STATISTICAL METHODS RESEARCH AT STATISTICS CANADA

David A. Binder, Business Survey Methods Division, Statistics Canada
11th Floor, R.H.Coats Building, Ottawa, Ontario, Canada, K1A OT6.

ABSTRACT

This paper reviews a number of the issues associated with conducting statistical methods research at Statistics Canada. The organizational setting is described briefly, and a list of research priorities is given. Some of the issues of managing a research program are discussed.

1. BACKGROUND

Statistical programs in national statistical offices are constantly evolving. Many of these offices are facing increasing pressure to improve the quality and timeliness of their outputs, while at the same time endeavouring to reduce costs and respondent burden. An active statistical research program is an important element in achieving these goals. However, organizational structures and funding arrangements for accommodating this research differ substantially among agencies. In this paper, the framework at Statistics Canada for conducting such a research program is described.

Statistics Canada comprises six major "Fields". The three Fields with primary responsibility for the agency's statistical programs and where most of the subject matter experts are situated are "Business and Trade Statistics", "Social, Institutions and Labour Statistics", and "National Accounts and Analytical Studies." The Divisions which provide the methodological services for these Fields, are centralized within the Methodology Branch, which is situated in the "Informatics and Methodology" Field. The three Divisions providing these services are Business Survey Methods Division, Social Survey Methods Division, and Time Series Research and Analysis Division. With some minor exceptions, staff in Business Survey Methods Division provide methodological services to the Business and Trade Statistics Field; similarly staff in Social Survey Methods Division provide services to the Social, Institutions and Labour Statistics Field. Both Divisions provide some services to the National Accounts and Analytical Studies Field. Time Series Research and Analysis Division provides services to all subject matter areas requiring expertise in seasonal adjustment and other time series issues.

In general, the statistical methods program aims to ensure that effective and efficient statistical methods are available to the Agency's statistical programs. The delivery of the program depends on the development and encouragement of well trained, highly motivated employees. This is partially achieved through an active survey methods research program.

Although the methodology function is centralized within the Methodology Branch, the staff associated with the statistical methods research program are decentralized within the Branch, in the sense that staff working on research projects are drawn from Business Survey Methods Division, Social Survey Methods Division, and Time Series Research and Analysis Division. For the most part, the research is conducted on a part-time basis by methodology staff, most of whose time is allocated to the provision of service to other projects. This is done deliberately to ensure the relevance of the research on the one hand, and also to ensure that the results of the research are applied when appropriate. As well, staff who are not research-oriented have easy access to those who have developed greater "specialist" knowledge through their research endeavours, because they work in close proximity.

One of the problems which many organizations experience is that research successes are slow to gain acceptance in the applications areas. It is felt that if those who perform the research are also working on applications, or are working closely with others who are providing methodology service, then the likelihood of applying the research results is greater. "Partnerships" between those working on the research and those providing the methodology service are strongly encouraged.

2. METHODOLOGY SERVICE FUNCTION

To better understand the context of the research program, we first enumerate the major methodological services offered by the statistical methods staff. These services include involvement in a variety of different activities. These include:

- total survey design,
- the development of survey strategies,
- the creation of survey frames,
- improving the design of questionnaires,
- developing sampling plans, including stratification and allocation,

- developing methods for estimation, including estimation of variances,
- testing data collection and data capture alternatives,
- developing techniques for edit and imputation and for handling non-response,
- ensuring confidentiality protection,
- developing methods for quality control, data quality measurement and data quality improvement,
- advising and performing data analyses, taking into account the sample design where appropriate,
- advising on appropriate use of seasonal adjustment methods,
- advising on best methods for record linkage.

Resources for providing these services vary widely across surveys, with those surveys undergoing development or redesign getting a higher level of methodological resources than the more stable surveys. Each year, the level of resources attached to each survey is carefully reviewed for all the statistical programs.

3. METHODOLOGY RESEARCH AND DEVELOPMENT PROGRAM

As a result of the ongoing delivery of methodology services, a number of problems and issues which need to be studied in greater depth are often identified. These form the basis for determining the research agenda. Some of the current research issues are as follows:

- Research on methods to **improve the design of questionnaires** has been actively pursued both at Statistics Canada as well as outside. Use of cognitive methods and focus groups is gaining popularity. At Statistics Canada, the Questionnaire Design Resource Centre was set up about six years ago and has been involved in the review and development of many questionnaires.
- There is renewed interest in the **design and analysis of repeated surveys**, including longitudinal surveys. This is becoming more important, as Statistics Canada is developing an increasing number of longitudinal surveys. This year, Statistics Canada's annual **international symposium** is on the Design and Analysis of Longitudinal Surveys. This is the ninth Symposium which has been organized at Statistics Canada. The tenth Symposium, which is a joint effort with a number of statistical organizations, to be held in June 1993, will be on Establishment Surveys.
- As part of a continuing effort to use generalized software to reduce overall survey costs, the development of a **Generalized Sampling System** (GSAM) is partially funded through the research program. This system will include a front-end analysis function for optimizing the sampling design. It will also include sample selection and maintenance functions.
- Computer assisted **collection and capture** is now technically more feasible with the advent of faster and smaller computing devices. As well, over the last few years we have been increasing our use of telephones for data collection. It is necessary, therefore, to better understand the impact on data quality resulting from these changes in methods of data collection. For example, choosing the most appropriate tools for editing the data needs further study.
- Since the use of telephone frames is becoming more important in household surveys, **multiple frame methods** will likely be used more frequently. Also, multiple frames are now commonplace in agricultural surveys which use both list and area frames. There are many issues relating to the use of dual or multiple frames. These include issues of design, processing, and estimation.
- **Modelling of non-response** behaviour is an important issue. Many of the techniques which are used for non-response adjustments can be justified under various model assumptions. Methods to improve non-response modelling are important to reduce the non-response bias. However, **reducing response errors** is often more important than non-response and sampling issues, since such reductions may require significant changes to the survey design. This is an area which tends to receive little attention in the literature, because the errors are difficult to quantify. One example which has received attention in the literature is the estimation of gross flows (period-to-period transitions) of labour force status. It is easy to show how the direct estimates based on the Labour Force Survey can be severely biased under small

errors in the classification of labour force status.

In the area of **quality assurance**, we are investigating the feasibility of using statistical process control methods, in addition to currently implemented statistical quality control, for assuring the quality of the repetitive manual operations in data processing. We would like to develop more techniques for total quality assurance. This has led to a whole range of new topics to be addressed.

Edit and imputation research has been active at Statistics Canada for several years. As a result, we have developed generalized software (GEIS - Generalized Edit and Imputation System) which is being used on a number of surveys. Most of the recent developments have been for quantitative data; more development is needed for qualitative and mixed data. Further developments in editing qualitative data are being considered for the Census of Population.

The emphasis of edit and imputation research is shifting towards reviewing the various editing practices at Statistics Canada. Alternatives such a selective editing and macro-editing methods are being considered as a result of studies that show that data tends to be overedited, editing is expensive, and data quality can be maintained even with a reduced level of inspection. Major changes in editing practices could lead to substantial resource savings, with only a minimal impact on the data quality.

Automated coding is a topic which has undergone much development over the last few years, particularly for the Census of Population and for household surveys. Current methods have performed well for variables such as language, religion, ethnic origin (ancestry), place of birth (country), mobility (city/township), and major field of study. In the processing of the 1991 Census of Population, over 200 person-years were saved by using automated coding for these variables. This represents a reduction of over 85% of the person-years required for coding these variables manually, and over $4M were saved. However, more research is needed to improve methods for industry and occupation coding.

- In the area of **estimation research**, there has been some work on alternatives to raking ratio estimation and other reweighting schemes. As part of the development of a Generalized Estimation System (GES), this research includes topics such as calibration estimators (generalized raking estimators) as well as estimating the variance when some data have been imputed. Other issues estimating the variance of seasonally adjusted data, and estimating the variance of the Consumer Price Index.
- **Outlier detection and treatment**, which has received much attention in the statistical literature in non-survey contexts, requires quite different procedures in survey contexts. At Statistics Canada, we have been studying this for a few years and we will incorporate some of our recommended practices in the Generalized Estimation System.
- There has also been developments in **benchmarking** methods, which are useful when adjusting the survey estimates when given a more reliable, but less frequent estimate of the same characteristic. Recent developments here include using improved models to explicitly account for the differences between the more frequent but less reliable statistics and their benchmarks.
- In the area of **time series modelling**, there has been recent work in topics such as calendarization, trading day variation, movable holiday effects, extrapolation methods, and diagnostics for the X11-ARIMA program. Time series modelling and seasonal adjustment methods are continually being updated and improved. The X11-ARIMA package is now in wide use at Statistics Canada, as well as at a number of other statistical agencies throughout the world. Research is being conducted on improving the trend-cycle filters and using smoothing for improving the seasonal adjustment procedures when the series is affected by long cycles or when the data has too much noise. Better diagnostics for the X-11/ARIMA program are being developed for testing for such phenomena as changing seasonal patterns. Recently some research has begun into

FARMA (Fractional Autoregressive Moving Average) models.

Demands for more detailed data without substantially increasing costs of data collection and respondent burden has led to continuing research **in estimation for small domains**. Theoretical developments need to be applied to real data to better understand their properties. Recent research has included combining both small area estimation methods and the use of time series models to improve estimates. The two are closely related in terms of their methods.

Recent developments in **analysis of data from complex surveys** include contingency table analysis, fitting proportional hazards models, and estimating measures of income inequality. The Data Analysis Resource Centre has been recently created to assist analysts within Statistics Canada to use the most appropriate and up-to-date methods. As well, the centre will evaluate hardware and software needs for data analysis.

Methods to **measure and improve survey coverage** will always be an important methodological concern for both census and non-census applications. For business surveys, this includes measuring classification errors and estimating for missing births.

Recent developments in **confidentiality protection** have led to the development of improved software. However, there are still a number of outstanding issues, particularly in the area of release of public use microdata files.

The theory of **record linkage** is still evolving. Recent developments need to be assessed for their practical applicability. CANLINK, our record linkage, software is continuing to be used for a wide variety of applications.

Some work has started in the **use of geographic data** to improve various aspects of survey methodology, such as data collection, small area estimation, frame improvements, use of auxiliary data for estimation models. Possible applications here include improving data collection using computer generated maps, improving sample designs by incorporating geographic variables in stratification, improving estimation methods by incorporating geographic auxiliary data, and estimating surface areas for farms.

- **Unstructured research:** Some time is allocated to accommodate research which is not explicitly funded. Such research often leads to new innovative research projects. Some of these resources are also used for other tasks such as preparation of presentations and refereeing papers.

The research program is also responsible for the development and delivery of various **statistical courses.** Courses are developed and presented on a variety of topics to Statistics Canada staff. These include courses in survey methodology, questionnaire design, data analysis, and quality control.

The Statistics Canada publication, ***Survey Methodology***, is an important vehicle for communicating recent research developments. The research program includes providing resources for editorial and production functions of this journal.

3. SETTING RESEARCH AND DEVELOPMENT PRIORITIES

Clearly the list of important topics is long and resources are not available to address all topics concurrently. Therefore, some priority-setting mechanism needs to be put into place. At Statistics Canada, statistical methods research is funded from a variety of sources. One of the major sources is the Methodology Block Fund which comprises a preallocated number of person-years. It is called a Block Fund because this "block" of resources is removed from the overall resource base available to meet operational needs. Each year, resources from this Block Fund are allocated to priority areas for the year. This allocation is determined by the Research and Development Committee. This committee also reviews Informatics Research and Development activities. Throughout the year, the researchers and senior mangers at Statistics Canada are consulted to help ensure that the research agenda remains relevant to Statistics Canada's needs.

However, not all the research activity is funded through this Block Fund. Research that is survey-specific tends to be funded through the individual survey budgets. For example, research into cluster analysis for stratification for an agriculture survey would not normally be funded through the Block Fund. However, since the results of such research could also be used by other program areas, it is important to communicate the results of such research through seminars and papers, where appropriate. At times, survey-specific research could lead to more general research and be supported through the Block Fund in a subsequent year.

The annual allocation to the Methodology Block Fund consists of 23 person-years, including specially earmarked resources for programs such as the Questionnaire Design Resource Centre. This represents close to 10% of the resources in the Methodology Divisions.

4. GENERALIZED SURVEY FUNCTION DEVELOPMENTS

In addition to the Methodology Block Fund, another important initiative at Statistics Canada is the Generalized Survey Function Developments (GSFD) project. This project was initiated in response to the need to have more efficient tools available for designing and redesigning survey functions and to expedite the conversion of surveys to a new systems hardware environment. Its main goal is to create generalized processing tools and software that are portable across various computing architectures.

The **Generalized Sampling System** (GSAM) is a modular set of programs which will support basic survey sampling functions, including stratification, sample size determination and allocation, and sample selection and maintenance. The **Generalized Data Collection and Capture System** (DC2) is a software product which provides specification, management, and operation of data collection processes at a production worksite. Its implementation is through a set of well-defined tasks and a management support system. The **Generalized Edit and Imputation System** (GEIS) is a collection of modules that ensure that the edit and imputation requirements for numeric, continuous and non-negative data are satisfied. The **Generalized Estimation System** is a modular set of programs which will support basic survey estimation functions, including variance estimation. It will be used for the design-based estimation of means, totals, ratios, and other more complex statistics under a variety of sampling designs.

There is a strong interaction between statistical methods research, informatics research, and the software being developed by GSFD. Much of the past research, such as application of coordinated sampling, edit and imputation for quantitative data, application of a unified estimation theory, and detection and treatment of outliers will be embedded in the GSFD products. As well, the development of the GSFD products has led to a number of new research problems which have been added to the research agenda. These are often related to computing and efficiency concerns. As well, hardware improvements have also led to evaluating the use of new technologies in data collection and capture.

5. COLLABORATIVE RESEARCH

In addition to our own research efforts, Statistics Canada appreciates the benefits from collaboration with other agencies facing similar research issues. We feel that we all can benefit by combining efforts in this way. This is particularly beneficial in areas such as data collection, data editing, and estimation research.

We also work closely with a number of university professors, either on a contract basis, or by collaborating on certain research problems. We have an active program encouraging eminent external researchers to spend time in Statistics Canada as Research Fellows.

6. MANAGEMENT ISSUES

We are constantly concerned with the relevance of our research agenda, and are continually looking for ways to improve the current process. We want to ensure that the research agenda is not only relevant for today's issues, but that there is also the opportunity for doing research which has potential relevance in the longer term.

As described earlier, much of the research is decentralized in the sense that persons working on research problems are also providing methodology services to other areas of the Bureau. This has the disadvantage that if there is a conflict between the research project and an operational project, in terms of the demands on a person's time, there is a tendency to drop the research and "put out the fire" which has arisen on the operational side. Therefore, it is necessary to develop a system for monitoring this tendency to reallocate resources away from the research program. In a centralized research organization, where staff work on research projects without conflicting demands on their time, this problem is less likely to occur. However, the disadvantages of such an organization, in terms of having less contact with those working on operational projects, can be much more detrimental to the eventual application of the research accomplishments.

Throughout the year, there are regular progress reviews of all the research projects. All staff involved in the research program participate in these reviews. Semiannual progress reports are discussed by the senior managers in the Methodology Branch.

A number of management groups review the plans for the methodology research programs each year. The Informatics and Methodology Research and Development Committee discusses the allocation of the research budget to projects each year. These are presented to various other Management Committees for their input. All these committees contain members from outside the Informatics and Methodology Field, and their input adds an important perspective to the statistical methods research program. Such presentations provide a mechanism to display our yearly accomplishments to other managers.

As well, we greatly benefit from the stimulus provided by an external Advisory Committee on Statistical Methods. This is a committee which consists of well known academic and research statisticians which reviews, criticizes, and stimulates not only our research, but also our applied methodological work.

7. STAFF DEVELOPMENT

One of the aims of a statistical methods research program is professional development of our own staff. We try to ensure that those who have the potential to do good research are given the opportunity. This should help develop the type of methodologists who can be more effective in the delivery of their service function. We believe that the research program contributes to overall professional development of our staff. Many of these benefits are intangible, since it is difficult to enumerate such benefits in measurable terms.

As well, we feel that such a research program is a contributing factor to attracting high calibre staff. By gaining recognition and respect from universities as a result of our own statistical research efforts, and by collaborating with them on some research projects, we hope that the university professors would encourage their better students to consider working at Statistics Canada.

8. CONCLUSION

Statistical research is an important function in national statistical offices. It should be structured in such a way that the important research accomplishments are accessible to and used by those who are not actively involved in the research program. Mechanisms are required to keep the research program relevant and free of conflicting demands. There is no shortage of important research problems which deserve study.

This paper has described how these requirements are addressed at Statistics Canada. Different agencies will use other frameworks, as there is no perfect structure which all agencies can emulate. One of our goals is to create and improve an environment in which relevant research can flourish.

ACKNOWLEDGEMENTS

I am grateful for the many comments which I have received to improve the content and readability of this paper. In particular, I wish to acknowledge contributions made by Gordon Brackstone, Nanjamma Chinnappa, Michael Hidiroglou, John Kovar, Georgia Roberts, and Jocelyn Tourigny.