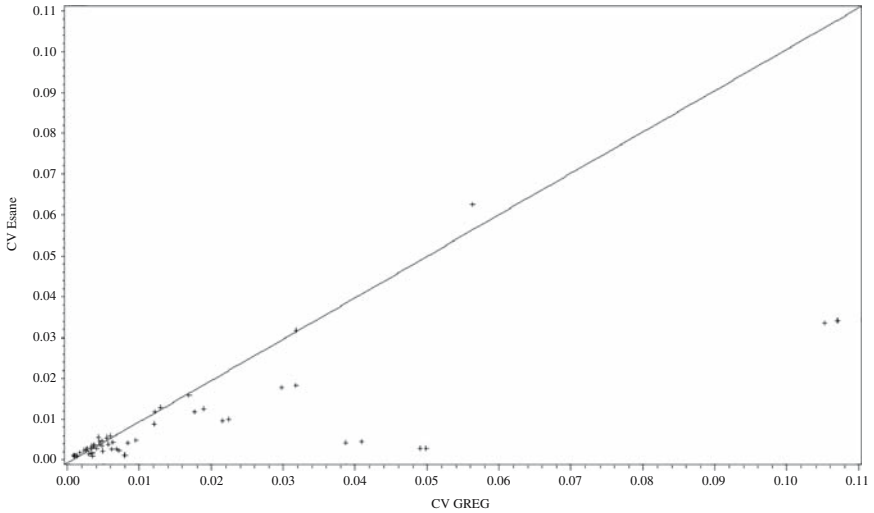

Fig. 3. Comparison of GREG and Esane coefficients of variation, estimations relating to the six main production sectors and the eight variables presented in Table 1

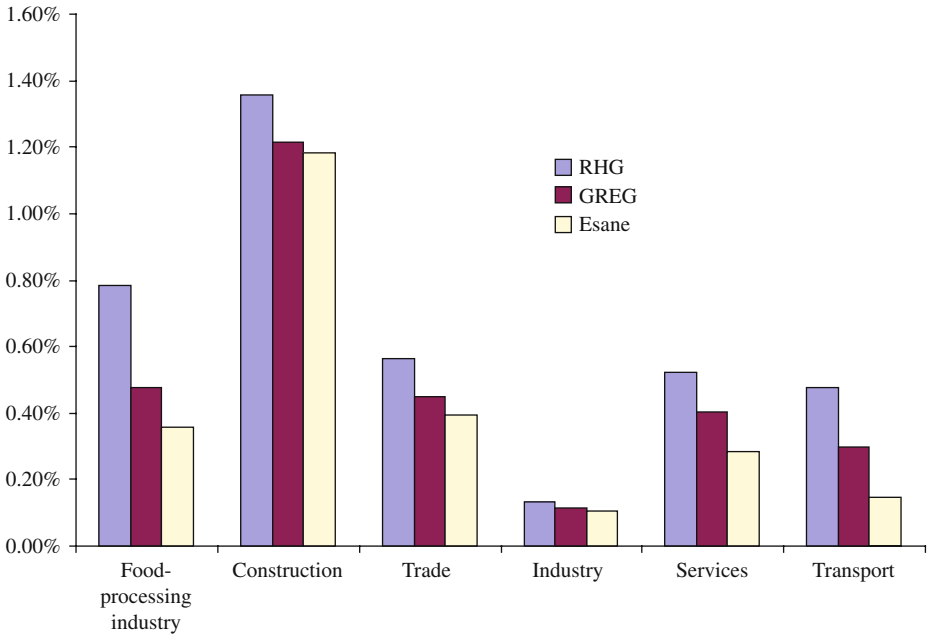

Fig. 4. Coefficients of variation of the three estimators (RHG, GREG, ESANE) for the estimation of the total of the value added by main economic sector

problematic, especially when it concerns variables for which negative aggregates make no economic sense, like turnover or salary.

In practice, this kind of problematic situation appears only when the estimation is relating to too small a domain, either because very few enterprises are concerned by the

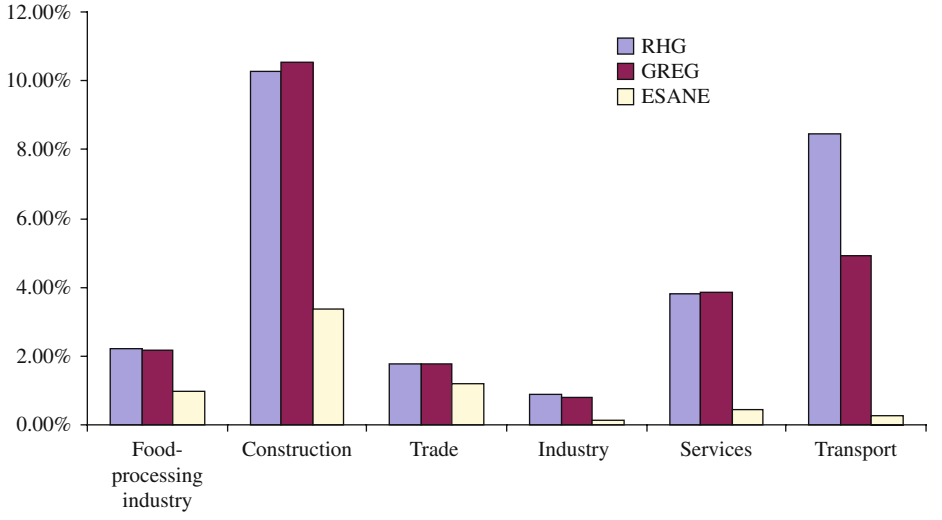

Fig. 5. Coefficients of variation of the three estimators (RHG, GREG, ESANE) for the estimation of the total of the gross investments in tangible goods by main economic sector

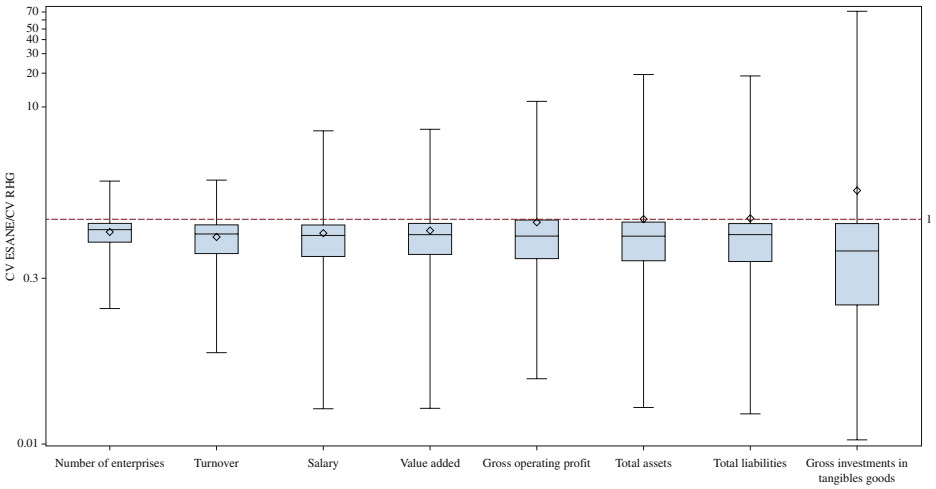

Fig. 6. Box plots of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the "group" (three-digit) level (using Esane's 2010 data)
Note: for a given variable, the diamond refers to the mean of the ratios.

variable of interest (like the variable "Sumptuary costs and expenses"), or mostly because the estimation is performed at fine levels of industry disaggregation. Indeed, as the industry disaggregation becomes finer, the amount of misclassification becomes larger, and simultaneously, the sample size available in finer-level cells to estimate this misclassification becomes smaller. Under these conditions, the difference estimators are not robust, and the change of the APE code of a single enterprise with a large value of one variable and/or a big sampling weight may create problems in the above formula, leading to negative values.

From a theoretical perspective, these negative estimates are not really problematic. Indeed, they merely reveal direct estimates' lack of precision when domain sample sizes are too small, a problem that would not necessarily appear so obviously when using classical methods: for such small domains, the RHG or calibrated estimators would have a very large variance, and when using administrative data directly with approximate values of the APE code coming from the register – available for all units but not necessarily up-to-date –, we would have a large bias.

However, these negative estimates constitute a practical drawback for the production of results at fine levels of industry disaggregation. To avoid being faced with a lot of potentially negative estimates for small domains, it has been decided to adjust the strategy concerning the estimators:

- For sector-based estimations at the "group" level (three digits of the NACE Rev.2 classification) and higher levels, the difference estimator presented in 3.3 is used. Indeed, at these relatively highly aggregated levels, we have very few "wrongly" negative estimates – less than 0,1% of all the group estimates – and they concern only variables of minor interest, such as the "Sumptuary costs and expenses". So we can deal with this problem by not publishing these rare negative estimates.

- For sector-based estimations at more detailed levels, we differentiate the "elementary" variables – that is, variables which are only components and never the result of accounting relationships – from the other variables:
  - For a given elementary variable Y, the group-level estimate is prorated to a finer level according to the structure of the elementary variable stemming from the survey. More precisely, for a group G and a finer area D $\subset$ G, the total of Y on the area D is estimated by:

$$\hat{Y}^{D}_{\text{prorated}} = \hat{Y}^{G}_{\text{diff}} \frac{\sum_{i \in R} w_i Y_i 1I_{\text{area\_survey=D}}(i)}{\sum_{i \in R} w_i Y_i 1I_{\text{Group\_survey=G}}(i)}$$

  - For the other variables, the estimates result from the accounting relationships applied to the appropriate elementary variables estimates (see Gros 2012a for more details).

By construction, such a strategy ensures both positive estimates and consistency between the different estimates in the ESANE device, and these "prorated estimators" remain asymptotically unbiased. On the other hand, they use the administrative data less intensively at an individual level than the difference estimators, so we can expect more mixed performances in terms of accuracy. This expectation is confirmed by the comparison of sector-based estimators' CV at the five-digit level of the French nomenclature, presented in Figure 7 (Table 3 in the Appendix gives the means and quintiles corresponding to these box plots).

As we can see, at this fine level of industry disaggregation, the prorated estimators indeed lead to mixed results in terms of accuracy: they perform better than the RHG estimators only half of the time. In fact, neither of the two estimators is statistically better than the other, but the prorated estimator has the advantage of preserving the consistency of group-level estimates and finer-level estimates.

4. Other Issues

The new system was implemented in 2009, and at the present time has produced results for five years. Besides the questions of estimators that have been presented above, some other issues were raised.

First, the data editing of this composite material is complex. It has been divided into subprocesses, each one dedicated to one source (administrative or survey): this choice was made mainly to keep some flexibility in case of changes in one source, for example if the content of the tax files is modified. Moreover, the calendar of the deliveries of the different files is not the same: the return of the statistical survey questionnaires is spread over a long period, between March and October $n + 1$ concerning data of year n, while for tax data there are only a few deliveries, each one containing a large number of enterprises. Then, a step comparing the survey and the administrative data helps to achieve a cross validation of each source. More precisely, the value of the turnover, as its "rough breakdown" (between production, sales and services), is available in the two sources

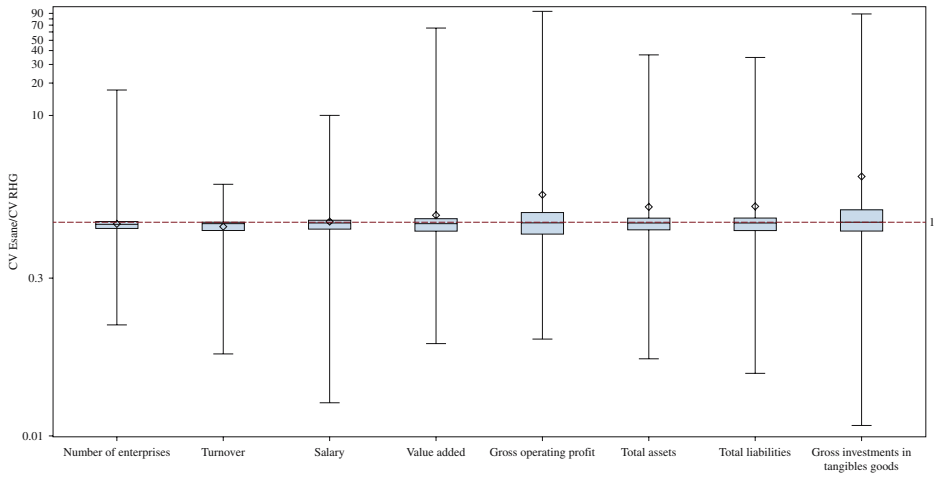

Fig. 7. Box plots of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the lower-class (5-digit) level (using Esane's 2010 data). Note: for a given variable, the diamond refers to the mean of the ratios

(survey and tax files), and the most important differences have to be checked by the clerks. This step is a very innovative part of the new system (Gros 2012b).

Questions were also raised concerning the scope of the business statistics. Using administrative and survey data jointly helped to revisit the choices made to define this scope. The scope is based on criteria available in the business register, such as the APE code and the legal status of the enterprise. Observing how the records of the tax files behaved relatively to the scope defined *a priori* helped to define choices concerning some specific categories of enterprises more precisely (Brion 2012b).

Mainly, the questions raised came back to the definition of the enterprise. In the European definition, an enterprise is the "smallest combination of legal units, that is, an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision making, especially for the allocation of its current resources". At the present moment, using a device mainly based on the legal units shows some limitations, and Insee is working to take the concept of enterprise into account better in the device: a second step concerning the renewing of the structural business statistics will consist in integrating these aspects, and some studies have shown that it will have general consequences for the significance of the statistics (Béguin et al. 2012). What is presented here concerns only the national part of the enterprise (sometimes named truncated enterprise) in the case of a multinational enterprise.

To conclude, we think that combining administrative and survey data leads to a strengthening of the quality of the produced statistics through the mutual improvement of the two kinds of sources. Moreover, in the presented device, the combined statistical estimators are intended to use every kind of information as much as possible. They show better statistical characteristics than other estimators, but in some cases this may go hand in hand with more complexity than in the case of the use of a single source.

Appendix

Table 2. Means and quintiles of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the "group" (three-digit) level (using Esane's 2010 data)

	Number of enterprises	Turnover	Salary	Added value	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Mean	0.78	0.68	0.65	0.78	0.94	1.00	0.98	1.70
Max	2.20	2.26	3.36	6.34	12.37	19.61	18.84	71.30
Q99	1.89	1.97	3.04	5.13	7.81	13.92	18.06	48.11
Q95	1.02	1.02	1.37	1.80	1.99	3.04	3.10	5.10
Q90	0.99	0.97	1.07	1.12	1.45	1.46	1.36	1.88
Q75	0.93	0.89	0.89	0.92	0.97	0.94	0.92	0.92
Median	0.81	0.73	0.68	0.72	0.78	0.71	0.72	0.52
Q25	0.63	0.47	0.41	0.46	0.45	0.43	0.41	0.17
Q10	0.46	0.24	0.20	0.19	0.20	0.17	0.12	0.05
Q5	0.37	0.11	0.13	0.08	0.03	0.08	0.05	0.02
Q1	0.17	0.00	0.03	0.00	0.00	0.02	0.00	0.01
Min	0.16	0.00	0.02	0.00	0.00	0.00	0.00	0.00

Table 3. Means and quintiles of the ratio between sector-based Esane estimators' CVs and CVs relating to sector-based RHG estimators at the "under-class" (five-digit) level (using Esane's 2010 data)

	Number of enterprises	Turnover	Salary	Added value	Gross operating profit	Total assets	Total liabilities	Gross investments in tangible goods
Mean	0.96	0.91	1.00	1.16	1.79	1.38	1.41	2.64
Max	17.34	2.26	10.01	65.83	93.75	37.01	34.79	88.32
Q99	1.50	1.72	3.56	4.43	19.09	14.10	17.32	41.70
Q95	1.11	1.14	1.50	1.60	3.57	2.82	2.96	7.74
Q90	1.06	1.06	1.15	1.35	2.32	1.70	1.72	3.40
Q75	1.01	1.00	1.04	1.07	1.23	1.09	1.10	1.31
Median	0.96	0.96	0.98	0.97	0.99	0.99	0.98	1.00
Q25	0.87	0.83	0.86	0.83	0.77	0.85	0.83	0.82
Q10	0.68	0.62	0.63	0.61	0.54	0.59	0.58	0.47
Q5	0.59	0.46	0.46	0.50	0.36	0.39	0.33	0.20
Q1	0.33	0.25	0.12	0.16	0.16	0.14	0.10	0.02
Min	0.11	0.06	0.02	0.07	0.08	0.05	0.04	0.01

5. References

Béguin, J.M., V. Hecquet, and J. Lemasson. 2012. "France's Economic Fabric More Concentrated Than it Seemed. New Definition and New Categories of Enterprises." *Insee-première*, 1399. Insee, Paris. Available at http://www.insee.fr/en/ffc/ipweb/ip1399/ip1399.pdf (latest access October 2015).

Brion, P. 2007. "Redesigning the French Structural Business Statistics, Using More Administrative Data." In Proceedings of the Third International Conference on Establishment Surveys, June 18–21, 2007, Montreal, Canada. 533–541. Alexandria, VA [CD-Rom]: American Statistical Association. Available at: https://www.amstat.org/meetings/ices/2007/Proceedings/ICES2007-000034.pdf (accessed October 2015).

Brion, P. 2011. "Esane, Le Dispositif Rénové de Production des Statistiques Structurelles D'entreprises." *Courrier des Statistiques* n°130, Insee, Paris. Available at http://www.insee.fr/fr/ffc/docs_ffc/cs130d.pdf (accessed October 2015).

Brion, P. 2012a. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28: 341–347.

Brion, P. 2012b. "The New French System of Production of Structural Business Statistics." In Proceedings of the Fourth International Conference on Establishment Surveys, June 2012, Montreal, Canada. Available at: http://www.amstat.org/meetings/ices/2012/papers/302161.pdf (accessed October 2015).

Chami, S. 2010. "Reengineering French Structural Business Statistics: an Extended Use of Administrative Data." In Proceedings of the Q2010 Conference, May 4–6, 2010, Helsinki. Available at: https://q2010.stat.fi/sessions/session-27 (accessed October 2015)

Costanzo, L. 2011. "An Overview of the Use of Administrative Data for Business Statistics in Europe." ESSnet Admin Data, workpackage 1, Eurostat. Available at: http://essnet.admindata.eu/Document/GetFile?objectId = 5358 (accessed October 2015)

Deroyon, T. 2013. "Missing Data Treatment in Administrative Fiscal Sources for the French Structural Business Statistics Production System." In Proceedings of the Third European Establishment Statistics Workshop, September 9–11, 2013, Nuremberg. Available at: http://enbes.wikispaces.com/file/view/Deroyon%202013.pdf/456103752/Deroyon%202013.pdf (accessed October 2015)

Deville, J.C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382. Doi: http://dx.doi.org/10.2307/2290268.

ESSnet on Administrative Data. 2011. "Main Findings of the Information Collection on the Use of Administrative Data for Business Statistics in EU and EFTA Countries." Deliverable 1.1, Eurostat. Available at: http://essnet.admindata.eu/WorkPackage/ShowAllDocuments?objectid=4251 (accessed October 2015)

Grandjean, J.P. 1997. "The System of Enterprise Statistics." *Courrier des Statistiques*. English series n°3, Insee, Paris. Available at: http://www.epsilon.insee.fr/jspui/bitstream/1/14403/1/csa3.pdf (accessed October 2015)

Gros, E. 2012a. "Esane ou les Malheurs de l'Estimateur Composite." In Proceedings of the *Journées de Méthodologie Statistique*, Insee, Paris. Available at: http://jms.insee.fr/files/documents/2012/936_2-JMS2012_S23-2_GROS-ACTE.PDF (accessed October 2015).

Gros, E. 2012b. "First Assessment of the Combined Use of Administrative and Survey Data in the New System of French Structural Business Statistics." In Proceedings of the Fourth International Conference on Establishment Surveys, June 2012, Montreal, Canada. Available at: http://www.amstat.org/meetings/ices/2012/papers/301882.pdf (accessed October 2015)

Haag, O. 2010. "Redesigning French Structural Business Statistics: Redesign of the Annual Survey." In Proceedings of the Q2010 Conference, May 4–6, 2010, Helsinki. Available at: https://q2010.stat.fi/sessions/session-14 (accessed October 2015)

Kovar, J. and P. Whitridge. 1995. "Imputation of Business Survey Data." In *Business survey methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York: John Wiley.

Kroese, A.H. and R.H. Renssen. 2000. "New Applications of Old Weighting Techniques – Constructing a Consistent Set of Estimates Based on Data from Different Sources." In Proceedings of the Second International Conference on Establishment Surveys, June 17–21, 2000, Buffalo, NY. 831–840. Available at: http://www.amstat.org/meetings/ices/2000/proceedings/INTRO.pdf (accessed October 2015)

Little, R.J. 2012. Rejoinder to the Discussion of his Paper: "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28: 367–372.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Received November 2012
Revised May 2015
Accepted August 2015