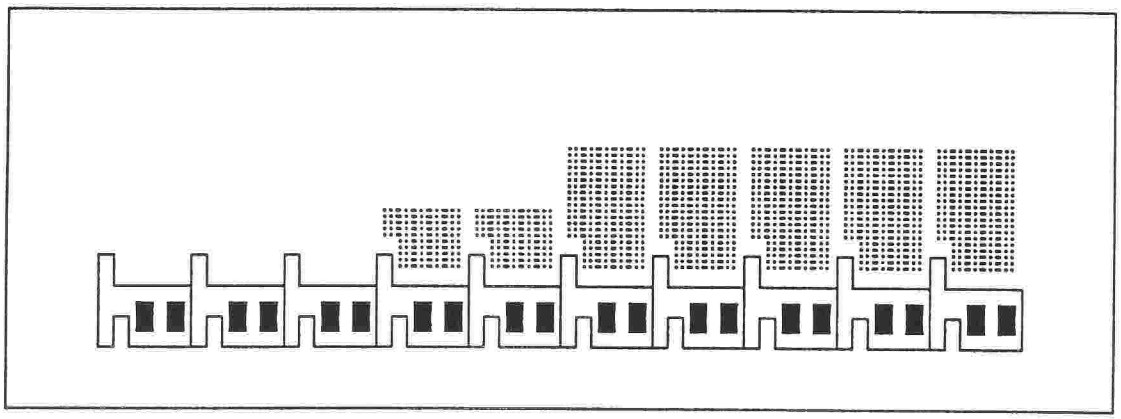
STATISTICS NETHERLANDS

Department of Statistical Methods
P.O. Box 959, 2270 AZ VOORBURG, The Netherlands

EDS, SAMPLING SYSTEM FOR THE CENTRAL BUSINESS REGISTER AT STATISTICS NETHERLANDS

Mila van Huis
Elly Koeijers
Jos de Ree *)



*) The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

BPA no.:4333-94-M1
April 18, 1994

Proj.: M1-89-104
First draft

EDS,
SAMPLING SYSTEM
FOR THE CENTRAL BUSINESS REGISTER
AT STATISTICS NETHERLANDS

Abstract

In January 1993 Statistics Netherlands introduced a new, comprehensive sampling system for business surveys. The system, called EDS, has been developed to spread the overall response burden as evenly as possible across enterprises. The enterprises are stratified on size class and economic activity. The sampling frame is extracted from the Central Business Register. The system is able to maintain rotating panels, taking into account the dynamic nature of the Register. The purpose of this paper is to describe the capabilities of EDS and to mention some aspects of the organization necessary to put the system into production for almost all business surveys at Statistics Netherlands.

Keywords: business surveys, response burden, co-ordinated sampling, sampling frame, stratification, rotating panels

1. Introduction

In January 1993 Statistics Netherlands introduced a new, comprehensive sampling system for almost all business surveys. The system is called EDS, which stands for its Dutch name 'EnquêteDrukSysteem'. The sampling frame is a file extracted from the Central Business Register. The sampling units are enterprises. EDS is a co-project of three departments within Statistics Netherlands. The methodology and a first prototype have been developed at the Department of Statistical Methods. The Automation Department has contributed to the software of the system. The maintenance of the sampling frame and the selection of samples are carried out by the Department of Economic Censuses.

EDS has been developed to spread the response burden as evenly as possible across the enterprises in the Central Business Register. This is achieved by co-ordinating samples and by accumulating a measure of response burden in the sampling frame. Besides, a central sampling frame combined with a central sampling system offers prospects of a better delimitation of the populations for the participating business surveys. Another advantage is the possible standardization of sampling units, which facilitates the comparison of survey results.

Flexibility of a general sampling system is an important condition. EDS allows users to have different designs regarding stratification and allocation. The system is mainly used to draw separate samples, but for several continuous surveys rotating samples are maintained as well. In case of a rotating sample the rotation fraction can be chosen freely by the user. Finally, it is possible to request samples at each time of the year.

Similar systems of co-ordinated sampling are in use at Statistics Sweden (Ohlsson, 1992), at INSEE in France (Cotton, 1989 and Cotton & Hesse, 1992), and at Statistics New Zealand (Templeton, 1990). The Dutch system is unique in the sense that the accumulated response burden is used in the selection procedure.

2. Response burden

Each year Dutch enterprises receive several questionnaires from Statistics Netherlands. The number of times that co-operation is requested can rise up to 40, depending on the size of the enterprise. The concept of 'response burden' is associated with the nuisance, cost, and time spent by an enterprise to complete a questionnaire. Some questionnaires are more time consuming than others. At Statistics Netherlands a substantial amount of data is available on completion times.

In an attempt to make response burden operational, we have introduced six classes. To each class a so called 'RB-value' is assigned to express the nuisance caused by a questionnaire in this class. Each questionnaire was classified according to its (estimated) completion time. We could have used the completion time itself instead of classes, but by choosing this less detailed classification it was easier to achieve consensus on completion times over the various departments of Statistics Netherlands. The classes are given in table 1.

Table 1. Classes according to completion time

class	completion time (min)	RB-value
1	1 - 30	1
2	31 - 60	2
3	61 - 120	4
4	121 - 180	6
5	181 - 240	8
6	241 -	10

In the sampling frame the RB-values are accumulated per enterprise. If an enterprise is selected, then the RB-value of the questionnaire is added to the RB-total of the enterprise in question. For some business surveys at Statistics Netherlands a single selection concerns more than one mailing of questionnaires, shortly called a 'sampling configuration'. In these cases the RB-totals of the selected enterprises are increased according to the

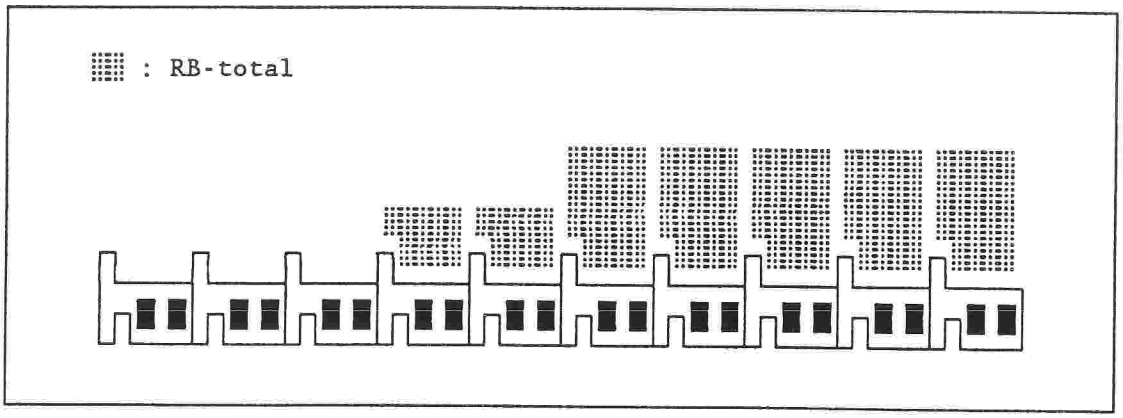
(estimated) total completion time. Summarized, the RB-value expresses the burden resulting from one single EDS sample session, whereas the RB-totals in the sampling frame are the accumulated RB-values per enterprise.

3. The sampling frame

The Central Business Register contains information on different kinds of units (see The Industrial Statistics of the Netherlands, 1992). The statistical units comprise enterprise groups, enterprises, and local units. The sampling frame for EDS is restricted to the enterprises in the Register. In the French system alternately samples of establishments and enterprises can be drawn (Cotton and Hesse, 1992), but EDS only handles enterprises. Each enterprise has been assigned a random number in the interval (0,1). The random numbers are permanent and unique. Per enterprise the following data are recorded in the frame: identification number, permanent random number, the values of size class and economic activity, and RB-total. The sampling frame is also divided into 'basic strata' (explained in more detail in section 4).

Before a sample is drawn, the enterprises are sorted per basic stratum in ascending order with respect to the RB-totals and, in case of equal RB-totals, on the permanent random numbers. Per basic stratum preference is given to enterprises at the beginning of the sequence. Figure 1 displays a basic stratum (after several samples) before the selection of a sample.

Figure 1. Enterprises sorted with respect to RB-totals



The control of response burden in EDS is explicitly connected with the selection stage. Whether an enterprise is willing to respond can not be taken into account. The frame does not act as an administrative system of the actual realized response burden per enterprise. Using the real response burden in the sampling procedure would lead to selection bias, because non-respondents would be overselected. The RB-totals are only instruments to spread the response burden as evenly as possible across the enterprises in the frame.

4. The basic stratification

The users of EDS at Statistics Netherlands generally apply some form of stratification based on size class and economic activity. The composition of their strata is not completely free. EDS uses a standard stratification, the 'basic stratification'. Basic strata are more or less homogeneous groups of enterprises classified by economic activity and by size class. The users build their strata with the basic strata as 'building-blocks'.

An example will clarify the necessity of the basic stratification. Suppose that the very first sample is a take-all sample of bakeries of some size class. Immediately after the selection the selected bakeries are assigned the RB-value of the corresponding survey. Let the second sample be drawn within the same size class from the group of bakeries and butcheries. Let the sampling fraction be small, say 10 percent. Bakeries and butcheries are now one, user built, stratum. If the stratum were sorted on RB-totals and ten percent of the enterprises at the beginning of the sequence were drawn, then obviously an underrepresentation of bakeries in the second sample would arise. For this reason EDS internally uses two separate basic strata, one for bakeries and another for butcheries.

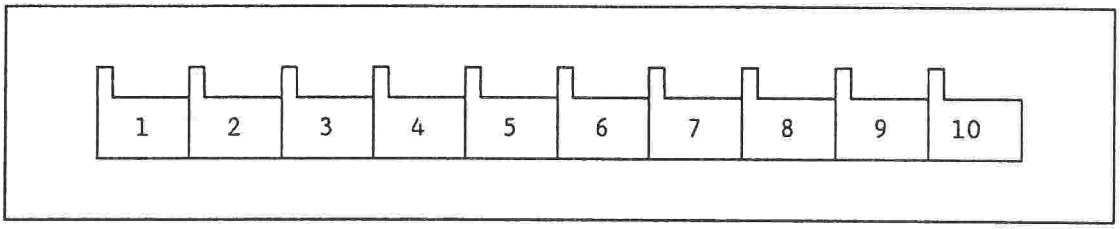
If a basic stratum is split into substrata the rule must be obeyed that the sampling fractions for these substrata are equal. This ensures that the response burden is equally distributed within the entire basic stratum. If users want to combine basic strata there are no specific rules.

5. Sampling

In this section the procedure of selecting separate samples is described. We start the description with the case that each stratum consists of only one basic stratum, then we consider substrata of basic strata and, finally, we combine basic strata. The procedure is in principle 'list sequential sampling' and results in simple random samples without replacement (Sunter, 1977 and Ohlsson, 1992). Rotating samples are treated in section 6.

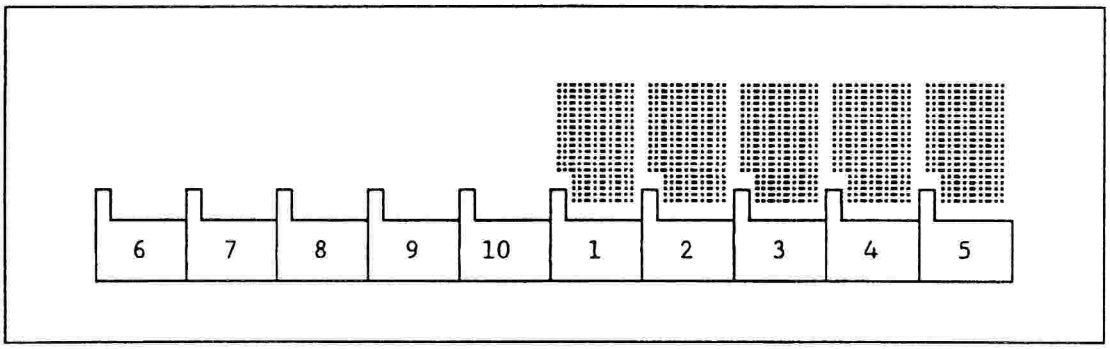
If each stratum consists of only one basic stratum, the sampling procedure is easily understood by a series of small examples. In the sampling frame each enterprise has a unique permanent random number. Before the selection of the very first sample with EDS, the enterprises are sorted according to their random number. Suppose, for example, that a basic stratum consists of 10 enterprises, see figure 2. Enterprise number 1 has the smallest random number and enterprise number 10 has the largest.

Figure 2. Enterprises in random order



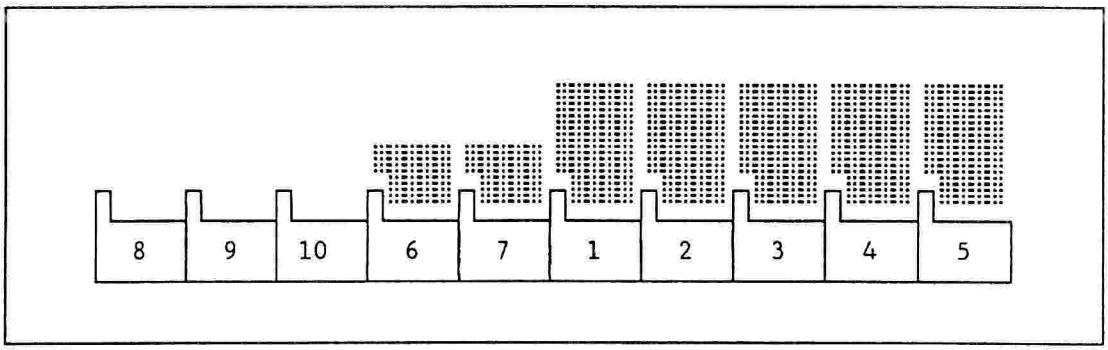
Let the sampling fraction for the first survey in this basic stratum be equal to 1/2. Now the first five enterprises, numbers 1 through 5, are selected. The selected enterprises are assigned the RB-value corresponding to the first survey; let this RB-value be 2. Before the sample for the second survey is drawn, the enterprises are sorted on their RB-totals and, within equal RB-totals, on the permanent random numbers. The enterprises with zero RB-values, 6 through 10, are now positioned at the beginning of the sequence, whereas 1 through 5 are at the end, see figure 3.

Figure 3. Before the second sample



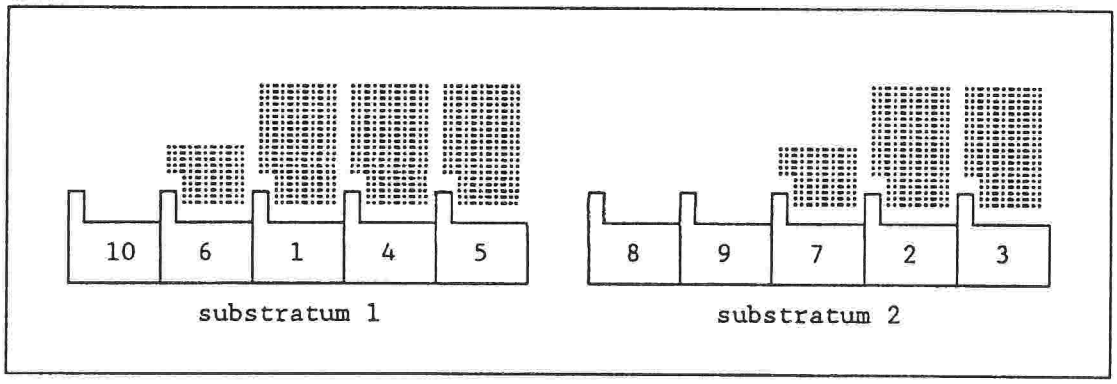
Let the sampling fraction for the second survey be 1/5 and the RB-value of the survey be 1. Now the first two enterprises, 6 and 7, are drawn. Their RB-totals become equal to 1. After this second sample, the enterprises are again sorted according to their RB-totals. Number 8, 9 and 10 are at the beginning of the sequence, because they have never been selected. Number 6 and 7 follow next, because they were selected for a survey with relatively low burden. Figure 4 shows the situation before the third sample.

Figure 4. Before the third sample



We now focus on substrata. As described in section 4, a basic stratum can be subdivided, provided that in the substrata the sampling fractions are equal. For simplicity, we continue the series of examples. Suppose that for the third sample the basic stratum, currently in the state of figure 4, is divided into two substrata, according to figure 5. Let the sampling fractions in both substrata be equal to 1/5. Then the enterprises 10 and 8 are selected. Thus, the sample is evenly spread over the two substrata.

Figuur 5. Division into substrata



Finally, if strata consist of more than one basic stratum, the ordering of the enterprises is not based on the RB-totals, but on the relative position of the enterprises within their basic stratum. If basic stratum h contains N_h enterprises (sorted with respect to RB-totals and random numbers), the relative position of the i-th enterprise is a real number randomly chosen in the interval

$$\left(\frac{i - 1}{N_h} , \frac{i}{N_h} \right] .$$

Per stratum the basic strata are merged and sorted with respect to the relative positions. The enterprises with the lowest relative positions are drawn. After selection the RB-values are accumulated, as usual, and the enterprises are reordered within their basic strata according to their RB-totals. This procedure applied on a stratum consisting of only one basic stratum gives the same result as described above.

6. Rotating samples

In this section we describe the method by which EDS handles rotating samples. We confine the description to the case that each stratum consists of only one basic stratum. For substrata and combinations of basic strata we refer to section 5.

Sample rotation is sometimes used as a tool for spreading the response burden. The main purpose, however, should be to provide samples which reflect the changing structure of the population. Partial replacement of the sample, instead of taking a completely new sample, is usually a good strategy to estimate both population totals and changes (Hidiroglou and Srinath, 1993). Another reference to recent research in methods of sample rotation for business surveys is Hidiroglou, Choudry and Lavallée, 1991.

For each survey using a rotating sample EDS creates a file containing the identification numbers of the enterprises that were in-scope for the survey at the time of the selection of the previous sample. In this section 'previous' refers to sampling for the same survey. In the file for each enterprise the inclusion probability at the previous occasion is recorded. A dummy variable indicates whether the enterprise was included in the sample.

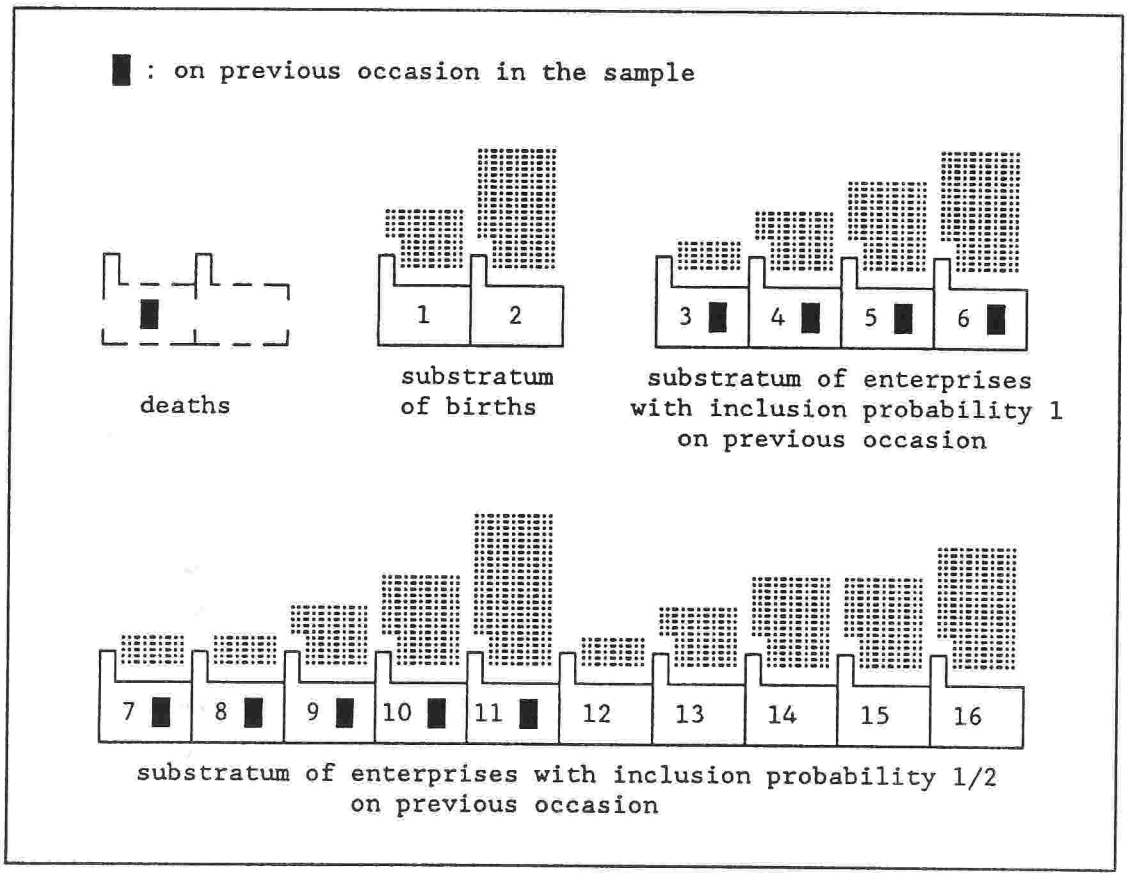
Before a rotating sample is drawn, the current version of the sampling frame is divided into substrata. Firstly, births or new enterprises are considered. A 'birth' is defined as an enterprise born in the period of time between the current selection and the previous one. Births are recognized by matching the current frame and the file created for the survey in question. Within each basic stratum of the current sampling frame a substratum of births is formed. Besides, a 'death' or defunct enterprise can be defined analogously to a birth. As deaths are not present anymore in the current frame, they are not considered in the selection process.

Apart from the births, each basic stratum is divided into substrata consisting of enterprises with equal inclusion probability at the previous occasion. Again the match of the sampling frame and the file is used. These substrata are created to prevent the current sample in this stratum from under or over-representation of enterprises entered from other basic strata in the period between the current and the previous selection.

An illustration of the division into substrata is given in figure 6. This example concerns a basic stratum containing 16 enterprises, numbered 1 through 16 for simplicity. The numbers 3 through 6 originate from one or

more basic strata in which the enterprises were selected with certainty on the previous occasion.

Figure 6. Example of substrata within a basic stratum



For each substratum within a basic stratum, including the substratum of births, the sample size is determined by multiplication of the sampling fraction and the size of the substratum. Subsequently, for each substratum the number of 'fresh' enterprises is determined by multiplication of the rotation fraction and the computed sample size. As the resulting numbers are usually broken, controlled random rounding (Fellegi, 1975) is applied to get fixed integer values for the sample size and the number of fresh enterprises. This sampling procedure is also described in De Ree (1983).

The rotation fraction determines the fraction of enterprises that will be refreshed with respect to the previous sample. Any rotation fraction in the

interval [0,1] can be chosen. If the rotation fraction is 0, the resulting sample will have a maximum overlap with the previous one, whereas the value 1 will result in a minimum overlap. In each substratum the enterprises with the lowest response burden are selected. Fresh enterprises are drawn from the group where the dummy variable in the file indicates that they were not in the sample on the previous occasion.

A rotation fraction of, for example, 1/5 means that, roughly, one out of five enterprises will be refreshed. Refreshing is carried out within the substrata and therefore sometimes restricted by practical constraints. For example, if the rotation fraction is 1/5 and a certain substratum does not contain fresh enterprises, then no refreshing can take place. This is the reverse of the situation in the substratum of births, where all enterprises were not in the sample on the previous occasion.

To illustrate the selection process, again the example from figure 6 is used. Suppose the sampling fraction is 1/2 and the rotation fraction is 1/5. Now, from the substratum of births number 1 is selected. In the substratum with inclusion probability 1 on the previous occasion the numbers 3 and 4 are selected. In this substratum refreshing is impossible. Finally, the numbers 7 through 10 are selected. They were already in the previous sample and according to their RB-totals, they should be selected. Number 11 is rotated out of the sample and is replaced by number 12.

After the selection of a rotating sample the RB-totals of the selected enterprises are adjusted, as described in section 5. Furthermore, a new file for the survey in question is created containing information on the in-scope enterprises. This file can be retrieved when the subsequent rotating sample is drawn.

7. The updating of the sampling frame

The Central Business Register continually changes due to births, deaths, classification changes etc. Therefore it is necessary to update the sampling frame regularly. This updating consists of two phases. First the Department of Economic Censuses determines which enterprises are born, dead or have changed classification. Then these changes have to be implemented in the sampling frame.

It should be noted that, using EDS as a sampling system, information from samples must be carefully handled to update the frame. Only information from strata that are completely enumerated, can be implemented in the sampling frame. Information from take-some strata can be used for updating the frame as soon as this information is confirmed by an independent source. In Hidiroglou, Choudry & Lavallée, 1991 and Hidiroglou & Srinath, 1993 this issue is also discussed.

EDS processes the changes in the sampling frame as follows. Deaths are eliminated from the sampling frame. Births are randomly inserted into the basic stratum to which they belong. Thus, births obtain a random position. They are assigned an artificial RB-total suitable to this position. This is done to prevent selection bias, otherwise each birth would be immediately selected into the subsequent sample.

When an enterprise changes from one basic stratum to another, it keeps its relative position, but its RB-total is adjusted. Otherwise, an enterprise moving from a basic stratum which is seldomly sampled to a stratum which is frequently sampled, will be at the beginning of the sequence in the new basic stratum due to its relatively low RB-total received in its former basic stratum. The RB-adjustment again prevents selection bias.

As regards births and deaths, the updating of the sampling frame takes place once every month. In conformity with the policy of the Department of Economic Censuses, changes in classification are implemented only once a year. At the end of the year two versions of the sampling frame are constructed, one without the implementation of classification changes and

one with. For quarterly surveys there are versions of the sampling frame every three months, containing all enterprises that have been active during the past period. Furthermore, for annual surveys, each year a version of the sampling frame is constructed which contains all enterprises that have been active during the past year. Altogether, each year there are eighteen versions of the sampling frame.

The various versions of the sampling frame succeed each other in time. The accumulated RB-values are passed from one version to the next. Therefore all samples from a certain version have to be drawn before the subsequent version is installed.

8. Organizational aspects

The Department of Economic Censuses maintains the sampling frame of EDS and selects the samples. Because of the successive versions of the sampling frame, strict planning is necessary. Each year an inquiry is made within the economic departments as to which samples are to be drawn with EDS. This investigation results in an annual scheme of successive samples.

A short time before a sample is planned, the user transmits the stratification and the sampling fractions via the internal PC-network of Statistics Netherlands to the administrator of EDS. To this aim a standard application form is available on every PC. Because of the various versions of the sampling frame, it is important that the administrator receives the information in time. Files of previous sampling sessions, necessary for rotating samples, are kept by the Department of Economic Censuses.

After the sample has been drawn, the user receives a file containing all enterprises that were in-scope for the survey. In this file for each enterprise two variables are recorded: the identification number and the value of a dummy variable, which indicates whether the enterprise was selected or not.

References

Cotton, F. and C. Hesse, 1992, Co-ordinated selection of stratified samples. Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys, November 1992.

Cotton, F., 1989, Use of SIRENE for enterprise and establishment statistical surveys. 4th International Roundtable on Business Survey Frames, Newport, Gwent. UK. 9th-12th October 1989.

De Ree, S.J.M., 1983, A system of co-ordinated sampling to spread response burden of enterprises. Contributed paper, 44th Session of the ISI, Madrid, pp 673-676.

Fellegi, I.P., 1975, Controlled random rounding. Survey Methodology, 1 (2), pp. 123-133.

Hidiroglou, M.A., G.H. Choudry and P. Lavallée, 1991, A sampling and estimation methodology for sub-annual business surveys. Survey Methodology, 17 (2), pp. 195-210.

Hidiroglou, M.A. and K.P. Srinath, 1993, Problems associated with designing subannual business surveys. Journal of Business & Economic Statistics, Vol.11, No. 4, pp. 397-405.

Ohlsson, E., 1992, SAMU, the system for co-ordination of samples from the business register at Statistics Sweden. Statistics Sweden, R & D Report 1992:18.

Sunter, A.B., 1977, List sequential sampling with equal or unequal probabilities without replacement. Applied Statistics, 26 (3), pp. 261-268.

Templeton, R., 1990, Poisson meets the New Zealand business directory. The New Zealand Statistician, 25(1), pp.2-9.

The Industrial Statistics of the Netherlands, 1992, Publication of the Department for Statistics of Manufacturing and Construction, Statistics Netherlands.