GENERALIZED SURVEY PROCESSING SYSTEMS: THE CANADIAN PERSPECTIVE

Robert Kozak, Claude Poirier, John Kovar
Statistics Canada, Tunney's Pasture, Ottawa, Ontario K1A 0T6

Key Words: generalized systems, processing tools, modular systems, foundation software

1. INTRODUCTION

Statistics Canada has a history of generalized systems development which dates back to the mid-1980s. During those nearly 20 years, Statistics Canada has developed several "core" generalized systems. These include systems for the survey functions of sampling, edit and imputation, estimation, autocoding, disclosure control, record linkage, and time series analysis. A brief description of each of these systems is given in Statistics Canada (1999). There was also a system built for the data collection and capture function, however this system is being gradually replaced by an externally developed system. There have not been any generalized systems built to date at Statistics Canada for the functions of survey mailout, tabulation or data analysis. More information on the overall strategy for generalized systems development at Statistics Canada can be found in Outrata and Chinnappa (1989).

There have also been numerous other systems designed at Statistics Canada which are more customized in nature. For example, the Nearest-Neighbour Imputation Methodology System (NIM) was initially designed to process the demographic data in the Canadian Census. Recently, the developers of NIM have been working on a more general approach to processing both qualitative and quantitative variables. Further details on NIM can be found in Bankier et al (1999).

2. DESCRIPTION OF PRODUCTS

For the purposes of this paper, we will focus on the three generalized systems which carry out the sampling, edit and imputation, and estimation functions. These are the Generalized Sampling System (GSAM), the Generalized Edit and Imputation System (GEIS), and the Generalized Estimation System (GES). Details on these systems are given by Faber et al (1998), Kovar et al (1991) and Estevao et al (1995) respectively . These three systems were the focus of the earliest development of generalized systems at Statistics Canada.

GSAM, GEIS, and GES have been in production mode for more than 10 years now. During that time, they have evolved through new requirements in survey processing, and by incorporating new methodologies.

Due to the manner in which they developed from their initial versions, these three systems are not integrated. For example, they employ different foundation softwares, such as SAS, Oracle, and C, with their associated data structures. The systems operate on different computing platforms, such as the mainframe computer, UNIX and personal computer, or a combination of these.

2.1 Functionality

In its current guise, GSAM meets the needs of a host of small- and large-scale sample surveys at Statistics Canada in the area of basic sample design and sample selection. It supports sample selection for periodic surveys, but can also be used for ad hoc surveys. The functions within the system have been designed in a modular fashion to facilitate ease of use. The four main functions in GSAM are stratification, sample allocation, sampling, and frame maintenance.

The current version of GEIS provides edit and imputation functions for primarily numerical survey data. It is also a modular system, although the modules are functionally linked to one another. GEIS is used for both periodic and one-time surveys. The system provides functionality for the specification and analysis of edits, univariate outlier detection, localization of errors, automatic imputation by deterministic, donor and estimator methods, and also a prorating function.

GES provides a framework to carry out domain estimation in business or social sample surveys, specifically to satisfy requirements for stratified one-stage cluster or element sample designs. Like the other systems, it can be used for periodic or ad hoc surveys. All of the functions in GES were designed as individual modules. The four main functions are the calculation of sample design weights, calculation of calibration factors, calculation of calibration estimates, and calculation of synthetic estimates, all for any specified domain of interest.

2.2 Data Structure

Both GSAM and GES operate on the Microsoft Windows platform. Also, either system can be run in a client-server configuration with a Windows client and a UNIX server. As they are both SAS-based, they accept SAS datasets as input and produce SAS datasets as output. In addition, GES accepts flat files as input.

The current version of GEIS, on the other hand, is Oracle-based. It will only accept and produce Oracle tables as input and output. GEIS resides on both the mainframe computer and UNIX platforms. Up until several years ago, there was also a PC-based version of GEIS, but Oracle stopped supporting their PC-DOS software, and Statistics Canada was obliged to stop supporting the PC version of GEIS as a consequence. Currently, a SAS-compliant version of GEIS is in development. This version will also be ported to the PC, along with UNIX. As it is based on SAS version 8, it is not possible to place this SAS-compliant GEIS on the mainframe computer at the present time, as SAS version 8 is not yet available on that platform at Statistics Canada.

Since our systems reside on various computing platforms and are based on varying foundation softwares, data transfers across the platforms and to/from different file types are not uncommon. For example, SAS is frequently used at Statistics Canada, and survey developers have to migrate SAS datasets or flat files into Oracle tables in order to use GEIS. To satisfy such requirements as well as others which are specific to the application concerned, custom-designed pre- and post-processor programs sometimes have to be written to work in concert with these three systems.

As for archiving of data, it is generally left to the users to develop their own customized tools to fulfill the function. An exception is the mainframe computer, where datasets are automatically archived after a certain period of time and can be retained indefinitely at a small cost to the user.

2.3 Users

Between 20 and 40 surveys per system use the generalized systems for their sampling, edit and imputation, and estimation needs. These surveys vary in size from very small (~1000 units) to very large (~1 million units). Some of these surveys use only one of the systems, while others may use two or all three of them.

These surveys are of various types, and can be simple or complex in nature. They generally have differing stratification schemes, may have multiple-phase samples, and deal with various sources of information. For example, they may process administrative or historical data in addition to current survey data. While they are mostly business and agricultural surveys, there are also a number of social surveys which use the systems.

3. GOALS AND OBJECTIVES

In 1984, there was a major redesign of the business surveys at Statistics Canada. This played a crucial role as the catalyst in generalized systems development. In principle, Statistics Canada always prefers to buy rather than develop tools for survey processing. However, at the time of the redesign such tools, especially of a generalized nature, were rare in the commercial sector and not suited to Statistics Canada's needs. If some did exist at Statistics Canada, they had been customized according to the needs of the particular application. Thus, development of generalized systems was initiated "from the top" for the above three survey functions.

When the development of these generalized systems did get underway, the main goal was to support the most common methods of sampling, edit and imputation, and estimation that were being used at Statistics Canada. At the same time, it was recognized that some flexibility also had to be built into the functions. In this way, the duplication of effort across the customized systems could be reduced. The requirements for satisfying this goal had already been gathered over a multi-year period during which this duplication of effort was occurring.

User-friendliness of the systems was one of the objectives to be strived for in reaching the overall goal. The use of graphical user interfaces, modular components, and efficient input-output functions were seen as the primary means of achieving this objective. In particular, graphical user interfaces were identified as an important tool to aid survey personnel in achieving a short learning curve in the use of the systems and thereby speed up development of survey applications.

Another objective was to focus on the use of a single data structure for all three systems. Early on, executive decisions steered development toward the use of Oracle as that data structure. It was seen as the most technically sound of the options investigated. It was soon realized, however, that despite the technical advantages of Oracle, it was not always suitable in the computing environment at Statistics Canada, in particular for computationally-intensive mathematical functions such as sampling and estimation.

In the mid-1990s, after the initial versions of the systems had been produced, a study of standardization of system architectures at Statistics Canada recommended that SAS be used as a common tool in survey processing. There were several reasons for this, including the fact that there was already widespread usage of SAS at Statistics Canada and thus a broad knowledge base already existed. SAS also has the advantage of portability across the different computing platforms in use at Statistics Canada, because of its availability on all of them. Finally, SAS is designed to be efficient with respect to the type of data processing associated with sampling and estimation functions. This has resulted in a policy of migrating automated survey functions to SAS when they are redeveloped.

The three generalized systems now appear to be mature enough in terms of the functionality they offer. This has prompted Statistics Canada to focus the current objectives on enhancement of the structure of the systems. The measures being pursued include the development of completely independent modules for all systems, adding flexibility in terms of the computing platforms and database formats, enhancing the user interfaces, and further integrating the existing customized tools into the generalized systems.

4. FIRST VERSIONS

4.1 Development

Behind the initial development of each of the systems was an individual who had great expertise in the methodology of the major functions of the system. While the enthusiasm of these experts was not the sole motivation for the decision to build the systems, they were the driving force behind the step-by-step development of the systems once that decision had been made, and their contributions were invaluable.

The general approach taken in the creation of the systems differed from one system to the other. GES and GEIS were developed to satisfy specific needs which could then be applied to a multitude of Statistics Canada surveys. For example, error localization with minimum change was one such specific need for edit and imputation. On the other hand, GSAM was built with the idea of satisfying some general needs for a specific business survey. Later, GSAM was generalized to meet the sampling needs of a much broader range of surveys.

For these initial versions, the period of development and testing comprised 2 to 3 years. For follow-up versions, up to 1 year was required for the actual development work, followed by 4 to 6 months of acceptance testing. Development of these later versions of the systems was usually undertaken to add some new functionality, which was identified through a mix of user support activities, user surveys, discussion groups, and advice from technical committees. Occasionally, technical requirements such as those resulting from changes to the foundation software or mainframe computer operating system necessitated changes to the systems themselves. Any modifications to the system which brought about development of a new version always had to be well justified using cost/benefit analyses before being approved by senior management.

The development teams for the systems were multidisciplinary, with mathematical statisticians coming from the Methodology Branch of Statistics Canada, and systems developers coming from the Informatics Branch. For the early development, subject matter personnel from various areas of Statistics Canada were also included. These teams were managed in a matrix management style rather than hierarchically. The methodology staff and the systems developers from the teams were never collocated during system development. There were several reasons for this, including a lack of common office space and some fear of loss of identity for these two rather distinct groups involved in the development.

As for the program code for the systems, it was and still is usual procedure for the mathematical statisticians to first write detailed methodological specifications. The systems developers then used these specifications in designing and implementing the data structure and program code, in close consultation with the statisticians to verify or clarify the methods specified. This was the procedure followed for both GEIS and GES; however the first version of GSAM was in fact programmed by statisticians.

If a prototype was used as the basis for development or for later additions to the system, it was usually written by mathematical statisticians. These prototypes were sometimes incorporated directly into the system, but more often than not they were modified to some degree or rewritten entirely by systems developers before being made part of the system.

4.2 Origins

The first versions of the systems which exist today vary with respect to their original source. While two of the three were based to some degree on pre-existing prototypes or simply stand-alone programs, the other was developed from scratch. It was preferred to start with some pre-existing system if possible to economize on development time and cost. However, this was not

always possible due to the non-availability of such prototypes or simply the fact that it was not feasible due to requirements for the generalized systems that were not compatible with the pre-existing system.

The first version of GSAM was in fact a prototype written by mathematical statisticians to satisfy the general needs of a specific business survey in the distributive trades area. While this worked well for that particular application, it was decided that the system could be used in a more generalized fashion for a broader range of surveys with some modifications made to it. This prototype was then developed further and enhanced by systems developers through the development of an improved user interface and improved input/output functions. GSAM was initially designed to operate in both the SAS and Oracle environments; it exists today only as a SAS-based product.

GEIS was based on a prototype system called the Numerical Edit and Imputation System, or NEIS (see Sande, 1979). The methodology behind NEIS was based on the Fellegi-Holt (1976) principles of minimum change, data integrity, and preservation of data distributions. GEIS maintained this same approach for its first version, and still does so presently. Due to a lack of detailed documentation for NEIS, it had to be reverse-engineered in order to produce GEIS. At the same time, the system was converted from a foundation software of FORTRAN to Oracle/C. The first few versions of GEIS only addressed the functions of error localization, imputation by estimator and donor imputation; the additional functionality of the current system was added later.

The first version of GES was based on a model-based approach to estimation (see Särndal et al, 1992) that was being implemented at Statistics Canada. There were no existing prototypes or programs which used this method upon which to further develop a system, and so the traditional Statistics Canada method of system development was followed, with methodological specifications written by statisticians, followed by the system code being written by systems developers from these specifications. The first version of GES had Oracle as the underlying software. However, due to some performance problems which were experienced due to the frequent writes and reads to/from the data tables, GES was later converted to a SAS-based product.

5. THE BENEFITS VS. THE COSTS

As Doucet (1992) said, there are many benefits in using generalized systems but there are also costs associated with their development. Before initiating the development of a new system or modifying an existing one, it is recommended to compare the benefits against the costs. At Statistics Canada, before releasing funds for the development of systems, senior management considers the potential users, the expected development costs, the impact on existing functions, and the chance of success - or the risk of failure. Their expectation is to attain functions that will be frequently used by survey methodologists and analysts.

From the users' point of view, advantages of generalized systems include the availability of good "error-free" functions, good related documentation, technical support, a growing expertise coming from an increasing number of users, an efficient implementation, the reduction of application development time, and the availability of complex methods that would not be developed otherwise.

On the other hand, anybody involved in the development or the use of a generalized system has to face disadvantages. Managers must deal with cost issues. Amongst other things, we know that generalized systems involve a long development schedule. The functions obtained from the development activities are generally complex and difficult to maintain. These two facts justify the need for good funding sources. From the developers' point of view, the complexity of the functions makes the job sometimes too challenging. Indeed, typical mathematicians do not have the technical knowledge to join such a team while systems developers are concerned with understanding the statistical concepts behind the systems. Furthermore, proposed changes must go through a series of management committees, and when approved, new ideas are often difficult to fit into existing system designs. As for the users' perspective, the use of generalized systems often means that pre- and post-processors must be developed. This is because the assumption behind the generalized systems is that data will "fit" them, as opposed to the customized systems which fit the data.

With the disadvantages listed above, there are also some risks in developing generalized systems. First, the dependency on third-party software adds a level of uncertainty. When old software versions become obsolete, systems have to be modified. That happened with GEIS at Statistics Canada. Its high dependency on Oracle was the first reason to stop its development on Windows, because of Oracle not being supported anymore on that platform. Another risk is the failure of a big project such as a generalized system. The decision to stop the development of the Data Collection and Capture system (DC2) was made knowing that the recovery of investments would not be complete. The

potential errors in methodological specifications or program coding, even if not likely, represent a third aspect of the risks. Such errors would impact many applications. Finally, a fourth aspect comes from the long development schedule. Some users may decide not to wait for a generalized system. A customized system with alternative methods would then be preferred.

6. DEVELOPMENT COSTS

We discussed cost considerations versus benefits in the previous section. The difficulty here is to quantify the cost. We already mentioned that systems are rarely developed from scratch in the Statistics Canada context. Every version is made possible by previous research activities, and for every successful idea, there are unsuccessful ones that had to be funded. As well, development ideas usually come from the experts only after years of experience have already been accrued. There is obviously a cost associated to this but it is difficult to quantify. For these reasons, any cost evaluation would not consider the time spent on pre-existing systems from which generalized systems are developed, nor the research activities, the past work or the development of expertise.

With these caveats, we can say the development of the first few versions of a generalized system roughly requires 2 to 4 mathematical statisticians and 3 to 5 systems developers (for an average of 7 persons) for a three-year period. This represents roughly 20 person-years for each system, but some systems require more resources, like GEIS, and others require less, like GSAM. After the initial development, system maintenance becomes mandatory because foundation software keeps changing. We also know that systems have to continuously evolve since otherwise, they would die in the user's perspective, with legitimate reasons. System maintenance and improvement are carried out at Statistics Canada at the cost of about 4 persons every year for each system. Globally, this means an average of 40 person-years have been spent on each system over the past 10 years to bring them to their current level. This excludes the 2 persons that are allocated every year to each system to provide users with technical and methodological support.

7. ARE GOALS AND OBJECTIVES MET?

The first goal of the generalized systems is to reduce the development costs required for customized systems. In order to evaluate this aspect, the support team tries to monitor the use of each system. Over the years, we notice that the number of surveys developed using generalized systems is increasing. Furthermore, the size and complexity of these surveys are also increasing. For instance, the Unified Enterprise Survey that integrates several business surveys was developed almost exclusively based on generalized systems.

A user database has been developed to store information about known users. It serves to monitor the number of users, to justify the support resources, to evaluate the benefits of generalized systems and to contact users for various reasons. This offers a good opportunity to identify the level of satisfaction and the needs regarding future development. In that context, the benefits are measured from a qualitative perspective rather than quantitative. The savings from the reduced development of customized systems would be too difficult to quantify.

Nevertheless, there are clear advantages in increasing the internal use of the systems to optimize the initial investment. This objective is achieved by making the system available at no cost to all Statistics Canada employees. The user support, whenever needed, is also provided internally at no cost to the users.

The users' satisfaction is another objective which helps achieve the main goal. When the product pleases the clients, they become more inclined to use it. The satisfaction and expectations are gathered through several vehicles like informal discussions during support activities, user discussion groups, and user surveys. These represent the main sources of information behind the continuous development.

8. HOW ARE THE SYSTEMS PROMOTED?

The development of generalized systems is justified only when there are users at the other end. Although nobody is ever forced to use generalized systems, senior management strongly encourages survey developers to use them. There are always cases where customized systems are preferable. In those cases, survey developers do not really have to justify their decisions but simply have to show their awareness of the existing corporate tools.

As mentioned before, a support team was put in place to offer both methodological and system support at no cost for every user. This is a great incentive to use the generalized systems. Other promotional activities are carried out through discussion groups, an intranet site, presentations to recruits, promotional seminars, demonstrations as part of general training programs, and free documentation.

The philosophy of the support team is that a good product reaches some levels of self-promotion. An important feature from the users' perspective is the availability of diagnostic reports to evaluate the choice of statistical method or the use of options. The reports also offer performance diagnostics when problems arise. For instance, in case of an unexpected interruption of GEIS processing, logs give information on what happened and jobs can be launched again from intermediate steps.

A second important feature is the availability of full system documentation. Unfortunately, the support team sometimes encounters staffing problems which cause delays in the delivery of up-to-date user guides, tutorials and on-line help documents.

9. LESSONS LEARNED AND FUTURE DIRECTIONS

There are always lessons to be learned from every project. In the case of our generalized systems, we realized that the scope of each decision was not always clear. It happened that decisions were made for some reason and had unexpected impacts on other aspects. Whether these impacts were good or bad, they represented lessons for everybody involved in such development projects. New directions have recently been adopted as a reaction (see Kovar, Jeays, and Poirier, 1999). Our lessons learned are the following.

Incremental development: The successful approach Statistics Canada adopted was to initially implement a few commonly-used methods, and then to gradually add other useful ones. This allowed the release of intermediate versions for users' benefit and facilitates management of the project.

A common foundation software: The choice of a common foundation software is important for a suite of systems. This approach was adopted initially with Oracle, mostly for performance reasons. Unfortunately, we realized later we could never make our users comfortable with this software, as opposed to SAS. Our current efforts target a slow migration of our systems to a SAS environment, involving custom-written SAS procedures using the C programming language.

Individual statistical functions: Developers may be tempted to develop a singular high-level statistical system where the various functions such as sampling, collection, editing, imputation, estimation, etc., are completely integrated, but the recommendation is to target individual components. The individual components are easier to maintain, to use, to transfer, and to support.

Modular development: Similarly to the individual functions, the creation of modules within the functions makes the maintenance easier. Power users would also benefit by changing the sequence of modules or replacing some with their own customized modules for special requirements. Some systems were not developed in a modular fashion and we are currently redesigning them with a modular approach. In some cases, the performance may suffer but this does not alleviate the benefits.

Graphical user interface: The development of a graphical interface, in the context of generalized systems, is time consuming. Statistics Canada has in certain cases dedicated more time to building interfaces than to developing the related statistical functions. The interface benefits do not justify their high costs at Statistics Canada, especially with the relatively small number of users.

These lessons drive the future directions of the systems' enhancement. While the short-term directions target the addition of functions to satisfy statistical programs, the longer-term directions aim at better integration of the suite of generalized systems. This targets the migration to SAS, a more appropriate foundation software. The Standard Economic Processing System (StEPS) from the United States may give a good example of such integration (see Ahmed et al, 2000).

As a second objective, a tool library is being prepared to offer functions and sub-functions to users outside of the rigid interface framework. This approach fits the need for modules, as described above.

10. CONCLUDING REMARKS

The development of generalized systems is a tremendous effort that requires good planning. While the goals and objectives focus on savings in the processing of statistical data, the scope and content have to grow slowly. With respect to this, only functions that are not already available should be targeted. At Statistics Canada, most generalized systems are the result of the redesign of pre-existing systems, with only a few changes being made every year. This ensures that usable functions are released to users while development is still ongoing.

ACKNOWLEDGMENTS

The authors would like to thank the referees for their insightful comments.

REFERENCES

AHMED, S., and TASKY, D. (2000). "An Overview of the Standard Economic Processing System (StEPS)". *Proceedings of the Second International Conference on Establishment Surveys*, 633-642.

BANKIER, M., LACHANCE, M., and POIRIER, P. (1999). "A Generic Implementation of the Nearest-Neighbour Imputation Methodology (NIM)". *Proceedings of the Survey Research Methods Section*, American Statistical Association, 548-553.

DOUCET, J. E. (1992). "General Survey Processing Software: The Argument for Investing in It". Statistics Canada Technical Report.

ESTEVAO, V., HIDIROGLOU, M. A., and SÄRNDAL, C.-E., (1995). "Methodological Principles for a Generalized Estimation System at Statistics Canada". *Journal of Official Statistics*, 11, 181-204.

FABER, G. B., LANIEL, N., and YEO, D. M. (1998). "Overview and Strategy for Version 1.2 of the Generalized Sampling System". Statistics Canada Methodology Branch Working Paper No. BSMD-98-007E/F.

FELLEGI, I. P., and HOLT, D. (1976). "A Systematic Approach to Automatic Edit and Imputation". *Journal of the American Statistical Association*, 71, 17-35.

KOVAR, J., JEAYS, M. and POIRIER, C. (1999). "Generalized Systems: Where Are We At and Where Are We Going?". Statistics Canada Technical Report.

KOVAR, J., MacMILLAN, J.H. and WHITRIDGE, P. (1991). "Overview and Strategy for the Generalized Edit and Imputation System". Statistics Canada Methodology Branch Working Paper No. BSMD-88-007E.

OUTRATA, E. and CHINNAPPA, N. (1989). "General Survey Functions Design at Statistics Canada". Statistics Canada Technical Report.

SANDE, G. (1979). "Numerical Edit and Imputation". Presented at the 42nd International Statistical Institute Meeting, Manila, Philippines.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

STATISTICS CANADA (1999). "Generalized Systems: Products and Services". Statistics Canada Technical Report.