EESW13 – September 2013

Missing data treatment in administrative fiscal sources for the French Structural Business Statistics production system

Thomas DEROYON (*)

(*) INSEE, Statistical Methodology Directorate

The French National Institute of Statistics (Insee) began in 2005 a reengineering of its structural business statistics (see [1] for more details), which resulted in a new production process, named Esane (for the French Élaboration des Statistiques Annuelles d'Entreprise), first used in 2009.

Esane rests on the implementation of administrative data (mainly enterprises' annual tax returns) as much as possible in order to lessen enterprises statistic burden. A survey is also carried out on a sample of firms to collect information on businesses turnover breakdown and some sector specific features not available in other sources. Esane uses specific estimators, named composite estimators, that aim at optimally combining these different sources.

Another important feature of the new Esane system is that the list of firms belonging to structural business statistics scope owing to their activity or their legal form is defined at the beginning of each annual campaign. However, no fiscal information is available at the end of each campaign for a significant part of the firms in the scope. Dealing with this fiscal missing data is therefore an important part of Esane, for a major part of the estimation process is based on fiscal information. As we will see, it involves the same steps as those needed to treat non response in a survey. First, we need to determine which among the missing firms ceased to exist and therefore could not answer and which should have answered but did not. Then, we need to treat for this missing information.

In part I, we will shortly describe annual income returns from firms to tax authorities and the way they are used to build composite estimators. Part II explains the possible sources of missing fiscal information and how we assess the reason why information is missing for each firm in Esane's scope. Part III depicts how fiscal missing data is treated in Esane with imputations. In part IV, we conclude by presenting some results on fiscal data imputation influence on structural business statistics' main aggregates.

1. Administrative fiscal sources and how they are used in the French structural business statistics

Annual income returns from enterprises to tax authorities (called IEG for the French Information Économique Générale) contain a large amount of information on firms balance sheets and income statements (for instance, firms turnover, its breakdown in three aggregated activities – production of goods, sales of services and trade -, wages and salaries, added values, investments…). This data can directly be used to produce Esane aggregates for they are based on the French Statement of Standard Accounting Practises to which structural business statistics also refer.

Moreover, this information is easy to use for statistical purposes because the firms identification number is almost always fixed and managed by Insee. Therefore, all these administrative sources are easy to match with each other and with surveys conducted by the French National Statistical Institute.

Fiscal sources however miss important data, especially on the detailed turnover breakdown which is needed to revaluate firms' principal activity. This information is collected through surveys on a sample of enterprises, with other sector specific features. The firms turnover is also collected, to guarantee its declared breakdown to be consistent with fiscal sources.

In Esane, sector-based aggregates are built using both administrative and survey data (see [2], [3] and [4] for more details on Esane composite estimators). For instance, for all fiscal variables that are not available in the survey (namely all fiscal variables except turnover and its aggregated breakdown), and for all sectors at the 3-digits level of the French NAF, the variable's sectoral total is estimated thanks to a difference estimator, that is as the sum of :

- The weighted sum of the variable values for all legal units in the survey sample whose principal activity, as evaluated in the survey, belongs to the sector;
- The sum of the variable values for all legal units in the complete field of Esane whose principal activity in the survey frame belongs to the sector;
- Minus the weighted sum of the variable values, for all legal units in the sample whose principal activity in the survey frame is in the sector.

which can be summed up by the formula :

$$\sum_{S} wi\, Y_i \mathbb{1}_{APE=X}(i) + \sum_{U} Y_i \mathbb{1}_{APEreg=X}(i) - \sum_{S} wiY_i \mathbb{1}_{APEreg=X}(i))$$

APEreg is the principal activity as evaluated in the sample frame, APE the principal activity revaluated thanks to the survey, Y is the fiscal variable, U is Esane's complete scope and S is the survey respondents sample.

That way, administrative data are used to evaluate the value of the fiscal variable for all units of the Esane's scope, and survey data are used to determine the correct sectoral classification.

The same formula is also used to estimate turnover aggregates, with values of the turnover in the first sum based on a confrontation between turnover values in the survey and in the firm's fiscal declaration.

It is therefore crucial that fiscal data be available for all units for which they are awaited.

2. The three sources of fiscal missing data and their identification

As in a survey, fiscal information is not available for all firms in Esane's scope. For instance, in 2009, no declaration to tax authorities was available for 1.12 million firms among the 3.3 million enterprises in Esane's scope. In 2010, information was missing for 1.6 million firms among almost 4 million enterprises. In 2011, there were 4.8 million of firms in Esane's scope and fiscal information was missing for 2.3 million of them.

Three reasons can explain the lack of fiscal reports to tax authorities by a firm :

- Fiscal sources, even if they cover a large portion of the structural business statistics scope, miss some of them, namely very little firms which can benefit from a specific fiscal status (called microentreprises in French) enabling them to pay a fixed tax and relieving them from any fiscal declaration as an enterprise. Their mixed incomes are indeed collected through the household income tax declarations and therefore harder to use in Esane production process. A new enterprise legal form and fiscal status, called autoentrepreneur, was moreover created in 2009 to ease individual firms creations. Fiscal informations for autoentrepreneurs is also more often collected through the household income tax declarations;

- The structural business scope heavily relies on the French business register, called SIRENE (for the French Système d'Identification au Repertoire National des Entreprises[1]), managed by Insee. Each new firm has to file for an identification number in order to open a bank account, but bankruptcy does not entail any declaration. No firm is also compelled to declare if its activity leaves the scope of productive sectors covered by Esane. Therefore, some firms alleged to still exist and belong to the productive sector do not send any information to tax authorities because they in fact no longer exist or belong to the non-profit sector;

- Finally, some firms in Esane's scope should send income returns declarations, still belong to Esane's scope but omit to do it. Other annual income reports are not transmitted by fiscal administration to Insee because they are still undergoing specific treatments. These firms are often rather small, for fiscal administrations are usually keen on getting actual information on regular sources of tax flows, that is large firms incomes.

Firms corresponding to the first and third cause of fiscal information non return can be regarded and have to be treated as non respondents in a census, the census being in fact income returns declarations; firms in the second case correspond to an error in the census frame and should be excluded ex post from the files before further statistical treatments. Identifying enterprises that cease to exist or that exit Esane's scope is therefore a first key step in the process of fiscal missing data treatment.

Tax administration provides Insee with a list of firms with microentreprise or autoentrepreneur fiscal status. Firms in the first case are therefore easy to identify. In 2009, 320 000 firms belonged to this list, 600 000 in 2010 and 1.1 million in 2011. The great increase in the number of firms concerned by this type of fiscal status is caused by the great success met by the autoentrepreneur status.

Other external sources enable us to identify a list of ceased firms, however far from being exhaustive. It leaves us with a number of firms for which no solid information on their actual status is available. We chose to classify them in two categories : the firms that did send an annual income report to fiscal administration in at least one of the two preceding years or that did send at least one monthly VAT declaration in the current year are presumed to be still active in Esane's scope, the other are supposed to be inactive, that is to be ceased or to be out of scope for structural business statistics. All firms identified as ceased or allegedly inactive are excluded from Esane's files and all further treatments. In 2009, 420 000 firms were excluded on the basis of this treatment, 600 000 in 2010 and 780 000 in 2011.

3. How to treat non response in administrative fiscal data ?

Fiscal data for microentreprises, autoentrepreneurs and supposedly active firms is still missing owing to a non response phenomenon and has to be dealt with accordingly. Two main strategies exist to treat for non response : imputation and reweighting procedures ([5]). The latter would have been difficult to implement in this case, for firms already have weights linked with the surveys carried out on a sample of firms to mainly get their turnover breakdown. It would have needed to work with a system of two weights for each firm. More important, fiscal sources are supposed to be an exhaustive source for call firms are supposed to communicate their income to tax authorities one way or the other.

That is why the decision was taken to treat non response in fiscal files with imputation procedures. For each non respondent firm, all fiscal variables have to be imputed so as to respect the numerous accounting principles and relationships that link fiscal aggregates with each other. Three imputation procedures are therefore applied to fiscal data non respondents :

- Microentrepreneur and autoentrepreneurs are very small firms whose fiscal variables have little impact on structural business aggregates, however numerous they may be. That is why their imputation is based on a raw cold-deck procedure : two typical fiscal

[1] Replaced starting from 2013 by a statistical register, called SIRUS (Systeme d'Identification au Repertoire des Unites Statistiques), which contains more information on each firm than SIRENE and is also able to deal with other statistical units than legal ones (for instance profiled enterprises or groups of legal units).

declarations are built thanks to external aggregate data given by fiscal administrations, one for very small firms operating on trade business, the other for the rest of the very small enterprises. Each microentrepreneur and autoentrepreneur receives one of these fiscal declarations based on its fiscal status. This method is called “micro-imputation” ;

- Firms that are not microentrepreneurs or autoentrepreneurs and for which a fiscal declaration was available the preceding year, either because they did actually send a declaration to tax authorities or because they were assumed to be active in Esane’s scope and already imputed, are treated according to a different method, based on their fiscal data. The median turnover growth rate for active firms (that is firms which sent a tax form for the current and preceding years) is computed in each stratum defined by crossings of principal activity (in three positions) and employment size (in five groups). For each firm to be imputed, the fiscal variables are calculated as the product of the median turnover growth rate of the stratum the firm belongs to and of the fiscal variable value the preceding year[2]. This method is called “n-1 imputation” ;

- The remaining firms were created or entered Esane’s scope during the current year. Averages of each fiscal variables for active firms in strata defined as crossings of principal activity (in three positions) and employement size (in fifteen groups) are computed and imputed to each of the supposedly active firms in the stratum that have not yet been imputed. This method is called “mean-imputation”.

In the first procedure, the typical fiscal declarations are set so as to respect the accounting relationships between accounting variables. The two other imputation procedures consist in imputing to each non respondent firm a linear transformation of real fiscal variables. The imputed fiscal data thus created respect accounting principles and relations.

4. Administrative fiscal data imputations’ influence on structural business statistics aggregates

As already mentioned, administrative fiscal imputations concern each year a large number of firms. However, these firms are often small, with very little turnover and number of employees. Thus, they do not weight much in structural business aggregates calculated on the whole scope covered by Esane, but have a greater influence as far as small enterprises are concerned (for instance firms with a small number of employees).
Table1 shows the main results on turnover and wages and salaries aggregates. Influence of imputations on aggregate growth rates are even smaller.

[2] corrected by the ratio of its exercise durations during the current and the preceding year.

Table 1 : Shares of imputations in turnover and wages and salaries aggregates

		2009 – whole scope	2010 – whole scope	2011 – whole scope	2009 – between 0 and 10 employees	2010 – between 0 and 10 employees	2011 – between 0 and 10 employees
Turnover	All imputations	3.7 %	4.5 %	4.1 %	12 %	14.9 %	13.8 %
	Micro imputations	0.1 %	0.2 %	0.3 %	0.2 %	0.6 %	0.6 %
	n-1 imputations	2.6 %	3.1 %	2.7 %	9 %	10.4 %	9.2 %
	Mean imputations	0.9 %	1.2 %	1.1 %	2.8 %	3.9 %	4 %
Wages and salaries	All imputations	4 %	5.3 %	5.1%	10 %	13.2 %	12 %
	Micro imputations	0 %	0.2 %	0.4 %	0.1 %	0.4 %	0.5 %
	n-1 imputations	2.8 %	3.8 %	3.4 %	7.3 %	9.6 %	8.1 %
	Mean imputations	1.2 %	1.3 %	1.3 %	2.6 %	3.2 %	3.4 %

5. Conclusion

Imputations of administrative fiscal data for structural business statistics are a major concern for they involve the treatment of a large number of legal units. Yet, it also concerns mainly small and very small firms, whose impact on aggregate results remains limited.

The imputation procedures we employ nevertheless show some problems; we indeed aim at using more micro-level fiscal information from individual income reports to build the cold-deck imputations of microentrepreneurs and autoentrepreneurs ; we also would like to use more robust imputation procedures for legal units now treated with mean imputation. Works are on progress at Insee on both aspects and these new methods should be implemented during the current or the future campaign.

References

[1] Brion Ph., “Redesigning the French Structural Business statistics, using more administrative data”, *Proceedings of the Third International Conference on Establishment Surveys*, Montreal, 2007

[2] Brion Ph., Gros E., “Methodological issues related to the reengineering of the French Structural Business Statistics”, *EESW*, 2009

[3] Gros E., « Esane ou les malheurs de l’estimation composite : comment gérer les valeurs négatives d’estimateurs par différence », *Journées de Méthodologie Statistique*, 2012

[4] Brion Ph., « L’utilisation combinée de données d’enquêtes et de données administratives pour produire les statistiques structurelles d’entreprise », *Journées de Méthodologie Statistique*, 2009

[5] Caron N., « Les principales méthodes de correction de la non réponse et les modèles associés », *Document de travail Insee*, 1996