The impact of profiling on sampling

EESW 2017

Emmanuel Gros & Ronan Le Gleut

August 30, 2017

National Institute of Statistics and Economic Studies (INSEE, France)

Introduction & context

Two main business surveys at INSEE: ESA and EAP

- Two main business surveys, part of the ESANE process, produce structural business statistics:
  - ESA, Annual Sectoral Survey:

    Scope: Activities of trade, construction, services and transport

    # 116 000 data collection units in Metropolitan France

  EAP, Annual Production Survey:

    Scope: Mining and quarrying, Manufacturing industry and energy

    # 35 000 data collection units in Metropolitan France

- Main purpose : to deduce the main activity of a company (sectoral classification) by breaking down its turnover into activities .
- Sample of LU drawn by stratified SRS.

The SBS European regulation

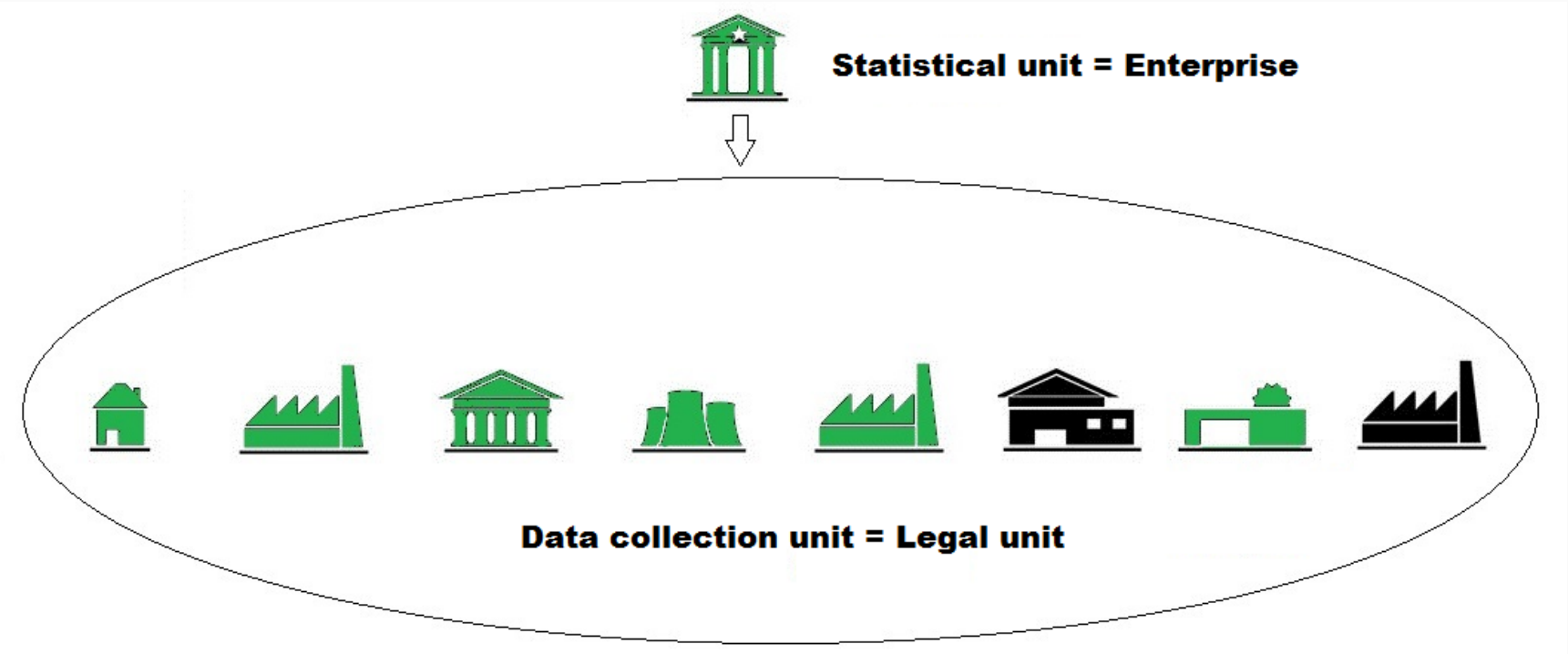
- In many countries of the European Union, business statistics are undergoing great changes.
- In France for instance, business surveys are currently based on the observation of legal units that have a juridical definition.
- From now on, in order to comply with the Structural Business Statistics (SBS) European regulation, business statistics will be more and more based on the economic notion of enterprise.

How to characterize an enterprise?

- In order to build a new sampling frame on the population of enterprises, we used administrative data on the financial links between legal units.
- Some of the attributes of an enterprise are given by these administrative data:
  - The main activity: depending on the economic weight of its legal units
  - The localization: the same as its “principal” legal unit (decision center)
- Others are obtained by consolidating or simply adding the attributes of its legal units:
  - The turnover
  - The number of employees
  - The total assets

Statistical unit *versus* data collection unit

- Since the statistical units (enterprises) are now different from the data collection units (legal units), the sample design can be seen as a stratified cluster sampling.



Optimization of the survey design

Definition of the take-all strata

- The rules are the following:
  - The biggest enterprises, in terms of turnover, number of employees or number of legal units, are automatically included in the take-all strata.
  - In order to decrease the number of legal units in these strata, we add a cut-off of the legal units achieving 95% of the turnover within each enterprise.
- In particular, this definition of the take-all strata permits to limit the variability of the total amount of legal units to be surveyed.

Stratification and domains of interest

- The take-some strata are defined by crossing:
  - The business sector of the French classification in 5 positions (activity);
  - The number of employees in each enterprise (9 strata).
- Two domains of interest are considered:
  - Activities (business sector in 5 positions);
  - Sectors $\otimes$ Employees (intersection between the business sector in 3 positions and the number of employees)

Neyman allocation under cost constraints

- Since data remain collected on **legal units**, the survey cost depends on the number of legal units to survey, which cannot exceed N_{LU}.

  $\Rightarrow$ Therefore, we introduce costs in the Neyman allocation to respect this constraint:

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] \\ u.c. \sum\limits_{h=1}^{H} C_h n_h = N_{LU} \\ u.c.\ n_h \leq N_h \end{cases}$$

  with $C_h = \bar{N}_{LU,h}$ the cost, i.e. the mean number of legal units per enterprise in stratum h.

Neyman allocation under local precision constraints

- The aim is also to respect a given local precision on the two domains of interest. To pursue this end while taking the cost constraint into account, we extend the algorithm presented by Koubi & Mathern (2009):

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] \\ u.c. \sum\limits_{h=1}^{H} C_h n_h = N_{LU} \\ u.c.\ n_h \leq N_h \\ u.c. \max\limits_{d \in D} CV_d \leq CV_{loc} \end{cases}$$

with D the whole range of domains of interest and CV_{loc} the maximal expected coefficient of variation.

Results

Precision of the total of turnover at the enterprise level

	Domains of interest					
	Activities (1)			Sectors × Employees (2)		
Levels	$n_{ent,(1)}$	$n_{ent,(2)}$	$n_{ent,mix}$	$n_{ent,(1)}$	$n_{ent,(2)}$	$n_{ent,mix}$
100% Max	5%	74,4%	**23,1%**	89,3%	11%	**43,1%**
90%	5%	9%	**6,3%**	20,8%	11%	**12,5%**
75% Q3	5%	4,9%	4,4%	9,2%	8%	8,9%
50% Median	2%	2%	2%	4,2%	4,6%	4,2%
25% Q1	0,9%	0,8%	0,8%	0,1%	0,2%	0,2%
10%	0,2%	0,1%	0,2%	0%	0%	0%
0% Min	0%	0%	0%	0%	0%	0%

Table 1: Distribution of local CVs of the total of turnover at the enterprise level depending on the allocation and the domain of interest (without the take-all strata of units with more than 200 employees for the second domain of interest).

Precision of the total of turnover at the legal unit level

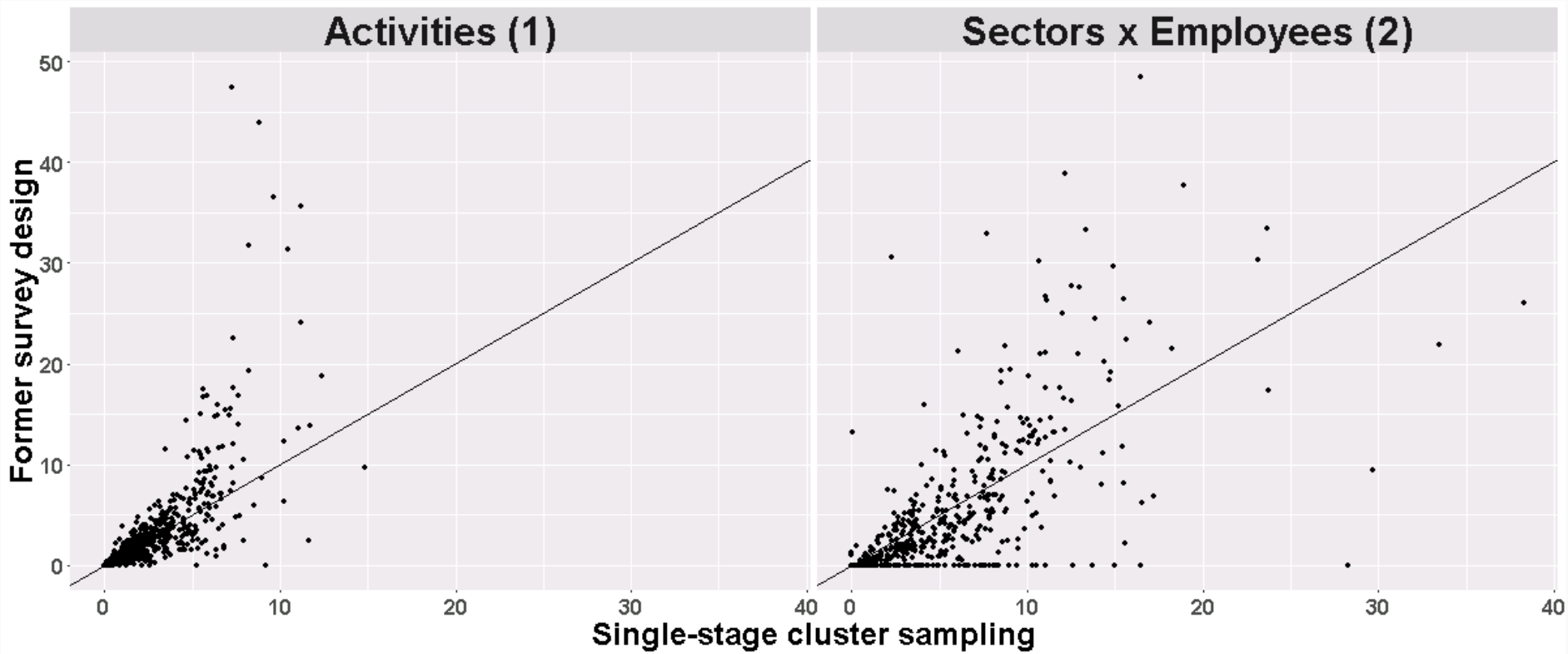


Figure 1: Local CVs of the total of turnover at the legal unit level depending on the domain of interest and the survey design.

Conclusion and perspectives

Conclusion

- The results presented here allow us to have an optimized sample design when considering the enterprises as the statistical units. This sample design respects the following constraints:
  - The number of legal units to be drawn for each survey is fixed.
  - The samples must be optimized at the enterprise level in order to have the best precision on total turnover.
  - For this optimization, two domains of interest are considered for the dissemination of results.
  - The surveys may be used for a dissemination of results at the legal unit level.

Perspectives

- In future analyses, one could:
  - Find optimal factors $(\alpha, 1-\alpha)$ instead of (1/2; 1/2) for the calculation of the “mixed” allocation;
  - Use an algorithm for sample allocation that allows an optimization on several domains of interest at the same time;
  - deal with the ”change of composition” of enterprises between the drawing of the sample and the dissemination of the results, thanks to the Generalised Weight Share Method;
  - Implement post-treatments such as non-response weight adjustment, calibration and winsorization of outliers.

Thank you for your attention