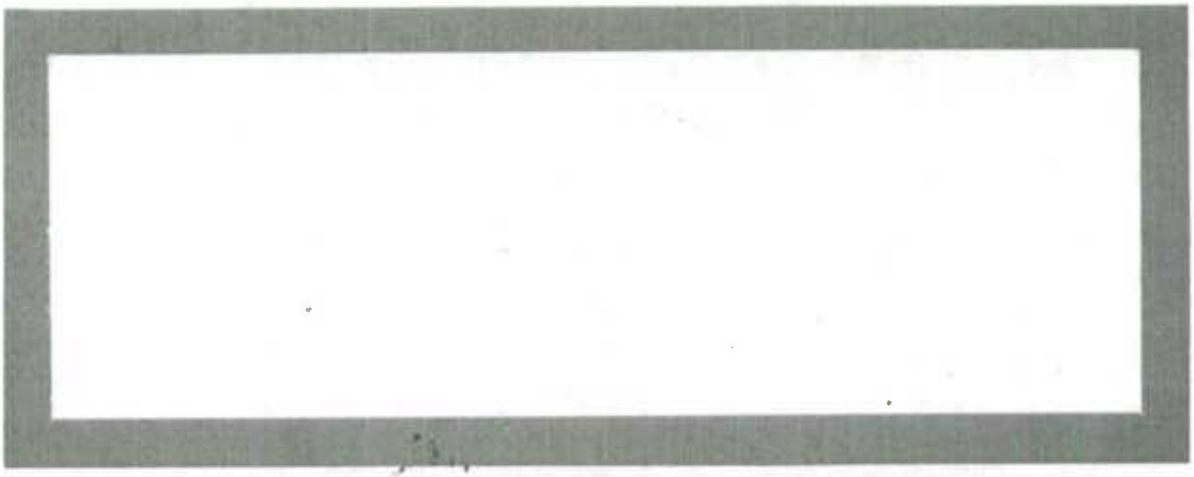
Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-98-007E/F
METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-98-007E/F
DIRECTION DE LA MÉTHODOLOGIE

Overview and Strategy for Version 1.2 of the Generalized Sampling System

G.B. Faber, N. Laniel and D.M. Yeo

June 1998

Overview and Strategy for Version 1.2 of the Generalized Sampling System

G.B. Faber, N. Laniel, D.M. Yeo

Sommaire

Le Système généralisé d'échantillonnage (SGECH) prévoit des fonctions fondamentales d'échantillonnage permettant de concevoir, de prélever et de mettre à jour des échantillons. Il s'agit d'un produit en base SAS conçu pour faciliter diverses fonctions de sondage dans le cadre d'enquêtes périodiques ou uniques. Comme le SGECH est interactif, on peut examiner différentes stratégies d'échantillonnage avant d'appliquer la meilleure. Grâce aux fonctions modulaires SGECH de stratification, de répartition, de sélection, etc., on peut réagir rapidement à l'évolution des besoins d'une enquête. Ajoutons que le SGECH fournit un cadre intégré d'actualisation de la base de sondage.

1 INTRODUCTION

Many steps are involved in the survey process. These include designing a questionnaire, creating a frame, selecting a sample, contacting and collecting information from respondents, editing and imputing data, weighting the data, and tabulating and releasing survey results. At Statistics Canada, a suite of generalized systems has been developed so that it can be used to handle several steps for a particular survey and so that a particular generalized system can be used for many surveys.

The Generalized Sampling System (GSAM) was developed to facilitate sample design, selection, and maintenance in periodic surveys or in surveys that are carried out only once. In addition to GSAM, systems have been developed for data collection and capture (DC2), automated coding (ACTR), edit and imputation (GEIS), estimation (GES), record linkage (GRLS), and time series analysis (GTSP—in development).

After the creation of a population data set (frame), GSAM can be used for each step of the sampling process (stratification, allocation, sampling, and frame maintenance). The sampling process can be viewed as having a design stage and a production stage. At the design stage, the user may determine the best method to stratify the frame and the best sample allocation for it. Both of these processes may take some experimentation. After allocation is done, regular production can occur and a sample will be drawn. For ongoing surveys, the frame may need to be updated. Units may have been added or removed from a population and information on units can change over time. Once a frame is updated, new stratification and allocation can be performed

if necessary and a new sample can be drawn. This sample is dependent on the previous sample if the user controls the sample overlap between sampling occasions. For instance, the user may wish to have units retained in sample for six months on a monthly survey, refreshing only one sixth of the sample each month. After a new sample has been chosen, the user may identify changes to the sample (e.g., births, new in-sample units, units rotated out of sample).

Section 2 of this paper gives an overview of GSAM. First, generalized systems are briefly discussed, along with their advantages. Then each major part of the system (stratification, allocation, sampling, and frame maintenance) is described. Definitions and concepts are provided, followed by a discussion of the main functions. Section 3 of this paper presents some sampling strategies that can be used when developing applications with GSAM.

2 OVERVIEW

2.1 Generalized Systems

A generalized system is a set of modular programs used for a variety of similar applications. By using common approaches and sound methods for surveys, generalized systems offer the user a "toolkit" that provides the individual application with considerable flexibility and a high degree of component reuse. Statistics Canada has developed generalized software packages to conserve resources and reduce the duplication of effort that arises from tailor-made systems.

Generalized systems are expensive to develop, but in the long term their use saves resources by reducing the need to develop customized systems and maintain many systems of a similar nature. Resources are also saved by concentrating effort on the process of developing and maintaining a common set of software. Additionally, it is easier and faster to update this single set of software and its documentation. Thus, better support is offered to the user.

GSAM has been built in a modular fashion. Each module corresponds to one of the major functions of the system. Users can specify all of the GSAM functions or only some of them (e.g., allocation only). Part of the reason SAS was chosen as the underlying software was because of its power, portability, and ease of use. SAS is also well known by many survey designers at STC, and offers many functions to make any statistical analysis easier.

2.2 Frame Maintenance

Initially a frame will need to be created. For a periodic survey, this frame occasionally needs updating. Units may be added or removed, and characteristics of units may change. For example, new information may show that a unit should be reclassified into a different stratum.

Many surveys have separate processes to handle their frame maintenance and sampling requirements. GSAM, however, has the ability to make changes to the existing frame. When only minor changes are required this can be more convenient than creating an entirely new frame and importing it into SAS.

Aside from the options that can be used to add units, delete units, and update information on units already on the frame, GSAM has two additional options. Using one option, a new frame can be created by altering an existing frame. The new frame contains either all or some observations from the input frame. To select only some observations, the user supplies a list of the ones to be selected. The list can be in the form of a data set or a subset of a data set. The other option, analytical in nature, allows the user to compare variables from two frames and identify changes such as new units.

2.3 Stratification

Prior to sampling, population units can be grouped into non-overlapping strata. Stratification will help ensure that all sectors of the population are represented in the sample and that estimates derived from each sector are accurate, especially for groups that are of particular interest. Stratification may also be done for the sole reason of generating the most precise estimates for a given sample size. Typically, estimates are more precise if homogeneous population units are grouped together, and if these groups differ significantly from each other. Units may also be stratified for administrative reasons, such as distributing the workload among regional offices.

With GSAM, users can develop a set of stratification rules to apply to their population file. Each rule defines one stratum and is formed by a compound set of SAS logical expressions. The rules are applied to each unit in the population to determine to which stratum each unit belongs. Each unit must belong to a single stratum. GSAM can be used to develop rules that will do this without error. However, some caution is required should users input their own rules. For instance a user could type in two rules and a unit could satisfy both of them. For example, a unit with a REVENUE value of 500 would satisfy both the expression "REVENUE LE 500" and the expression "REVENUE GE 500 and REVENUE LT 10000". (In this case the unit would be assigned to the stratum represented by the rule that was defined first.)

In making these rules, the user can explicitly specify the SAS expressions to be used or can have GSAM determine possible combinations of user-specified variables (e.g., province by industry by size). Also, the user can have GSAM determine appropriate stratum boundaries based on any numeric variable of interest. These boundaries can be chosen using either the Dalenius-Hodges cumulative$\sqrt{f}$ method (Cochran, 1977, pp. 128-130) or a clustering method. The first method gives an approximately minimum variance stratification under Neyman allocation based on a known auxiliary variable. The latter method creates strata containing similar units by forming clusters of units with minimum dispersion (based on the target variable). The user will specify the maximum number of clusters that can be formed for a given set of units. Using this information and the data, the system will determine how many clusters should be formed, each cluster comprising one stratum.

Rules can be edited: expressions can be added to or removed from existing expressions, and strata can be collapsed. GSAM reports can be generated anytime, showing what will result if the stratification rules, as specified, are applied to a particular data set. Statistics such as sums, counts, means, variances, and coefficients of variation (CVs) can be generated for variables and strata selected by the user. These reports help the user decide which rules should be changed or whether the current stratification rules are correct.

2.4 Allocation

Allocation is the determination of the number of population units that should be sampled from each stratum. Often, stratum sample sizes are allocated in a certain way to meet a particular user requirement. Common choices include meeting a fixed cost, fixed sample size, or fixed CV level for some variable of interest. Allocation takes into account the within-stratum variation and the cost of obtaining information in each stratum. If the unit cost in each stratum is constant, then strata with more variation will get a larger allocation. If the cost varies per stratum but the within-strata variations are about the same, then it is generally more efficient to sample more from strata where it costs less to do so (Murthy, 1977).

In GSAM, there are two options available for a stratified one stage element design under simple random sampling without replacement (SRSWOR): minimizing a weighted variance function or minimizing the total sampling cost. Both allocation problems are solved using an algorithm proposed by Bethel (1989) and extended by Estevao (1993). Different parameters and constraints are specified under each option.

To minimize the weighted variance function, a sample is allocated subject to a fixed total sampling cost or a fixed sample size. The allocation requires one user-specified variable and optionally a second, auxiliary, variable to be specified. If an auxiliary variable is specified it should be highly correlated with the first specified variable. The general method used under this option is a power allocation; proposed by several authors including Bankier (1988). Depending on how the parameter of the power allocation is set, one may obtain a Neyman allocation, an approximately equal CV allocation, or a compromise between these two. It is also possible to set the auxiliary variable so that the allocation is proportional to the population size or to a variable at a certain power. GSAM either finds the optimal allocation or determines that no allocation is possible within the given constraints.

Constraints, such as the unit sampling cost and the minimum and maximum allowable sample size in each stratum, can be specified. If the unit cost information is not available then it is set to one in each stratum. The objective of trying to meet a fixed cost is then equivalent to trying to fix the total sample size. Default values, which the user can modify, are used when the stratum-level minimum and maximum allowable sample sizes are not specified.

For minimizing the overall sampling cost, an optimal allocation method is used. This method uses constraints such as limits on CVs for user-specified variables. The user can choose a different CV limit for each variable. GSAM either finds the optimal allocation or determines that no allocation is possible within the given constraints. As with the previous option, the unit cost and the minimum and maximum allowable sample sizes for each stratum can be specified. Defaults will be used where this information is not supplied.

For both allocation methods, users have the option of supplying their own statistics file instead of using the frame as the input. The system will use any statistics supplied, such as stratum variable counts/totals/variances, instead of those calculated from the population data set. In fact, if all the statistics are supplied for all the strata then a population data set does not have to be specified for Allocation.

2.5 Sampling

Sample surveys are often taken, rather than a census, for a number of reasons, including reduced costs, improved timeliness, and lowered respondent burden. Many methods of selecting a random sample have been conceived; currently in GSAM stratified simple random sampling designs can be handled. Other sampling methods such as two-phase, Bernoulli sampling, and probability proportional to size will be added in the future.

The sampling function offers several useful features for selecting a simple random sample without replacement and maintaining it over time. Users can update their sample for additions or deletions to the frame as well as for changes in the stratification of the population. They can also introduce rotation to reduce respondent burden and rebalance the sample when too many changes have occurred in the population. The user has a choice of using the stratified population file and the allocation file from GSAM, or independently generating in SAS either or both files.

GSAM uses collocated sampling (GSAM Development Team, 1996). Units are randomly assigned selection numbers that are equally spaced on a [0, 1] interval within each stratum. A sampling window is determined for each stratum; units with selection numbers that fall within the window are sampled. Should the stratum be a take-all stratum, the whole interval is covered by the sampling window.

To maximize sample overlap, GSAM uses a method that closely mirrors the Kish-Scott approach (Kish and Scott, 1971), but in the collocated sample context. Sample overlap between two survey periods can be controlled using rotation. Selection numbers remain the same between sampling occasions while the sampling window shifts a certain distance to the right. This distance is based on user-specified parameters. For instance, a user can specify that units should be in sample for twelve months (with rotation one twelfth of the sample should be refreshed each month) and out of sample for at least twenty-four months, while also respecting the desired sampling fractions. In this way, a user can specify as high a sample overlap between periods as desired. This overlap is controlled at the stratum level. The following rule is applied to determine the distance, $shift_h$, a window should be shifted for stratum h which uses rotation:

$$shift_h = min\left(\frac{f_h}{time\text{-}in_h}, \frac{(1-f_h)}{time\text{-}out_h}\right)$$

where f_h is the sampling fraction for stratum h (the number of sample units in stratum h divided by the number of population units in stratum h).

To minimize respondent burden, GSAM attempts to keep units in sample the exact number of periods specified in the time-in constraint and out of sample at least as the number of periods as specified in the time-out constraint. For example, consider a monthly survey where a given stratum has a sampling fraction of .3, a time-in constraint of 24 months and a time-out constraint of 12 months. This results in the shift calculation, min (0.3/24, (1-0.3)/12), or 0.0125; i.e., each month the sampling window for this stratum will be shifted only by this very small amount,

0.0125. A small shift ensures that a sampled unit will remain in sample for the specified 24 months and then will be kept out of the sample for a long period of time (in this instance, for 56 months).

The above formula accounts for the possibility that the specified time-in and time-out parameters cannot both be met. In GSAM the time-out constraint is always met or exceeded. For instance, all units cannot be in sample for no more than 12 months and then out of sample for at least 12 months in a stratum where 90% of the units are to be sampled. In this situation, the unit will be kept in sample for longer than 12 months, but once it is rotated out of sample, it will remain out for the specified 12 months. It is considered less burdensome to stay in sample for a long period of time then to be rotated out and then back in very quickly.

GSAM allows users to update their sample for changes in sampling rates or in stratification. The approach used to handle these changes consists of using the previous-period selection numbers, based on the old stratification, to determine which units will be in the current-period sample with the new stratification. The goal of the approach is to retain as many sample units as possible from the previous period. When in-sample units of the previous period are dropped, those that have been in the sample for the longest time are the first ones to be dropped. This approach also allows for rotation while resampling.

For example, if the sampling rate has increased, and rotation is turned off, the new sample will consist of all the previously-sampled units, and some new ones, thus maximizing sample overlap. Should rotation be specified, the sampled units that have been in the sample the longest will be replaced by units that have been out of sample the longest. If some units have changed strata, or if a restratification has been done, GSAM moves each reclassified unit into the new stratum, placing it in the same location relative to the sampling window of that stratum. In this way GSAM ensures maximum sample overlap for restratified units.

The selection of new units on the frame is done by giving each unit a selection number that is equally spaced within each stratum. Then they are selected, independently of old units, using the same method as for selecting an initial sample.

One drawback to adding units, dropping units, and having units change strata is that the units' selection numbers may no longer be equally spaced within each stratum. Although $f_h \times 100\%$ of the units within a particular stratum are to be sampled, more or fewer units may be in the selection window. GSAM provides a rebalancing option that can be specified to stabilize the sample size. When rebalancing is specified, the units' selection numbers are adjusted to make them equally spaced within each stratum once again. Then the proper number of units will be selected.

3 STRATEGY

This section of the paper provides a general outline of how users can take advantage of the various modules of GSAM to suit their various application needs. To this end, a description of the computing environment is given, followed by specific tips for using the stratification, allocation, sampling, and frame maintenance functions in GSAM. For a more detailed description of the functions, please refer to GSAM Development Team, 1996.

Once an appropriate methodology has been identified, GSAM can be used simply by providing the necessary information. Usually, all GSAM functions will be used together; in this case, the output from one function will become the input for the next function. However, due to the modular nature of GSAM, users can choose only the functions they need. To skip a function, the user must create information that would have been generated by the skipped function and pass it on to the next function. Whether the information comes from outside the system or is generated by GSAM, GSAM always checks that the information supplied to a module is always checked for consistency.

3.1 Computing Environment

GSAM is a SAS®-based application for IBM®-compatible microcomputers. The current version, GSAM Version 1.2, runs under SAS® (version 6.08, 6.10, or 6.11) for Windows™ version 3.1, Windows for Workgroups™ version 3.11, or Windows NT ™. Access to SAS/BASE® software, SAS/FSP®, SAS/IML®, and SAS/STAT® is required to run GSAM. GSAM functions are accessed in an interactive SAS session using a set of menus. In each session the regular SAS program editor, log, and output windows are available. GSAM reports are displayed in the output window and can be saved using standard SAS commands. Users will benefit from a basic familiarity with SAS. Please see SAS Institute Inc. (1985) for a good reference to SAS.

A 486 or better microcomputer with at least 8 MB of memory is recommended. The software requires about 70 MB of hard disk space, mostly for SAS. These programs and all user files, the total size of which will vary for different applications, can be installed on the microcomputer itself or on a file server for access through a local area network.

The GSAM modules are invoked though the use of a menu system. While GSAM performs a verification of the input parameters, the actual sequencing of the functions and their parameters is at the user's discretion. As such, it is assumed that the user not only understands the subject matter, but also is familiar with the basic concepts of stratification, allocation, sampling, and frame maintenance. Although some steps have been taken to warn the user of potential pitfalls, it is impossible to ensure that the system will not be used incorrectly. Therefore we would suggest that subject matter staff create their GSAM applications with the aid of methodologists.

3.2 Design Stage

Stratification

As stratification rules are being made, GSAM reports can be produced frequently to examine the strata that would be created if the rules were applied to a specific population file. Statistics such as sums, counts, means, variances, and expected CVs are generated for variables and strata specified by the user. These reports may help a user decide which rule changes, if any, should be made. For instance, a user may discover that some defined strata are too large or that some strata have too few units and need to be combined with other strata.

Users often stratify on relevant size measure variables to increase the efficiency of their sample or to reduce costs dramatically. Thorough stratification by size will generally achieve many of the benefits of Probability Proportional to Size (PPS) sampling and is generally the recommended approach. Individual size measures are often not extremely stable (especially for business surveys) and stratification by size will often produce more efficient samples than PPS sampling.

The clustering method is recommended for grouping similarly-sized units. For example, consider a population that is initially grouped into strata by province and industry. Then using the clustering method each province by industry group can be further stratified using a relevant size measure, such as revenue. Some experimentation involving the user-defined maximum number of clusters to be formed for each province by industry group is then required. The optimum stratification can vary depending on the CV or total sample size specified.

Large units are often grouped into take-all strata. While this is often done to reduce the variance, these units are typically large enough that they should be self-representing. If a large unit was placed in a take-some stratum, it may or may not be sampled. If it were not sampled the final estimates may be low and if it was sampled its characteristics probably are different enough from other units that the complete data set is hard to model. More detailed information on specifying take-all strata is given below.

Stratification procedure to simplify the construction of an Allocation constraints file

After stratifying by size, it will often be desirable to specify some strata as take-all. This is done by providing Allocation with a constraints file. The use of the clustering algorithm in Stratification, together with the specification of take-all strata for Allocation, can result in a greatly reduced CV or sample size.

The constraints file contains one observation per stratum; user-defined default values will be applied to strata not specified on the file. Four variables are stored: the stratum identifier, the unit cost for sampling, and the minimum and maximum sample sizes.

It is easier to make a constraints file if stratification is done first by size (e.g., take-all/take-some by province by SIC). Then the stratum identifiers assigned to take-all strata will be together at the start or at the end of all those identifiers generated. If the user's only restrictions are that

particular strata be take-all, specifications are required only for these strata; knowing the stratum codes for these strata facilitates the creation of the constraints file.

To allocate all units from a stratum, a very large number should be assigned to the minimum sample size variable in the constraints file for each stratum to be made take-all. GSAM will automatically convert this number to the population size of each stratum specified. If a specific number of units are to be sampled from a particular stratum, say 76, then the minimum and maximum sample sizes are both set to 76.

Only two units need to be sampled from each stratum to calculate a variance. However, it may be more appropriate to take at least five units from each stratum (even; for example, in a take-some stratum of a business survey where units have little revenue) to reduce the effects of nonresponse. The user can ensure that the minimum sample size stored on the constraints file is five for each take-all stratum, but it is simpler to specify a default value in Allocation.

Allocation

Allocation may need to be run several times. When the sample cost is minimized subject to fixed CV constraints, the user may find, for instance, that the calculated sample size is too large. At this point the user must relax some CV constraints, rerun the allocation function, and see if the sample size is now acceptable. The user can rerun this function as often as required to obtain satisfactory results. In fact, the allocation function can be used to test different strategies, e.g., the resulting sample sizes if a specified CV limit is set at 2%, 5%, etc.

The allocation can also be run with various power function exponent values. Specifically, Neyman allocation is realized by using the target variable as the auxiliary variable and by setting the power exponent to one. This allocation should be used if the user wishes to minimize the overall CV, while leaving the stratum CVs uncontrolled. To achieve approximately equal CVs across strata, given that the ratio of the variance to the mean does not vary significantly over strata and the sampling fractions are small, the power exponent should be set to zero for any auxiliary variable. This allocation will minimize stratum CVs, but at the cost of a higher overall CV. Compromise allocations can also be specified; exponent values of 0.50 (the square root allocation) or of 0.33 (the cube root allocation) are not unusual. These compromise allocations can achieve significant reductions in stratum CVs with relatively little increase in the overall CV (Bankier, 1983).

When minimizing the sample cost subject to fixed CV constraints, the constraints are set at the population level (e.g., 5% CV for 'revenue' at the Canada level). However, it is possible to set marginal CV constraints, (e.g., a 6% CV for 'revenue' for Ontario and a 3% CV for 'revenue' for a stratum made up of large enterprises in Atlantic provinces). To obtain a 6% CV specifically for 'revenue' in Ontario, a new variable is created on the stratified population file. The corresponding 'revenue' value for each observation in Ontario is copied to this new variable. For each observation outside Ontario, the value stored is zero. A 6% CV constraint is then specified for this new variable. Currently the user can specify up to 22 CV variables in GSAM. Users should contact the Generalized System Methods Section, BSMD, to obtain a macro should it be required to specify more than 22 variables.

Because of budgetary constraints, the total sample cost or sample size is often fixed. The option that minimizes a variance function subject to a fixed cost is then the proper choice. When working with a fixed sample size, as opposed to a fixed sample cost, the unit cost for sampling in each stratum should be set to one (this makes the option of fixing the sample cost equivalent to fixing the total sample size).

If the allocation algorithm does not converge, it may generate a solution that does not meet the target specification. For instance, the resulting sample size may be smaller than that fixed when using the minimum variance option. Alternatively, when using the minimum cost option the resulting sample size may be inappropriate because targeted CV levels were not met. Should such problems occur, the parameter that controls the number of iterations used in the Bethel algorithm should be increased. Also, the strata should be checked for homogeneity. Further stratification may be needed, perhaps by size; there may be very large units that should be grouped separately and specified as take-all.

Rotation

Rotation generally should be used for ongoing surveys, to prevent response burden from always being placed on the same set of respondents and to sample a greater part of the population over time. Users should also examine the rate of rotation, for several reasons. Slower rotation maximizes the sample overlap, allowing for a more accurate measurement of trends. The effort required from respondents is often less on subsequent sampling periods. Also, collection costs are often less on subsequent occasions. On the other hand, faster rotation will reduce the length of time a unit spends in the sample, reducing respondent fatigue.

Note that the time-in and time-out parameters can be set to different values for each stratum. In this manner, for example, the user can regulate the respondent burden placed on small companies or on rural households. The user can determine how long each sampled unit (business, household, etc.) is expected to be in the sample and then inform the respondent of the maximum amount of time they will be in sample.

3.3 Production

Initial period in GSAM

A population data set, also known as a frame file, must exist prior to sample selection. The user may import this file directly into SAS. If this file covers a greater population than desired, GSAM can be used to create a subset of this larger file. Another benefit of using the *create* file option available in Frame Maintenance is that a variable is then created to indicate which units were sampled in the previous period, allowing rotation to be implemented if desired. For this reason, the *create* file option is recommended when it is known that rotation will be implemented.

To select a sample for a new survey, the *new* sample selection option must be specified. If an ongoing survey is being brought into GSAM, the user has two choices. If the user wishes to draw a sample independently of any previously drawn samples, then no previous information is required and again the user must specify the *new* sample selection option. Alternatively, the user may wish

to control the sample overlap between sampling occasions. To do this the user must create, in GSAM or outside the system, a parameters file and a frame file for the previous period. Further details on this approach are found in Faber, 1997.

Subsequent periods

Frame Maintenance

Frame maintenance is needed for most periodic surveys, whether they are conducted monthly, quarterly, annually, or at irregular intervals. GSAM options (*add*, *delete*, and *update*) provide an alternative to creating and importing an entirely new frame when changes are required. If only minor changes are required to a large frame file, such as some units need to be added, then the *frame maintenance-add* option can be used to save time. Here, just a list of the new units with their corresponding information is needed.

In an ongoing survey, new information on the frame may show that units would change strata if the stratification rules that were used in the previous sampling round were reapplied. The user must decide if new stratum values should be assigned to previously existing units or if the original values should be retained. If it is expected that these units could change strata again on the next sampling occasion, it may be wise to leave them in their original strata.

Caution should be used before restratifying or removing units from a frame. A user may know that a unit is misclassified or bankrupt, for example. To avoid bias, only sources of information independent of the sample process should be used to update the frame. If the user's knowledge comes from the attempt to survey a sampled unit, then the unit should normally remain, unchanged, on the frame. Such a unit can represent other units that have been misclassified or have "died" but are unknown to the user because they were not selected for sampling.

Sample Updating

If the user wishes to control the sample overlap between sampling occasions then the *update* sample selection option should be specified. Using rotation, users can specify for each stratum the number of survey occasions sampled units remain in sample and the number of survey occasions they should be kept out once they have rotated out. To maximize sample overlap, a relatively long time-in constraint should be specified. The choice of time-in and time-out constraints should be tempered, however, by the need to minimize respondent burden. Further, with the *update* option GSAM will automatically maximize sample overlap for units that change strata, whether because of reclassification or because of a change in the stratification rules (e.g., collapsed strata). Should a user want instead to draw a sample independently of any previously drawn samples, then the *new sample selection* option should be used.

Note that if units have been added, dropped, or have changed strata, units will no longer be equally spaced—a 17% sampling window may not cover 17% of the units within a particular stratum. If the frame has changed greatly then rebalancing should be specified to ensure that the correct proportion of units within each stratum is sampled. For a monthly survey where the frame

is updated once a year, the user would only want to rebalance the one month each year when frame updates are made.

Rebalancing only occasionally is recommended because rebalancing allows units to enter and leave the sample in a less-controlled manner. For example, a user may want a sampled unit to be in the sample for up to twelve continuous months and once this unit has been rotated out to remain out for at least twenty-four continuous months. Rebalancing will cause these constraints to not be met for some units. However, if it is more important for the user to have a fixed sample size each month, then the rebalancing option should be used.

Analysis of frame and sample changes

The *compare* option can be used for ongoing surveys. With this option, different survey occasions can easily be compared. Criteria may be chosen to identify, for example, units added to or removed from the frame, units that have changed strata, and units added to or dropped from the sample.

4 CONCLUDING REMARKS

The Generalized Sampling System provides sampling functions for survey developers. GSAM, a SAS-based microcomputer product, can be used for periodic surveys and for surveys that are carried out only once. Further, GSAM modular functions allow for quick response to changes in survey requirements between survey periods.

Development work continues on GSAM and more features will be added based on budgetary constraints and users' requirements. For instance, client-server and batch processing capabilities will be included in the next version. As other enhancements are considered, the GSAM Development Team will review development plans and will implement changes as time and resources permit.

REFERENCES

Bankier, M. (1983). *Comparison of Neyman Allocation to Power Allocations*. Statistics Canada, internal memorandum.

Bankier, M. (1988). *Power Allocations: Determining Sample Sizes for Subnational Areas*. The American Statistician, August 1988, Vol. 42, No. 3.

Bethel, J.W. (1989). *Sample Allocation in Multivariate Surveys*. Survey Methodology, June 1989, Vol. 15, No. 1, pp 47-57.

Cochran, William G. (1977). *Sampling Techniques*. Third Edition. John Wiley & Sons, Inc.

Estevao, V. (1993). *Optimum Allocation for Stratified One-Stage SRSWOR Designs*. Statistics Canada, internal report.

Faber, G.B. (1997). *Using GSAM the first time for an ongoing survey*. Statistics Canada, internal report.

GSAM Development Team. (1996). *Generalized Sampling System Version 1.2 User Guide*. Statistics Canada Technical Report.

Kish, L. and Scott, A. (1971). *Retaining Units after Changing Strata and Probabilities*. Journal of the American Statistical Association, pp 461-470.

Murthy, M.N. (1977). *Sampling Theory and Methods*. Second Impression. Indian Press Private Ltd.

SAS Institute Inc. (1985). *SAS Language Guide for Personal Computers, Version 6 Edition*. SAS Institute Inc.

Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-98-007E/F
METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-98-007E/F
DIRECTION DE LA MÉTHODOLOGIE

Aperçu de la version 1.2 du Système généralisé d'échantillonnage et stratégie employée

G.B. Faber, N. Laniel et D.M. Yeo

Juin 1998

STATISTICS CANADA LIBRARY

Aperçu de la version 1.2 du Système généralisé d'échantillonnage et stratégie employée

G.B. Faber, N. Laniel et D.M. Yeo

Abstract

The Generalized Sampling System (GSAM) provides basic sampling functions for designing, selecting, and maintaining samples. GSAM is a SAS-based product designed to facilitate various sampling functions in periodic surveys or in surveys that are carried out only once. The interactive nature of GSAM permits different sampling strategies to be examined before implementing the most appropriate one. GSAM modular functions, including stratification, allocation, and sampling, allow for a quick response to changes in survey requirements over time. Furthermore, GSAM provides an integrated environment for updating the frame.

1 INTRODUCTION

La démarche d'enquête comporte un grand nombre d'étapes : élaboration d'un questionnaire, création d'une base de sondage, tirage d'un échantillon, communication avec les enquêtés, collecte, vérification, imputation, pondération et mise en tableaux des données et diffusion des résultats de l'enquête. Statistique Canada s'est doté d'un ensemble de systèmes généralisés lui permettant de franchir plusieurs étapes d'une enquête, tout comme d'appliquer un même système généralisé à une foule d'enquêtes.

Le Système généralisé d'échantillonnage (SGECH) vise à faciliter la conception, le prélèvement et la mise à jour d'échantillons dans le cadre d'enquêtes périodiques ou uniques. Outre le SGECH, on a conçu les systèmes suivants : Système généralisé de collecte et de saisie de données (CSD), Système de codage automatisé par reconnaissance de textes (CART), Système généralisé de vérification et d'imputation (SGVI), Système généralisé d'estimation (SGE), Système généralisé de couplage des enregistrements (SGCE) et Système généralisé d'analyse de séries chronologiques (SGASC en cours d'élaboration).

Une fois constitué un ensemble de données sur la population (base de sondage), le SGECH peut servir à chaque étape de l'échantillonnage (stratification, répartition, tirage et mise à jour de la base de sondage). On peut considérer la démarche d'échantillonnage comme comportant un stade de conception et un stade de production. Au premier de ces stades, l'utilisateur peut choisir la meilleure méthode de stratification de la base de sondage et de répartition de l'échantillon correspondant. Ces deux activités peuvent exiger une certaine expérimentation. Après répartition des unités d'échantillonnage, on peut passer à l'étape de la production et prélever l'échantillon.

Dans le cas d'enquêtes permanentes, on aura peut-être à mettre à jour la base de sondage. On peut ainsi avoir à ajouter ou à retrancher des éléments d'échantillonnage. Il est également possible que les renseignements sur des unités changent dans le temps. Après l'actualisation de la base de sondage, on peut refaire la stratification et la répartition et tirer un nouvel échantillon. L'échantillon dépend de l'échantillon antérieur si l'utilisateur contrôle le chevauchement entre les deux périodes d'échantillonnage. Ainsi, dans une enquête mensuelle, il pourra vouloir conserver les unités pendant six mois en renouvelant chaque mois seulement le sixième de l'échantillon. Après le tirage d'un nouvel échantillon, il peut enfin vouloir constater les changements (créations, nouvelles unités échantillonnées, unités retranchées par renouvellement).

À la section 2 du présent document, nous donnons un aperçu du SGECH. Il est d'abord brièvement question des systèmes généralisés et de leurs avantages. Nous décrivons ensuite chaque grand volet du système (stratification, répartition, tirage et mise à jour de la base de sondage). Nous présentons définitions et concepts, puis examinons les principales fonctions. À la section 3, nous exposons certaines stratégies d'échantillonnage utiles lorsqu'on élabore des applications avec le SGECH.

2 APERÇU

2.1 Systèmes généralisés

Un système généralisé est un jeu de programmes modulaires servant à un éventail d'applications semblables. Par leurs orientations communes et leurs solides méthodes d'enquête, les systèmes généralisés sont une sorte de «boîte à outils» permettant à l'utilisateur de conférer une souplesse considérable à une application et de ménager un haut degré de réutilisation des éléments. Statistique Canada a élaboré des progiciels généralisés pour économiser les ressources et atténuer le double emploi que comporte l'adoption de systèmes sur mesure.

L'élaboration de systèmes généralisés est coûteuse, mais avec de tels systèmes on économise à long terme puisque on a moins besoin de se doter de systèmes sur mesure et d'entretenir une foule de systèmes de même nature. On épargne aussi en se concentrant sur l'élaboration et l'entretien d'un ensemble commun de logiciels. Ajoutons qu'il est plus facile et rapide de mettre à jour cet ensemble unique de logiciels et sa documentation. L'utilisateur jouit ainsi d'un meilleur soutien.

Le SGECH est d'une conception modulaire où chaque module correspond à une des grandes fonctions du système. L'utilisateur peut spécifier toutes les fonctions SGECH ou une partie seulement (la fonction de répartition, par exemple). Si on a choisi SAS comme logiciel de base, c'est à cause de sa puissance, de sa portabilité et de sa facilité d'exploitation. Il est aussi bien connu d'un grand nombre de concepteurs d'enquêtes à Statistique Canada et offre de nombreuses fonctions qui viennent faciliter toute analyse statistique.

2.2 Mise à jour de la base de sondage

Il faut d'abord créer une base de sondage. Dans une enquête périodique, celle-ci devra être actualisée à l'occasion. On peut ajouter ou retrancher des unités ou encore les caractéristiques des unités peuvent évoluer. De nouveaux renseignements pourraient nous dire, par exemple, qu'une unité devrait appartenir à une autre strate.

Dans une foule d'enquêtes, on dispose de mécanismes distincts aux fins de la mise à jour de la base de sondage et de l'échantillonnage. Le SGECH permet cependant de modifier la base de sondage en place. Si les changements apportés demeurent légers, il peut être plus commode de modifier la base existante que d'en créer une de toutes pièces et de l'intégrer à SAS.

Il n'y a pas que les options d'adjonction et de suppression d'unités et de mise à jour des renseignements sur les unités faisant déjà partie de la base de sondage, le SGECH offre deux autres possibilités. Par la première de ces options, on peut créer une base en modifiant la base existante. La nouvelle contient alors en tout ou en partie les observations de la base d'origine. Pour choisir une partie des observations seulement, l'utilisateur fournit la liste des observations à retenir sous forme d'ensemble ou de sous-ensemble de données. Par l'autre option à caractère analytique, il peut comparer les variables de deux bases et relever des changements comme l'adjonction d'unités.

2.3 Stratification

Avant l'échantillonnage proprement dit, on peut grouper les unités de la population en strates non chevauchantes. Par cette stratification, on s'assure que tous les secteurs de la population sont représentés dans l'échantillon et que les estimations tirées de chaque secteur sont exactes, surtout dans le cas des groupes auxquels on s'intéresse en particulier. On peut également stratifier uniquement pour obtenir les estimations les plus précises pour une taille donnée d'échantillon. D'ordinaire, les estimations seront d'une plus grande précision si on regroupe les unités homogènes de population et que les groupes ainsi constitués diffèrent nettement les uns des autres. On peut enfin stratifier pour des motifs d'ordre administratif comme la volonté de bien répartir la charge de travail entre les bureaux régionaux.

Avec le SGECH, l'utilisateur peut concevoir diverses règles de stratification d'une population. Chaque règle définit une strate et est formée d'un ensemble complexe d'expressions logiques SAS. On applique les règles à chaque unité de la population pour juger de la strate à laquelle celle-ci appartient. Chaque unité doit se ranger dans une seule strate. On peut recourir au SGECH pour concevoir des règles infaillibles de stratification. La prudence sera de mise si les utilisateurs y vont de leurs propres règles. Ainsi, l'utilisateur pourrait verser deux règles dans le système et une unité pourrait satisfaire aux deux exigences. L'unité avec une valeur de RECETTES de 500 pourrait répondre aux critères respectifs des deux expressions «RECETTES $\leq$ 500» et «RECETTES $\geq$ 500 et RECETTES $<$ 10000». (Dans ce cas, l'unité serait attribuée à la strate représentée par la règle qui a été définie en premier lieu.)

Dans l'élaboration de ces règles, l'utilisateur peut directement spécifier les expressions SAS à utiliser et peut faire déterminer par le SGECH des combinaisons possibles de variables spécifiées

par lui (province par industrie selon la taille, etc.), tout comme des limites appropriées de strates en fonction de toute variable numérique d'intérêt.

On peut procéder à cette délimitation par la méthode cumulative $\sqrt{f}$ Dalenius-Hodge (Cochran, 1977, p. 128-130) ou par une méthode de mise en grappes. La première permet une stratification à variance approximativement minimum par répartition de Neyman en fonction d'une variable auxiliaire connue. La seconde crée des strates d'une composition homogène en unités par mise en grappes à dispersion minimum (en fonction de la variable cible). L'utilisateur précise le maximum de grappes pouvant être constituées pour tel ou tel nombre d'unités. À l'aide de ces données et indications, le système établit combien de grappes devraient être formées à raison d'une grappe par strate.

On peut corriger les règles, c'est-à-dire ajouter ou retrancher des expressions ou combiner des strates. On peut aussi produire en tout temps des rapports SGECH qui indiqueront ce qu'il adviendra si les règles spécifiées de stratification sont appliquées à un ensemble particulier de données. On peut établir des statistiques (sommes, comptes, moyennes, variances et coefficients de variation (CV)) pour les variables et les strates choisies par l'utilisateur. Grâce à ces rapports, l'utilisateur pourra mieux juger des règles de stratification à changer ou voir si les règles employées sont les bonnes.

2.4 Répartition

La répartition est la détermination du nombre d'unités de population à échantillonner dans chaque strate. Souvent, on fixe les tailles d'échantillonnage de strates de manière à répondre aux besoins particuliers de l'utilisateur. Comme choix courants, mentionnons notamment un coût, une taille d'échantillon ou un coefficient de variation (CV) fixes. La répartition tient compte de la variation à l'intérieur des strates et du coût d'obtention de l'information dans chaque strate. Si le coût unitaire par strate est constant, il y aura plus d'unités échantillonnées dans les strates où la variation est supérieure. Si le coût diffère selon les strates, mais que les variations sont à peu près les mêmes à l'intérieur des strates, il sera généralement plus efficient de tirer plus d'unités dans les strates où le coût unitaire est moindre (Murphy, 1977).

Dans le SGECH, il existe deux options pour un plan stratifié à un degré de tirage d'unités par échantillonnage aléatoire simple sans remise (EASSR) : on peut minimiser une fonction de variance pondérée ou le coût total d'échantillonnage. On résout les deux problèmes à l'aide d'un algorithme proposé par Bethel (1989) et développé par Estevao (1993). On spécifie des paramètres et des contraintes différents pour chaque option.

Pour minimiser la fonction de variance pondérée, on répartit l'échantillon en fonction d'un coût total d'échantillonnage ou d'une taille d'échantillon qui sont fixes. Pour répartir, on a besoin d'une variable spécifiée par l'utilisateur et, à titre facultatif, d'une autre variable auxiliaire à préciser. Si on spécifie une variable auxiliaire, elle devrait être en étroite corrélation avec la première variable. La méthode générale employée à cette fin est une répartition de puissance proposée par plusieurs auteurs, dont Bankier (1988). Selon la façon dont on fixe le paramètre de la répartition de puissance, il est possible d'obtenir une répartition de Neyman et une répartition CV à peu près égale ou un moyen terme entre ces deux répartitions. Il est également possible de fixer la variable auxiliaire pour que la répartition soit proportionnelle à la taille de la population ou à une variable portée à une

certaine puissance. Le SGECH trouve la répartition optimale ou juge si une répartition est impossible compte tenu des contraintes formulées.

On peut spécifier des contraintes comme le coût unitaire d'échantillonnage et les valeurs minimale et maximale de taille admissible d'échantillon dans chaque strate. Si on ne connaît pas le coût unitaire, on fixe sa valeur à un dans chaque strate. Dans ce cas, essayer de s'en tenir à un coût fixe est comme essayer de fixer la taille totale de l'échantillon. On recourt à des valeurs implicites (par défaut) que peut modifier l'utilisateur si les valeurs minimale et maximale de taille admissible d'échantillon par strate ne sont pas précisées.

Pour minimiser le coût total d'échantillonnage, on applique une méthode de répartition optimale avec des contraintes comme des valeurs limites de CV pour les variables spécifiées par l'utilisateur. Ce dernier peut varier les valeurs limites de CV selon les variables. Le SGECH trouve la répartition optimale ou dit si une répartition est impossible compte tenu des contraintes formulées. Comme dans l'option précédente, on peut préciser le coût unitaire et les valeurs minimale et maximale de taille admissible d'échantillon. On prendra des valeurs implicites si ces indications ne sont pas fournies.

Pour l'une et l'autre des méthodes de répartition, les utilisateurs peuvent fournir leurs propres fichiers de statistiques au lieu d'utiliser la base de sondage comme source de données. Le système exploitera toutes les statistiques procurées (comptes, totaux et variances des variables de stratification, etc.) au lieu des valeurs calculées à partir de l'ensemble de données sur la population. En fait, si on fournit toutes les statistiques voulues sur toutes les strates, on n'a pas à fournir d'ensemble de données sur la population pour faire la répartition.

2.5 Échantillonnage

On procède souvent par sondage plutôt que par recensement ou dénombrement pour diverses raisons, qu'il s'agisse d'abaisser les coûts, de réduire les délais ou d'alléger le fardeau des répondants. On a conçu de nombreuses méthodes de tirage d'un échantillon aléatoire. À l'heure actuelle, on peut appliquer des plans d'échantillonnage aléatoire simple stratifié dans le SGECH. D'autres méthodes de sondage s'ajouteront un jour : échantillonnage double, bernoullien ou avec probabilité proportionnelle à la taille.

La fonction d'échantillonnage offre plusieurs caractéristiques utiles pour le tirage d'un échantillon aléatoire simple sans remise, ainsi que pour son actualisation. Les utilisateurs peuvent mettre à jour leur échantillon en fonction des adjonctions ou des suppressions dans la base de sondage, tout comme des changements de stratification de la population. Ils peuvent aussi renouveler l'échantillon pour alléger le fardeau des répondants et rééquilibrer l'échantillon si la population a trop changé. Ils peuvent soit utiliser le fichier de stratification de la population et le fichier de répartition SGECH ou produire l'un ou l'autre de ces fichiers indépendamment dans SAS.

Le SGECH procède par échantillonnage cospatial (Équipe de développement du SGECH, 1996). Il y a attribution aléatoire aux unités de numéros de sélection espacés également sur un intervalle [0,1] dans chaque strate. On établit une fenêtre d'échantillonnage pour chaque strate et les unités dont le numéro de sélection tombe dans la fenêtre sont tirées. S'il s'agit d'une strate à tirage complet, tout l'intervalle est couvert par la fenêtre.

Pour maximiser le chevauchement entre échantillons, le SGECH emploie une méthode qui ressemble fort à la technique Kish-Scott (Kish et Scott, 1971), mais dans un contexte de cospatialité d'échantillon. On peut contrôler le chevauchement des unités entre deux périodes d'enquête par le renouvellement d'échantillon. Les numéros de sélection ne changent pas entre périodes de sondage, mais la fenêtre d'échantillonnage se déplace sur une certaine distance vers la droite en fonction de paramètres spécifiés par l'utilisateur. Ainsi, celui-ci peut indiquer que les unités seront échantillonnées pendant 12 mois (en renouvellement, le douzième de l'échantillon change tous les mois) et déséchantillonnées pendant au moins 24 mois, et ce, sans qu'on s'écarte des taux d'échantillonnage visés. L'utilisateur peut prévoir de la sorte un chevauchement entre périodes de sondage aussi important qu'il le désire. Ce chevauchement se contrôle au niveau des strates. On applique la règle suivante pour déterminer la distance de déplacement d'une fenêtre pour la strate h en renouvellement :

$$\textit{déplacement } h \;=\; \textit{min}\left(\frac{f_h}{\textit{échantillonnage } h}\,,\ \frac{(1-f_h)}{\textit{déséchantillonnage } h}\right)$$

où f_h est le taux d'échantillonnage de la strate h (quotient du nombre d'unités de l'échantillon et du nombre d'unités de la population dans la strate h).

Pour réduire au minimum le fardeau des répondants, le SGECH essaie de garder les unités dans l'échantillon pour le nombre exact de périodes qu'indique la contrainte d'échantillonnage et hors de l'échantillon pendant au moins le nombre de périodes qu'indique la contrainte de déséchantillonnage. Ainsi, prenons le cas d'une enquête mensuelle où une strate a un taux d'échantillonnage de 0,3, une contrainte d'échantillonnage de 24 mois et une contrainte de déséchantillonnage de 12 mois. Le calcul du déplacement est le suivant : min (0,3/24, (1-0,3)/12), ou 0,0125. Ainsi, chaque mois, la fenêtre d'échantillonnage de cette strate se déplacera d'une très faible valeur seulement, soit 0,0125. Un léger déplacement garantit qu'une unité sera échantillonnée pendant les 24 mois prévus et déséchantillonnée pendant longtemps (56 mois en l'occurrence).

Cette formule tient compte de la possibilité que les paramètres spécifiés d'échantillonnage et de déséchantillonnage ne soient pas tous les deux respectés. Dans le SGECH, la contrainte de déséchantillonnage est toujours respectée, voire dépassée. Ainsi, les unités ne peuvent être toutes échantillonnées pendant 12 mois au plus et déséchantillonnées pendant au moins la même période dans une strate où 90 % des unités doivent être tirées. Dans ce cas, l'unité restera dans l'échantillon plus de 12 mois, mais si elle en sort par renouvellement, elle restera déséchantillonnée pendant les 12 mois prévus. On juge moins fastidieux d'échantillonner sur une longue période que de sortir une unité par renouvellement pour ensuite la rééchantillonner très rapidement.

Le SGECH permet à l'utilisateur d'actualiser son échantillon en fonction de changements dans les taux de tirage ou dans la stratification. Pour tenir compte de ces changements, on se reporte aux numéros de sélection de la période précédente - d'après l'ancienne stratification - pour juger quelles unités feront partie de l'échantillon de la nouvelle période compte tenu de la nouvelle stratification, le but étant de conserver le plus d'unités possible de la première période. Si des unités échantillonnées de la période précédente sont supprimées, ce sont les unités échantillonnées le plus

longtemps qui seront retranchées en premier, méthode qui permet aussi de renouveler pendant qu'on refait l'échantillon.

Ainsi, si le taux de tirage s'est élevé et qu'on cesse de renouveler, le nouvel échantillon consistera en toutes les unités déjà échantillonnées avec quelques nouvelles, ce qui maximisera le chevauchement entre échantillons. Si on prévoit un renouvellement, les unités échantillonnées le plus longtemps feront place aux unités déséchantillonnées le plus longtemps. Si des unités ont changé de strate ou qu'il y a eu restratification, le SGECH range chaque unité reclassée dans sa nouvelle strate, la mettant au même endroit par rapport à la fenêtre d'échantillonnage de cette strate. C'est ainsi que le SGECH assure un chevauchement maximum dans le cas des unités restratifiées.

On choisit de nouvelles unités dans la base de sondage en attribuant à chacune un numéro de sélection espacé également dans chaque strate. On les tire ensuite indépendamment des anciennes unités par la méthode employée pour le prélèvement d'un échantillon initial.

Un inconvénient de l'adjonction, de la suppression et de la restratification d'unités est que les numéros de sélection peuvent ne plus être espacé également dans chaque strate. Bien que $f_h \times 100\%$ des unités d'une strate doivent être tirées, il pourrait s'en trouver plus ou moins dans la fenêtre de sélection. Le SGECH comporte une option de rééquilibrage que l'on peut spécifier en vue d'une stabilisation des tailles d'échantillon. En cas de rééquilibrage, on rajuste les numéros de sélection des unités pour qu'ils soient à nouveau espacés également dans chaque strate. On tire ensuite le bon nombre d'unités.

3 STRATÉGIE

Dans cette section, nous exposerons généralement comment les utilisateurs peuvent exploiter les différents modules du SGECH selon leurs divers besoins en matière d'applications. C'est ainsi que nous décrirons le cadre informatique, puis donnerons des indications précises sur l'utilisation des fonctions de stratification, de répartition, d'échantillonnage et de mise à jour de la base de sondage dans ce système. Pour se renseigner plus en détail sur les fonctions, on consultera Équipe de développement du SGECH (1996).

Après avoir choisi une méthode appropriée, on peut simplement utiliser le SGECH en procurant l'information nécessaire. D'ordinaire, toutes les fonctions du système s'emploient ensemble, auquel cas la sortie d'une fonction devient l'entrée de la suivante. Toutefois, en raison du caractère modulaire du SGECH, l'utilisateur peut retenir seulement les fonctions dont il a besoin. Pour sauter une fonction, il doit créer l'information qui aurait été produite par celle-ci et la transmettre à la fonction suivante. Que l'information vienne de l'extérieur ou de l'intérieur du système, ce dernier vérifie toujours la concordance des données fournies à un module.

3.1 Cadre informatique

Le SGECH est une application SAS®- pour les micro-ordinateurs IBM® et compatibles. La version actuelle 1.2 du SGECH est exploitée en SAS® (version 6.08, 6.10 ou 6.11) pour la version 3.1 Windows™, la version 3.11 Windows for Workgroups™ ou Windows NT ™. Il faut avoir accès à SAS/BASE®, SAS/FSP®, SAS/IML® et SAS/STAT® pour faire tourner le SGECH, dont les fonctions sont accessibles en session interactive SAS par jeu de menus. À chaque séance, l'éditeur ordinaire, le journal et les fenêtres de sortie de SAS sont disponibles. Les rapports SGECH s'affichent en fenêtre de sortie et peuvent être sauvegardés par commande de type SAS. L'utilisateur gagnera à connaître les rudiments du système SAS. On trouvera de bonnes indications sur ce système en consultant SAS Institute Inc. (1985).

Nous recommandons un micro-ordinateur 486 ou supérieur doté d'au moins 8 mégaoctets de mémoire. Le logiciel demande environ 70 mégaoctets de disque dur, en majeure partie pour SAS. On peut installer ces programmes et tous les fichiers utilisateur, dont la taille totale variera selon les applications, dans le micro-ordinateur même ou dans un serveur de fichiers en cas d'accès par réseau local.

On appelle les modules SGECH par menu. Le système vérifie les paramètres d'entrée, mais il appartient à l'utilisateur d'ordonner les fonctions et leurs paramètres. On suppose que celui-ci connaît non seulement son domaine spécialisé, mais aussi les rudiments de la stratification, de la répartition, de l'échantillonnage et de la mise à jour de bases de sondage. Des mises en garde indiquent les embûches éventuelles, mais il est impossible de garantir que le système sera utilisé sans erreur, aussi recommandons-nous que les agents des programmes spécialisés créent leurs applications SGECH avec l'aide de spécialistes de la méthodologie.

3.2 Stade de la conception

Stratification

Pendant que l'on élabore les règles de stratification, le SGECH peut fréquemment produire des rapports où seront examinées les strates qui seraient créées si les règles étaient appliquées à une population. On obtient ainsi des statistiques comme des sommes, des comptes, des moyennes, des variances et des coefficients prévus de variation (CV) pour les variables et les strates spécifiées par l'utilisateur. Ces rapports peuvent aider ce dernier à juger des modifications à apporter le cas échéant aux règles. Ainsi, l'utilisateur pourrait découvrir que les strates sont trop importantes ou encore comptent trop peu d'unités et devraient être combinées à d'autres strates.

L'utilisateur stratifie souvent par rapport à des variables pertinentes de mesure de taille pour accroître l'efficience de l'échantillonnage ou abaisser considérablement les coûts. Une stratification systématique selon la taille permettra généralement de réaliser les nombreux avantages d'un échantillonnage avec probabilité proportionnelle à la taille (PPT). C'est généralement la méthode recommandée. Souvent, les mesures individuelles de taille ne sont pas extrêmement stables (surtout dans le cas des enquêtes auprès des entreprises) et la stratification selon la taille produira fréquemment des échantillons d'un meilleur rendement que ceux d'un échantillonnage PPT.

On recommande la méthode de mise en grappes pour le regroupement d'unités de taille homogène. Prenons le cas d'une population initialement stratifiée par province et industrie. Par cette méthode, on peut stratifier davantage chaque province par groupe d'industries en faisant intervenir une mesure de taille utile comme les recettes. Il faut ensuite expérimenter quelque peu pour évaluer le nombre maximal - défini par l'utilisateur - de grappes à constituer dans chaque province par groupe industriel. La stratification optimale peut varier selon ce qu'on spécifie comme CV (coefficient de variation) ou taille totale d'échantillon.

On groupe souvent les unités importantes en strates à tirage complet. On agit souvent ainsi pour réduire la variance, et ces unités sont d'ordinaire assez grandes pour être considérées comme autoreprésentatives. Si une grande unité est mise dans une strate à tirage partiel, elle peut être échantillonnée ou non. Si elle ne l'est pas, les estimations définitives seront basses et, si elle l'est, ses caractéristiques seront probablement assez différentes de celles des autres unités pour que l'ensemble de données soit difficile à modéliser. On trouvera ci-après plus de détails sur la spécification de strates à tirage complet.

Procédure de stratification permettant de simplifier la construction d'un fichier de contraintes de répartition

Après une stratification basée sur la taille, il sera souvent souhaitable de spécifier les strates à tirage complet, ce que l'on fera en prévoyant un fichier de contraintes de répartition. Avec un algorithme de mise en grappes dans le cadre de la stratification et une désignation de strates à tirage complet dans le cadre de la répartition, on peut largement abaisser les valeurs de CV ou de taille d'échantillon.

Le fichier de contraintes contient une observation par strate. Des valeurs implicites définies par l'utilisateur seront appliquées aux strates non spécifiées au fichier. On stocke quatre variables, à savoir l'identificateur de strate, le coût unitaire d'échantillonnage et les valeurs minimale et maximale de taille d'échantillon.

Il est plus facile d'établir un fichier de contraintes si la stratification se fait d'abord selon la taille (constitution de strates à tirage complet ou partiel par province et par catégorie CTI). Les numéros de strate attribués aux strates à tirage complet seront groupés au début ou à la fin de toute la liste des codes produits. Si la seule restriction énoncée par l'utilisateur est que certaines strates doivent être à tirage complet, on ne spécifie que ces strates. Si on en connaît les codes, la création du fichier de contraintes s'en trouve facilitée.

Pour la répartition de toutes les unités d'une strate, un nombre très élevé devrait être assigné à la valeur minimale de taille d'échantillon dans le fichier de contraintes pour chaque strate devant être à tirage complet. Le SGECH fera automatiquement la conversion par rapport à la taille de population de chaque strate indiquée. Si un certain nombre d'unités doivent être tirées d'une strate, disons 76, les valeurs minimale et maximale de taille d'échantillon sont toutes deux fixées à 76.

Pour qu'une variance puisse être calculée, deux unités seulement doivent être tirées de chaque strate, mais il convient peut-être plus de prélever au moins cinq unités dans chacune des strates (même dans le cas d'une strate à tirage partiel dans une enquête entreprises où les unités ont de faibles recettes) si on veut diminuer les effets de la non-réponse. L'utilisateur peut veiller à ce que la valeur minimale de taille d'échantillon versée au fichier des contraintes soit de cinq pour chaque strate à tirage complet, mais il est plus simple de spécifier une valeur implicite (par défaut) de répartition.

Répartition

On peut avoir à exécuter plusieurs fois l'opération de répartition. Si on minimise le coût d'échantillonnage en fonction d'un niveau fixe de coefficient de variation (CV), l'utilisateur pourra constater, par exemple, que la taille d'échantillon calculée est trop grande. Il peut alors relâcher certaines contraintes CV, réexécuter la fonction de répartition et voir si la taille d'échantillon est maintenant acceptable. Il peut réexécuter la fonction seulement s'il le faut pour obtenir des résultats satisfaisants. En fait, cette fonction peut servir à essayer diverses stratégies (taille d'échantillon résultante si la valeur limite de CV est fixée à 2 %, 5 %, etc.).

On peut aussi procéder à la répartition avec diverses valeurs exponentielles de fonction de puissance. Plus précisément, on réalise la répartition de Neyman en utilisant la variable cible comme variable auxiliaire et en fixant l'exposant de la fonction de puissance à l'unité. L'utilisateur devrait effectuer cette répartition s'il désire minimiser le CV global, tout en ne contrôlant pas les CV des strates. Pour que les CV soient approximativement égaux entre strates dans une situation où le rapport variance-moyenne ne varie pas amplement selon les strates et où les taux d'échantillonnage sont petits, l'exposant de fonction de puissance devrait être fixé à zéro pour toute variable auxiliaire. Dans cette répartition, on minimise les CV des strates, mais en augmentant le CV global. On peut aussi spécifier des répartitions intermédiaires où des valeurs exponentielles de 0,50 (répartition racine carrée) ou de 0,33 (répartition racine cubique) n'ont rien d'inhabituel. Grâce à ces répartitions intermédiaires, on peut largement diminuer les coefficients de variation des strates sans augmenter outre mesure le CV global (Bankier, 1983).

Lorsqu'on minimise le coût d'échantillonnage par rapport à des contraintes de valeurs fixes de CV, on établit ces contraintes au niveau de la population (coefficient de variation de 5 % des recettes à l'échelle du Canada, par exemple), mais il est possible de fixer des valeurs marginales de contrainte (sur le plan des recettes, coefficient de variation de 6 % pour l'Ontario et de 3 % pour une strate formée des grandes entreprises des provinces de l'Atlantique, par exemple). Pour obtenir un CV de 6 % expressément pour les «recettes» de l'Ontario, on crée une nouvelle variable au fichier de la population stratifiée et reporte sur cette variable la valeur correspondante des «recettes» pour chaque observation dans cette province. La valeur stockée est nulle pour chacune des observations hors Ontario. On spécifie ensuite une contrainte de CV de 6 % pour la nouvelle variable. À l'heure actuelle, l'utilisateur peut spécifier jusqu'à 22 variables CV dans le SGECH. S'il veut en spécifier plus de 22, il doit obtenir une macro-instruction de la Section des méthodes des systèmes généralisés de la Division des méthodes d'enquêtes-entreprises.

Pour des considérations d'ordre budgétaire, on fixe souvent le coût total ou la taille globale de l'échantillon. L'option consistant à minimiser une fonction de variance par rapport à un coût fixe est alors la meilleure. Si on a arrêté une taille d'échantillon par opposition à un coût total d'échantillonnage, le coût unitaire d'échantillonnage devrait être fixé à l'unité pour chaque strate (fixer le coût d'échantillonnage équivaut à fixer la taille globale de l'échantillon).

S'il n'y a pas convergence de l'algorithme de répartition, on peut se retrouver avec une solution qui ne respecte pas la spécification cible. Ainsi, la taille d'échantillon obtenue peut être inférieure à la taille fixée lorsqu'on prend l'option de la variance minimum. Autre possibilité, si on prend l'option du coût minimum, la taille d'échantillon obtenue peut ne pas convenir parce que les niveaux visés de CV n'ont pas été atteints. Si un tel problème se pose, on devrait accroître le paramètre de contrôle du nombre d'itérations dans l'algorithme de Bethel. On devrait également vérifier l'homogénéité des strates. Une nouvelle stratification pourrait s'imposer, peut-être selon la taille. Il peut y avoir de très grandes unités qui devraient être groupées séparément en une strate à tirage complet.

Renouvellement

On devrait généralement prévoir un renouvellement dans le cas des enquêtes permanentes pour que le fardeau de réponse ne pèse pas toujours sur le même groupe de répondants et aussi pour qu'une plus grande partie de la population soit échantillonnée avec le temps. L'utilisateur devrait également s'attacher au taux de renouvellement pour plusieurs raisons. Un renouvellement plus lent maximise le chevauchement entre échantillons, d'où la possibilité de mesurer plus précisément les tendances. Dans ce cas, les enquêtés ont souvent moins d'efforts à fournir dans les périodes ultérieures d'échantillonnage. Ajoutons que les frais de collecte diminuent fréquemment dans ces périodes subséquentes. En revanche, un renouvellement plus rapide réduira le séjour des unités dans l'échantillon, et donc la fatigue des répondants.

Il convient de noter que les paramètres d'échantillonnage et de déséchantillonnage peuvent prendre des valeurs diverses dans chaque strate. C'est ainsi que, par exemple, l'utilisateur peut régler le fardeau de réponse imposé aux petites entreprises ou aux ménages ruraux. Il peut établir combien de temps chaque unité échantillonnée (entreprises, ménages, etc.) devrait demeurer dans l'échantillon et ensuite informer l'enquêté de la durée maximale de son échantillonnage.

3.3 Stade de la production

Période initiale dans le SGECH

Un ensemble de données sur la population, aussi appelé fichier de la base de sondage, doit exister avant tout prélèvement d'échantillon. L'utilisateur peut importer directement ce fichier dans SAS. Si celui-ci porte sur une population plus grande que prévu, le système peut servir à tirer un sous-ensemble de ce fichier trop ample. Autre avantage de l'option de *création* dans le cadre de la mise à jour de la base de sondage, le système crée une variable qui indique quelles unités ont été échantillonnées dans la période précédente, ce qui permet de procéder au renouvellement de l'échantillon si on le désire. C'est pourquoi l'option de *création* est recommandée si on sait qu'un renouvellement aura lieu.

Pour prélever un échantillon en prévision d'une nouvelle enquête, on doit spécifier l'option de tirage de *nouvel* échantillon. Si une enquête permanente est introduite dans le SGECH, l'utilisateur a un double choix. S'il désire tirer un échantillon qui soit indépendant de tout échantillon antérieur, il n'a pas besoin de renseignements préalables et, une fois de plus, il doit spécifier l'option de tirage de *nouvel* échantillon. Il peut aussi souhaiter contrôler le chevauchement entre périodes d'échantillonnage. Il doit créer à cette fin, à l'intérieur comme à l'extérieur du SGECH, un fichier paramétrique et un fichier de la base de sondage pour la période précédente. On trouvera d'autres détails sur cette méthode dans Faber (1997).

Périodes ultérieures

Mise à jour de la base de sondage

On a besoin d'une mise à jour de la base de sondage pour la plupart des enquêtes périodiques (mensuelles, trimestrielles, annuelles, etc.) ou occasionnelles. Les options SGECH *d'addition, de suppression et de mise à jour* représentent une solution de rechange à la création et à l'intégration d'une base de sondage entièrement nouvelle lorsque des changements sont nécessaires. S'il s'agit de modifier légèrement un grand fichier de base de sondage (e.g. unités à ajouter, etc.), on peut prendre l'option d'addition en mise à jour de la base de sondage pour épargner du temps. Dans ce cas, il faut seulement la liste des nouvelles unités avec les renseignements correspondants.

Dans une enquête permanente, de nouveaux renseignements sur la base de sondage peuvent nous dire que des unités changeraient de strate si les règles de stratification appliquées dans la période précédente d'échantillonnage l'étaient à nouveau. L'utilisateur doit décider si de nouvelles valeurs de strate doivent être attribuées aux unités appartenant à cette période précédente ou si les valeurs initiales devraient être conservées. Si on prévoit que les unités changeront encore de strate dans la prochaine période d'échantillonnage, il pourrait être sage de les laisser dans leur strate initiale.

On doit user de prudence avant de restratifier des unités ou de les retrancher d'une base de sondage. L'utilisateur peut savoir, par exemple, qu'une unité est mal classée ou qu'il s'agit d'un établissement en faillite. Pour prévenir tout biais, on devrait puiser uniquement à des sources d'information

indépendantes du cadre d'échantillonnage pour mettre à jour la base de sondage. Si l'utilisateur a obtenu les renseignements en tentant d'interroger l'unité échantillonnée aux fins de l'enquête, celle-ci devrait normalement demeurer inchangée dans la base de sondage. Elle peut représenter d'autres unités qui ont été mal classées ou ont disparu à l'insu de l'utilisateur, n'ayant pas été tirées.

Mise à jour de l'échantillon

Si l'utilisateur désire contrôler le chevauchement entre périodes d'échantillonnage, il devrait spécifier l'option de *mise à jour* de l'échantillon. Par la fonction de renouvellement, il peut préciser pour chaque strate le nombre de périodes d'enquête où les unités doivent rester dans l'échantillon et où elles doivent en être exclues par renouvellement. Pour maximiser le chevauchement, on devrait fixer une contrainte de durée d'échantillonnage relativement longue. Le choix des contraintes d'échantillonnage et de déséchantillonnage devrait cependant être tempéré par la nécessité de minimiser le fardeau des répondants. De plus, avec l'option de mise à jour, le SGECH se trouvera automatiquement à maximiser le chevauchement pour les unités qui changent de strate par reclassement ou modification des règles de stratification (combinaison de strates, par exemple). Si l'utilisateur veut constituer un échantillon qui soit indépendant de tout échantillon antérieur, il devrait prendre l'option de *sélection de nouvel échantillon*.

Il convient de noter que, si des unités ont été ajoutées ou retranchées ou ont changé de strate, elles ne seront plus espacées également, auquel cas une fenêtre d'échantillonnage de 17 % pourrait ne pas couvrir 17 % des unités d'une strate. Si la base de sondage a beaucoup changé, on devrait spécifier la fonction de rééquilibrage pour que la bonne proportion d'unités soit tirée dans chaque strate. Dans une enquête mensuelle dont la base de sondage est actualisée une fois l'an, l'utilisateur pourrait vouloir rééquilibrer pour le seul mois de l'année où se fait la mise à jour de la base.

Le rééquilibrage ne devrait se faire qu'à l'occasion, car on exerce alors un moindre contrôle sur les unités qui entrent dans l'échantillon ou en sortent. Ainsi, l'utilisateur pourrait vouloir qu'une unité soit dans l'échantillon jusqu'à 12 mois consécutifs et qu'elle en sorte par renouvellement pour au moins 24 autres mois consécutifs. À cause du rééquilibrage, cette contrainte ne sera pas respectée pour certaines unités. Toutefois, si une taille fixe d'échantillon chaque mois importe plus à l'utilisateur, l'option de rééquilibrage devrait être prise.

Analyse des changements au niveau de la base de sondage et de l'échantillon

On peut employer l'option de comparaison pour les enquêtes permanentes. On peut ainsi facilement comparer des périodes d'échantillonnage. On peut fixer des critères pour reconnaître, par exemple, les unités ajoutées ou retranchées dans la base de sondage, celles qui ont changé de strate ou celles qui se sont ajoutées à l'échantillon ou en ont été retirées.

4 EN GUISE DE CONCLUSION

Le Système généralisé d'échantillonnage (SGECH) prévoit des fonctions d'échantillonnage pour les concepteurs d'enquêtes. Produit de micro-informatique en SAS, il peut servir aux enquêtes tant uniques que périodiques. Ajoutons que ses fonctions modulaires permettent de réagir rapidement à l'évolution des besoins des enquêtes entre les périodes de sondage.

On continue à développer le SGECH et d'autres caractéristiques pourraient s'ajouter compte tenu des contraintes budgétaires et des exigences des utilisateurs. Ainsi, la nouvelle version comportera des capacités de traitement client-serveur et de traitement par lots. Comme d'autres améliorations sont envisagées, l'Équipe de développement du SGECH examinera les plans d'élaboration et apportera les changements dans les limites des délais et des ressources disponibles.

BIBLIOGRAPHIE

Bankier, M. (1983). *Comparison of Neyman Allocation to Power Allocations*. Statistique Canada, note interne.

Bankier, M. (1988). *Power Allocations: Determining Sample Sizes for Subnational Areas*. The American Statistician, août 1988, vol. 42, n° 3.

Bethel, J.W. (1989). *Répartition de l'échantillon dans les enquêtes à plusieurs variables.* Techniques d'enquête, juin 1989, vol. 15, n° 1, p. 47-57.

Cochran, William G. (1977). *Sampling Techniques,* troisième édition, John Wiley & Sons, Inc.

Estevao, V. (1993). *Optimum Allocation for Stratified One-Stage SRSWOR Designs*, Statistique Canada, rapport interne.

Faber, G.B. (1997). *Using GSAM the first time for an ongoing survey,* Statistique Canada, rapport interne.

Équipe de développement du SGECH (1996). *Système généralisé d'échantillonnage, version 1.2 : guide de l'utilisateur,* rapport interne de Statistique Canada .

Kish, L., et Scott, A. (1971). *Retaining Units after Changing Strata and Probabilities*, Journal of the American Statistical Association, p. 461-470.

Murthy, M.N. (1977). *Sampling Theory and Methods*, deuxième impression, Indian Press Private Ltd.

SAS Institute Inc. (1985). *SAS Language Guide for Personal Computers, Version 6 Edition,* SAS Institute Inc.