Working Papers

06/2012

Use Of Administrative Sources To Reduce Statistical Burden And Costs In Structural Business Surveys (Ufaes). (A reduced english version).

Jorge Saralegui, Cristina González e Ignacio Arbués

The views expressed in this working paper are those of the authors and do not necessarily reflect the views of the Instituto Nacional de Estadística of Spain

First draft:

This draftl: 5 Mars

Use Of Administrative Sources To Reduce Statistical Burden And Costs In Structural Business Surveys (Ufaes)

Resumen

The use of administrative sources with statistical purposes is part of the current activity of the National Statistics Institute (INE, Spain), in various fields. UFAES project provides a new qualitative impulse to these activities, with objectives oriented to significantly reduce the sample size of the major INE annual structural business surveys.

The core idea is to maintain, for a rotating subpopulation, the estimates from a complete direct observation (in the sense of being obtained from a survey of the same sample size as at present) in the first year of a biennium, while in the following year the sample for direct observation will be drastically reduced in size, compensating the efficiency losses with the gains provided by an additional subsample extracted from the previous year effective sample, updated with model assisted estimates based on the change at the microdata level in the tax variables associated with target statistical variables.

Palabras clave

Integration of tax microdata in enterprise surveys;

Indirect estimation of change in enterprise structural variables

Use of tax declarations for reducing business surveys sample sizes

Autores y Afiliaciones

Jorge Saralegui

Cristina González

S.G. de Estadísticas Estructurales y del Medio Ambiente

Ignacio Arbués

D.G. de Metodología, Calidad y Tecnologías de la Información y las Comunicaciones

Instituto Nacional de Estadística

Use Of Administrative Sources To Reduce Statistical Burden And Costs In Structural Business Surveys (Ufaes)[1]

Saralegui Jorge[1], Gonzalez Cristina[2] Arbués Ignacio[3]
[1]Ine (Spain) jorge.saralegui.gil@ine.es
[2]Ine(Spain) cristina.gonzalea@ine.es
[3]Ine(Spain) ignacio.arbues.lombardia@ine.es

Abstract

The use of administrative sources with statistical purposes is part of the current activity of the National Statistics Institute (INE, Spain), in various fields. UFAES project provides a new qualitative impulse to these activities, with objectives oriented to significantly reduce the sample size of the major INE annual structural business surveys.
The core idea is to maintain, for a rotating subpopulation, the estimates from a complete direct observation (in the sense of being obtained from a survey of the same sample size as at present) in the first year of a biennium, while in the following year the sample for direct observation will be drastically reduced in size, compensating the efficiency losses with the gains provided by an additional subsample extracted from the previous year effective sample, updated with model assisted estimates based on the change at the microdata level in the tax variables associated with target statistical variables.

Keywords:
Integration of tax microdata in enterprise surveys;
Indirect estimation of change in enterprise structural variables
Use of tax declarations for reducing business surveys sample sizes

1. Introduction

The use of administrative sources is, obviously, part of the ordinary business of the National Statistical Institute (INE) of Spain in various areas of its activity, particularly in comprehensive statistics of administrative origin, frame building and other tools of statistical infrastructure (especially for censuses), synthesis operations (national accounts, indicator systems), as well as a support for data editing, imputation and analysis. More recently, direct use of administrative microdata has been implemented in household surveys (wages, labour force, living conditions...). This development has been greatly influenced by the provisions of the 2002 INE-Tax Department institutional agreement, in addition to those general EU statistical legislation and specific regulations which refer specifically to the need to facilitate the use of administrative sources in statistical operations carried out for compliance with Member State reporting obligations.

[1] A reduced english version of this paper, annexed, was presented by Saralegui J. at the NTTS2013, 5-8th Mars , Brussels.

The use of accounting information from the Register of Enterprise Public Accounts, web searches and other sources for the editing stages of structural business surveys has significantly reduced the response burden on enterprises particularly by reducing the number of further contacts for data updating. Nonetheless, the project described in the paper, with its two stages of development, simulation (SIMFAES) and use (UFAES) of administrative sources in structural business surveys, gives a new qualitative boost to that approach, with aims geared towards significantly reducing the sample size of the two main structural business operations at INE: the Annual Survey on Manufacturing Industries (EIAE) and the Annual Business Survey in the Services Sectors (EAS). There exists a cumulated experience on the problems related to some shortcomings of administrative sources, and particularly tax sources, in meeting statistical information needs currently provided by structural business surveys, specifically, the existence of structural variables required by national and/or EU regulations for which there is no correspondence with variables included in the administrative source and the limited coverage of the business subpopulation with the legal status of natural person (this subpopulation is particularly significant in the EAS).

The project described here was devised with a view to overcoming as far as possible the above limitations and at the same time obtaining a significant reduction in the burden and costs of structural surveys. It was formulated with a preliminary simulation phase to replicate the planned procedure, applied to sub-samples of microdata extracted from the effective samples of the annul surveys which were already available.

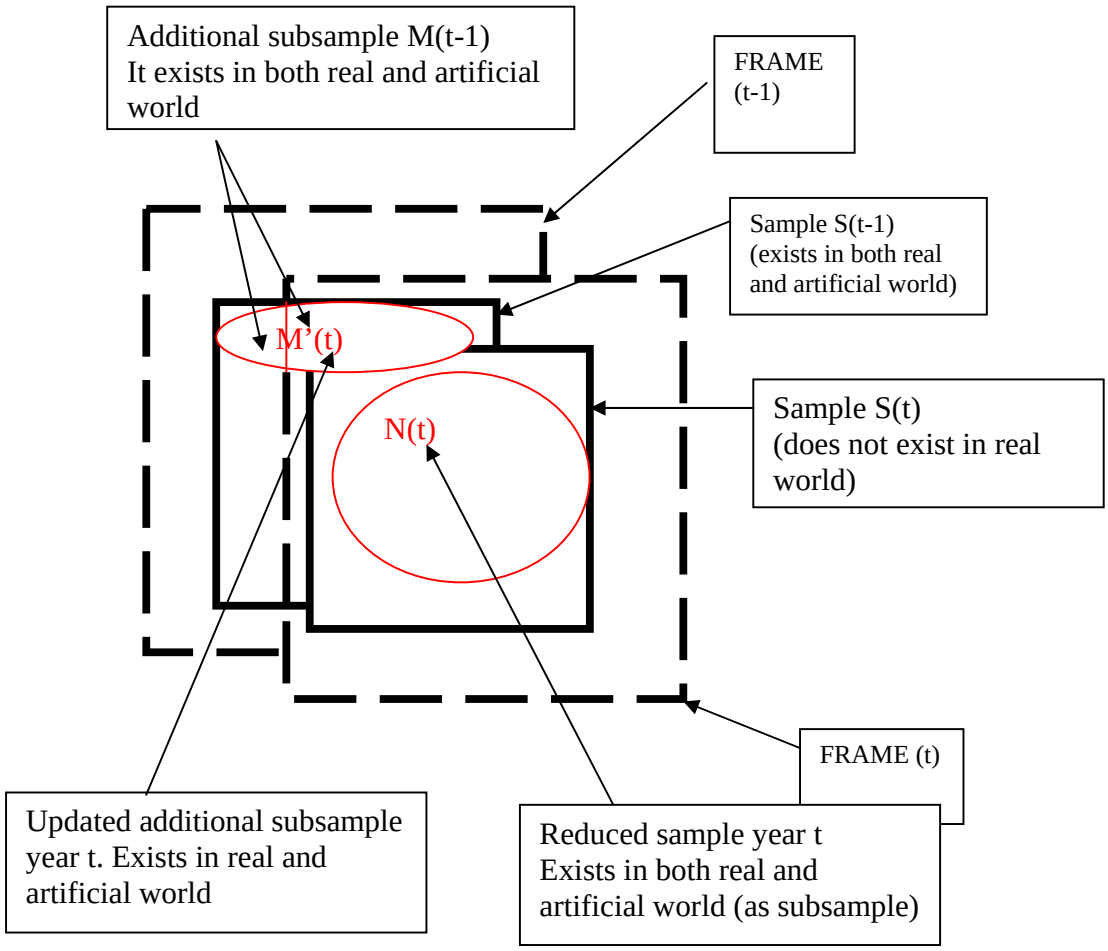
2. Simulation of the use of administrative sources in structural business surveys (SIMFAES)

UFAES project was designed on a two-year basis taking profit of the use of the panel component of tax information given that primary information is available for the same units over consecutive periods. The aim is to produce estimates based on a complete direct observation (in the sense that they are obtained from a survey of the same sample size as the current one) in the first year of the biennium, while in the following year the sample for direct observation will be drastically reduced in sample size. Losses in sample sizes are compensated by efficiency gains provided by an additional subsample extracted from the previous year final sample, updated with model assisted estimates and imputations based on the change at the microdata level observed in the tax declarations, for the same enterprise (sample unit) .

Through this approach, SIMFAES simulations were implemented over the two-year period for which both final survey data and tax microdata were already available at the time (2008-2009). In the simulation for the second year the sample size was artificially reduced to the limit of efficiency for the national level, replicating four artificial subsamples of the effective total. In real world the ‘reduced’ sample will constitute the only collected sample in that year. In addition, four supplementary subsamples were artificially selected as subsets of the effective sample in t-1.

The project team decided to work primarily with the Annual Business Survey on the services sectors (EAS), which has the greatest sample reduction potential, postponing the processing of the Annual Survey on Manufacturing Industries (EIAE) until the evaluation

of the results achieved on EAS had been carried out. The tests were conducted across the full scope of NACE activities in the EAS, although the aim of the UFAES implementation phase during its first two-year real-world reference period (2012-2013) is that the term 'complete sample' or 'reduced sample' only be applied to a subset of activities (covering complete NACE divisions representing aprox. 50% of the target universe), rotating symmetrically with the rest of the divisions over the next two-year UFAES period to ensure that the sample sizes to undergo real-world direct observation remain fairly stable every year.



C.1. Selection chart

The above chart represents the planned design for UFAES, applied in the simulation stage. The notation refers to effective samples, observed in fieldwork, although the selection is initially performed with theoretical samples (not shown in the chart) associated with the latter in order to log unit deaths, which must be kept track in order to calculate the sampling weights of integrated samples. First, the theoretical sample used to obtain the effective sample N(t) is selected in advance with the allocation and sizes needed for a satisfactory degree of efficiency in the variables and strata required by EU SBS regulation as well as by the objectives of the survey at the national level. As the

results of the simulations are assessed, the sample sizes may be reduced below this standard level. The artificial effective sample N(t), simulated for this phase as a subsample of S(t), is the one that would provide direct observation information in the real world in year t, in which what is designated here as S(t) will not exist (instead, the entire frame will be available in t). The administrative variables available for t and t-1 will be loaded for each elementary record of N(t).

Subsequently, an effective subsample M(t-1) will be obtained from the EAS theoretical sample in (t-1) and updated with the effective sample S(t-1). The theoretical supplementary sample in t-1, from which the effective sample M is derived, is selected with the necessary allocation and sizes to ensure that the integration of N(t) and M (t-1) – after the estimation processing designed in UFAES–meets the national and international demand for results for all sub-national domains in the reference year t. Successive size reductions of the subsample N(t) will normally be associated with increases in M(t-1).

The effective subsample M(t-1) will be updated to the frame of the year t, producing deaths of units when switching from t-1 to t, which will generate the subsample M'(t). The latter will load the complete record of M(t-1) as well as the set of associated variables for each sampling unit extracted from the administrative sources in t and t-1.

3. Central statistical variables and associated administrative variables

An essential part of the project is the allocation of a typology to the different variables involved in the estimators, using a specific coding system in order to distinguish the estimation process that they undergo in the different UFAES development phases.

Only the *associated* category is considered for administrative variables, with the standard notation *a* in the project documents. They are present in the tax forms and should have a very high correlation with a variable of the same name or that is conceptually close to it, obtained from the statistical source by means of the survey, referred to as a 'central variable'. UFAES uses the information on the associated variable to estimate the change in the central variable so that, should it arise, the impact of the lack of an exact conceptual match between the two would be attenuated to an extent. When there are central variables with associated variables present in several administrative sources, as is the case of the different tax returns, the information will be uploaded from all donor sources. After performing validity checks on the associated value, a principal associated value is defined with an order of priority. For tax sources, this order is 1) Corporate Income Tax (IS); 2) VAT; 3) Personal Income Tax (PIT).

Some of the variables present in the administrative source do not correspond to any single variable in the questionnaire, but they can be associated to variables that can be obtained as a combination of several of them. We call these combinations composite central variables.

When it comes to non central variables (with no primary or composite associated variable) we distinguish:

i) non-central variables for direct estimation. These variables do not have associated variables in external sources and are estimated using combined estimators of change that use as ancillary information its observed value in t-1, and the estimation of the central

variables ('explanatory') in t for the same unit as well as structures of a nearest neighbour observed in the sample N in t (see general estimator below).
ii) Non-central variables for indirect imputation, based on central and non-central variables with direct imputation. And
iii) Non-central variables for identical imputation. These variables repeat the data in t-1 due to their structural stability.

4. Integration of associated variables and matching quality checks

The INE file for loading administrative variables for UFAES simulation (SIMFAES) was built by integrating the theoretical samples t and t-1 referring to the years 2009 and 2008 surveys. The microdata include the basic identifiers of each business unit common to those available in the administrative source as well as other identifiers in use internally for survey processing.

Coverage checks are initially performed at aggregate level in order to detect the frequencies with which the sample units appear in the different sources used. This allows us to confirm that the matching success levels for the various sources meet the coverage requirements theoretically expected by UFAES preliminary studies. A second type of coverage check is performed for each central variable at a much more detailed level by controlling the frequency of their presence in the three sources (IS, VAT and PIT), in both years. The results were analysed to asses the contribution of each donor source and the quality of the matching within strata in relation to the project expectations.

Control indicators were massively produced through tables showing distance measures of mean values within strata of the principal associated variables *a* in the reference year t for sample units S(t-1) versus the mean of the central variable *x* referenced in the year t-1 for sample units S(t-1), and similarly for the rest of the combinations of reference periods of the tax files and survey data. The means of the central variable or its principal associated variable within the stratum are understood to be calculated by correcting weights for non response (missing in the associated variable). Subsequently, for each central variable, the quartiles of the distributions within the stratum of the ratio of the central variable in respect of its principal associated variable in t; t-1 were also analysed.

Besides detecting suspicious matching results for variables within strata these controls can also be used to fine-tune decision-making on the design of validity checks on a particular associated variable value just before the imputation of its central variable. Once the received files with integrated administrative and statistical variables are approved in respect of the matching output at the aggregate level, it is possible to link the associated variables of each survey unit with the data on the corresponding central variables. Then a validity check on the principal (when more than one is available) associated variable is performed in a similar way to an outliers control. In cases when the associated administrative value was found invalid, a nearest neighbour (NN) search is implemented on the set S(t-1), to impute to the non-associated central variable the change in the valid associated values of the donor variable. The nearest neighbour capture is performed on the sample in t-1, for which all questionnaire variables exist. A selection is made to form the search vector with the main structural variables of the enterprise in t-1. NN search dominium is restricted by stratum and legal form (natural person and others), given the

varying availability of associated variables in each case. A maximum number of recipients is set for each donor.

5. Estimation of change in variables of the additional subsample M'(t)

We estimate first the central variables with a valid associated variable. Subsequently the elementary central variables with valid composite variables, the non-central variables with direct estimation and the non-central variables with indirect estimation are estimated. The estimation of the central variables is carried out by using an estimator of the change in the validated tax associated variable (ASOC):

$$\left|a_{i(t-1)}\right| \geq 10(mil€) \,\&\, (sig(x_{t-1}) = sig(a_{t-1}) = sig(a_t)) \Rightarrow x_{i(t)} = x_{i(t-1)} \frac{a_{it}}{a_{i(t-1)}}; \quad \textbf{(1)}$$

Otherwise:

$$sig(x_{t-1}) = sig(x_{i(t-1)} + (a_{i(t)} - a_{i(t-1)})) \Rightarrow x_{i(t)} = (x_{i(t-1)} + a_{i(t)} - a_{i(t-1)});$$
$$sig(x_{t-1}) \neq sig(x_{i(t-1)} + (a_{i(t)} - a_{i(t-1)})) \Rightarrow x_{i(t)} = x_{i(t-1)};$$

a: administrative source associated values, valid ; x: values of statistical central variable in t; t-1For all other variables, the estimators are of the structure-preserving type, combining a nearest neighbour component and the observed values in t-1 for the same unit. Formulation of the general estimator, which have some exceptions described in project documentation, follows. Other non-central variables with indirect estimation or identical estimation (that maintain the t-1 level or structure % in respect to its explanatory variable) are estimated once the main estimation is complete, in the phase immediately prior to the integration of the final UFAES M'U N file:

$$\hat{y}_{iht} = ((1 - a) * \hat{x}_{ih(t)} * \left(\frac{y_{hjt}}{x_{hjt}} \right) + a * y_{ih(t-1)} \left(\frac{\hat{x}_{hit}}{x_{hi(t-1)}} \right) \qquad (2)$$

y: non-central variable or central non-associated variable, real in t-1, imputed in t.
x: central variable with valid associated variable, explanatory.
t-1; t; reference years; h stratum.
i: survey unit.
j: nearest neighbour unit.
0<= a <=1; Value for general case: ½ (other values under evaluation) .

Distance vector component variables, general: Turnover; Wages; Employment; Total Consumption; NN used for investment variables, includes total investment and excludes wages.

The donor search stratum is constructed as a partial aggregation of design strata to obtain a sufficient number of donors.

Here, the definition of nearest neighbour is different to the one used to impute the change in the associates, referred to formerly, since the search in this case is performed in t, i.e. the variables must be central with an associated variable. In parallel to the first implementation of UFAES, research will continue into the potential of estimators assisted by multivariate regression models using central variables with valid associated variables as explanatory .Given the need to implement UFAES as soon as possible, due to the significant savings in the fieldwork budgets of the INE for 2013 (with reference to 2012), it was considered necessary to postpone the diagnosis on the eventual introduction of estimators with these techniques in subsequent UFAES cycles.

6. Iteration of artificial samples.

The work for the SIMFAES sub-project, including all preparatory work, was performed part-time by the UFAES project team over a year along 2011 and 2012.

Selection of samples N(t)**:** The overall sample of the EAS-09 totalled approximately 139,000 units (aprox. 9% global sample rate) . Sample sizes were calculated to obtain accurate national level estimations, resulting in an overall theoretical size of approximately 79,000 units. Then four subsamples were selected at random that were cross-referenced with the overall effective sample producing effective subsamples of variable sizes: 59,502, 59,549, 59,528 and 59,502. On average, the size of these subsamples is roughly that expected in the real world, i.e. a mean reduction of -35.9% of the theoretical EAS sample. UFAES will be implemented in its first year (2013, referring to 2012) on a subpopulation of approximately half the frame of the whole services sector (i.e. EAS activities excluding trade and transport, NACE div: 45 to 53) to allow for the rotating effect needed in UFAES biennial design. Hence, the expected direct and indirect net real-world savings in the first year of implementation are estimated at €900,000, approximately –25% of the EAS budget of the previous year (before UFAES).

M'(t) samples: Four subsamples M'(t) of around 39,000 units each (size needed to cope with the whole set of current EAS objectives) were extracted as follows: The sizes of the additional sample M (08) were allocated to strata, with an approximate increase of 15% to take into account subsequent updating for frame unit deaths

From the theoretical sample of EAS-08, four subsamples M(08) were selected, with negative coordination with each subsample N(09). These were then cross-referenced with the effective sample S(08) and subsequently updated with the frame in 09, eliminating unit deaths, to obtain the four final subsamples M'(09) of sizes 39,075, 39,021, 39,133 and 39,084 enterprises. Sample files contain the information inputs needed for calculating the weights to be applied in estimators both for the MUN sample and for N alone (necessary for estimators that use the change in the means between periods).Units >20 employees in sample M, included in SIMFAES on experimental basis only, given their observed impact on certain SIMFAES evaluation indicators and their small numbers, will not be included in the real UFAES subsamples M which will be thus restricted to units <20 employees only .

7. Evaluation

The evaluation was conducted in two phases:
i) An assessment based on a set of Montecarlo indicators of bias and MSE in respect of the target real value (EAS 09), calculated by using the four simulated samples, for a subset of central variables, determinants of the enterprise accounts structure:

X: central variable
B: indicators of bias.
EMC: indicators of Mean Squared Error.
M U N: integrated SIMFAES samples; 4 replications.
N: subsample observed on the field in t; 4 replications; id.
M: UFAES supplementary subsample; 4 replications;
EAS09; EAS08; EAS10: estimated values of the real world EAS survey.

$$\hat{B}_{M\cup N}\% = \frac{100}{4\hat{X}_{\acute{E}AS09}}\sum_{S}^{4}\hat{X}_{\cdot S(M\cup N)} - \hat{X}_{EAS09} = \frac{100}{\hat{X}_{\acute{E}AS09}}(\overline{\hat{X}}_{(M\cup N)} - \hat{X}_{EAS09});(3)$$

$$\hat{EMC}_{M\cup N}\% = \frac{100}{\hat{X}_{\acute{E}AS09}}\sqrt{\frac{1}{4}\sum_{S}^{4}(\hat{X}_{\cdot S(M\cup N)} - \hat{X}_{EAS09})^2};(4)$$

$$\hat{EMC}_{N}\% = \frac{100}{\hat{X}_{\acute{E}AS09}}\sqrt{\frac{1}{4}\sum_{S}^{4}(\hat{X}_{\cdot S(N)} - \hat{X}_{EAS09})^2};(5)$$

$$DifTV_{09/08} = (\frac{\overline{\hat{X}}_{(M\cup N)}}{\hat{X}_{\acute{E}AS08}} - \frac{\hat{X}_{EAS09}}{\hat{X}_{EAS08}})*100 = \frac{\hat{B}_{(M\cup N)}}{\hat{X}_{\acute{E}AS08}}\cdot\cdot*100;(6)$$

$$\hat{B}^{rel1}_{(M\cup N)} = \frac{\hat{B}_{(M\cup N)}}{\hat{EMC}_{N}};(7)$$

$$\hat{EMC}^{rel1}(\%) = \frac{\hat{EMC}_{M\cup N}}{\hat{X}_{EAS09}}*100;(8)$$

$$\hat{EMC}^{rel2} = \frac{\hat{EMC}_{M\cup N}}{\hat{EMC}_{N}};(3);(9)$$

And ii) A detailed analysis on an extended set of variables, in particular those needed to meet the requirements of SBS EU Regulation, performed on the simulated final files integrated with all survey variables as currently produced by EAS annual procedures. The outputs of this analysis may be iterated after the detection and correction, where applicable, of outliers impact on the subsample M' (this effect is considered neutralised in the subsample N by the usual editing processing to produce EAS final files). This detailed analysis was performed for parsimony purposes with a single simulated sample only, so other indicators rather than Monte Carlo statistics were used with the four simulated samples, as described below.

8. Comments on the results

Generally speaking, NACE classes with higher biases normally remain within the interval defined by the MSE of the sample N (actual sample without UFAES estimated component). The exceptions detected to date relate to the presence of a very influential observation or because they appear in classes and/or NUT II regions with very small sample sizes for which the Montecarlo MSE of N obtained in SIMFAES is presumably not a very efficient estimation. For the purpose of detecting whether an increase in the size of the subsample M in % has any effect on the SIMFAES estimated errors the relation bias vs. MSE of the sample N and the weight M% in MUN within each subpopulation, was analysed.

Again, the deviations observed do not seem to depend on the weight of M in the sample MUN within the NACE class concerned, which confirms the importance of the impact on SIMFAES bias of the Montecarlo MSE of the sample N, which is free from UFAES effects.

9. Some conclusions

The UFAES procedure is ready for use in EAS 2012 (fieldwork in 2013), since all the necessary developments are complete.

Further research is needed into stochastic estimators with lower deterministic components than those used in SIMFAES09, namely those offering multivariate methods of imputation (such as IVEware, of the University of Michigan) used in the INE Living Conditions Survey, although implementation difficulties are expected with this type of model in a scenario of complex variables relating to the accounting structure of enterprises. The performance of simulation tests is also recommended with the Annual Business Survey on the Manufacturing Sectors for its possible incorporation into UFAES techniques. It is nonetheless to be expected that the sample reduction obtained with this operation will not be of the size applicable to the EAS.

References

Brackstone G.J. (1987) Issues in the Use of Administrative Records

EC (2011) Regulation concerning structural business statistics

Eurostat (2012) Framework for the Integration of Business Statistics (FRIBS)

Eurostat (2011) Daniel Lewis on behalf of work package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics. Using administrative data to estimate survey variables not directly available from administrative sources

INE (2011) Annual Business Survey on the Services Sectors. Methodology www.ine.es

Saralegui, J. (2003) "Integration of external data from Tax and Public Accounts in the Central Business Register". 54th ISI session (Berlin).

Saralegui, J. (2004) "Use of tax data for sample design under confidentiality restrictions" Q2004 conference. Mainz (Germany)