Integration of external data from Tax and Public accounts in the Central Business Register

Jorge Saralegui-Gil
Instituto Nacional de Estadística (Ine)
Paseo Castellana 183
28047, Madrid, Spain
jsaralegui@ine.es.

1. Introduction

In recent years, the number of official statistical operations in the business sector has increased sharply in Spain impelled by new national and EU requirements, in particular in the services sector and short term statistics. At the same time, other public administrations and private market researchers are requiring more and more information from companies, which produces heavy workloads on reporting units for their administrative procedures and form filling and a risk of negative impact on quality.

A clear example of the increasing workloads on enterprises can be seen in the following data: In 1995 Ine addressed 162,300 reporting units in structural annual statistics and 680,400 in short term statistics, while in 2002 they rose to 280,211 and 1,174,808 respectively, which means an increase of 73% during that period. Around 40,000 entities report for more than one survey in the same year. It has to be understood that such scenario is the unavoidable workload once the procedures of sample co-ordination at the selection stage are systematically implemented, using permanent random numbers and rotation techniques.

In this context, SAVIA project (Auxiliary System of Variables of Administrative Origin) has been implemented within Ine wit the aim of generating an information system, taking the Central Business Register (CBR) as a base, for developing a cost effective use of external information related to the CBR universe . The final product should help improving the general quality of the different phases of the statistical operations using the CBR as a frame. SAVIA devotes an important part of its resources to administrative metadata analysis and archiving, to facilitate quality diagnosis on coverage and reference periods as well as the precise cell identification in administrative forms an files, in order to derive administrative variables comparable to homonymous statistical target variables. Other two blocks of activities given priority for development have been the improvement of the structural information kept by the Central Business Register (CBR), and modelling the use of external information in estimation .

2. Turnover in a continuous format in CBR

One of the more recent steps towards improving the quality of CBR has been the inclusion in continuous format of the turnover variable in the 100% of its entities. The CBR, active in Ine for more than a decade, makes intensive use of Tax Administration and Social Security microdata files, fully identified. A basic element for identification is the tax identification number –TIN- which defines the legal business entity. TIN is in use by all administrative departments in Spain and is well accepted by society. Up to the year 2000 turnover was provided by the Tax Administration to Ine in discrete format excluding the small entities (less than 1 million € turnover). This CBR variable was in use for comparison purposes mainly, since Ine structural annual surveys provide the turnover as continuos value for the medium/big size units, as they are included systematically within the take all sample strata.

In 2001 , through an special agreement with Tax Adm. authorities, a modelled procedure to assign turnover to the whole CBR population was designed to provide Ine yearly with two types of ‘products’ from the Tax Administration. On the one hand, turnover size class is now available for the whole CBR population. In addition, turnover in continuous format is also available for the same microdata file but in this latter case, the file is unidentified (no TIN). In addition, this file contains all the discrete variables which define the stratum (activity, region, employees and turnover size classes). To select the tax declaration source, from the different available options, we gave priority to Annual Company Tax (CT). We took the Value Added Tax (VAT) data in the case of non-detected units and the Personal Income Tax (PIT) data as ultimate option. The source with the most units detected was VAT (52.6%), followed by CT (28.2%) and PIT (19.2%). Yearly, a process of statistical matching of CBR and the referred unidentified file is implemented to assign turnover in continuous format taking into account the distribution of the number of employees (a pre-existing CBR variable) and within 4,455 cells defined by the common partition criteria present in both sources. It is possible now to allocate a monetary size variable in continuous format originated in tax declarations to the total population of the active entities in the CBR (2,645,018 units), which altogether gave a figure of 1.3 $*10^{12}$ € (2001). In relation to the source of origin of this data, we found that 92% of sales came from CT, 7.5% from VAT and the remaining 0.5% from PIT. Results show high – acceptable for SAVIA objectives- consistency with statistical results estimated by current structural surveys. Subject matter units are in the process of making an intensive use of the CBR turnover variable in the editing and data analysis survey production phases. These activities produce the feedback for improving particular aspects of the matching methodology implemented. The turnover variable is being in use for stratification and allocation, and it is already possible to identify some strata with potentials for optimisation of sample sizes in respect of former allocation .

3. Use of external accounting data for sampling design.

It is well known that regression and calibration techniques can produce substantial efficiency gains when adjustment variables are highly correlated with target variables. Calibration can have an important impact in reducing sample sizes, in the second phase of two phases sampling when the adjustment variables are known for the first phase elementary sample units. This ‘potential’ decrease in sample sizes can be achieved in sampled strata but also in take all strata, where the problem of burden is particularly demanding. Performance indicators of preliminary simulation studies should help to make decisions on to what extent it is feasible to implement rotation procedures in take all strata too, alleviating so the burden/costs of the operation. Obviously, it is necessary to have access at the microdata level to the set of auxiliary variables for the theoretical sample as well as their aggregates for the ‘regression groups’ where the calibration is applied. The selected variables, of administrative origin, nominally and conceptually respond to the same concept as an statistical target variable although technically should be considered as ‘different’ from its statistical homonymous which they are supposed to ‘explain’ in the underlying model for calibration. The set of auxiliary variables under analysis at present are turnover, payroll, purchases, other operating costs and other operating income. For the following example, the former two have been tested. As target variables in our example we have selected Turnover(statistical) and Gross Value Added .

4. External sources

The envisaged external source is the already mentioned Tax on Companies declaration (TC) used annually to assign the (external) turnover variable to CBR entities with partnership as legal status. TC accounts for approximately 90% of total turnover in the CBR. Another source used by the CBR systematically for some applications as units building and enterprise groups profiling is the Yearly Accounts Public Register (YAPR). This external source is compulsory to Companies , and it has a similar institutional coverage as TC (natural persons are not obliged as well as some non market activities). The selected auxiliary variables are also accessible in YAPR, whose supporting files can be bought in the market , linked to the Tax Identification number (TIN). On the contrary, tax monetary variables are not generally speaking accessible to Ine linked to TIN. Nevertheless, one of the interesting advantages of calibration with standard software is that it can be implemented in situ, in secure posts within the Tax Administration and by its own staff. Ine can receive back the calibrated weights linked to identification as far as it is compatible with the tax regulations and statistical confidentiality rules. The simulation example presented here uses YAPR as external source. In the medium term special agreements with the tax administration is envisaged for this particular model, similarly to a procedure already implemented for household surveys (Wealth Survey), what will imply an improvement of quality and saving of YARP acquisition costs.

5. Simulations

As artificial population for simulation has been selected the sample of the structural surveys 2001, covering the NACE sectors corresponding to services and manufacturing industries with initial total size n=170000. For that period, sample m <n of units in scope of YAPR, matching positive, amounted to 92000 units. The set of (n-m) units includes natural persons (out of YARP scope, not obliged to elaborate standard accounts), delays in declaring YARP, small units for which YAPR legislation is more permissive and newly born entities.

Of course, not all the strata of the initial sample can be considered as target for subsampling and calibration. Sample size reduction is particularly aimed at in strata with a substantial (i.e,. over 100) number of units in scope of YARP. In the following example, subsamples of size s <m are replicated 100 times for calibration to sample subtotals estimated by the 'first phase' sample , within NACE –2 groups. Macro SAS Calmar (Insee) is used for calibration to Turnover and Payroll external continuous variables. As performance indicator we have selected in our example the relative mean square error:

$$(1)\ \ \mathrm{RMSE}_{\mathrm{d}} = \frac{\sqrt{\dfrac{1}{K}\sum_{m=1}^{K}\left(\hat{\bar{Y}}_d^{(k)} - \bar{Y}_d\right)^2}}{\bar{Y}_d}\ \ (\%)\ ;\ \text{where}$$

k: simulation index ; d: NACE-2 group ;$\hat{\bar{Y}}_d^{(k)}$: estimator of the mean of variable Y for each simulated subsample in NACE group d, for simulation k ; .$\bar{Y}_d$: estimation based on the sample units of the original sample, in scope of YARP .

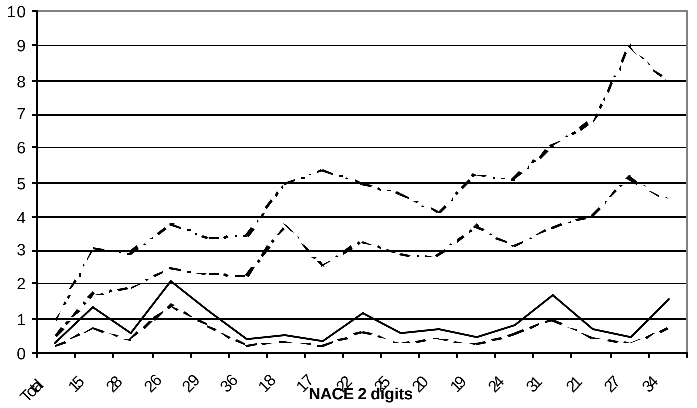
6.Results

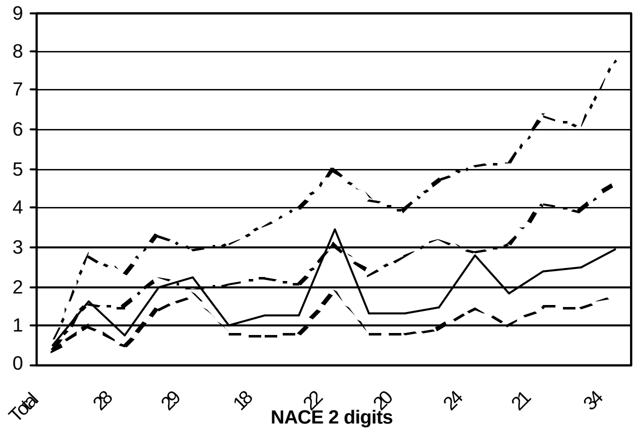
Figures 1 and 2 illustrate results for a subset of 13830 units of size 20-99 employees in scope of YAPR belonging to manufacturing industries original strata, of which 11968 gave positive matching with YARP, forming the 'frame' for simulating subsamples. RMSE of the direct and calibrated estimator respectively for every NACE –2 calibration group are represented in four lines, two of them for the subsampling rate 80% and analogous two for the subsampling rate 60%.

Figure 1 presents results for the target variable turnover (statistical concept from surveys). As it is to be expected, substantial gains in precision are achieved in this variable, for all the regression groups, which are ordered decreasingly by its original sample size. The same has been observed in other statistical target variables when its concept is comparable to a calibration variable (as Payroll). For these structural variables, particularly within former take all strata, calibration will surely permit reduction in costs and burden with losses of precision substantially lower than s.r.s.

Figure 2 illustrates the relative efficiency of the calibrated estimator versus the direct, when the 'administrative homonymous' of the target variable (in the example, the gross value added as collected in the field), is not used in calibration. Here also all groups present efficiency gains too, and in nearly all the performance of the calibrated estimator is better in the smallest subsampling rate (60%) versus the highest subsampling rate (80%) with direct estimator. The same cannot be said for other variables (of much higher variability in the population) being tested, where the gains of precision in respect of the s.r.s. are not so evident. Next steps of the project are oriented to optimisation of subsamplig within original strata and to agree on the organisational issues related to the availability in time of the external files for data processing.

Relative Mean Square Error (%), Direct versus Calibrated. 100 Replications. Calibration Variables: Turnover, Payroll. Sampling rates 60% and 80%.





(-..- 60% direct; -.- 80% direct;__60% calibrated; --80% calibrated)

fig.1 Target variable :Turnover

fig 2.Target variable :Gross Value Added

REFERENCES

Amstrong,J (1993). Greg. Estimation for a Two Phase sample of Tax Records.
Brackstone G.J. (1987) Issues in the use of Adm. Records.
Deville, J.C. , Sändarl (1992). Calibration Estimators in Survey sampling.
EC Regulation concerning structural business statistics (R 58/97),

RÉSUMÉ

Le project SAVIA (Système Auxiliaire de Variables d'Origine Administrative) est un système d´information associé au Registre Central d'Entreprises (DIRCE) qui a a été mis en oeuvre par l´INE (Espagne) avec l´objectif d´ à améliorer la qualité générale des différentes phases des enquêtes économiques, en diminuant les coûts et la charge de réponse des interviewés. Les premières activités ont permis de disposer pour toutes les unités DIRCE d´une variable monétaire continue d'origine fiscale (chiffre d'affaires) comme base aditionnelle pour l'optimisation des échantillons. Un deuxième ensemble d'activités est orienté à atteindre des gains d'éfficacité dans les estimations, en introduisant le calibrage des variables externes, disponibles dans la première phase d'un échantillon théorique obtenu en deux phases au moyen d'accords institutionnelles ad hoc.