35th Voorburg Group meeting on service statistics

Virtual meeting
September 24th – 25th, 2020

Cross-cutting Topic part two
SPPI sampling method and sources

“The French way of SPPI sampling method”

Antonia Bertin,
Swann-Emilien Maillefert
INSEE – France

The views expressed in this paper are those of the authors alone and do not necessarily represent the position of INSEE or any other organization with whom the authors may be affiliated.
The authors would like to thank Jean-Marie Fournier (INSEE) and Anne-Cécile Argaud (INSEE) for their explanations and proofreadings.

Table of content

Introduction

The purpose of the present paper is to specify procedures implemented in France for the sampling of firms and product services for the calculation of the Service Producer Price Indices.
It contains a summary of the different sampling methods implemented and highlights the particularities of the sample refreshing system in its entirety. Then, it focuses on the challenges encountered by notably illustrating with examples of industries which required additional and specific statistical treatments.

1. General aspects

1.1 Data description

The INSEE's survey for services' price indices (Observation of Prices in Industry and Services: Opise) monitors the evolution of production prices for services from French services companies to all markets (BtoAll), to French companies (BtB), to households (BtoC), and to foreign markets (BtoX). For this last indicator, we distinguish if the foreign market is based in the European Union or outside it according to the European regulations. The prices measured are basic prices (including intra-group transactions, subsidies and excluding taxes), but for contract escalation purposes, market prices are also released.

When it comes to indices publication, France is already compliant with FRIBS regulation which extend PPI perimeter for services. The most troubling branch of activity 5110 for air travel is currently under investigation as part of a partnership with the supervisory authority (General Direction for Civil Aviation). In terms of precision, we release price indices with a lower level of granularity than the one required by STS and the FRIBS regulations. Indeed, because our indices are used for volume indices calculation, as deflators for National Accounting (resources-use balance) and for Service Production Indices (SPI) also required by the FRIBS regulation, INSEE disseminates SPPI until the NACE at 4 digits level (classes) instead of a 2 digits level (divisions)

1.2 A sampling system rather than pure statistical sampling

The renewal of the service industries is a continuous process, so that each industry is refreshed once every five years on average, although this interval may be adjusted to take account of technological, product or market swift changes.

This process is subdivided into several phases, with manual interventions, making it more accurate and less reductive to consider it as "sample elaboration" rather than "sampling" to describe our work.

For French SPPIs in general, samples are determined at two different levels: first, firms and secondly service products.

- For firms, there is a two-step process:
  - a statistical "cut-off" sampling,
  - and then, a "manual method" or "well-informed choice" method, that tries to determine firms that would add appreciable data to the initial sample, for example in order to provide better coverage of one of the different markets.

- For service products, engineers-surveyors from INSEE, specialized in services, visit the sample of firms to define or re-define services that will be followed in a customized quarterly questionnaire. These visits allegedly provide standard products whose prices are easy to follow through time which guarantee a good quality for the second level of the sample.

2. Sampling frames

2.1 Pre-determination of the list of units to be surveyed

The survey system is based on a sample of firms that is completely renewed over a five-year cycle. INSEE proceeds in two different ways, with distinct treatment for the sampling of prices of building maintenance and improvement works (part of the NACE division 43 – specialised construction activities). Indeed, this specific survey (out of the scope of indicators required by European regulations) does not follow the same procedure as other surveys for service or even industry. Particularly, it is not the same network of surveyors that it is used; it is the INSEE's network of households' investigators which is required to pursue this survey, whereas for industry and other service industries we rely on a dedicated team of engineers-surveyors.

Except for NACE division 43, each year and prior to any processing, the list of renewals of industries is drawn up. However, since 2017, INSEE has been developing sampling frames and sampling over the entire scope of the FRIBS regulation. First of all, this makes it possible to dispose of a sampling immediately ready for use in the event that an industry has to be renewed unexpectedly. But more importantly, it allows to identify and process multi-branch companies, potentially surveyed several times during the renewal of different industries by different engineers-surveyors.

In the case of maintenance and improvement of buildings, renewal is not carried out according to industry, but according to the Siren number of the enterprises (id code handled by SIRENE, the french business register identification system). Each year, one-fifth of the total sampling is renewed on the basis of a draw established on the last digit of the Siren number of the enterprises: 0 and 5 in a given year, 1 and 6 the following year, etc[1]. Thus, the samples from 5 consecutive years are severed and the global sample is renewed every 5 years, according to INSEE objectives.

2.2 Setting up databases from data disseminated by the annual scheme of companies (SBS)

The information used for sampling comes from an annual scheme, the "Elaboration of annual statistics of companies" (Esane) which notably identifies for each enterprise its sales revenues, with 2 years apart, in every industry of activity (according to the French Nomenclature of Activities – NAF – linked to the NACE). The use of these data provides us with exhaustive information as this scheme synthesizes tax data and annual sectoral surveys in a coherent framework.

From these data, INSEE aim to sample productive firms to compute services producer price indices (SPPIs) by destination of output: domestic business to business (BtoB), domestic business to consumer (BtoC), domestic business to export (BtoX) dissociated in two destination markets (Eurozone excluding France and rest of the world), according to the FRIBS regulation.

From a methodological angle, two consequences follow:

- it is necessary to collect representative services for each of the observed indicators listed above. This suggests ideally the drawing of specific samples for each of them;
- it is crucially important to reliably identify the productive unit. Indeed, the exhaustive Esane bases contain legal units, profiled enterprises and units involved in restructurings. Before setting up the sampling frames, one must therefore eliminate double counts related to restructurings and integrate descriptive information on contours of the profiled enterprises to bring back to the unit level corresponding to the ones under investigation.

1 If the sample is made in 20XN, their last SIREN number must be either N-2 or N+3 modulo 10. In 2020, the last SIREN number must be 8 or 3.

2.3 Setting up basic sampling frames

For the BtoAll indicators, the sampling frames are immediately created: we keep the units whose turnover in the industry of the level considered, CPA at 4 or 5 digits levels, is at least 4 million €. The firms are then classified in descending order of this turnover, so that a cut-off selection can be made.

For the sampling frames corresponding to indicators BtoB and BtoX, the “Esane” data also provide a split of sales by customer category. We dispose of a nationality of customers variable (national customers, foreign customers inside European Union, foreign customers outside European Union) and a kind of customers variable (households and enterprises, including civil service and local and regional authorities)

The firms are sorted in descending order of BtoB (resp. BtoX) turnover, calculated as the sum of the variables customer kind (resp. customer nationality). But these variables correspond to the whole activity of the enterprises, and not only to the activity in the industry concerned. In order to bypass out of scope, the bases are restricted to enterprises whose first 4 characters of the APE code (NAF code of the main activity carried out by a unit) match the NAF (equivalent to the NACE code at 4 digits level) industry in consideration. Finally, enterprises whose turnover in the NAF industry is less than 4 Million € are also removed from the sampling frame.

Note that there is no specific sampling frame set up for the indicator BtoX dissociated between two destination markets (“E1” Eurozone and “E9” rest of the world) and for indicator BtoC: we use the information from the frames obtained previously. This last point is even less critical, as for most SPPI, BtoC indices can be duplicated from CPIs.

For the sampling in prices of buildings' maintenance and improvement works, we also rely on the annual sectoral structural survey, especially on one of its specific construction modules with an exhaustive stratum and a sampled stratum. This module provides the decomposition of revenues depending on the type of buildings the work is done on: individual houses, collective housing, offices, and other type of building (factory, commercial establishments,…), which comes in handy to determine whether the enterprise is doing buildings' maintenance and improvement works. Notably, this module is not specific for the branch 43, but it is used in order to estimate the revenues in the field of study by industries and types of buildings. The aim is to distinguish residential and non residential buildings). For that purpose, some peculiar assumptions are necessary:

- the works in ancient buildings in this module are only works in the field of study;
- for each industry in the 43 division, buildings' maintenance and improvement works concern only a part of activity. Thus, for each enterprise sampled, their revenues in buildings' maintenance and improvement works are disaggregated according to the industries in 43 as all of their revenues in the 43 NACE division.

In order to be sampled, the enterprises have to meet the criteria below:

- Their revenues in one of the 10 branches covered by buildings' maintenance and improvement works must be higher than 500 000 €, in order to limit interrogation of small business (cut-off sampling).
- In order to reduce the out of domain’s risk, it has also been chosen to delete some enterprises on the basis of their partition of revenues. For example, if a company, whose main code activity (APE) is not either 4399A (waterproofing work) or 4399C (general masonry work and building shells), only verifies the condition of revenues on branch 4399 (other specialised construction activities), we delete such business from the sample.
- And if there are small businesses already interrogated on another survey, they are removed from the sampling frame.

3. Sampling methods

3.1 Sample of firms

3.1.1 First step: statistical "cut-off" sampling

The first sampling of an industry is a succession of elementary samples, each one corresponding to the crossing of an indicator and a CPA level.

The method used is a cut-off procedure: for a given industry and a particular indicator, the firms are ranked by decreasing turnover within the branch in consideration.

Firms are being kept up to a certain limit, which is defined by 3 parameters:

- a coverage rate to ensure sufficient representativity;
- a minimum number of firms to be compliant with statistical confidentiality down-stream;
- a maximum number of firms to prevent a plethoric sample.

The table below lists the parameters selected for the 2020 sampling for each elementary sample. These parameters may slightly change from one year to the next.

Indicator	CPA level	Maximum number of firms	Minimum coverage rate (%)	Minimum number of firms
BtoAll	CPA4	35	70	5
	CPA5	35	55	5
BtoC	CPA4	5	0	0
btoX	CPA4	5	0	0

Then, we get a unique sample of firms for each industry (NACE4), by concatenating the elementary samples. Thus, this means that a firm selected for any one of the four elementary samples will be retained in the overall NACE4 sampling. Consequently, the real coverage rate can be significantly higher than the nominal 70%.

This automatic procedure ends with additions made by matching with other databases:

- the business and establishment register to update the location data;
- the database for restructurings in order to indicate if the company has recently changed its business conditions;
- the financial links between companies to identify and outline the groups of enterprises operating in France.

For the sampling of prices of building maintenance and improvement works, we make an early sample based on the Annual Sectorial Survey of Y-3 and another later sample with the Annual Sectorial Survey of Y-2.

The early sample is used to start in advance the investigators' visits. Notice that the these surveys providing an exhaustive and a sampled stratums, businesses already have an initial sampling weight even before the sampling specific to building maintenance and improvement works. Since the drawing do not rest on a probabilistic method (only on the last Siren number), for both sample (early and later samples), initial weights do not require additional treatment. But because both of these samples are used to make a definitive one, a shared weighted method is used to determine the final weight of each enterprise.

The general formula is shared weight = initial weight/ number of connections, which is how many times the enterprise appears in both of these samples, so it can be 1 or 2. The rules are as follows:

- for the enterprises only present on the early sample, shared weight = Max(1, Weight_early/2);
- for the enterprises only present on the last sample and if its creation was before Y-2, shared weight =Max(1, Weight_last/2);
- for the enterprises only present on the last sample and if its creation was on year Y-2, shared weight =Weight_last;
- for the enterprises present in both samples, shared weight = (Weight_early+Weight_last)/2.

Since 2017, in order to reduce indices variance, the shared weights are truncated at 100, meaning that if they are higher than 100, their definitive value will be 100.

Once the definitive sample is done, the companies of which Siren ID code last with N-2 or N+3 (if the sample is made in 20XN, their last SIREN number must be either N-2 or N+3 modulo 10) and which are not in the definitive sample, are no longer kept in the survey.

3.1.2 Second step: "well-informed choice" method

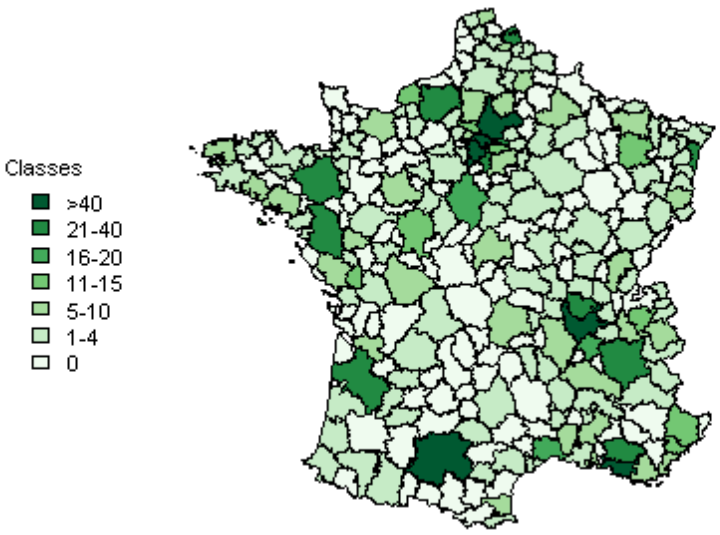
First, a research carried out on the Internet aims at identifying firms that have been forgotten with the previous automatic processes. This is possible because some firms may be misclassified. Besides, meetings with one or several industry federation of employers or professional trade unions are organized to discuss about industry concerns, and in particular about sampling difficulties.

Lastly, if further businesses are needed because of bad respondents, poor coverage or to make up for out of perimeter enterprises, an other sample is drawn, called Top100 with the first 100 companies in terms of sales revenues in the industry and new units are picked from it. We might also draw new samples if necessary to provide an even better coverage.

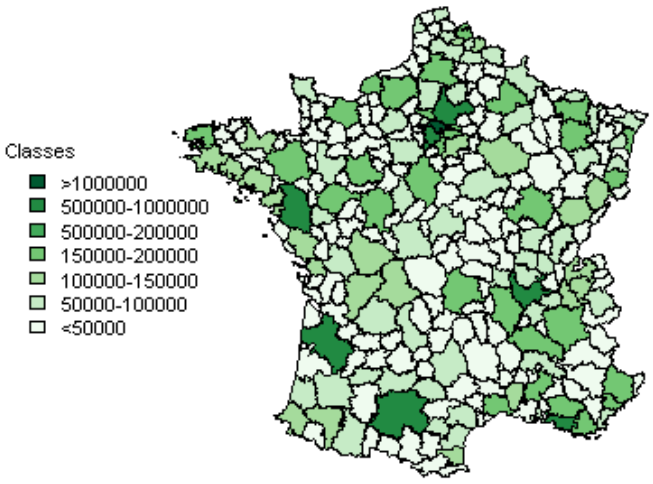
3.1.3 Assessment of these methods

Even if we use a non-probabilistic method of sampling, we are still representative of how activity is located all over France territory as underlight the following maps.

Map 1: distribution of companies sampled by employment zone



Map 2: distribution of employment in France



3.2 Sample of services product

Once the sampling of firms is complete, engineers-surveyors from INSEE visit the sample of enterprises to define or re-define services that will be followed in a customized quarterly questionnaire. During this process, they select price indicators that can reflect the realizations of the firms, without being too complex to follow, so that they can be provided in a quarterly basis. Engineers-surveyors choose the appropriate price method according to the situation of the firm.

The services have to be representative of the price variation of a product family (kind of product x kind of market). In practice, the products with the biggest turnover within each family are most often chosen.

During the initialization visit, engineers-surveyors usually begin their interview with the turnover part, comparing the data collected by INSEE from structural business statistics, with the actual activities of the enterprise. In that part of the visit, export activities are reported as precisely as domestic activities if possible.

Within the industry level, transactions prices are weighted by the corresponding turnover according to the distribution of the turnover collected by the INSEE engineers-surveyors in the firm, dissociated by destination (BtoB / BtoC / BtoX in eurozone / BtoX with the rest of the world). Sometimes, just an estimation is available for this distribution of the turnover, especially for isolating activity with the Eurozone.

4. Difficulties and Challenges

From a theoretical point of view, since we use a non probabilistic cut-off technique to select enterprises, there is no sampling error as such. Because the cut-off conditions are based on sales revenues, in every NACE4 commodity-group, only main firms are surveyed. Thus small businesses do not get sampled and their evolution of prices for services are not included in calculation. Consequently it could be seen as a source of bias. Nevertheless, if one assume that main firms are price-maker and that the others are price-taker, the bias should be limited. Furthermore, since 2016, in order to reduce the statistical burden of surveys on small businesses (especially the ones with less than 10 employees), they are limited to one mandatory survey each year by the law. Thus if we exclude micro-businesses and *autoentrepreneur*, there are only 0,2

million companies left from which we also have to count out industrial and commercial businesses. And because the French system of interrogation rests on face to face interrogation, we have to sample the main firms for each branch. Indeed, the disposable interview capacity can't afford to interrogate more companies especially the ones with a small amount of sales revenues in the branch which will not much affect the evolution of prices in the entire industry.

Notice that on a more practical level, for some industries, the conditions for cut-off sampling give a too small sample to cover a sufficient amount of the sales revenues for the branch and even if one takes the Top100 sample it can encounter the same issue. Take for example the architects' branch ; the first 100 companies in terms of sales revenues cover only 16 % of all the industry's incomes. Even with the first 1000 companies, one can only reach a 43 % coverage. Furthermore, we noticed that the coverage for sales revenues from businesses to households (BtoC) was really low. Indeed one can assume that households only use services from small architects' firm. Consequently we had to produce an additional sample which will notably allow a better coverage for BtoC and to consolidate BtoB.

We used proportional sampling method to do so with a 4 times higher weight for BtoC than for BtoB (we did not take in account the BtoX, already judged sufficient). This choice of weight was determined to obtain sample weights which were not too far from 1 in order not to distort sales revenues distribution (which is after used to determine the weights for representative services provided by the company). Indeed, the initial sample coming from a cut-off method, there is no weight on the initial businesses. The risk incurred is then that the sales incomes for weighted companies get higher than the most important units from the initial sample, which will led to issues about calculation: we do not want that small businesses make the results for the all branch. But in the mean time, we have to cover for a massive part of missing revenues with only a few new sampled companies, thus it is hard to get sampling weights really closed to 1.

The sampling of prices of building maintenance and improvement works presents the opposite problem. The method gives the engineers-surveyors a longer list of businesses than necessary. Because of that, they only contact a small part of the sample. There was a concern that the companies chosen were presenting some specific characteristics such as proximity with the investigator's location. Thus a bias can be induced.

That is why in 2017 a prioritization method was implemented. The companies are separated on the basis of their quantile of sales revenues, their stratum and the number of activity branches they work in. There are 4 scales of prioritization and rules are given to the investigators to collect: every businesses on scale 1, about 1 on 4 for scale 2, 1 on 20 for scale 3 and no one for scale 4.

Furthermore, we encounter some product coverage issues. At the beginning of the renewal process of an industry, professional unions are requested by INSEE to assess under-coverage and over-coverage, so that both are limited when an activity is "renewed" or just "refreshed". But with time, under-coverage is able to grow: new firms are created and they are not canvasses, new products are sold, and they are seldom surveyed. While companies are able to "switch products" when responding to the questionnaire, in practice they do not do so very frequently.

Lastly, we have classical non-response problems. Primarly, some actions to speed up or increase the rate of response have been implemented:

- The list of representative products are discussed face-to-face between firms and field surveyors: this helps improving the response rate and the quality of the survey. If necessary, engineers- surveyors are likely to maintain contact long after the visit.
- Web-data collection is widely favored.

But there will always be some residuals of non-response to be processed before each dissemination. The reasons for this are multiple: representative products unperfectly defined, no sale during the period under review, difficulties to estimate prices because of an unsuitable information system, change of contact inside firms, momentary mere oversight. To limit their impact on aggregates, for non-respondents, we estimate indices through close groups of products.