

STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 27  
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing  
(Prague, Czech Republic, 14-17 October 1997)

Item 5 of the provisional agenda

**OCEAN: TOWARD A GENERAL SYSTEM FOR SAMPLE SELECTION AND  
COORDINATION**

Submitted by INSEE, France <sup>1</sup>

---

<sup>1</sup> Prepared by Pascal Rivière (INSEE, France) and Guy Laflamme (Statistics Canada).

## I. INTRODUCTION

1. Sampling procedures are one of the most crucial operations in business surveys: sample coordination is a prerequisite for limiting the "response burden"—the target of recurrent complaints by enterprises. Hence the need to establish a common, proven methodology for ensuring a consistent, homogeneous survey sample. Such a methodology was developed for the Annual Enterprise Surveys (EAEs), which provide one of the most important sources of information on the French production system. Since 1989, the EAEs have been using the annual-survey coordination tool called OCEAN (for « Outil de Coordination des Enquêtes Annuelles »), which enables the response burden to be more evenly distributed—or actually reduced—while ensuring greater consistency in the System of Enterprise Statistics.

2. In an article published in *Courrier des Statistiques* in 1989,<sup>2</sup> F. Cotton gave a detailed description of the OCEAN mechanism used by the Annual Enterprise Surveys. The article was written before the OCEAN came into use. This second discussion of the subject will give a brief assessment of its actual use—past, present, and future.

## II. TWO FUNCTIONS: SAMPLE SELECTION AND FACILITATING THE INFORMATION FLOW

3. OCEAN serves two main purposes: to select a sample and to facilitate the flow of information between system partners. OCEAN, therefore, is neither a survey management software package nor a register of statistical units. It must be used jointly with other tools for the successful completion of a survey.

4. Unlike the sample-selection function, the application to **handle information flows** between partners was *specifically developed for the Annual Enterprise Surveys*. It cannot be used "as is" for another survey. Applying this methodology elsewhere would thus require an investment that would only be justified for a major survey with a large sample. These conditions have been met only once, when an environment resembling OCEAN was set up for several Labor Ministry surveys. However, while based on the same concepts, the two applications remain distinct, for they use different sample-selection methods.

5. The **sample-selection** function comprises two stages: the construction of a sampling frame and, from that frame, the actual selection.

### Constructing the sampling frame

6. The sampling frame gives access to target-population units for which we want to produce estimates. In business surveys, the sampling frame is a list of target-population units from which the sample is drawn.<sup>3</sup>

---

<sup>2</sup>F. Cotton, "OCEAN, outil de coordination des enquêtes annuelles," *Courrier des Statistiques* no. 52, Dec. 1989.

<sup>3</sup>In household surveys, it is often impossible to compile the list of individuals to be contacted; in such cases, other techniques are used.

7. To identify the units in the population, we need to define the **survey field** based on the known characteristics of the units contained in the initial list. One could describe this as a filter applied to an initial list in order to extract the sampling frame. In OCEAN, the characteristics that can be used for this purpose are: *principal activity*, *legal status*, and *number of employees in the unit at December 31*.

8. To make sure that the sampling frame created by OCEAN will work for several different surveys, we define a very wide field. The field covers all active and non-singular units<sup>4</sup> engaged in a market activity. This so-called "overall" sampling frame is created once a year and contains about 3.7 million units.

9. No survey makes direct use of the overall sampling frame. In practice, the field of a given survey is defined by means of a file containing the three variables mentioned above. As the overall sampling frame includes both enterprises and local units, it can be exploited for a wide range of surveys.

10. The main qualities of an effective list-type sampling frame are:

- **exhaustiveness**: the list must contain all the units belonging to the target population;
- **exclusiveness**: the list contains only the units belonging to the target population;
- **lack of duplicates**: each unit is listed only once;
- **high information quality**, in particular its recency.

### Comparing information sources

11. OCEAN uses three information sources to create the overall sampling frame: the SIRENE register, a portion of the Annual Enterprise Survey results, and the previous year's overall OCEAN frame. The initial list of all local units and enterprises is extracted from SIRENE. The overall sampling frame thus meets the first three requirements listed earlier, as all local units active in France must register with SIRENE. Moreover, SIRENE is the only entity responsible for issuing the single identification number (SIRET) for each local unit. All this greatly reduces the risk of duplication.

12. The use of Annual Enterprise Survey data improves the quality of OCEAN information. A change in unit characteristics may not be immediately recorded in SIRENE, whereas the survey contacts many enterprises each year. It is entirely possible, therefore, that a more recent item of information may be present in the survey but not in SIRENE. That is why, when setting up the overall sampling frame, we compare the SIRENE and survey information on units included in both files. The comparison rules can be highly elaborate for some variables, but, as a rule, the more recent information takes precedence. The year-earlier

---

<sup>4</sup>The "active/non-singular" code, managed by the SIRENE register, makes it possible to identify the local units engaged in an economic activity of their own. By eliminating such categories as units that lease facilities, non-operating units, etc., the code prevents double counting.

sampling frame is used to check replacements in the survey sample.

13. The overall sampling frame obtained in this manner contains few variables, so as to avoid storing information already present in other files. However, it does centralize the information needed to coordinate the samples for two surveys. At the end of the processing sequence, the overall sampling frame contains identification variables, stratification variables, a brief statistical history of the unit, and a random number between 0 and 1. The samples are selected and coordinated on the basis of this unit's random number.

14. Thanks to its solid theoretical foundations, OCEAN is capable of selecting samples stratified according to three variables. The user must specify the survey field and the sampling plan. The selection is performed stratum by stratum, in keeping with the sampling plan supplied. The basic selection principle is very simple: to select  $n$  units in a stratum, the user simply chooses the units corresponding to the  $n$  smallest random numbers.<sup>5</sup>

### III. OCEAN IN PRACTICE

15. Many OCEAN functions were developed to meet the specific requirements of the Annual Enterprise Survey, which will doubtless remain the principal user. The survey therefore provides a convenient illustration of OCEAN in actual use.

16. The **Annual Enterprise Survey** is the main source of structural statistics; it supplies many estimates needed for preparing the national accounts. The estimates are produced from data gathered from a sample of enterprises concerning the financial year just ended. The survey covers six broad sectors of economic activity: industry (i.e., manufacturing and related); food and agriculture; construction and public works; transportation; wholesale/retail trade; services. The survey field is defined on the basis of principal activity, number of employees, and legal status. The fields are coordinated to make sure that all the economic activities are covered and that each enterprise is included in just one survey field, without omissions or double-counts.

17. Each sector is handled by a separate survey unit. Four are managed by the survey department of a ministry with authority in that sector; the surveys of the wholesale/retail trade and services are managed by INSEE.

18. **The survey sample is selected** from a sub-set of the overall sampling frame every December. The survey universe contains about 1.8 million enterprises; the sample, about 190,000 enterprises. The sample comprises two groups: (1) an "exhaustive" group of 80,000 enterprises that are surveyed each year without sampling, and (2) a sampled group. Each year, **one-half of the sampled group is replaced.**

19. The entire survey sample selection takes about two weeks. There are three main reasons for the December timing:

---

<sup>5</sup>In reality, the OCEAN sample selection is more complex, for it includes a random-number management procedure and a method for controlled rounding of sample size in each stratum.

- to take advantage of the mass updates of the SIRENE population, which occur in the fourth quarter;
- to obtain a maximum amount of information on enterprises contacted by the survey;
- to meet the scheduling requirements of the surveying units, which want to be able to start their survey in early January.

20. The list of enterprises to contact is forwarded to each surveying unit and also loaded into a data base containing all the enterprises surveyed in EAEs since 1989. This data base is the central node for the circulation of information between EAE partners, i.e., the surveying units and SIRENE. The base itself is inappropriately referred to as OCEAN: in fact, it only stores the information needed to manage the sample, along with a series of key-event dates.

21. At every stage of the survey processing, the OCEAN data base is updated by the surveying units (for example, to record the death of an enterprise listed as living in SIRENE). The information received from respondents may entail an update of the principal activity code. In some cases, this update shifts the enterprise to the field of another surveying unit. When this happens, OCEAN sends a transfer-alert signal to the surveying unit concerned.

22. The updates can create discrepancies between SIRENE and OCEAN information on a given enterprise. An OCEAN procedure identifies these discrepancies and reports them to SIRENE. The SIRENE staff can then conduct an investigation to resolve the inconsistency between the two sources. While this procedure can be applied to several variables, only the discrepancies on the principal activity have been handled so far.

23. Updating the OCEAN data base requires a considerable effort by the surveying units. This investment is fully justified by the resulting benefits, as the surveying units gain a simple, effective access to SIRENE resources. In exchange, SIRENE gains a complementary information source enabling it to identify hitherto undocumented changes. Lastly, the OCEAN data base is the channel for the EAE data used to construct the overall sampling frame, whose quality is improved as a result.



## SAMPLE COORDINATION

24. Thanks to a suitable "management" of the random numbers,<sup>6</sup> OCEAN allows the coordination of (a) consecutive samples of a single survey; (b) samples of different surveys in a given year; and even (c) local-unit samples and enterprise samples. This "management" is sufficiently flexible to meet the needs of most enterprise surveys. The three types of coordination are described more fully below.

(a) **The time coordination of samples of a single survey** is designed to strike a balance between the need to lighten the response burden and the need for robust estimates of change. Obtaining good estimates of changes between two periods requires a large overlap between the samples. However, when the overlap rate rises, so does the response burden, although it becomes more concentrated. Hence the need for a compromise that will minimize the burden while making it possible to calculate reasonably accurate estimates of change. As mentioned earlier, one-half of each survey sample is replaced every year. In fact, this does not mean that only 50% of the enterprises are surveyed again. The selection in the replaced part is not linked to the previous year's half-sample. Consequently, some of the units are re-selected.

(b) The number of surveys that a respondent must answer in a given period is the respondent's "response burden" at that date. **The coordination of samples of different surveys** is aimed at finding the right tradeoff between reducing the burden and obtaining full information on each respondent. By combining information from two surveys, we will obtain a fuller, more detailed picture. This gain does, however, increase the burden for respondents. If the combination of the surveys is shown to be non-essential, every effort will be made to avoid sample overlaps. This is done by renumbering the units: as the selected units carry the lowest random numbers, we change the numbers so as to place the units just surveyed at the bottom of the list, without violating the indispensable stochastic (or probabilistic) properties.

(c) Some parameters also enable us to **coordinate a local-unit sample with an existing enterprise sample**.<sup>7</sup> If the coordination is positive, the local units of a selected enterprise are themselves more likely to be selected. By combining the data on the enterprise with the data from its local units, the user will obtain a broad range of information. But there is the usual downside: a heavier response burden for units of a single enterprise.

25. Two points are worth making before we conclude this section:

- First, the OCEAN methodology exhibits a particularly valuable feature: **it allows the coordination of samples of surveys with different fields and different stratifications.**

- Second, one should always bear in mind the **unavoidable limitations of coordination methods**—whatever the method used. For example, one can never separate two samples selected from the same population if the sum of the sampling rates exceeds unity. Specifically, there is no point in coordinating the "exhaustive" portions of samples. But the presence of an

---

<sup>6</sup>For details of this method, see F. Cotton and C. Hesse, "Tirages coordonnés d'échantillons," INSEE working paper E9206.

<sup>7</sup>Conversely, OCEAN allows the coordination of an enterprise sample with an existing local-unit sample.

exhaustive portion is vital to obtaining economic statistics of acceptable quality. Lastly, the proper treatment of deaths and births of statistical units—always a delicate operation in enterprise surveys—can reduce the user's room for maneuver.

## **V. TOWARD A GENERAL SYSTEM?**

26. OCEAN is not yet a truly all-purpose tool capable of selecting coordinated samples for any type of enterprise survey. Indeed, the selection of samples for certain labor force surveys on local units required the development of a new tool based on rather different theoretical principles.

27. In view of the growing pressure from enterprises—particularly from small and medium-sized businesses—to lighten the response burden, it will be necessary to take account of the "overall" burden. To meet this challenge, OCEAN needs to be remodeled into a general tool for sample selection. The first step will be to reassess the methodology used to select and coordinate samples. Reducing the burden will also require the development of a tool that enables users to prepare efficient sampling plans. Ultimately, OCEAN would become a "toolbox" containing all the resources needed to select samples for most enterprise surveys.

28. As this article has tried to show, the goals are ambitious, and it will certainly take much time to reach the desired degree of generality. We invite the reader to a fresh progress report in seven years' time.