**35th Voorburg Group Meeting on Services Statistics**

September 24th – 25th, 2020

Session: Cross-cutting Topic (2)

SPPI sampling method and sources

# SPPI Sampling Methods in México

Elizabeth García Flores

Rafael Posse Fregoso

**National Institute of Statistics and Geography**

**México (INEGI)**

## Purpose of this Paper

The main purpose of this paper is to provide to the participants of the 35th Meeting of the Voorburg Group an overview of the statistical layout for the sampling of establishments[1] and product services for prices collection for the calculation of the Service Producer Price Indices (SPPI).

In this paper we emphasis on explaining the various sampling methods used as a result of the differentiated characteristics of establishments from which the SPPI prices are quoted, such as:

i) Probabilistic sampling - used for the selection of businesses in industries where the number of them is vast and justifies a random selection.

ii) Cut-off sampling - used in non-probability sampling of large enterprises that lead the industry.

iii) Probabilistic and non-probabilistic samples of the consumer price indices - that are taken into account as approximations for the calculation of SPPI since they concern to small producers of final consumption services (BtoC).

iv) Sampling of specific services or varieties offered by the economic units.

A brief description of the price quotation, which is carried out through administrative records, is also included; these records provide data generated for operational purposes as part of the functions of government agencies, in order to provide an alternative for on-field data collection. The document dedicates a specific section, by its importance, to the description of the Statistical Business Register of Mexico, (RENEM by its name in Spanish). The main information source of all the establishments or economic units existing in the country. This source is used as the foundation for the construction of sampling frames necessary for the selection of economic units within each industry.

For the design of sampling frames, the usefulness of The North American Industry Classification System (NAICS) is also described. As an economic activity classification system built on the production function, it states a unique, consistent and updated framework. It also facilitates the calculation of weights and integration of services for the SPPI. The NAICS standardizes the presentation of results; this allows correspondence and comparability with other national statistics that follow the classification system. Lastly, the document refers to the main challenges and difficulties over the sampling of the SPPI.

[1] Establishment (Economic unit). Entity that produces goods or services.

# 1. SPPI Conceptual Elements

## 1.1 Definition of SPPI

The National Producer Price Index (PPI) is a family of price indices whose purpose is to provide measurements on the variation in prices of a fixed basket of goods and services representative of national production.

The SPPI are the price indices corresponding to each of the 110 service products[2] covered by the PPI regarding the service industries. They follow the conceptual framework of the national index. The structure of this conceptual framework is based on the Mexican System of National Accounts (SCNM by its acronym in Spanish).

The SCNM splits national production into supply and destination sides. The supply side are the root sectors of goods and services produced (BtoAll), while the destination side are the sectors, where the products are further processed until a final stage. The destination side is divided into intermediate (BtoB) and final goods or services (BtoC + BtoCF + BtoE). In order to comply with this, the INEGI uses the Supply and Use Tables as well as the Input – Output Matrix of the SCNM, these instruments allow identifying the destination of goods and services, and their participation in intermediate and final goods. These mentioned structures grant a conceptual agreement between the PPI and the SCNM.

## 1.2 SPPI: Main Characteristics

The current national indicator is the PPI based on the month July 2019 = 100. Its basket is formed by 560 goods and service products, of which 110 are services. To make up the SPPI baskets, the services of the industrial sectors are identified according to the NAICS. The service products that are included in the basket are those whose participation is equal to or greater than 0.005% of the National Production Value. Prices are collected monthly, except for industrial premises rent, quoted every six months. The SPPI have a national geographic coverage.

[2] **Service products:** A set of services with similar characteristics and with a common end use. They represent the primary category of aggregation or **elemental aggregate** that provides weights for the calculation of the Product Price Indices according to the formula for the Laspeyres Price Index.

According to NAICS 2013, the PPI reaches a sectorial coverage about 75%. There are 560 goods and services in the PPI basket of which 68 belong to the Primary Sector, 382 belong to the Secondary Sector and 110 belong to the Tertiary Sector. The 110 service products of the Tertiary Sector represent 33.5 % of the participation in the PPI basket. Even though NAICS differs from other economic-activity classification systems, the statistics compiled in it are comparable to those based on the International Standard Industrial Classification (ISIC) of the United Nations Statistical Office.

## 2. Information Sources for Sampling Establishments

According to the characteristics of each service product, there are different sources of information, therefore differentiated sample designs. Overall, three types of information sources are used to price the 110 service products:

i) Industrial establishments: Economic units that produce goods and services for collecting prices.

ii) Sources of information or economic units that are used to collect prices by the Consumer Price Index (CPI): Small providers of final consumption services that provide services whose prices are collected by the CPI, some of them are cinemas, restaurants, medical consultation, beauty salons, sport shows, among others.

iii) Administrative records or databases: These data are collected by government agencies and shared with the PPI. For example, administrative records from the Ministry of Communications and Transportation that help INEGI to quote prices for bus and truck tolls. Another example are the tariffs of commercial electricity from the Federal Electricity Commission. In Table 1 you can find the types of information sources and sampling methods used for the selection of establishments.

**Service products classified by Information Source and Sample Design Table 1**

| Sample Design | Price Information Source | Number of Service Products |
|---|---|---|
| Probabilistic sampling for SPPI | SPPI Establishments | 34 |
| Non-probabilistic sampling for SPPI | SPPI Establishments | 40 |
| Probabilistic sampling for CPI | CPI / SPPI establishments | 10 |
| Non-probabilistic sampling for CPI | CPI / SPPI establishments | 24 |
| Non-probabilistic | Administrative records | 2 |
| **Total** | | **110** |

## 3. Sampling Methods

Regarding to the nature of the information sources, there are two types of sampling methods to select establishments where prices are quoted in the field: Probabilistic and Not Probabilistic, for the last one using the Cut-off Sampling (See Table 1). For both types of sampling, the selection of the most representative specific services or varieties within the establishments, is made according to the information provided by the informant regarding to the sell value per product.

### 3.1 Probabilistic Sampling

Probability sampling is applied to 34 service products out of 110 referring to SPPI, since there are a significant number of establishments with a low participation rate in relation to the total sales turnover of the industry. This justifies the random selection. With this sampling, the economic units are selected impartially and it is possible to measure the quality of the sampling through the estimation of the variance or the sampling error.

Specifically, this probability sampling allows:

- To calculate the uncertainty or sampling error associated with the estimates.
- To perform statistical inference to generalize the results.
- To ensure the objective and impartial selection of observation units.
- To make the study more economical.

Some of the most important service products, where probabilistic sampling is applied, are freight truck transportation, rental of offices, hotels, accounting services, medical services, among others.

### 3.2 Non-Probabilistic Sampling (Cut – off Sampling)

Cut-off sampling is applied in 40 service products. It is used to select establishments that lead the industry in which there are a few service providers, such as air passenger transport, internet services, advertising, maritime freight transport, among others. This is also the case of some government-regulated services, such as bus tolls and public transport. In this sampling, all units above a threshold are included in the sample. The threshold selection is

judged based on the value of the control or size variable (sales turnover). Cut-off Sampling is done in two stages. In the first stage, the threshold is defined for the population units that will be included in the sample. In the second stage, the elements of the sample are selected until the defined threshold is reached (See section 5.4.2).

### 3.3 Service products quoted through Administrative Records

For services that are obtained from administrative records, the data is systematically collected and permanently updated by public or private organizations for non-statistical purposes. Such as the case of truck tolls services and electricity for public service.

### 3.4 CPI / PPI Sampling

Regarding 34 CPI services for SPPI, they are ejected to the PPI data base as a reliable sample for the same purpose, using the price quotations for each service and variety shared between them. The sample in the CPI has a design for the same aim, use the prices for do indices. This is a multipurpose sample design, with a national coverage, and the selection is probabilistic for services like beauty salons, dry cleaning services, cinemas, funeral services, restaurants and cafes, medical services, accountants, etc.

The Non-probabilistic sampling includes services whose fees are regulated by the government, there is no sampling frame, or there are only few service providers, therefore it is not necessary to carry out a probability sampling (foreign bus, metro, taxis, mobile telephony, water supply services, etc.). The prices are collected using the regulated prices published by the local government, the state or federal government.

### 3.5 Sampling of Specific Service Products

The specific service or variety to be drawn within each economic unit is the most sold or representative in the opinion of the informant, an average of three varieties per economic unit are selected. Although the varieties have been selected during the initial contact phase, once the information available is updated in the Economic Census or administrative records, the varieties are replaced.

There are some specific varieties that the producer is updating and must be replaced when it is observed. When the attributes of the services or products change, and the service in sample has been out production, it is replaced immediately for the new one. In this way, the sample of varieties is updated according to the dynamics of the national production plant, ensuring that the sample contains the most representative services or those with the highest sales turnover.

## 4. Establishment´s Sampling Framework

The SPPI sampling frames are built from the Statistical Business Register of Mexico (RENEM by its acronym in Spanish). It is designed to select probabilistic and non-probabilistic samples of economic units for each of the SPPI. The framework is the foundation for developing the sample design.

### 4.1 Mexico´s Statistical Business Register

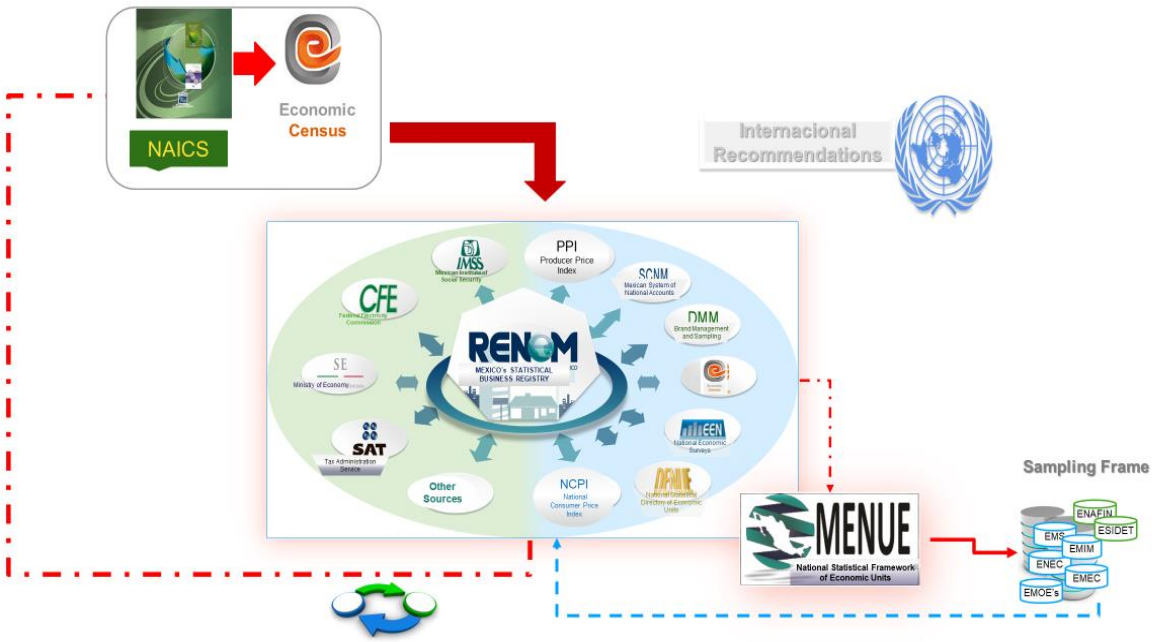
The RENEM is a data registry that includes current and historical basic information of all the economic units that produce goods and services in the country for domestic consumption and for export. This registry is managed and periodically updated by INEGI.

RENEM is the main source of information and it is the foundation of all statistical projects related to economic units for the construction of sampling frames. It is based on information from the Economic Censuses and other sources, according to the structure of the NAICS, which allows to identify the correspondence between the economic activity of the establishment and the goods or services it provides. It is a registry that links statistical information with digital cartography also, and it is the only business registry that has a public version. Its main variables are:

- Identification variables: business name.
- Location variables: address and geographic coordinates.
- Economic variables necessary for statistical design: employed personnel, sales turnover and economic activity.
- Demographic variables: starting year for activities, closure date, historical record of address changes and corporate to which it belongs (if applicable).
- Control variables: unique CLEE identification key, update date, source of information for update, among others.

RENEM is periodically updated with the results of the Economic Censuses and field information provided by economic surveys and price indices, as well as the information provided by administrative records of government agencies such as: The Tax Administration Service, The Ministry of Economy, Mexican Petroleum, etc. (See image 1). The periodic update allows to identify the births and closures of economic units, changes in their production activity and other information.

**RENEM Update Cycle Image 1**




RENEM´s main uses are:

- Extracts sampling frames for National Economic Surveys, special surveys and price indices.
- Links the administrative records of foreign trade with information captured in the census and economic surveys to obtain the Profile of Export Manufacturing Companies and Exports by Federal Entity.
- Used as input for the publication of updates to the National Statistical Directory of Economic Units (public version of the RENEM).
- Links administrative records between government agencies and RENEM statistical records for update and advantage of the statistical potential of administrative records.

To build the sampling framework of each service product, the NAICS activity class is identified from the description of the service, subsequently a sampling frame for each SPPI is developed for the selection of independent samples.

### 4.2 Frame Coverage

Regarding the SPPI 2019, the framework is made up by the domestic economic units whose main activity correspond to one of the service industries of 2013 NAICS classes of economic activity in which one of the 74 service products in the basket can be listed, according to RENEM information. Each service corresponds to only one class of NAICS activity, it may be that the economic unit repeats itself in several classes because there are companies that produce various products or services. Around 338.7 thousand economic units make up the framework of service providers for 74 service products.

One or more service products in the basket are related for each NAICS activity class or industry, therefore the total number of establishments in the activity class was considered to form the framework of each service. There is a one-to-one correspondence between the sampled service and the NAICS activity class in about 85% of the industries, therefore it is always verified that there is a consistent relationship between the description of the activity class and the service product assigned during sampling.

### 4.3 Outdated Frame and Attrition

Sampling frames may include economic units that have stopped providing the desired service, units that have closed operations definitively, or exclude newly created establishments. Due to the high cost of verifying and updating the information of the frameworks and that economic census events in Mexico are carried out every five years, it is necessary to recognize that the sample framework has some degree of error or that is outdated in order to define strategies to control it. These include the use of the information resulting from the field operations of various statistical programs, as well as the annual update of closures and births of economic units and another changes derived from the RENEM.

## 5. Sample Design

The objective of the statistical design is to obtain a sample of economic units to get the prices variation for service products that establishments provide. The economic units are selected in the first stage and within them, the most representative varieties are selected based on the informant's criteria.

### 5.1 Target Population

All the domestic economic units whose main activity corresponds to one of the NAICS activity classes that produces some of the service products in the basket. Producer prices are quoted in these economic units.

### 5.2 Study Domains

They are defined by each of SPPI for which a statistical design is carried out with national coverage. The domains are made up of the domestic economic units whose main activity correspond to the production of the service product, thus results are generated with statistical representativeness at the national level for 34 services of probabilistic sampling.

The results of 40 SPPI with cut-off sampling, are not statistically representative to the national service product since they are considering a non-probability sampling. Nevertheless, it is taken into account that they have adequate coverage, since the economic units are selected “most representative” according to their participation in the economic activity class, achieving good coverage.

### 5.3 Sampling Unit

The sampling unit is the establishment that provides the services in nine of the 10 NAICS Service Sectors, on the other hand for the one corresponding to the Transport, Mail and Storage Sector, the corporation it is the sampling unit.

## 5.4 Sampling Methods

A probabilistic design was considered for services that are provided in small economic units with atomized economic activity that are numerous in the industry. The non-probabilistic design was foreseen for the companies with a high concentration of productive activity and a high participation within the economic activity class regarding the size measure (sales turnover). The sample allocation for both kind of sampling is explained below.

### 5.4.1 Sample Size and Allocation in Probabilistic Design

Simple Random Sampling is applied to services with probabilistic design where the sampling units have a known and non-zero probability of entering the sample; the results of the sample are generalized to the entire population.

The sample size of each service product was determined considering sales turnover as the design variable with a confidence level of 95%, a non-response rate of 15% and a relative error of 15%. The percentage of non-response is derived from the analysis of samples prior to the base year change, while the relative error is established based on the tolerance in the estimation. This non-response cushion seeks to compensate the eventual refusal of the economic unit to provide prices, as well as the possible closure of operations of any establishment or even the cancellation of a variety required, and preserve the optimal sample size.

The sample size of 34 service products with a probabilistic design is 1,507 economic units distributed in 31 NAICS classes or industries in which some may quote one or more services. In the probabilistic design, the economic units were selected randomly without replacement in each study domain and since the design is Simple Random Sampling for each service product, all economic units have the same probability of belonging to the sample. The formula to determine the sample size is the one used for an interval estimation of a total.

### 5.4.2 Non-Probability Sampling

The institutional coverage indicators, defined by the INEGI's Quality Assurance Committee, the following cut-off ranges were defined for any survey with non-probabilistic sampling (Table 2):

**Coverage indicators for non-Probability sampling** **Table 2**

| Coverage range | Coverage |
|---|---|
| Greater than or equal to 80% | High |
| Greater than or equal to 60% and less than 80% | Medium |
| Less than 60% | Low |

For this design, the economic units are selected according to their sales turnover contribution with respect to the total industry turnover in which the services are produced. These units are ordered from highest to lowest based on their percentage share of sales turnover, and the first economic units are taken until the fixed coverage is reached. Hence, the Cut-off Sampling allows a high degree of coverage with few representative units. This is possible due to the distribution of the control variable, which is concentrated in a reduced number of large economic units within the corresponding economic activity class.

## 5.5 Sampling Error

The statistical precisions of the SPPI with probabilistic sampling are estimated through the variances and coefficients of variation. Considering that the estimator of the index for each probabilistic service product is not a linear estimator but a geometric mean, the variance is estimated using the sample variance of the logarithms of the relative prices used to calculate the price index. The coefficient of variation of the index is calculated as the quotient of the square root of the variance between the index estimate.

Regarding non-probabilistic service products, their quality sampling is valued based on coverage according to the ranges defined by the Quality Assurance Committee, mentioned in Table 2. Before the pandemic all non - probabilistic SPPI had high and medium coverage thresholds according to institutional standards.

## 5.6 Replacement of Economic Units in the Sample

Once the economic unit is selected, it remains in the sample until the following base change period or the unit closes operations, stops providing the service, or refuses to continue providing prices. The price collectors report the non–response on the field: the lack of the product, closure of the establishment, change of business, etc. When the economic unit stops producing a variety, the service is replaced by a similar one in the same establishment

according to its importance in sales, or the establishment must be replaced looking for the same service. It is planned to launch a rotation scheme to avoid the attrition of the probabilistic sample.

In the event that some of the above situations prevent further capturing prices, the establishment is replaced in the case of cut-off sampling. Central offices send to the price collectors the number of substitute units necessary to maintain the cut-off coverage.

## 6. Sampling Tips

According to our experience, one of the main advantages of sampling the economic units of the SPPI is to rely on sampling frames for each service product that are based on the RENEM. Since this Register is updated annually through field visits, administrative records from government agencies, and every five years the Economic Censuses carry out their updating in a comprehensive manner, this constant update helps to mitigate the problems of over and under coverage of the frameworks derived from the closure and birth of new establishments.

Our experience in this regard tells us that the use of the NAICS classifier with a six-digit level serves not only to select the samples of economic units by service product, but it is also an important advantage because it makes it easier to identify the NAICS class of economic activity to the economic unit according to its main service. This facilitates also the construction of sampling frames.

Another recommendation gathered from the SPPI sample design is the convenience of having independent samples for each of the service products. This makes the selection and maintenance of these samples easier and provides the possibility of conducting a scheduled rotation by industry without affecting the coverage and statistical precision of the other industries.

## 7. Difficulties and Challenges

### 7.1 Sample Rotation of Establishments

The selected economic unit remain in the sample until the following base period change or

until a change of PPI weights occurs. This is done regularly every five years. The aforementioned causes wear on the informant and lose of representativeness of the sample, due to the fact that the establishments stopped providing the service or closed operations definitively. The main issue for non-response from the PPI is the informant's refusal to continue collaborating.

It is planned to launch an annual or biannual rotation of the sample based on RENEM updates, this will allow the sampling frames to be kept up to date; while the frames are updated, the samples will remain fixed during the period, at the end of that period a part of the sample will be replaced.

In regard to the outcome of the RENEM update at the end of the year, the percentage of the sample that will remain fixed will be defined, the types of economic activity or industries where the rotation of establishments and services will take place, the period of rotation, as well as the implications on the weights of the indices.

### 7.2 Sampling in the new Normal

Covid 2019 imposed severe restrictions on the production and operation of companies and on the collection of prices for the calculation of indices. In addition, consumption and prices of products and services linked to non-essential activities were badly affected. The economic units most affected have been precisely those related to the provision of services. The airlines severely restricted their available flights, the lodging services reduced their offer, the cinemas, shows and all kinds of sporting events were canceled, while the restaurants closed for the most part, providing only in some cases delivery or take-out service. Consumption patterns are also changing as in the case of purchasing airline tickets, since consumers now buy tickets on dates very close to departure.

This has led to the introduction of new methods of price collection or to intensify their use, such as websites, email and telephone interviews, the use of which increased after confinement. The non-response, by temporary or definitive closure of the service providers, as well as the decrease in the variety of services, forced the application of the established imputation procedures.

Even though these adjustments are still considered temporary, the new normality looms with great challenges and difficulties after confinement. One of the main problems is the eventual under-coverage of the sampling frames derived from the definitive closure of operations of economic units. This poses the challenge of the necessary updating of the sampling frames, and consequently, of the establishment samples and their selection probabilities. The other big problem is the changes and representativeness of the samples of varieties of services, given the decrease in the offer or their disappearance in the market, as well as changes in the quality of the services, which will also impact the comparability of prices.

Some options to solve the previous problem would be the replacement of the non-existent varieties by others with the same characteristics in the same economic unit, as an economic consequence of the pandemic, if possible. The other alternative will involve the replacement of the economic units with definitive closure, which is affected by the non-timely updating of the sampling frame.

In Mexico, a Business Demography Study has been launched to measure the impact of the pandemic by making field visits to a probabilistic sample of economic units. It is estimated to obtain the survival and mortality rates of the establishments, as well as the number of them that were born nationwide, by state and by economic activity sector. It is expected that the results of the study will allow the economic sectors and regions most affected to be dimensioned in order to determine the replacement or resampling requirements for service products.