

12-001

C.3



Statistics
Canada

Statistique
Canada

STATISTICS STATISTIQUE
CANADA CANADA

AUG 5 2010

LIBRARY
BIBLIOTHÈQUE

SURVEY METHODOLOGY

TECHNIQUES D'ENQUÊTE

December - 1977 - Décembre

VOLUME 3

NUMBER 2 - NUMÉRO 2

A Journal produced by
Statistical Services Field
Statistics Canada

Publié par Le Secteur des
Services Statistiques,
Statistique Canada

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

C O N T E N T S

Confidentiality of Statistical Information	
D.A. WORTON	127
Synthetic Estimation In Periodic Household Surveys	
P.D. GHANGURDE and M.P. SINGH	152
L'enquête sur la profession des salariés (E.P.S.)	
M.A. HIDIROGLOU	182
Some Factors Affecting Non-Response	
R. PLATEK	191
Survey of Canadian Residents Returning By Land	
J.H. GOUGH and P.D. GHANGURDE	215
An Investigation of the Properties of Raking	
Ratio Estimators: II With Cluster Sampling	
H.R. ARORA and G.J. BRACKSTONE	232

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1977

Vol. 3

No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Editorial Board/ Comité de rédaction:	R. Platek	- Chairman/Président
	M.P. Singh	- Editor/Rédacteur en chef
	G. Brackstone	
	P.F. Timmons	

Assistant Editors/ Rédacteurs adjoints	J.H. Gough
	P.D. Ghangurde

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed, however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department. Copies of papers in either Official Language will be made available upon request.

Politique de la rédaction:

La revue Techniques d'enquête veut donner aux personnes qu'intéressent les aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration des méthodes d'enquête: les problèmes de conception causé par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada. On pourra se procurer sur demande des exemplaires d'un article dans l'une ou l'autre langue officielle.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1977

Vol. 3

No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, 10th Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division de l'élaboration d'enquêtes ménages, Statistique Canada, 10^e étage, Edifice Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi. Les auteurs des articles publiés dans cette revue sont libres de les faire paraître dans d'autres revues statistiques.

CONFIDENTIALITY OF STATISTICAL INFORMATION¹

D.A. Worton

Assistant Chief Statistician, Marketing Services Field

The paper first identifies some of the factors which have recently made it more difficult for statistical agencies to satisfy society's growing needs for information, while at the same time reassuring respondents that their privacy is adequately protected.

The conceptual basis of privacy is then discussed, as well as the privacy provisions of the new Canadian Human Rights Act. The paper next reviews the confidentiality provisions of Canada's Statistics Act by which the privacy rights of respondents are protected. There then follows an account of the circumstances under which the confidential treatment of corporate information is being challenged, and the way in which Statistics Canada is endeavouring to meet governmental needs for access to individual corporate returns in a foreign ownership context without prejudicing traditional confidentiality practices in mainstream statistical reporting.

Finally, the paper notes two subjects which are likely to feature in future discussions of confidentiality: first, scholarly access to historical statistical records; and second, the possibility of future freedom of information legislation in Canada.

1. INTRODUCTION

The most basic prerequisite for the effective operation of an official statistical agency has always been its ability to protect the confidentiality of the information which it acquires about individual respondents. The collection of such information appears as a potential threat to the privacy of the individuals or legal entities involved, but when the scope and content of the inquiries are perceived to be a reasonable expression of

¹ Adapted from a paper prepared for the XIIIth Session of the Committee on Improvement of National Statistics (COINS) of the Inter-American Statistical Institute, November 1977.

society's informational needs, and when privacy is seen to be protected by unassailable safeguards against the unauthorized disclosure of identifiable information, public trust in the agency can be maintained.

During the past two or three decades, however, the increasingly complex economic and social problems which beset most modern societies have enormously increased the needs for statistical information, leading to a corresponding increase in the number and complexity of statistical surveys, many of which now seek information in non-traditional, and therefore potentially sensitive, areas. While users can invariably present convincing arguments for such new statistical requirements, the onus of allaying respondent concerns about the increasing threat to privacy and, in the case of business respondents, the sheer burden of response, rests largely with the statistical agency. This task is not made easier by growing public disenchantment with government and scepticism about the contribution of statistics to decision-making processes.

Again, maintenance of confidence in the statistical process can no longer be considered in isolation from concerns about the broader informational developments of recent years. The growing interventionist roles of governments in economic affairs and their assumption of responsibility for the provision of social services have generated very large bodies of administrative information on the subjects or potential beneficiaries of these activities. With major exceptions such as income tax records, safeguards against the prejudicial use of this information have seemed much less certain than those relating to statistical information, giving rise to a spillover effect by which the traditional relationship between the statistical agency and its respondents has been clouded by larger concerns about the volume and potential abuses of the total information system. Indeed, many of the subjects of this information appear to find it progressively more difficult to distinguish between statistical and non-statistical purposes and applications.

In both the narrower statistical and broader informational contexts, developments in computer technology have aggravated public concerns about privacy and confidentiality. Computers have made it possible for users of statistics to manipulate data on a scale hitherto not practicable, and refinements in analytical techniques increasingly require that these data be furnished in a highly disaggregated form. This, in turn, has shifted the pattern of statistical production from pre-planned tabulations of discrete surveys towards a data base approach in which information can be flexibly retrieved from files of microdata. The mere existence of such files and the technical ease of retrieval from them cannot but make the public apprehensive about the potential for abuses which would have been much more difficult, if not impossible, under earlier technology. One very real and pressing problem is that the greater volume and more detailed kinds of aggregate statistical information now being put into the public domain as a result of the developments referred to increase the risk of inadvertent direct or residual disclosure of individual information [1].

The implications of computer technology have probably appeared more directly threatening in the area of administrative records not protected by confidentiality safeguards. Two or more records of the same individual can be brought together and merged into a single record by linkage techniques, in such a way as to provide a more revealing, and therefore potentially more prejudicial, picture than would be evident from the separate study of individual records. While record linkage can be effected by a number of techniques, the increasing frequency with which personal identifying numbers are called for in various administrative contexts tends to be seen by the public as prima facie evidence of intent to engage in record linkage. These generalized, but nonetheless real, apprehensions have almost certainly inhibited statisticians from pursuing as vigorously as they might record linkage for their legitimate purposes, particularly through the use of personal identifying numbers.

Privacy and confidentiality issues are being extensively studied in the United States. One recent discussion paper [2] documents twelve separate studies, recently completed or still in process, under various auspices. The situation in the United States is complicated by the fact of a decentralized statistical system with substantial variation in the standards for protection of statistical data across the component agencies, and the most basic thrust of current recommendations is directed towards the provision of unambiguous legislative protection for the confidentiality of all information collected solely for statistical purposes.

This paper will review Canadian experience with respect to confidentiality issues, primarily from the standpoint of Statistics Canada which, for almost sixty years, has carried out the greatest part of official statistical collection and production within Canada under extremely rigorous statutory provisions for the protection of confidentiality. The largely centralized nature of the Canadian statistical system differentiates it importantly from the U.S. system, but Canadian problems with respect to confidentiality should not on that account, be regarded as inherently less complex than those of the United States, since there has been a growing involvement of other departments of government, both federal and provincial, in the collection of statistics, with the protection (or lack of protection) of confidentiality being much less explicit than that enjoined on Statistics Canada. To some extent, these problems are addressed in subsequent discussion of Statistics Canada's co-ordinating responsibility vis-à-vis the statistical activities of other departments, but their broader implications lie beyond the scope of this paper.

2. PRIVACY: CONCEPTUAL BASIS

The importance attached to privacy has varied historically within particular societies and similarly varies at any one time between different societies, depending upon prevailing models of social organization. As the introductory section of this paper shows, indications of a heightened

concern with privacy seem unmistakably clear, even though documented evidence is scarce. In this connection, the findings of a recent survey conducted by the Swedish Central Bureau of Statistics, dealing with Public attitudes towards the furnishing of statistical information, are of considerable interest [3]. Of particular relevance here is the fact that, in a list of ten national problems covering such topics as unemployment, inflation, sex discrimination, etc. "protecting people's privacy" was rated third in importance. Furthermore, almost one third of the respondents expressed the view that they currently enjoyed less privacy than they did five years previously.

In clarifying this paper's usage of the term privacy in the informational context as a prelude to the discussion or provisions for its protection, three related definitions put forward in the U.S. paper referred to above [2] may be quoted as follows:

- "1) the right to be left alone, to be spared from unauthorized oversight and observations, and from searching inquiries about oneself and one's business;
- 2) the ability to control the use of information about oneself, whether to give it free circulation, limited circulation, or no circulation at all; and
- 3) the right to participate in a meaningful way in decisions about what information will be collected and how that information will be used."

These definitions recognize privacy as an inherent right of the individual, analogous to other generally accepted expressions of the concept, such as the privacy of the physical person, territorial privacy, etc. They further acknowledge the conflict between public and private interests referred to at the outset of this paper, but also the individual's rights in determining the privacy threshold, i.e. the point beyond which he should not be asked to furnish information. Only outside the bounds of this threshold is privacy potentially at risk.

If no information were to be collected from the individual or generated about him, his privacy would obviously not be at risk. The extreme position in defence of privacy, therefore, is not to collect or generate any information at all. But, in modern societies, the public interest requires such information, and the individual sanctions its creation, albeit tacitly, through the political process. However, society then has a corresponding obligation to guarantee, first, that intrusions across the individual's privacy threshold will be kept to a minimum and, second, to the extent that information needs to be collected or generated, that the individual's privacy will be protected to the maximum extent possible.

The individual's concern about privacy is a function not only of the quantity of the information which is on record about him, but also of its quality. Statistical experience, at least, indicates that there is a reasonable acceptance of the need to furnish certain kinds of "basic", or relatively uncontroversial, kinds of personal information, e.g. demographic, occupational, educational, etc., although this in no way lessens concerns about the need for its proper handling. Other kinds of information, such as that relating to income and housing characteristics, to name only two which have given rise to public relations problems for Statistics Canada, are still much more sensitive.

A formal violation of privacy can be said to have occurred when identifiable information relating to an individual is communicated to others by the collector or generator without that individual's permission. That is what a statistical agency seeks to guard against through its confidentiality rules governing the handling and use of the information it collects directly from individuals or acquires from third parties. However, before outlining and commenting on these rules, as expressed in Canada's Statistics Act [4], it will be appropriate to discuss the legal expression of basic privacy principles in Canada.

3. PRIVACY: LEGAL BASIS IN CANADA

Part IV of the Canadian Human Rights Act [5] seeks to give effect to the principle that "the privacy of individuals and their right to access to records containing personal information concerning them for any purpose including the purpose of ensuring accuracy and completeness should be protected to the greatest extent consistent with the public interest".

The most visible provisions in support of this principle are very similar to those embodied in the U.S. Privacy Act of 1974[2] and may be summarized as follows. First, an annual index is to be published identifying each "federal information bank", i.e. any store of governmental records used for administrative purposes relating to individuals, together with a description of the type of records stored and of the uses deemed to be consistent with the use for which the records were compiled. Second, every individual is entitled to ascertain what records concerning him are contained in these banks, as well as the uses to which the records have been put. The individual has the further right to examine such records, regardless of whether he actually furnished the information they contain, to request correction of errors or omissions in the information, and to require a notation of the requested correction where such correction is not made. These two provisions, for indexing and access, are subject to a number of designated exceptions, in such areas as international relations, national defence or security, federal-provincial relations, and investigations pertaining to the detection and suppression of crime.

Third, every individual is entitled to be consulted and must give consent to the use of information furnished by him to a government institution for a particular purpose before that information can be used or made available for so-called "non-derivative" uses for an administrative purpose¹, i.e., uses going beyond those which might be considered consistent

¹ This appears to exempt statistical uses.

with the uses for which the information was compiled. This provision is expressly not applicable to non-derivative uses authorized by law.

Finally, the Act provides for the designation of a Privacy Commissioner who is authorized to receive, investigate and report on complaints from individuals who allege that they are not being accorded the rights to which they are entitled under the Act.

The strong emphasis upon individual rights in the provisions so far described will do much to reassure the public that Orwell's 1984 is not, after all, just around the corner. Both the actual and potential use of these rights is likely to induce greater responsibility and discrimination on the part of government in adding to the numbers and uses of administrative records. Any attenuation of privacy concerns in a major area such as the latter will probably also be indirectly helpful to better public understanding and support of the statistical system through elimination of the spillover effects referred to earlier.

The statistical system is however, touched upon quite explicitly in the co-ordination and control provisions of Part IV, for the purposes of which information banks are defined so as to include statistical data banks as well as administrative data banks.

These provisions require, first, that the Minister with designated responsibility for the Act "shall cause to be kept under review the utilization of existing information banks and proposals for the creation of new information banks or the substantial modification of existing ones and shall make such recommendations as he considers appropriate to appropriate Ministers with regard to information banks that, in his opinion, are under-utilized or the existence of which can be terminated". Second, "no new information banks shall be established and no existing information bank shall be substantially modified without the approval of the designated Minister".

3

•

The first point that can be made with respect to these provisions is that, given the mandate already assigned by the Statistics Act [4] to the central statistical agency, both for its own direct collection activities and for collaboration with and co-ordination of other parts of the national statistical system, they give rise to overlaps and possible conflicts of jurisdiction with respect to statistical data banks, since two separate lines of ministerial responsibility are now involved. It is reasonable to suppose, however, that some kind of practical accommodation to this situation will develop as the new provisions become operational.

A more difficult problem might arise if it were to be thought that the control provisions could be implemented by considering both statistical and administrative data banks in more or less the same way. Criteria for assessing the usefulness and extent of use of existing administrative data banks, and for creating new ones, seem likely in principle to be simpler than those for statistical data banks since they will presumably derive from the specific objectives of the programs, activities, etc., which the particular data bases are meant to serve. Statistical data banks, by contrast, are usually general-purpose systems which can serve a variety of end-users in a variety of ways. Further, because of the long lead times involved in putting useable and useful data into the public domain, there has to be a considerable element of anticipation in statistical planning. Thus, criteria which seek to protect privacy by placing heavy emphasis on the immediacy and specificity of data use, while appropriate for administrative data banks, would not in general tend to do justice to most statistical proposals.

What the control provisions seek to protect in the case of both administrative and statistical data banks is the privacy threshold. The subsequent protection of the privacy of data once collected rests on different premises in the two cases. In the case of administrative data, the possibility has always existed of their being used to make administrative decisions of kinds different from those for which they were ostensibly collected, and this is what the consultation and consent provisions seek

to protect the individual against. In the case of statistical data, however, the possibility of disclosure prejudicial to the individual simply does not arise, because the data are collected under guarantee of absolute confidentiality. These different circumstances as to the subsequent use of data also suggest the need for different sets of considerations in the determination of the privacy threshold.

Statisticians are, of course, no less concerned than other segments of society about the need for restraint in data collection activities, since the effectiveness of their operations depends upon the maintenance of public confidence. They exercise this restraint through the use of sampling and data sharing arrangements, for example, and are, of course, subject to budgetary constraints like any other operation of government. Nevertheless, their official role as spokesmen for a wide variety of user interests inevitably puts them on the defensive where the privacy threshold is concerned. It is to be hoped, therefore, that the application of the new legislation as it relates to statistical data banks can take into account the factors mentioned, so as to ensure that reasonable proposals for new statistical programs in response to society's demonstrated informational needs are thoughtfully considered.

To this end, it will be even more necessary for Statistics Canada to continually emphasize its independent role as an agency which seeks to serve all clienteles impartially, and to foster a wider and more profound understanding of the protection which it affords to the privacy of the information it collects and holds. This then is an appropriate point at which to review the confidentiality provisions of the Statistics Act [4], in the light of the previously expressed proposition that violations of privacy occur only when identifiable information is communicated to others without the respondent's permission.

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

4. CONFIDENTIALITY: LEGAL AND OPERATIONAL ASPECTS

Canada's Statistics Act [4] authorizes Statistics Canada to collect data from individual respondents, and also to use the individual administrative records of other government departments, to carry out the purposes for which the agency was set up. However, as a counter-vailing guarantee that no harm will result to an individual respondent in consequence of his compliance with the Act, it provides protection against the publication or dissemination of his information in individually identifiable form.

The most basic element of this protection is expressed in the defined mandate of the agency "to collect, compile, analyse, abstract and publish statistical (author's emphasis) information ...". While this broad statement of purpose allows a great deal of discretion as to how the mandate might be carried out, such discretion is exercised within the traditional conception of statistics as a science which describes or characterizes defined populations on the basis of their aggregate rather than their individual properties. Thus the association of information with particular respondents is of no substantive interest to the agency, although it necessarily plays a part in the collection, editing and other procedures leading to the end product. By definition then, the objectives of the agency represent no threat from the standpoint of improper use of collected information, although the public may well consider that the agency pursues its mandate over-vigorously by collecting too much information in the first place.

These objectives are operationally buttressed by three important provisions of the Statistics Act. The first of these is that data collected and stored by the agency may only be handled by employees of the agency, each of whom is required to swear an oath of secrecy under sanction of legal penalties. The second provision protects copies of statistical returns in the hands of respondents by designating them as privileged information which cannot be used as evidence in legal proceedings

of any kind other than prosecutions under the Statistics Act itself. The third provision is one which formally defines unauthorized disclosure by prohibiting disclosure of information obtained under the Act "in such a manner that it is possible from any such disclosure to relate the particulars obtained from any individual return to any identifiable individual person, business or organization".

The first of these provisions is essentially a security provision and is supported by appropriate arrangements for the physical security of statistical records, which increasingly emphasize computer security. This paper will not attempt to exemplify the many kinds of breaches of data security which can potentially occur. It may be noted, however, that the statistical agency's concerns about security are not confined to individual information. The disclosure of aggregate information before its authorized date or time for general release is equally a breach of security.

The third provision imposes a quite different requirement upon the statistical agency, namely for the continuous scrutiny of its publication programme, through whatever medium of dissemination, to guard against the inadvertent disclosure of identifiable information. The problems in this area are numerous and complex and, as noted earlier, are compounded by the more detailed breakdowns of data currently being furnished to users. In their technical aspect, these problems lie outside the scope of the present paper, although they have been closely addressed elsewhere [1].

The risks of inadvertent disclosure can only be avoided completely at the expense of publication policies which would be so conservative as to fall far short of user expectations. The approach must therefore be one of balancing the risks to privacy against legitimate needs for the enhancement of knowledge through a careful comparison of the risks and benefits of each unique situation or class of situation. Progress with the technical problems has resulted in a gradual shift of this

balance in favour of the user. For instance, the dissemination of sample tapes of individual data which are purged of direct identification and detailed geographic codes, although not completely devoid of risk, makes possible types of analysis that cannot be carried out with aggregate data. As with Caesar's wife, however, who was fatally tainted by mere suspicion, public confidence in the statistical agency is extremely vulnerable to exaggerations or imperfect understandings of the risks involved in these trade-offs.

5. CONFIDENTIALITY OF SHARED INFORMATION

The duplicative collection of information is of particular relevance to privacy concerns since, the more times a given item of information is furnished, the greater in general is the potential for its improper use. Again, duplicative collection is also considered by business respondents to be a major factor in the "paperwork burden" [6]; when conducting surveys, statistical agencies are frequently told that "you" (meaning any agency of any level of government) already have the information. On the other hand, respondents are properly apprehensive about the sharing of information unless it is clear to them that this can be done without violating their privacy.

The Statistics Act [4] specifically enjoins Statistics Canada to "promote the avoidance of duplication in the information collected by departments of government", and the kinds of arrangements by which this responsibility is carried out are next illustrated, with particular emphasis on the related confidentiality safeguards.

The statistical exploitation of the administrative records of other government departments is an important means of avoiding duplicative collection when the definitions, content and coverage of the records are a sufficiently close approximation to what the statistical agency needs. Statistics Canada is empowered by Section 12 of the Statistics Act [4] to require access to the records of any department from which information sought in respect of the objects of the Act can be obtained. Information

the first of these is the fact that the
 second of these is the fact that the
 third of these is the fact that the
 fourth of these is the fact that the
 fifth of these is the fact that the
 sixth of these is the fact that the
 seventh of these is the fact that the
 eighth of these is the fact that the
 ninth of these is the fact that the
 tenth of these is the fact that the

CONCLUSION

The first of these is the fact that the
 second of these is the fact that the
 third of these is the fact that the
 fourth of these is the fact that the
 fifth of these is the fact that the
 sixth of these is the fact that the
 seventh of these is the fact that the
 eighth of these is the fact that the
 ninth of these is the fact that the
 tenth of these is the fact that the

APPENDIX

The first of these is the fact that the
 second of these is the fact that the
 third of these is the fact that the
 fourth of these is the fact that the
 fifth of these is the fact that the
 sixth of these is the fact that the
 seventh of these is the fact that the
 eighth of these is the fact that the
 ninth of these is the fact that the
 tenth of these is the fact that the

furnished under Section 12 may frequently have been originally collected under conditions of confidentiality less stringent than those of the Statistics Act and, for disclosure purposes, Section 16 of the Act accords such information the same degree of confidentiality as when collected, so that its availability for future dissemination does not suffer. When the question of disclosing identifiable information to third parties arises, the original collector has control over the manner and extent of disclosure, and Statistics Canada insists that the actual disclosure be undertaken by the collecting department, so that any discretion with respect to disclosure is seen to have been exercised by the latter.

The most important examples of the use of administrative records by Statistics Canada relate to income tax records and customs documents. The former are a major source for the agency's financial statistics and are also used as proxies for small business returns to industrial censuses, while the latter form the exclusive basis for the system of external trade statistics. Neither of these sources is in fact accessed under the authority of Section 12, but rather under the specific authority of a separate section of the Statistics Act. In the case of income tax records, this is because they are subject to extremely rigorous confidentiality safeguards in the hands of the collecting department, and it was felt to be in the best interests of both parties that access in such a sensitive area should be explicitly authorized.

Arrangements for the sharing of purely statistical information, in contrast with those relating to administrative records, have the objective of avoiding duplicative collection by either party. A first kind of arrangement is authorized by Section 10 of the Statistics Act which permits agreements with a statistical agency of any province for the exchange or transmission of replies to original inquiries. Such agreements can only be entered into when the provincial statistical office has statutory authority to collect the information in question, including the power to compulsory response, and when it also operates under

essentially the same confidentiality restrictions as Statistics Canada. Since the rights of the respondent are protected by this latter requirement, his permission for the exchange or transmission of his information under Section 10 is not required. He is, however, notified of the names of the statistical agencies with which agreements exist in respect of that particular collection. Section 10 agreements, then, provide an illustration of what the U.S. paper [2] refers to as an "exchange of information between protected enclaves".

A second kind of arrangement is authorized by Section 11 of the Statistics Act and covers joint collection "with any department or any municipal or other corporation". Because such other parties cannot generally provide the respondent with statutory guarantees of confidentiality, he must be informed and has the right to object to the sharing of the information. Most typically, Section 11 agreements call for each party to maintain strict confidentiality of information furnished by the respondents, with the breach of this or any other term by one party being grounds for immediate termination of the agreement by the other party. A sharing department may, however, wish to publish individual identifiable information and can do so provided the respondent is notified at the time of collection, and can thus take this possibility into account when deciding whether to exercise his right to object to sharing.

The ability of other departments of the provincial and federal governments to collaborate with Statistics Canada in implementing Section 10 and 11 agreements is greatly facilitated by affording them access to lists of the names and locations of individual establishments, which may be characterized by size and detailed industrial classification. This constitutes an authorized exemption from the general prohibition of the disclosure of identifiable information. Although widely applicable in principle, this exemption is used very conservatively - primarily in fact for provincial and federal governments which must undertake, in the case of lists based on taxation records, not to transmit the lists to third parties. This situation comes about because all new accessions to

the agency's business register derive from income tax records of new payroll deduction accounts. Until these new accessions have actually been surveyed by Statistics Canada, even the names are subject to the confidentiality restrictions generally applicable to tax records.

While in principle, Section 10 and 11 agreements are applicable to both business and household surveys, they have in practice been used almost exclusively to share information from business surveys. Statistics Canada has frequently been urged by provincial governments to enter into agreements for the sharing of identifiable microdata from Censuses of Population and Agriculture, but has always rejected such requests on policy grounds, since the viability of Census collections would be extremely vulnerable to public concerns about the sharing of these sensitive data.

6. NEW PRESSURES AGAINST CONFIDENTIALITY

Canada's Statistics Act [4], described in Section 4 above, makes no distinction between natural and legal persons in the protection it affords to the anonymity of statistical respondents. During the last decade, however, the view has been expressed with increasing frequency that the public interest requires a great deal more disclosure of, or access to, corporate information, and that it is inappropriate for large corporations, at least, to enjoy the same rights of privacy as individuals. One recent Canadian study [7] puts the point as follows:

"There is an important difference between the government requesting information from a corporation and from a private individual. A corporation is an aggregation of power with indefinite life and, in most cases, limited liability. It is also a creation of the state; enjoying certain rights and having certain obligations in law. It is to be expected that its creator should wish to be kept informed about its activities, particularly when the decisions of the corporations have a widespread public impact. Firms that help to determine the salary and work conditions of 10,000 or 20,000 people, that play a leading role in our trade, or in a particular industry, are in a very different position than individuals. Thus, what might be regarded as unjustifiable "snooping" in the case of an individual may not apply in the case of a corporation".

The context of the quotation - the problem of foreign ownership and control of economic activity in Canada - is fertile ground for commentary of this kind. An earlier Canadian study in the same area [8], remarking on the common situation whereby wholly-owned subsidiaries of foreign corporations operating in Canada are registered as private companies, clearly considered this to be an abuse of the privacy rights intended for small individually or family-owned firms. "... wholly-owned subsidiaries of foreign corporations can avoid disclosure¹ in spite of the fact that their parents are typically public companies in the foreign jurisdiction, and in spite of the fact that they may be large and important firms dominating important industries".

The perceived adequacy of corporate information depends very much on the purposes of the potential users. The first study referred to [7] suggested four broad purposes, as follows, for which information is required in order to achieve better control of the domestic economic environment:

- 1) identification of the foreign controlled companies;
- 2) economic analysis as a basis for policy formulation;
- 3) implementation of a review process for foreign direct investment; and
- 4) better public disclosure to improve operation of capital markets.

Purposes of this kind probably describe the informational requirements of almost any economic problem in which the characteristics and performance of the corporate sector are a major factor. Their interest from the standpoint of the present paper is that three of them 1), 3), and 4) - can only be addressed on an individual company or enterprise basis. Indeed, arguments can also be put forward that, in situations where particular industries are dominated by a few large firms, the second purpose of framing effective policies can only be fully

¹ Subsequent amendments to the Canada Corporations Act now oblige federally incorporated private companies of economic significance to file with the incorporation authority the same kinds of information as for public companies.

achieved by analysis of the characteristics and performance of the individual firms. A review of the availability of information at the time of the study (1971/2) indicated - and the situation has not greatly changed since that time - that a good deal of information relevant to these purposes already existed, but that much of it had been collected under conditions of strict confidentiality by the federal income tax authority and Statistics Canada, and was therefore inaccessible to the interested government departments.

The question therefore arose, and is still an active one, as to how the necessary degree of access and disclosure might be brought about. In particular, should there be any departure from traditional confidentiality practices on the part of the statistical agency? Or should the various departments and agencies resort to duplicative collection in order to carry out their particular responsibilities? The notion of a number of governmental purposes being simultaneously satisfied through a single collection of data is an attractive one, but against this must be set the loss of confidence in the integrity and impartiality of the statistical agency that would occur if the business community thought that the statistical system was being indiscriminately used for non-statistical purposes.

In considering what scope actually exists for Statistics Canada to facilitate a greater degree of access to business data, but at the same time maintain the confidence of business respondents, account must be taken of the two distinct statutory authorities under which the agency collects business statistics, and of the separate traditions associated with these authorities. In respect of the longest-standing and most far-reaching in scope of these, the Statistics Act [4], the point has already been made that the continuing co-operation of respondents depends upon unqualified assurances about the confidential treatment of the information they furnish. The Act sanctifies a long tradition of mainstream statistical collection in which corporations have enjoyed the same rights or privacy as individuals. Any movement towards the diminution of these rights would almost certainly result in measurably less co-operation from the

business community, legal sanctions notwithstanding. And arguments on the part of the statistical agency to the effect that it would guard against harmful kinds of disclosure would carry little weight. The damage would have been done simply by conceding that any kind of disclosure was admissible.

Statistics Canada's second role as a collector of business statistics is in a tradition much newer than and quite different from that of the Statistics Act, and derives from the responsibility assigned to it in 1962 for administering the newly enacted Corporations and Labour Unions Returns Act (CALURA) [9]. The intent of the Act was to provide information about the influence on the Canadian economy of foreign controlled corporations and international unions. It set an important precedent with respect to disclosure by requiring that the individual non-financial returns of corporations covered by the Act were to be made available to the public and furthermore, as originally enacted in 1962, that officials of other government departments were to be permitted access to individual returns of financial information "for any purpose relating to the determination of policy in connection with the formulation of any law of Canada or the ascertainment of any matter necessarily incidental thereto". This latter provision was withdrawn when the Act was later amended to authorize the use of tax returns in lieu of certain of the financial reporting requirements. This appears to have been a matter of expediency rather than of principle, however. Outside access was incompatible with the confidentiality requirements attaching to income tax records, and the operational advantages of using the latter source greatly outweighed the prospective benefits of an access provision for research purposes which, in fact, had been little used up to that time.

To repeat then, CALURA was enacted with a view to meeting the kinds of informational requirements outlined earlier in this section through an instrument other than the Statistics Act. It might be added here that this latter intention could have led to the vesting of responsibility

for CALURA in some agency other than Statistics Canada, and this would no doubt have reinforced the public's perception of the logical distinction between the two different kinds of purpose. Be that as it may, the precedent set by CALURA makes it possible to respond to the needs of the late 1970's in an evolutionary manner rather than by breaking completely new ground.

Growing pressures by policy departments of the federal government for better and more detailed information on corporations have led Statistics Canada, as the custodian of CALURA, to develop a set of proposed changes to Part I (Corporations) of the Act¹ which will consolidate and improve in the one reporting vehicle a set of hitherto diffused and imperfectly integrated reporting requirements for large corporations, while at the same time appreciably diminishing the burden of response on the corporations in question. The precise details of these changes lie outside the scope of this paper. What is of interest, however, is their impact on confidentiality and this will be briefly assessed in relation to the kinds of informational purposes which require individualized data.

First, and most important, the changes are not intended as a means of furnishing information for any regulatory purposes. Even though CALURA has always had, and will continue to have, access and disclosure features, these are conceived of as marginal extensions of its primarily statistical character. Its use for regulatory purposes would therefore be a clear breach of this principle. Again, the changes do not (and future changes probably never will) seek to satisfy any of those public disclosure purposes which are more properly addressed by incorporation and securities legislation. However, through conceptual changes, additional reporting requirements, and elimination of the reporting exemption for certain classes of corporation, the quality and completeness of the non-financial information available to the public will be greatly improved.

¹ In fact, changes are being proposed to Part II (Labour) also.

The major positive thrust of the changes will be the re-introduction of an original feature of the Act through the provision of "open access with exceptions" to individual financial reports of corporations by federal agencies. Agencies or arms of agencies that engage in regulation or in program administration will be denied access and, more generally, access for any purpose other than policy analysis is not likely to be permitted. The re-introduction of access will, it should be emphasized, be accompanied by a divorce of CALURA from access to income tax records as a source of information.

Should these proposed changes be sanctioned by Parliament, Statistics Canada's customary role of intermediary between supplier and user will be more than usually difficult, since it alone is likely to be held accountable by suppliers for responsible use of the new access provisions. Any perceived abuse of this access would put in jeopardy not only current information flows, but also future prospects for extended reporting and access to meet new needs.

7. TWO FUTURE TOPICS OF DISCUSSION

Rather than concluding with a recapitulation or synthesis of the foregoing, brief mention will be made of two topics which are likely to feature importantly in future discussions of confidentiality in Canada. The topics are not new ones, but in neither case have they been pushed to the forefront of public discussion by official legislation or administrative action, as has happened in the United States.

The first of these topics concerns access to historical census records. Up to the present time, Statistics Canada, which retains custody of all national census records from 1881¹ onwards, has taken the view that such records are confidential in perpetuity, and has therefore refused

¹ Records of the 1871 Census, Canada's first national census following Confederation in 1867, are in the Public Archives of Canada and have been made publicly available on the authority of that agency.

access to historians and other interested researchers. Although nineteenth century censuses offered no statutory protection for the confidentiality of individual information, the argument has been that the prohibitions against disclosure of identifiable information contained in the present Statistics Act [4] apply retroactively.

Whatever the legal status of these older records, it is clear that archival-type access to information collected under the present Statistics Act and its direct antecedents would require an amendment to the Act. In contemplating such an amendment, however, far in the future it might take effect, a major concern would be the effect on current response rates and the accuracy of responses. This, of course, is the substantive argument that has been most frequently used in the United States where the Privacy Act is seen as obliging the Bureau of the Census to inform respondents in 1980 of the fact that their questionnaires will be made accessible to researchers in 2052 if the Bureau's 1952 agreement with the National Archives is allowed to stand [10].

On the other hand, in the major documented argument put forward by the Canadian historical research community, the point is made that access after a period such as seventy-two years offers little threat to the privacy of living individuals [11]. In this connection, the outcome of a pilot study by the Committee on National Statistics of the National Academy of Sciences to determine the response effects of guaranteeing various levels of confidentiality, should be of considerable interest [10].

Representations to Statistics Canada for changes in policy and law to permit access to historical census documents will undoubtedly continue to mount. From Statistics Canada's point of view, the debate could be most profitably conducted, not in "all or nothing" terms, but on the basis of concrete proposals, in which the appropriateness of particular kinds of research applications could be determined, the qualifications and standings of researchers established, protocols for the publication of findings agreed upon, and security procedures for the handling of records set out.

The second topic arises from the possibility of future freedom of information legislation in Canada. A recently published Green Paper [12] seeks to set out for public debate the issues and legislative options in furnishing public access to government documents. This document is, understandably, not very explicit about implications for the statistical system. For purposes of discussion, a list of obvious or possible exceptions from the government's obligation to make documents available is put forward, two of which are: financial or commercial information which, among other things, 'would result in significant and undue financial loss or gain by a person, group, organization or government institution, or would adversely affect a person, ... institution in regard to its competitive position'; and, information the disclosure of which might be prohibited by any federal enactment.

The second of these possible exemptions would appear to protect all operations carried out by Statistics Canada under the authority of the Statistics Act [4], although its application to information collected under CALURA [9], in the form in which it is proposed to be amended, would be less certain since accessibility of individual information to other government departments could be read as a presumption of similar availability to the general public. It may be, however, that CALURA information would be protected from public disclosure by an exemption of the first kind.

Thus, the area of statistical information that would be affected by freedom of information legislation would appear to be business information not collected under rigid guarantees of confidentiality such as are provided by the Statistics Act, and where in turn, eligibility for exemption from disclosure on account of the likelihood of "significant and undue financial loss or gain", or adverse effect effect on competitive position is not readily apparent. In this connection, it may be salutary, if not reassuring, to quote the American Statistical Association Report [10] which, in commenting on a similar set of exempted categories in the U.S. Freedom of Information Act, noted that "various legal cases ...

the first of these is the fact that the system is not a simple one, and that the results are not always the same.

The second of these is the fact that the system is not a simple one, and that the results are not always the same.

The third of these is the fact that the system is not a simple one, and that the results are not always the same.

The fourth of these is the fact that the system is not a simple one, and that the results are not always the same.

The fifth of these is the fact that the system is not a simple one, and that the results are not always the same.

The sixth of these is the fact that the system is not a simple one, and that the results are not always the same.

The seventh of these is the fact that the system is not a simple one, and that the results are not always the same.

The eighth of these is the fact that the system is not a simple one, and that the results are not always the same.

The ninth of these is the fact that the system is not a simple one, and that the results are not always the same.

The tenth of these is the fact that the system is not a simple one, and that the results are not always the same.

The eleventh of these is the fact that the system is not a simple one, and that the results are not always the same.

The twelfth of these is the fact that the system is not a simple one, and that the results are not always the same.

The thirteenth of these is the fact that the system is not a simple one, and that the results are not always the same.

The fourteenth of these is the fact that the system is not a simple one, and that the results are not always the same.

The fifteenth of these is the fact that the system is not a simple one, and that the results are not always the same.

suggest that demonstrating that data fall into these exempted categories is not routine".

RESUME

Le présent document commence par dégager certains des facteurs qui ont récemment rendu plus difficile la tâche des organismes statistiques, à savoir satisfaire aux besoins croissants de la société en matière de renseignements et assurer en même temps aux répondants une protection convenable de leur vie privée.

Le document s'arrête ensuite au concept de la vie privée, puis examine les dispositions s'y rattachant que comprend la nouvelle Loi canadienne sur les droits de la personne. Le document étudie aussi les dispositions sur la confidentialité que contient la Loi canadienne sur la statistique et en vertu desquelles les droits à la vie privée des répondants sont protégés. Vient ensuite un exposé sur les faits à cause desquels le traitement confidentiel des renseignements sur les corporations est contesté et sur la façon dont Statistique Canada s'efforce de satisfaire aux besoins d'accès du gouvernement aux rapports des corporations, en particulier dans le contexte de l'appartenance étrangère, sans nuire aux usages traditionnels en matière de confidentialité qui sont en vigueur dans le cadre général de la déclaration statistique.

Le document aborde enfin deux sujets susceptibles de faire l'objet de débats futurs sur la confidentialité: l'accès des savants aux archives statistiques chronologiques et la possibilité d'une éventuelle législation canadienne sur la liberté de l'information.

REFERENCES

- [1] I.P. Fellegi, "On the Question of Statistical Confidentiality", Journal of the American Statistical Association, Volume 67, Number 337, March 1972.
- [2] Statistical Policy Division, Office of Management and Budget, "Confidentiality of Statistical and Research Data" (draft excerpt from "A Framework for Planning U.S. Federal Statistics, 1978-1989"), Statistical Reporter, January 1977.

the first of these is the fact that the
the second is the fact that the

the third is the fact that the

the fourth is the fact that the
the fifth is the fact that the

the sixth is the fact that the

the seventh is the fact that the
the eighth is the fact that the

the ninth is the fact that the

the tenth is the fact that the
the eleventh is the fact that the

the twelfth is the fact that the

the thirteenth is the fact that the
the fourteenth is the fact that the

the fifteenth is the fact that the

the sixteenth is the fact that the
the seventeenth is the fact that the

the eighteenth is the fact that the

the nineteenth is the fact that the
the twentieth is the fact that the

- [3] Survey Research Institute, National Central Bureau of Statistics, Sweden, "The National Central Bureau of Statistics and the General Public (the findings of an interview survey taken in Sweden during the Spring of 1976)", Stockholm, June 1977.
- [4] "Statistics Act 1971", Chapter 15, 19-20 Elizabeth II.
- [5] Bill C-25, "The Canadian Human Rights Act", as passed by the House of Commons of Canada, June 2, 1977.
- [6] D.A. Worton, "The Reporting Burden and the Expansion of Data", 34th North American Conference on Labor Statistics, Minneapolis, Minn., June 1977.
- [7] "Foreign Direct Investment in Canada" (Gray Report), Information Canada, Ottawa, 1972.
- [8] "Foreign Ownership and the Structure of Canadian Industry", (Watkins Report), Information Canada, Ottawa, 1968.
- [9] "Corporations and Labour Unions Returns Act, 1962", Chapter 26, 10-11 Elizabeth II.
- [10] American Statistical Association, "Report of Ad Hoc Committee on Privacy and Confidentiality", American Statistician, May 1977.
- [11] Minutes of Proceedings and Evidence of the Standing Joint Committee on Regulations and Other Statutory Instruments, 1st Session, 30th Parliament, Issue No. 79, June 15, 1976; Supply and Services, Canada, 1976.
- [12] Honourable John Roberts, Secretary of State, "Legislation on Public Access to Government Documents", Supply and Services Canada, 1977.

SYNTHETIC ESTIMATION IN PERIODIC HOUSEHOLD SURVEYS

P.D. Ghangurde and M.P. Singh
Household Surveys Development Division

In periodic household surveys, area samples are usually selected in geographic strata with probability of selection of areal units proportional to population size in these units. The design-based estimates for areas composed of domains within strata can have poor precision due to cluster sampling with a few primary sampling units per stratum. In this paper, synthetic estimates are investigated as an alternative to these estimates. An empirical evaluation based on the design of the Canadian Labour Force Survey is given.

1. INTRODUCTION

For many large scale periodic household surveys like the Canadian Labour Force Survey (LFS), Current Population Survey and Health Interview Survey in the U.S., data are obtained by sample designs involving geographic stratification and multi-stage area sampling. The objective of these household surveys is estimation of total (or proportion of) population with a certain attribute (e.g. labour force status, health status). Due to the lack of availability of updated frames for the ultimate stage sampling units (dwellings or clusters of dwellings) and for reducing costs, multi-stage area samples are drawn within geographic strata. Since characteristic totals are known to be correlated to population within areal units, the population is used as a size measure in probability proportional to size (pps) sampling of areal units and also as an auxiliary variable in ratio estimation.

These surveys can provide reasonably accurate estimates at the national, regional or provincial level due to large sample sizes (in terms of total number of persons in the sample) available for these areas. Smaller areas composed of complete strata do not pose a problem in

estimation even though the reliability of estimates is low. If these small areas cut across design strata and are thus composed of areal domains within strata, the reliability of design-based domain estimates is severely reduced due to cluster sampling.

This paper investigates an alternative method of obtaining estimates for small areas which are composed of domains within strata, in clustered sample designs. The method which has often been called synthetic estimation in the literature, (see e.g. Gonzales [4], and [11]), uses knowledge of population structure. Since conventional estimates based on probability samples are inefficient due to cluster sampling, knowledge of population structure is used in formulation and evaluation of these estimates. The knowledge of population structure has been used in sample surveys in design (as in pps sampling) and estimation (as in ratio and regression estimation) where a relationship between a variable of interest and an auxiliary variable is assumed. The super-population models used in this paper are appropriate for categorical data, i.e. for characteristics which are counts of persons with a certain attribute.

Super-population models have recently been used in the literature for proving optimality of non-random sampling procedures (see Royall [14], Royall and Herson [15]). Unlike these papers, the use of these models in this paper is more in the traditional sense (see page 203, Smith [16]).

2. SYNTHETIC ESTIMATION

It is well-known (e.g. page 8, Kish and Frankel [9]) that in cluster sampling ratio estimates for total or mean per unit for areal domains are biased due to the large variance of the number of units in the domain from sampled clusters. The unbiased estimates for totals have large variances for the same reason. In particular cases where a domain has no sampled clusters, both the estimates become impractical.

The alternative method of synthetic estimation evaluated in this paper exploits the homogeneity of characteristic counts in the sub-groups (e.g. demographic, occupation groups) in the population. In household surveys these sub-groups are used as post-strata in estimation. For simplicity of

exposition, we consider the case of a single stratum and introduce synthetic estimation for the estimates in cluster sampling with pps with replacement. The empirical results for this case are obtained for domains of various relative sizes and are applicable to the practical case of a set of domains in several strata by averaging over sizes. The extension of results to sub-groups and strata is considered later.

2.1 Synthetic Estimates

A stratum consists of N clusters of which n are drawn with pps with replacement. The unbiased estimate $\hat{j}X$, the characteristic total in domain j , which is assumed to be composed of a group of clusters, is given by

$$\hat{j}X = P \cdot \sum_{i=1}^n \frac{jX_i}{nP_i}, \quad (2.1)$$

where P_i is population in the i th cluster, X_i is the characteristic total in the i th cluster, jX_i is X_i if the i th cluster is in the j th domain and is zero otherwise and $P = \sum_{i=1}^N P_i$. The synthetic estimate $j\tilde{X}$ is defined as

$$j\tilde{X} = WP \sum_{i=1}^n \frac{X_i}{nP_i} \quad (2.2)$$

where $W = jP'/P'$, jP' and P' being population totals in the j th domain and the stratum as of last census. A justification for $j\tilde{X}$ can be given by the optimality criterion of minimum discrimination information (see Section 7). When the domain is equal to the stratum $W = 1$ and hence $\hat{j}X = j\tilde{X}$. The bias and variance of $j\tilde{X}$ and variance of $\hat{j}X$ are given by

$$\begin{aligned} B(j\tilde{X}) &= (WX - jX) \\ V(j\tilde{X}) &= \frac{W^2}{n} \left[\sum_{i=1}^N \frac{P_i X_i^2}{P_i} - X^2 \right] \\ V(\hat{j}X) &= \frac{1}{n} \left[\sum_{i=1}^N \frac{P_i^j X_i^2}{P_i} - jX^2 \right] \end{aligned} \quad (2.3)$$

the first of these is the fact that the
 second of these is the fact that the
 third of these is the fact that the
 fourth of these is the fact that the
 fifth of these is the fact that the

the first of these is the fact that the

the second of these is the fact that the
 the third of these is the fact that the
 the fourth of these is the fact that the

the fifth of these is the fact that the

the sixth of these is the fact that the
 the seventh of these is the fact that the
 the eighth of these is the fact that the

the ninth of these is the fact that the
 the tenth of these is the fact that the

the eleventh of these is the fact that the
 the twelfth of these is the fact that the
 the thirteenth of these is the fact that the
 the fourteenth of these is the fact that the

the fifteenth of these is the fact that the

the sixteenth of these is the fact that the
 the seventeenth of these is the fact that the
 the eighteenth of these is the fact that the

where $X = \sum_{i=1}^N X_i$. The gain, i.e. reduction in mean square error due to

synthetic estimation, is given by

$$\begin{aligned} V(\hat{j}_X) - V(j_X) - [B(j_X)]^2 \\ = \frac{P}{n} \sum_{i=1}^N \frac{[j_{X_i}^2 - W^2 X_i^2]}{P_i} - \frac{(j_X^2 - W^2 X^2)}{n} - [WX - j_X]^2 \end{aligned} \quad (2.4)$$

2.2 One and Two-Balance

If $j_X = WX$

and
$$\sum_{i=1}^N \frac{j_{X_i}^2}{P_i} = W \cdot \sum_{i=1}^N \frac{X_i^2}{P_i}, \quad (2.5)$$

synthetic estimate j_X is more efficient than \hat{j}_X . Relations similar to (2.5) in the case of simple random sampling were called one-balance and two-balance by Jones and Coopersmith [7]. The concept of balancing has been discussed by Royall and Herson [15] from the viewpoint of prediction theory for obtaining optimal sampling procedures under super-population models. We consider balance as a structural property of the population which can be used in evaluation of alternative estimators in cluster sampling. The deviation from conditions (2.5) can be more formally expressed by assuming a model for super-population from which the finite population can be assumed to be drawn as a random sample. Such models have been traditionally used in the literature for evaluation of alternative sample designs and estimators.

3. EVALUATION OF EFFICIENCY UNDER MODELS

We consider the following model. Let N clusters be drawn as a random sample from a super-population and for the i th cluster let

$$X_i = \beta P_i + e_i, \quad i = 1, 2, \dots, N \quad (3.1)$$

where β is a regression coefficient and

$$\epsilon(e_i | P_i) = 0, \epsilon(e_i^2 | P_i) = \sigma^2 \cdot P_i^t, t > 0, \sigma^2 \geq 0$$

$$\epsilon(e_i e_{i'} | P_i, P_{i'}) = 0, i \neq i'; i, i' = 1, 2, \dots, N.$$

where ϵ stands for expectation over all finite populations. The model is appropriate for characteristics which are defined as counts of persons with a certain attribute. The regression coefficient can be considered as the proportion of population in the stratum with the attribute. The heterogeneity between clusters is shown by a non-zero value of σ^2 , with dependence of variance on cluster size P_i being appropriate for categorical data. Many studies show that for socio-economic characteristics t lies between 1 and 2.

The efficiency gain due to synthetic estimation under (3.1) is given by

$$\frac{\epsilon[V(\hat{J}_X) - \text{MSE}(\hat{J}_X)]}{\epsilon[\text{MSE}(\hat{J}_X)]}$$

$$= \frac{\frac{1}{n} [J_P(P - J_P)] - J_P^2 \left[\frac{P}{J_P} W - 1 \right]^2 + \frac{\sigma^2}{n\beta^2} \left[\sum_{i=1}^N P_i^{t-1} [(P - P_i)(J_{A_i} - W^2) - nP_i(J_{A_i} - W)^2] \right]}{J_P^2 \left[\frac{P}{J_P} W - 1 \right]^2 + \frac{\sigma^2}{n\beta^2} \left[\sum_{i=1}^N P_i^{t-1} [(P - P_i) W^2 + nP_i(J_{A_i} - W)^2] \right]} \quad (3.2)$$

where $\text{MSE}(\hat{J}_X)$ is mean square error of \hat{J}_X and $J_{A_i} = 1$ if the i th sampled cluster is in the domain and $J_{A_i} = 0$ otherwise.

It can be seen that when the domain is equal to the stratum, $W=1$ and the efficiency gain is zero. If $\sigma^2 = 0$, i.e. there is homogeneity in the stratum, the gain depends on the number of clusters in the sample, relative size of the domain J_P/P and inaccuracy of weights shown by $(\frac{P}{J_P} W - 1)$ due to uneven growth since the last census.

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

2019年10月10日

It may be noted that the synthetic weight W is based on the last census and may not be equal to true synthetic weight jP/P . In continuous household surveys like the Canadian Labour Force Survey (LFS) size measures, though based on population counts as of the last census, are updated periodically in areas with large growth (see e.g. Kish and Scott [10], Platek and Singh [12]). However, since population projections are not available for small areas for intercensal years, synthetic weights will have to be based on population counts as of the last census. In practice, synthetic weights would be needed for subgroups like age-sex groups (see Section 5 for extension to sub-groups) rather than for gross population. The effect of inaccurate synthetic weights, therefore, needs to be explicitly considered in evaluation. It may be noted that in the above expressions P_i , $i=1, 2, \dots, N$ are considered as population totals.

If $\sigma^2 > 0$, i.e. there is heterogeneity between clusters, it can contribute to gain if n is small and synthetic weights are correct. The term σ^2/β^2 in (3.2) is related to the coefficient of variation of domain estimates under the model (3.1) (see Section 4).

We now consider an alternative model which takes account of the possibility of different regression coefficients in the domains. We assume that the population consists of two domains j and its complement with the model of the form (3.1) but with different regression coefficients. Let

$$\begin{aligned} X_i &= \beta P_i + e_i, \quad i \in j \\ X_i &= \delta \beta P_i + e_i, \quad i \notin j, \quad \delta \neq 1 \end{aligned} \tag{3.3}$$

where e_i , $i = 1, 2, \dots, N$ have properties similar to those in (3.1). Under this model, the relative gain in efficiency is given by

$$\epsilon[V(\hat{J}_X) - \text{MSE}(\hat{J}_X)]/\epsilon[\text{MSE}(\hat{J}_X)] = A/B,$$

where

$$\begin{aligned}
 A = & \frac{j_p [p - j_p]}{n} - j_p^2 \left[\left(1 - \delta + \delta \frac{p}{j_p} \right) w - 1 \right]^2 \\
 & - \frac{w^2 p \left[\sum_{i=1}^N p_i [j_{A_i} + \delta (1 - j_{A_i})]^2 - p \left[\frac{j_p}{p} + \delta \left(1 - \frac{j_p}{p} \right) \right]^2 \right]}{n} \\
 & + \frac{\sigma^2}{n\beta^2} \left[\sum_{i=1}^N p_i^{(t-1)} [(p - p_i) (j_{A_i} - w^2) - n p_i (j_{A_i} - w)^2] \right], \\
 B = & j_p^2 \left[\left(1 - \delta + \delta \frac{p}{j_p} \right) w - 1 \right]^2 \\
 & + \frac{w^2 p \left[\sum_{i=1}^N p_i [j_{A_i} + \delta (1 - j_{A_i})]^2 - p \left[\frac{j_p}{p} + \delta \left(1 - \frac{j_p}{p} \right) \right]^2 \right]}{n} \\
 & + \frac{\sigma^2}{n\beta^2} \left[\sum_{i=1}^N p_i^{(t-1)} [(p - p_i) w^2 + n p_i (j_{A_i} - w)^2] \right]. \quad (3.4)
 \end{aligned}$$

When $\delta=1$ (3.4) reduces to (3.2). For an outline of the derivation of (3.4) see the Appendix. The additional terms in the numerator and denominator of (3.4), as compared to (3.2), represent loss in efficiency due to heterogeneity in the stratum as represented by $\delta \neq 1$.

The parameters in (3.2) and (3.4) introduced by the model are β, σ^2 and δ . The best linear unbiased estimates by generalized least squares under model (3.1) are given by

$$\hat{\beta} = \frac{\sum_{i=1}^N x_i p_i^{(1-t)}}{\sum_{i=1}^N p_i^{(2-t)}},$$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N p_i^{-t} (\hat{e}_i - \bar{e})^2,$$

where

$$\hat{e}_i = X_i - \hat{\beta} P_i, \quad i = 1, 2, \dots, N.$$

$$\bar{e} = \frac{1}{N} \sum_{i=1}^N \hat{e}_i.$$

The parameter δ can be estimated as a ratio of estimates of β in the complement of the domain and that in the domain. However, selected values of δ are used in the empirical evaluation given in section 6.

The expressions (3.2) and (3.4) can be computed for any characteristic if data on X_i and P_i , $i = 1, 2, \dots, N$ are available. In the present paper, empirical evaluation of efficiency gains of \hat{y}_X over \hat{y}_X under the two models is done for various values of sampled clusters n , domain sizes $W' = \frac{J_p}{p}$, and ratio δ for a sample design with strata and cluster delineations the same as strata and primaries in the rural areas of the LFS. The results of the empirical evaluation based on 1971 Census data on counts of unemployed, employed and population are given in Section 6.

In general, any realistic values of the ratio σ^2/β^2 for the characteristic of interest in a survey could be substituted in (3.2) and (3.4) for evaluation of efficiency and such an empirical evaluation would need only values of population counts P_i in the clusters in strata defined by the survey.

In household surveys many characteristics are homogeneous within sub-groups (e.g. demographic groups) and due to lack of availability of frames for these sub-groups this homogeneity cannot be exploited in the design by stratification; however, it can be exploited in estimation by post-stratification. These post-strata provide natural 'sub-groups' in which synthetic estimation can be done. These estimates and appropriate models for their evaluation are considered in Section 5.

the first of these is the *Journal of the American Medical Association*.

the second is the *Journal of the American Medical Association*.

the third is the *Journal of the American Medical Association*.

the fourth is the *Journal of the American Medical Association*.

the fifth is the *Journal of the American Medical Association*.

the sixth is the *Journal of the American Medical Association*.

the seventh is the *Journal of the American Medical Association*.

the eighth is the *Journal of the American Medical Association*.

the ninth is the *Journal of the American Medical Association*.

the tenth is the *Journal of the American Medical Association*.

the eleventh is the *Journal of the American Medical Association*.

the twelfth is the *Journal of the American Medical Association*.

the thirteenth is the *Journal of the American Medical Association*.

the fourteenth is the *Journal of the American Medical Association*.

the fifteenth is the *Journal of the American Medical Association*.

the sixteenth is the *Journal of the American Medical Association*.

the seventeenth is the *Journal of the American Medical Association*.

4. EVALUATION OF BIAS UNDER MODELS

Though the expressions for efficiency gains take account of bias of synthetic estimates, it is important to have analytical expressions for bias and evaluate the same empirically. For simplicity, we obtain the results for synthetic estimate $j\hat{X}$ under models (3.1) and (3.2). It is assumed that \hat{X} is an unbiased estimate of the characteristic total, X , in the stratum. For the reasons given in section 3, the synthetic weights are assumed to be based on the last census. Consider the bias of $j\hat{X}$ defined as

$$\begin{aligned} B(j\hat{X}) &= E(j\hat{X}) - jX \\ &= XW - jX, \end{aligned} \quad (4.1)$$

where $W = jP'/P'$, jP' and P' being population counts in the domain and the stratum as in the last census. Let

$$\begin{aligned} Q &= W \frac{P}{jP} - 1 \\ &= \frac{jP'}{jP} \cdot \frac{P}{P'} - 1, \end{aligned} \quad (4.2)$$

where jP and P are correct population counts in the domain and the stratum and

$$\frac{P}{P'} = 1 + g \text{ and } \frac{jP}{jP'} = 1 + jg, \text{ where } jg \text{ and } g \text{ denote growth rates in}$$

the domain and the stratum respectively. Assuming $|jg| < 1$ and taking terms up to the first degree in jg in the expansion of $(1 + jg)^{-1}$, we have $Q \approx (g - jg)$. Thus Q is approximately equal to the difference of growth rates in the stratum and the domain. It is, therefore, appropriate to express W in terms of Q and the population counts. From (4.2) we have

$$W = \frac{jP}{P} (1 + Q).$$

1910年10月10日武昌起义爆发，结束了清朝的统治。

一、武昌起义的背景与经过

武昌起义是辛亥革命的重要组成部分。其背景是清政府腐败无能，社会矛盾激化。起义经过如下：

1. 1910年10月10日晚，新军中的革命党人首先发难。
2. 起义军迅速占领了武昌城内的主要据点。
3. 清军调集重兵镇压，但起义军英勇抵抗。
4. 起义最终失败，但革命火种得以保存。

武昌起义的成功，标志着中国两千多年封建帝制的终结。

武昌起义后，全国各省纷纷响应，掀起了辛亥革命的高潮。

二、意义

武昌起义的意义在于：

1. 结束了清朝的统治，建立了中华民国。

2. 推动了中国的民主化进程。

3. 为中国的现代化奠定了基础。

4. 激发了中国人民的民族意识。

5. 为中国的统一和独立创造了条件。

武昌起义是中国近代史上的重要事件。

Hence substituting in (4.1), the relative bias is given by

$$\frac{B(j_X^y)}{j_X} = \frac{(X \frac{j_P}{P} - j_X)}{j_X} + Q \cdot \frac{X}{P} \cdot \frac{j_P}{j_X} \quad (4.3)$$

The first term on the right hand side of (4.3) represents relative bias as of census time and the second term represents contribution of uneven population growth. Under the model (3.2), the average relative bias is given by

$$\begin{aligned} E\left[\frac{B(j_X^y)}{j_X}\right] &= W \frac{P}{j_P} \delta + W(1-\delta) - 1 \\ &= (1-\delta)(W - 1) + Q\delta \end{aligned} \quad (4.4)$$

when $Q = 0$, $\frac{WP}{j_P} = 1$ and then

$$E\left[\frac{B(j_X^y)}{j_X}\right] = (1-\delta)(W-1). \quad (4.5)$$

Thus, in order to reduce bias due to synthetic estimation, it is desirable that strata be homogeneous and that growth rates be the same within strata. In practice, this can be achieved if synthetic estimates are formed in sub-groups which are homogeneous with δ close to 1. Since important characteristics in the household surveys are known to be homogeneous in sub-groups (e.g. age-sex groups) which are used as post-strata in estimation, these are appropriate sub-groups for controlling bias. The bias due to uneven population growth as shown by terms involving Q can be controlled by stratification in rural and urban areas.

In order to obtain the super-population variance of the relative bias, we consider for simplicity the model (3.1) and the case $t=1$ and have by the usual approximation the super-population variance given by

$$\begin{aligned} V\left[\frac{B(j_X^y)}{j_X}\right] &\doteq \frac{\sigma^2}{j_P^2 \beta^2} [W \cdot j_P Q + (1-W)j_P] \\ &= \frac{\sigma^2}{j_P \cdot \beta^2} [1-W] \text{ if } Q = 0. \end{aligned} \quad (4.6)$$

1. The first part of the paper is devoted to the study of the

properties of the function $f(x)$ defined by

$$f(x) = \sum_{n=0}^{\infty} \frac{a_n}{n!} x^n$$

where a_n are the coefficients of the power series

$$\sum_{n=0}^{\infty} a_n x^n = \frac{1}{1-x}$$

It is well known that the function $f(x)$ is analytic in the

region $|x| < 1$ and that it satisfies the differential equation

$$x f'(x) = f(x) - 1$$

which can be solved by the method of variation of constants.

The general solution of this equation is

$$f(x) = \frac{1}{x} \int_0^x \frac{1-t}{t} dt + C$$

where C is an arbitrary constant. Since $f(0) = 1$, we have

$$C = 1$$

and therefore the function $f(x)$ is given by

$$f(x) = \frac{1}{x} \int_0^x \frac{1-t}{t} dt + 1$$

It is easy to see that this function is analytic in the

region $|x| < 1$ and that it satisfies the differential equation

$$x f'(x) = f(x) - 1$$

which is the same as the one obtained by the method of variation of constants.

It is also easy to see that the function $f(x)$ is analytic in the

region $|x| < 1$ and that it satisfies the differential equation

$$x f'(x) = f(x) - 1$$

which is the same as the one obtained by the method of variation of constants.

It is also easy to see that the function $f(x)$ is analytic in the

region $|x| < 1$ and that it satisfies the differential equation

$$x f'(x) = f(x) - 1$$

which is the same as the one obtained by the method of variation of constants.

It is also easy to see that the function $f(x)$ is analytic in the

region $|x| < 1$ and that it satisfies the differential equation

$$x f'(x) = f(x) - 1$$

Since $\sigma^2/jp.\beta^2 = \sigma^2jp/jx^2$ is approximately equal to the coefficient of variation of jx , the characteristic total in the domain under the model (3.1), the relative bias is expected to be more stable for characteristics for which variation between clusters (as shown by σ^2) is small. Also, the relative bias is expected to be more stable for domains of large sizes.

5. EXTENSION TO POST-STRATIFICATION

As pointed out before, the sub-groups in the population, which are used as post-strata in estimation are appropriate for forming synthetic estimates. In periodic household surveys, ratio estimates based on population within the sub-groups as auxiliary variable, are used to further improve efficiency of design-based estimates within post-strata. We give below domain and synthetic estimates for this extension.

The unbiased domain estimate is given by

$$j\hat{x}' = \sum_a \frac{j\hat{x}_a}{\hat{p}_a} \cdot p_a \quad (5.1)$$

where $j\hat{x}_a$ is the design-based estimate of the ath sub-group in the jth domain, \hat{p}_a is the unbiased estimate of population in sub-group a in the stratum and p_a is the projected population total in sub-group a based on the last census. The synthetic estimate for the case of post-strata is given by

$$j\hat{y}' = \sum_a \frac{\hat{x}_a w_a}{\hat{p}_a} \cdot p_a \quad (5.2)$$

the first of these is the fact that the
the second is the fact that the
the third is the fact that the
the fourth is the fact that the
the fifth is the fact that the
the sixth is the fact that the
the seventh is the fact that the
the eighth is the fact that the
the ninth is the fact that the
the tenth is the fact that the

the first of these is the fact that the
the second is the fact that the
the third is the fact that the
the fourth is the fact that the
the fifth is the fact that the
the sixth is the fact that the
the seventh is the fact that the
the eighth is the fact that the
the ninth is the fact that the
the tenth is the fact that the
the eleventh is the fact that the
the twelfth is the fact that the
the thirteenth is the fact that the
the fourteenth is the fact that the
the fifteenth is the fact that the
the sixteenth is the fact that the
the seventeenth is the fact that the
the eighteenth is the fact that the
the nineteenth is the fact that the
the twentieth is the fact that the

the first of these is the fact that the
the second is the fact that the
the third is the fact that the
the fourth is the fact that the
the fifth is the fact that the
the sixth is the fact that the
the seventh is the fact that the
the eighth is the fact that the
the ninth is the fact that the
the tenth is the fact that the
the eleventh is the fact that the
the twelfth is the fact that the
the thirteenth is the fact that the
the fourteenth is the fact that the
the fifteenth is the fact that the
the sixteenth is the fact that the
the seventeenth is the fact that the
the eighteenth is the fact that the
the nineteenth is the fact that the
the twentieth is the fact that the

where $W_a = jP'_a/P'_a$ is the synthetic weight for sub-group a based on the population count in the last census. In practice, a separate ratio estimate, combined over a number of geographic strata in a region or province is formed, the implicit assumption being that the proportion of characteristic total to population within sub-groups is the same in all strata. This estimate can be defined as

$$j\hat{y}'_X = \sum_a \frac{(\sum_{h \in h} \hat{x}_{ha} W_{ha})}{\hat{P}_{ha}} \cdot P_{ha} \quad (5.3)$$

where h is a group of strata and other symbols have obvious meanings. Also, the population totals within sub-groups are not known for small areas which are of interest in synthetic estimation. Hence the ratio adjustment P_{ha}/\hat{P}_{ha} is done at a larger area level than the level h at which estimates are required. The effect of this ratio adjustment on the bias and variance is investigated in Gray and Ghangurde [5].

The analytical results on the bias of $j\hat{y}'_X$ and the efficiency of $j\hat{y}'_X$ over $j\hat{x}'_X$ under simple extensions of models (3.1) and (3.3) are given in Ghangurde and Singh [2].

6. EMPIRICAL EVALUATION

6.1 Use of Census Data

The analytical results on bias and efficiency presented in sections 3 and 4 show the effect of population growth and heterogeneity of population explicitly in terms of model parameters. However, an empirical evaluation of bias and efficiency for a given set of strata, clusters and given characteristic needs data on population totals in the clusters, and values of parameters β , δ , σ^2 , $W' = jP/P$ and Q . The value of Q , which represents the difference between growth rates in the domain and the stratum can be obtained from census data on population

counts in clusters. Also, various domain sizes W' can be used. Estimates of parameters β , δ and σ^2 can be obtained from census data on the characteristic or a related characteristic.

Since efficiencies under the models (3.1) and (3.3) are functions of $\sigma^2/\beta^2 j_P$ and δ , the assumption in the use of census data as proxy for survey true values for a characteristic is that these are approximately the same for the survey and census. Both $\sigma^2/\beta^2 j_P$, representing squares of the coefficient of variation of the domain estimate under model (3.1), and δ , representing the ratio of the proportion of the characteristic total to the population in the domain, are relative measures and hence the assumption seems to be valid. Alternatively, any realistic values of $\sigma^2/\beta^2 j_P$ and δ , as in the case of W' and Q , can be used in empirical evaluation.

6.2 Sample Design Based on the LFS

The evaluation was done for a stratified cluster sample design with strata and cluster delineations identical to rural strata and primaries of the LFS. Details of the LFS design and method of estimation are given in Platek and Singh [13]. The domains were assumed to be composed of groups of clusters. In the 1971 Census, labour force data were collected for a systematic sample of dwellings with a sampling ratio of 1/3, except in the case of collectives all of which were enumerated. The collectives cover approximately 1.4% of the total population in the 1971 Census in Ontario. By assuming the systematic sample as a random sample of all persons, the counts of 'unemployed' and 'employed' were weighted within each cluster (i.e. primary of the LFS design) with weight equal to the inverse sampling ratio within these clusters.

The empirical study was confined to the province of Ontario which has ten economic regions, rural areas of which are divided into 20 strata in the LFS. The results on bias and efficiency of synthetic estimates were obtained for domains of approximate sizes $W' = 0.25, 0.50$ and 0.75 and number of sampled clusters $n = 2, 3, 4$. The synthetic estimates j_X^v and j_X^u were evaluated in 17 strata of these economic regions; the region 51 with three strata was excluded from the study.

6.3 Analysis of Results

Table 1 gives the number of clusters in the strata N , the number of clusters in the domain N_D , the proportion of population of the stratum in the domain W' and the average relative size for strata in an economic region W_h , used in the study. Table 2 gives % relative bias of j^y_X and % relative efficiency of j^y_X over $j^y_{\hat{X}}$ for domains within strata. Table 3 gives % relative bias for the post-stratified synthetic estimate j^y_X and % relative efficiency of this estimate over domain estimate $j^y_{\hat{X}}$ based on census data in economic regions 52 and 58. In Table 3, the relative efficiency of the combined ratio estimate j^y_X is given in the last column. The numerical values of efficiency gains under models (3.1) and (3.2) and their extensions for various values of n , W' , δ and Q are given in Tables 5 to 7.

The % relative bias of synthetic estimate j^y_X of unemployed lies between $\pm 8\%$ for 23 of 29 domains considered. It can be seen that % relative bias increases as W' decreases. The % relative bias of the post-stratified synthetic estimate, $j^y_{\hat{X}}$, can be seen from Table 3 to be not significantly different from that of j^y_X . The values of relative bias are as of census time and in the case of j^y_X , these are values of the first term in (4.3). The evaluation of relative bias of j^y_X for intercensal years requires knowledge of parameters Q and δ . It was found that for areas of rural strata (called NSR strata in the LFS) in 40 counties out of 50 counties in Ontario the value of Q lies in the range of ± 0.06 . The values of Q for domains within strata could not be calculated since 1976 Census data were not available for these areas.

The efficiency of j^y_X decreases as domain size W' is increased and the number of sampled clusters n is increased from 2 to 4. The same is true for efficiency of the post-stratified synthetic estimate $j^y_{\hat{X}}$. The efficiency gains under models (with estimates of β , σ^2 from census data) are shown in Table 5 for three domains each in three strata of economic regions 52 and 58. These show that as n and W' are increased, efficiency gain decreases. The value of Q is assumed to be 0.10, since this was close to the average absolute value of Q for intercensal years obtained for counties.

Table 6 gives the efficiency gains under the model for $Q = 0.0$ and $t = 1.0$ for economic regions 53 to 57 consisting of 8 strata. The δ and n values considered are the same as in Table 5. The last column in Table 6 gives efficiency gains for $j\hat{y}$ based on census data and these gains compare well with those obtained for models with $\delta = 1.0$ except in economic region 53 and stratum 1 of economic region 57 indicating that δ could be higher than 1.20 in these strata. It may be noted that it is appropriate to assume $Q = 0$ in the model since synthetic weights are as of census. The results on bias and efficiency of employed in economic regions 52 and 58 (see Table 4) show that % relative bias is of the order of 2% even for $W = 0.2$ and efficiency is much higher than that in the case of unemployed. Other results on employed data (not shown) indicate that efficiency gains based on census data are close to those based on models with $Q = 0$, $\delta = 1.0$ or 1.10 and $t = 1.0$. These results provide a rough validation of models used in the evaluation.

In Table 7, the results on efficiency gains of post-stratified synthetic estimate, $j\hat{y}'$, are compared with efficiency gains of $j\hat{y}$ under models with $Q_a = 0$, $\delta_a = 1$ and $Q = 0$, $\delta = 1$. The former are two to three times greater than the later due to homogeneity within age-sex groups.

The efficiency gains for various values of t studied ($t = 1.0, 1.5, 1.7, 2.0$) were quite close. The % change in efficiency of synthetic estimate over domain estimate for various values of parameters can be seen from various tables.

7. CONCLUDING REMARKS

This paper attempts to provide a framework for evaluation of bias and efficiency of synthetic estimates in household surveys. The basic sample design assumed for derivation of efficiency results is stratified cluster sampling with pps with replacement. However, the results on bias are valid for any sample design and an unbiased estimate. Though in practice, sample selection is done without replacement in these

surveys, the assumption of independent selection is made in variance estimation based on the Keyfitz method [8] and its extensions (see Woodruff [18], Tepping [17]) for computational simplicity; the with replacement design assumed in this paper is, therefore, of practical interest. Though this paper assumes a single stage design for simplicity, in practice multi-stage designs are used in these surveys. Since the total variance is of interest in appraising reliability of estimates and since domains in the study are assumed to be composed of groups of clusters, the results on efficiency gains are also valid for multi-stage designs.

A measure of reliability of synthetic estimates has to take account of their bias. Relative mean square error, defined as mean square error divided by the square of the estimate, is an appropriate extension of relative variance as a measure of reliability of synthetic estimates. The ratio of relative mean square error of the synthetic estimate for a set of domains in a group of strata to relative variance of the estimate for the group of strata can be defined as synthetic estimation effect. This concept and its practical use in the LFS is explained and numerical values of the measure of relative precision are given in Ghangurde and Singh [3].

Synthetic estimates can be proved to be optimal using the criterion of minimum discrimination information under some constraints; the estimates can be obtained by a single iteration of the Iterative Proportional Fitting (IPF) algorithm (see Freeman and Koch [1] for this approach to the problem and Ireland and Kullback [6] for the theoretical basis of the IPF algorithm). However, for practical application of synthetic estimation in sample surveys, it is important to investigate the bias and efficiency of synthetic estimates and to evaluate the estimates empirically by relaxing the ideal conditions assumed by constraints. This is the problem considered in the paper.

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to Dr. I.P. Fellegi, Assistant Chief Statistician and R. Platek, Director, Household Surveys Development Division for helpful discussions and to referees for their comments. They wish to acknowledge the valuable assistance of A. Scott, Systems Development Division, in writing programs and K. O'Grady, Data Dissemination Division in producing census data.

RESUME

Dans les enquêtes ménages périodiques, les échantillons d'aires sont normalement tirés dans des strates géographiques, la probabilité de sélection des unités aréolaires étant proportionnelle à la taille de la population dans ces unités. Les estimations basées sur le plan d'échantillonnage pour des régions qui sont constituées de domaines à l'intérieur des strates peuvent être très imprécises en raison de l'échantillonnage par grappes lorsqu'on a tiré peu d'unités primaires par strate. Dans cet article, on étudie des estimations synthétiques qui peuvent servir comme alternative à ces estimations. On discute aussi une étude empirique basée sur le plan de l'Enquête sur la population active canadienne.

Table 1

Domain Sizes and Weights

Economic Region	Stratum	N	N _D	W _i	W _h
52	1	13	3	0.22	0.24 0.49 0.81
			6	0.46	
			10	0.77	
	2	11	3	0.25	
			6	0.53	
			9	0.83	
	3	15	4	0.25	
			7	0.48	
			12	0.82	
58	1	18	4	0.27	0.23 0.48 0.71
			9	0.48	
			14	0.76	
	2	17	4	0.21	
			9	0.50	
			14	0.72	
	3	13	3	0.27	
			6	0.46	
			10	0.63	
53	1	13	6	0.46	0.53
54	1	15	8	0.55	
55	1	17	8	0.47	
	2	14	7	0.50	
56	1	14	7	0.50	
	2	13	6	0.39	
57	1	14	7	0.49	
	2	17	8	0.69	
50	1	14	6	0.44	0.42
59	1	14	6	0.45	
	2	19	7	0.39	

Table 2

% Relative Bias and Efficiency

Economic Region	Stratum	W ¹	% Rel Bias	% Relative Efficiency		
				n = 2	n = 3	n = 4
52	1	0.22	10.98	5,789	4,824	4,135
		0.46	-3.68	3,940	3,770	3,613
		0.77	1.39	1,130	1,123	1,116
	2	0.25	2.00	10,861	10,702	10,549
		0.53	1.34	3,770	3,366	3,344
		0.83	.08	892	890	883
	3	0.25	5.33	4,684	4,513	4,354
		0.48	5.48	1,563	1,484	1,412
		0.82	3.94	403	394	386
58	1	0.27	2.69	5,272	5,219	5,167
		0.48	-3.70	1,908	1,868	1,830
		0.76	-2.15	602	597	593
	2	0.21	2.82	1,415	1,093	891
		0.50	-8.15	1,609	1,470	1,353
		0.72	-6.85	689	645	606
	3	0.27	1.24	3,780	3,116	2,650
		0.46	8.43	2,249	1,978	1,768
		0.63	2.23	1,755	1,731	1,707
53	1	0.46	27.32	416	337	282
54	1	0.55	2.43	2,924	2,867	2,812
55	1	0.47	8.02	1,452	1,351	1,262
	2	0.50	-7.06	2,228	1,992	1,802
56	1	0.50	-6.40	1,653	1,563	1,482
	2	0.39	-4.09	5,180	4,919	4,683
57	1	0.49	1.59	7,953	6,790	5,924
	2	0.69	-3.16	2,908	2,739	2,588
50	1	0.44	9.11	1,530	1,399	1,289
59	1	0.45	4.25	2,769	2,669	2,575
	2	0.39	-7.68	1,311	1,261	1,215

Table 3

% Relative Bias and Efficiency (Post-Stratified)

Economic Region	Stratum	W'	% Relative Bias ($\frac{j\bar{y}'}{\bar{y}}$)			W _h '	n	% Rel Eff $\frac{j\bar{y}'}{\bar{y}}$
			Male	Female	Overall			
52	1	0.22	7.24	17.96	10.85	0.24	2	4008
		0.46	-2.98	-3.19	-3.06		3	3428
		0.77	0.87	2.19	1.39		4	2994
	2	0.25	3.02	6.36	4.81	0.49	2	2095
		0.53	-5.92	-16.80	16.13		3	2036
		0.83	2.36	-1.58	0.83		4	1982
	3	0.25	12.61	-2.20	5.70	0.81	2	556
		0.48	2.53	13.63	7.15		3	540
		0.82	6.03	2.00	4.22		4	524
58	1	0.21	-1.75	6.64	1.75	0.23	2	1814
		0.48	-1.03	-6.35	-3.44		3	1390
		0.76	1.89	-5.03	-1.28		4	1126
	2	0.21	17.85	4.40	13.64	0.48	2	1939
		0.50	-1.03	-5.19	-8.31		3	1857
		0.72	-7.21	-4.36	-6.09		4	1815
	3	0.27	20.23	1.82	12.06	0.71	2	761
		0.46	2.49	18.26	8.28		3	740
		0.63	2.29	2.48	0.24		4	721

the first of these is the fact that the

the second is the fact that the

the third is the fact that the

the fourth is the fact that the

the fifth is the fact that the

the sixth is the fact that the

the seventh is the fact that the

the eighth is the fact that the

the ninth is the fact that the

the tenth is the fact that the

the eleventh is the fact that the

the twelfth is the fact that the

the thirteenth is the fact that the

the fourteenth is the fact that the

the fifteenth is the fact that the

the sixteenth is the fact that the

the seventeenth is the fact that the

the eighteenth is the fact that the

the nineteenth is the fact that the

the twentieth is the fact that the

Table 4
% Relative Bias and Efficiency
 (Employed)

Economic Region	Stratum	W ⁱ	% Rel Bias	% Relative Efficiency n=		
				2	3	4
52	1	0.46	0.28	77293	76888	76487
	2	0.53	0.86	11816	11710	11606
	3	0.48	-2.03	12339	11805	11315
58	1	0.48	- .28	65259	20313	31231
	2	0.50	1.30	64948	19646	30213
	3	0.46	-1.12	64641	19021	29260
58	1	0.21	1.39	177600	83169	61753
	2	0.21	-0.95	162486	81580	55375
	3	0.27	-0.23	149743	80051	50192

Table 5
% Efficiency Gains Under Model

Q = 0.10, t = 1.00

Dom	Stratum	n	ER 52			ER 58		
			$\delta=1.00$	$\delta=1.10$	$\delta=1.20$	$\delta=1.00$	$\delta=1.10$	$\delta=1.20$
1	1	2	4853	2813	1632	3013	2123	1408
		3	3848	2063	1143	2518	1647	1030
		4	3182	1620	869	2159	1339	804
	2	2	4512	2590	1494	3004	2136	1428
		3	3548	1888	1042	2499	1655	1045
		4	2917	1477	790	2135	1344	815
	3	2	2116	1839	1217	3678	2208	1307
		3	2177	1422	888	2963	1639	922
		4	1861	1153	691	2475	1295	702
2	1	2	1900	1191	725	1024	766	534
		3	1555	896	711	895	619	404
		4	1309	516	392	792	515	318
	2	2	1555	996	615	935	708	499
		3	1260	747	590	817	574	379
		4	1055	437	330	724	478	298
	3	2	1064	788	545	1759	1132	702
		3	920	631	521	1451	859	502
		4	808	409	320	1230	685	383
3	1	2	493	379	272	283	238	187
		3	400	289	196	246	196	144
		4	332	228	146	216	162	113
	2	2	348	277	206	375	309	238
		3	275	207	145	328	255	185
		4	222	159	105	290	214	185
	3	2	209	179	143	906	643	429
		3	179	145	109	750	494	311
		4	154	118	83	635	396	237

THEORY OF THE EARTH

BY J. H. VAN DIJK

OF THE UNIVERSITY OF AMSTERDAM

AND OF THE UNIVERSITY OF ROTTERDAM

AND OF THE UNIVERSITY OF GRIFFITH

OF THE UNIVERSITY OF SYDNEY

AND OF THE UNIVERSITY OF MELBOURNE

AND OF THE UNIVERSITY OF ADELAIDE

AND OF THE UNIVERSITY OF PERTH

AND OF THE UNIVERSITY OF BRISBANE

AND OF THE UNIVERSITY OF QUEENSLAND

AND OF THE UNIVERSITY OF AUSTRALIA

AND OF THE UNIVERSITY OF NEWCASTLE

AND OF THE UNIVERSITY OF WOLLONGONG

AND OF THE UNIVERSITY OF SYDNEY

AND OF THE UNIVERSITY OF MELBOURNE

AND OF THE UNIVERSITY OF ADELAIDE

AND OF THE UNIVERSITY OF PERTH

AND OF THE UNIVERSITY OF BRISBANE

AND OF THE UNIVERSITY OF QUEENSLAND

AND OF THE UNIVERSITY OF AUSTRALIA

AND OF THE UNIVERSITY OF NEWCASTLE

AND OF THE UNIVERSITY OF WOLLONGONG

AND OF THE UNIVERSITY OF SYDNEY

AND OF THE UNIVERSITY OF MELBOURNE

AND OF THE UNIVERSITY OF ADELAIDE

AND OF THE UNIVERSITY OF PERTH

AND OF THE UNIVERSITY OF BRISBANE

AND OF THE UNIVERSITY OF QUEENSLAND

AND OF THE UNIVERSITY OF AUSTRALIA

AND OF THE UNIVERSITY OF NEWCASTLE

AND OF THE UNIVERSITY OF WOLLONGONG

AND OF THE UNIVERSITY OF SYDNEY

AND OF THE UNIVERSITY OF MELBOURNE

AND OF THE UNIVERSITY OF ADELAIDE

AND OF THE UNIVERSITY OF PERTH

AND OF THE UNIVERSITY OF BRISBANE

AND OF THE UNIVERSITY OF QUEENSLAND

AND OF THE UNIVERSITY OF AUSTRALIA

AND OF THE UNIVERSITY OF NEWCASTLE

AND OF THE UNIVERSITY OF WOLLONGONG

AND OF THE UNIVERSITY OF SYDNEY

AND OF THE UNIVERSITY OF MELBOURNE

Table 6

% Efficiency Gains Model vs Census Data

Economic Region	Stratum	n	Q = 0.0, t = 1.00			Census Data
			$\delta=1.00$	$\delta=1.10$	$\delta=1.20$	
53	1	2	1030	959	789	416
		3	946	864	681	337
		4	873	786	597	282
54	1	2	2759	2253	1483	2924
		3	2116	2082	1212	2867
		4	2486	1866	1044	2812
55	1	2	1670	1488	1115	1452
		3	1567	1356	956	1351
		4	1476	1244	834	1262
	2	2	1994	1729	1226	2228
		3	1865	1563	1039	1992
		4	1750	1425	899	1802
56	1	2	1309	1190	930	1653
		3	1219	1083	804	1563
		4	1141	993	705	1482
	2	2	4202	3328	2031	5180
		3	3798	2857	1615	4919
		4	3464	2499	1335	4683
57	1	2	1322	1202	939	795
		3	1231	1093	811	679
		4	1150	1001	711	592
	2	2	2784	2200	1333	2908
		3	2709	2052	1166	2739
		4	2638	1922	1034	2588

1914

1914-1915

1914		1915		1916		1917		1918		1919		1920		1921		1922		1923		1924		1925		1926		1927		1928		1929		1930		1931		1932		1933		1934		1935		1936		1937		1938		1939		1940		1941		1942		1943		1944		1945		1946		1947		1948		1949		1950		1951		1952		1953		1954		1955		1956		1957		1958		1959		1960		1961		1962		1963		1964		1965		1966		1967		1968		1969		1970		1971		1972		1973		1974		1975		1976		1977		1978		1979		1980		1981		1982		1983		1984		1985		1986		1987		1988		1989		1990		1991		1992		1993		1994		1995		1996		1997		1998		1999		2000		2001		2002		2003		2004		2005		2006		2007		2008		2009		2010		2011		2012		2013		2014		2015		2016		2017		2018		2019		2020		2021		2022		2023		2024		2025		2026		2027		2028		2029		2030		2031		2032		2033		2034		2035		2036		2037		2038		2039		2040		2041		2042		2043		2044		2045		2046		2047		2048		2049		2050		2051		2052		2053		2054		2055		2056		2057		2058		2059		2060		2061		2062		2063		2064		2065		2066		2067		2068		2069		2070		2071		2072		2073		2074		2075		2076		2077		2078		2079		2080		2081		2082		2083		2084		2085		2086		2087		2088		2089		2090		2091		2092		2093		2094		2095		2096		2097		2098		2099		2100		2101		2102		2103		2104		2105		2106		2107		2108		2109		2110		2111		2112		2113		2114		2115		2116		2117		2118		2119		2120		2121		2122		2123		2124		2125		2126		2127		2128		2129		2130		2131		2132		2133		2134		2135		2136		2137		2138		2139		2140		2141		2142		2143		2144		2145		2146		2147		2148		2149		2150		2151		2152		2153		2154		2155		2156		2157		2158		2159		2160		2161		2162		2163		2164		2165		2166		2167		2168		2169		2170		2171		2172		2173		2174		2175		2176		2177		2178		2179		2180		2181		2182		2183		2184		2185		2186		2187		2188		2189		2190		2191		2192		2193		2194		2195		2196		2197		2198		2199		2200		2201		2202		2203		2204		2205		2206		2207		2208		2209		2210		2211		2212		2213		2214		2215		2216		2217		2218		2219		2220		2221		2222		2223		2224		2225		2226		2227		2228		2229		2230		2231		2232		2233		2234		2235		2236		2237		2238		2239		2240		2241		2242		2243		2244		2245		2246		2247		2248		2249		2250		2251		2252		2253		2254		2255		2256		2257		2258		2259		2260		2261		2262		2263		2264		2265		2266		2267		2268		2269		2270		2271		2272		2273		2274		2275		2276		2277		2278		2279		2280		2281		2282		2283		2284		2285		2286		2287		2288		2289		2290		2291		2292		2293		2294		2295		2296		2297		2298		2299		2300		2301		2302		2303		2304		2305		2306		2307		2308		2309		2310		2311		2312		2313		2314		2315		2316		2317		2318		2319		2320		2321		2322		2323		2324		2325		2326		2327		2328		2329		2330		2331		2332		2333		2334		2335		2336		2337		2338		2339		2340		2341		2342		2343		2344		2345		2346		2347		2348		2349		2350		2351		2352		2353		2354		2355		2356		2357		2358		2359		2360		2361		2362		2363		2364		2365		2366		2367		2368		2369		2370		2371		2372		2373		2374		2375		2376		2377		2378		2379		2380		2381		2382		2383		2384		2385		2386		2387		2388		2389		2390		2391		2392		2393		2394		2395		2396		2397		2398		2399		2400		2401		2402		2403		2404		2405		2406		2407		2408		2409		2410		2411		2412		2413		2414		2415		2416		2417		2418		2419		2420		2421		2422		2423		2424		2425		2426		2427		2428		2429		2430		2431		2432		2433		2434		2435		2436		2437		2438		2439		2440		2441		2442		2443		2444		2445		2446		2447		2448		2449		2450		2451		2452		2453		2454		2455		2456		2457		2458		2459		2460		2461		2462		2463		2464		2465		2466		2467		2468		2469		2470		2471		2472		2473		2474		2475		2476		2477		2478		2479		2480		2481		2482		2483		2484		2485		2486		2487		2488		2489		2490		2491		2492		2493		2494		2495		2496		2497		2498		2499		2500		2501		2502		2503		2504		2505		2506		2507		2508		2509		2510		2511		2512		2513		2514		2515		2516		2517		2518		2519		2520		2521		2522		2523		2524		2525		2526		2527		2528		2529		2530		2531		2532		2533		2534		2535		2536		2537		2538		2539		2540		2541		2542		2543		2544		2545		2546		2547		2548		2549		2550		2551		2552		2553		2554		2555		2556		2557		2558		2559		2560		2561		2562		2563		2564		2565		2566		2567		2568		2569		2570		2571		2572		2573		2574		2575		2576		2577		2578		2579		2580		2581		2582		2583		2584		2585		2586		2587		2588		2589		2590		2591		2592		2593		2594		2595		2596		2597		2598		2599		2600		2601		2602		2603		2604		2605		2606		2607		2608		2609		2610		2611		2612		2613		2614		2615		2616		2617		2618		2619		2620		2621		2622		2623		2624		2625		2626		2627		2628		2629		2630		2631		2632		2633		2634		2635		2636		2637		2638		2639		2640		2641		2642		2643		2644		2645		2646		2647		2648		2649		2650		2651		2652		2653		2654		2655		2656		2657		2658		2659		2660		2661		2662		2663		2664		2665		2666		2667		2668		2669		2670		2671		2672		2673		2674		2675		2676		2677		2678		2679		2680		2681		2682		2683		2684		2685		2686		2687		2688		2689		2690		2691		2692		2693		2694		2695		2696		2697		2698		2699		2700		2701		2702		2703		2704		2705		2706		2707		2708		2709		2710		2711		2712		2713		2714		2715		2716		2717		2718		2719		2720		2721		2722		2723		2724		2725		2726		2727		2728		2729		2730		2731		2732		2733		2734		2735		2736		2737		2738		2739		2740		2741		2742		2743		2744		2745		2746		2747		2748		2749		2750		2751		2752		2753		2754		2755		2756		2757		2758		2759		2760		2761		2762		2763		2764		2765		2766		2767		2768		2769		2770		2771		2772		2773		2774		2775		2776		2777		2778		2779		2780		2781		2782		2783		2784		2785		2786		2787		2788		2789		2790		2791		2792		2793		2794		2795		2796		2797		2798		2799		2800		2801		2802		2803		2804		2805		2806		2807		2808		2809		2810		2811		2812		2813		2814		2815		2816		2817		2818		2819		2820		2821		2822		2823		2824		2825		2826		2827		2828		2829		2830		2831		2832		2833		2834		2835		2836		2837		2838		2839		2840		2841		2842		2843		2844		2845		2846		2847		2848		2849		2850		2851		2852		2853		2854		2855		2856		2857		2858		2859		2860		2861		2862		2863		2864		2865		2866		2867		2868		2869		2870		2871		2872		2873		2874		2875		2876		2877		2878		2879		2880		2881		2882		2883		2884		2885		2886		2887		2888		2889		2890		2891		2892		2893		2894		2895		2896		2897		2898		2899		2900		2901		2902		2903		2904		2905		2906		2907		2908		2909		2910		2911		2912		2913		2914		2915		2916		2917		2918		2919		2920		2921		2922		2923		2924		2925		2926		2927		2928		2929		2930		2931		2932		2933		2934		2935		2936		2937		2938		2939		2940		2941		2942		2943		2944		294	
------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	------	--	-----	--

Table 7

% Efficiency Gains Under Models

Economic Region	Stratum	n	$j\chi$ $\delta = 1.0$ $Q = 0.0$	$j\chi'$ $\delta = 1.0$ $Q_a = 0.0$
53	1	2	1030	4153
		3	946	3843
		4	873	3575
54	1	2	2759	3439
		3	2116	3263
		4	2486	3103
55	1	2	1670	4785
		3	1567	4500
		4	1476	4247
	2	2	1994	3342
		3	1865	3135
		4	1750	2951
56	1	2	1309	4791
		3	1219	4486
		4	1141	4216
	2	2	4202	5701
		3	3798	5164
		4	3464	4719
57	1	2	1322	3703
		3	1231	3458
		4	1150	3243
	2	2	2784	4747
		3	2709	4620
		4	2638	4500

REFERENCES

- [1] Freeman, D.H. and Koch, G.G. (1976), "An Asymptotic Covariance Structure for Testing Hypothesis on Raked Contingency Tables from Complex Surveys", Paper presented at the annual ASA meeting.
- [2] Ghangurde, P.D. and Singh, M.P. (1976), "Synthetic Estimation in the LFS", Technical Report, Household Surveys Development Staff.
- [3] Ghangurde, P.D. and Singh, M.P. (1977), "Evaluation of Synthetic Estimation in the LFS", Technical Report, Household Surveys Development Staff.
- [4] Gonzales, M.E. (1973), "Use and Evaluation of Synthetic Estimates", Proceedings of Social Statistics Section, American Statistical Association, 33-36.
- [5] Gray, G.B. and Ghangurde, P.D. (1975), "On A Ratio Estimate With Post-Stratified Weighting", Survey Methodology, Vol. 1,2, 134-142.
- [6] Ireland, C.T. and Kullback, S. (1968), "Minimum Discrimination Information Estimation", Biometrics, 24, 707-713.
- [7] Jones, D.H. and Coopersmith, L.A. (1976), "Ratio Estimator of the Total of a Subpopulation", Communications in Statistics, Theory and Methods A5(3), 251-260.
- [8] Keyfitz, N. (1957), "Estimates of Sampling Variance When Two Units Are Selected From Each Stratum", Journal of the American Statistical Association, 52, 503-510.
- [9] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples (with discussion), Journal of the Royal Statistical Society, B, 36, 1-37.
- [10] Kish, L. and Scott, A. (1971), "Retaining Units After Changing Strata and Probabilities", Journal of the American Statistical Association, 66, 461-70.
- [11] National Centre for Health Statistics (1968), Synthetic State Estimates of Disability, Technical Report.
- [12] Platek, R. and Singh, M.P. (1975), "A Strategy for Updating Continuous Surveys", Survey Methodology, Vol. 1,1, 16-26.
- [13] Platek, R. and Singh, M.P. (1976), "Methodology of Canadian Labour Force Survey", Technical Report, Household Surveys Development Staff.

- [14] Royall, R.M. (1971), "Linear Regression Models in Finite Population Sampling Theory", In Foundation of Statistical Inference, Holt, Rinehart and Winston.
- [15] Royall, R.M. and Herson, J. (1973), "Robust Estimation in Finite Populations I", Journal of the American Statistical Association, 68, 880-889.
- [16] Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review", Journal of the Royal Statistical Society, A, 139, 183-204.
- [17] Tepping, B.J. (1968), "Variance Estimation in Complex Surveys", Proceedings of the Social Statistics Section, ASA, 11-18.
- [18] Woodruff, R.S. (1971), "Simple Method for Approximating Variance of Complicated Estimate", Journal of the American Statistical Association, 411-414.

APPENDIX

Consider the model (3.3) given by

$$X_i = \beta P_i + e_i, \quad i \in J$$

$$X_i = \delta \beta P_i + e_i, \quad i \notin J, \quad \delta \neq 1$$

where $\epsilon(e_i | P_i) = 0$, $\epsilon(e_i^2 | P_i) = \sigma^2 \cdot P_i^t$, $t > 0$, $\sigma^2 \geq 0$

$$\epsilon(e_i e_{i'} | P_i, P_{i'}) = 0, \quad i \neq i'; \quad i, i' = 1, 2, \dots, N.$$

Now

$$B(JX) = (XW - JX)$$

$$= [W\beta^J P + W e + W\delta\beta(P - J_P) - \beta^J P - J e]$$

$$= [\beta^J P W (1-\delta) + W\delta\beta P - \beta^J P + W e - J e],$$

$$\epsilon[B(JX)]^2 = \beta^2 J_P^2 \left[(1-\delta + \delta \frac{P}{J_P}) W - 1 \right]^2 + \sigma^2 \cdot \sum_{i=1}^N P_i^t (W - J_{A_i})^2,$$

$$\epsilon \left[\frac{J_X^2 - W^2 X^2}{n} \right] = \epsilon \left[\frac{(\beta^J P + J e)^2 - W^2 (\beta^J P + \delta \beta (P - J_P) + e)^2}{n} \right]$$

$$= \frac{\beta^2 J_P^2 - W^2 \beta^2 [J_P + \delta (P - J_P)]^2}{n} + \frac{\sigma^2}{n} \sum_{i=1}^N P_i^t (J_{A_i} - W^2),$$

$$\epsilon \left[\frac{P}{n} \sum_{i=1}^N \frac{[J_X^2 - W^2 X^2]}{P_i} \right]$$

100

100

100

100

100

100

100

100

100

100

100

100

100

100

$$\begin{aligned}
 &= E \left[\frac{P}{n} \sum_{i=1}^N \frac{\beta^2 P_i^2 J_{A_i} - W^2 [\beta P_i J_{A_i} + \beta \delta P_i (1 - J_{A_i})]^2}{P_i} \right] \\
 &+ \frac{\sigma^2}{n} \left[P \sum_{i=1}^N P_i^{t-1} (J_{A_i} - W)^2 \right] \\
 &= \frac{\beta^2 P}{n} \left[J_P - W^2 \sum_{i=1}^N P_i [J_{A_i} + \delta(1 - J_{A_i})]^2 \right] + \frac{\sigma^2}{n} \left[P \sum_{i=1}^N P_i^{t-1} (J_{A_i} - W^2) \right].
 \end{aligned}$$

Hence

$$\begin{aligned}
 &E[V(\hat{J}_X) - \text{MSE}(\hat{J}_X)] \\
 &= E \left[\frac{P}{n} \sum_{i=1}^N \frac{[J_{X_i}^2 - W^2 X_i^2]}{P_i} - \frac{(J_X^2 - W^2 X^2)}{n} - [W_X - J_X]^2 \right] \\
 &= \frac{\beta^2 P}{n} \left[J_P - W^2 \sum_{i=1}^N P_i (J_{A_i} + \delta(1 - J_{A_i}))^2 \right] \\
 &- \frac{1}{n} [\beta^2 J_P^2 - W^2 \beta^2 [J_P + \delta(P - J_P)]^2 - \beta^2 J_P^2 [(1 - \delta + \delta \frac{P}{J_P}) W - 1]^2] \\
 &+ \frac{\sigma^2}{n} \left[P \sum_{i=1}^N P_i^{t-1} (J_{A_i} - W^2) - \sum_{i=1}^N P_i^t (J_{A_i} - W^2) - n \sum_{i=1}^N P_i^t (W - J_{A_i})^2 \right] \\
 &= \frac{\beta^2}{n} \cdot J_P (P - J_P) - \frac{\beta^2}{n} W^2 P \left[\sum_{i=1}^N P_i (J_{A_i} + \delta(1 - J_{A_i}))^2 - P \left[\frac{J_P}{P} + \delta(1 - \frac{J_P}{P}) \right]^2 \right] \\
 &- \beta^2 J_P^2 [(1 - \delta + \delta \frac{P}{J_P}) W - 1]^2 \\
 &+ \frac{\sigma^2}{n} \left[\sum_{i=1}^N P_i^{t-1} [(P - P_i) (J_{A_i} - W^2) - n P_i (J_{A_i} - W)^2] \right].
 \end{aligned}$$

...
...
...

...
...

...
...

...

...
...

...
...

...
...

...
...

...
...

...
...

...
...

...
...

The relative gain in efficiency is given by

$$\epsilon[V(\hat{J}_X) - \text{MSE}(\hat{J}_X)]/\epsilon[\text{MSE}(\hat{J}_X)].$$

Now

$$\epsilon[\text{MSE}(\hat{J}_X)] = \epsilon\left[\frac{W^2}{n} \left[P \sum_{i=1}^N \frac{X_i^2}{P_i} - X^2 \right] + [B(\hat{J}_X)]^2\right],$$

$$\epsilon\left[\frac{W^2}{n} \left[P \sum_{i=1}^N \frac{X_i^2}{P_i} - X^2 \right]\right]$$

$$= \frac{W^2}{n} \epsilon\left[P \cdot \sum_{i=1}^N \frac{\beta P_i J_{A_i} + \delta \beta P_i (1-J_{A_i})}{P_i} - [\beta J_P + \delta \beta (P-J_P)]^2\right]$$

$$+ \frac{\sigma^2}{n} W^2 \left[\sum_{i=1}^N P_i^{t-1} (P-P_i) \right].$$

$$= \frac{W^2}{n} \beta^2 P \left[\sum_{i=1}^N P_i [J_{A_i} + \delta(1-J_{A_i})]^2 - P \left[\frac{J_P}{P} + \delta \left(1 - \frac{J_P}{P}\right) \right]^2 \right]$$

$$+ \frac{\sigma^2}{n} \cdot W^2 \left[\sum_{i=1}^N P_i^{t-1} (P-P_i) \right]$$

Hence by simplifying

$$\epsilon[V(\hat{J}_X) - \text{MSE}(\hat{J}_X)]/\epsilon[\text{MSE}(\hat{J}_X)] = A/B$$

where

$$A = \frac{J_P(P-J_P)}{n} - J_P^2 \left[\left(1-\delta + \delta \frac{P}{J_P}\right) W-1 \right]^2$$

$$- \frac{W^2 P}{n} \left[\sum_{i=1}^N P_i [J_{A_i} + \delta(1-J_{A_i})]^2 - P \left[\frac{J_P}{P} + \delta \left(1 - \frac{J_P}{P}\right) \right]^2 \right]$$

the first of these is the fact that the

second of these is the fact that the

third of these is the fact that the

fourth of these is the fact that the

fifth of these is the fact that the

sixth of these is the fact that the

seventh of these is the fact that the

eighth of these is the fact that the

ninth of these is the fact that the

tenth of these is the fact that the

11

twelfth of these is the fact that the

thirteenth of these is the fact that the

$$+ \frac{\sigma^2}{n\beta^2} \left[\sum_{i=1}^N p_i^{t-1} [(p-p_i)(j_{A_i}-w)^2] - np_i(j_{A_i}-w)^2 \right]$$

$$B = j_P^2 \left[\left(1-\delta + \delta \frac{P}{j_P}\right) w-1 \right]^2$$

$$+ \frac{w^2 P}{n} \left[\sum_{i=1}^N p_i [j_{A_i} + \delta(1-j_{A_i})]^2 - P \left[\frac{j_P}{P} + \delta \left(1 - \frac{j_P}{P}\right) \right]^2 \right]$$

$$+ \frac{\sigma^2}{n\beta^2} \left[\sum_{i=1}^N p_i^{t-1} [(p-p_i) w^2 + np_i(j_{A_i}-w)^2] \right]$$

which is (3.4).

L'ENQUÊTE SUR LA PROFESSION DES SALARIÉS (E.P.S.)

M.A. Hidiroglou
Division des méthodes d'enquêtes "entreprises"

Cet article décrit brièvement la méthodologie reliée à l'Enquête sur la profession des salariés, qui s'effectue en principe tous les deux ans depuis 1973. L'article présente entre autre l'étendue de l'enquête, le plan d'échantillonnage et l'estimation.

1. INTRODUCTION

L'enquête sur la profession des salariés (E.P.S.) a été conçue par le ministère de la Main-d'Oeuvre et de l'Immigration, et par Statistique Canada dans les années soixante-dix. Le but de cette enquête était d'évaluer la répartition des professions parmi les salariés de la population active au Canada et de présenter cette répartition par province et par secteur d'activité économique.

Ce genre de données avait toujours été fournies par le recensement décennal. Cependant, entre les années 1961 et 1971, un nouveau dictionnaire d'occupations fut développé, et il annula par conséquent la possibilité de comparer le recensement de 1971 au recensement de 1961 vis-à-vis du changement d'orientation professionnelle. Il aurait donc fallu attendre jusqu'au recensement de 1981 pour obtenir les résultats de ces changements. Ceci aurait été un retard inacceptable. Un des moyens pour remplir ce vide aurait été d'ajouter au recensement de 1976 une question sur l'occupation professionnelle. Après beaucoup de discussions il fut jugé que cette solution serait non seulement trop chère, mais imposerait aussi, à cause du manque de temps pour planifier efficacement, des tensions trop sévères au recensement de 1976.

D'autres sources possibles furent étudiées. En voici quelques-unes: des listes de membres d'associations, un élargissement de l'enquête sur la Main-d'oeuvre et une enquête postale auprès des familles. Toutes ces possibilités furent rejetées pour des raisons diverses, parmi lesquelles on peut énumérer une base de sondage incomplète, des coûts excessifs, une complexité peu maniable. Donc, les choix se réduisirent très rapidement à une enquête sur les employés. L'avantage le plus important de ce choix était le coût. On peut expliquer ce fait par deux raisons. Primo, on peut amasser beaucoup de données lors d'une entrevue auprès d'un établissement: en effet, la plupart des grands établissements fournissent des listes préparées par ordinateur. Secondo, une partie de l'enquête, plus précisément le secteur ES-2 ou le secteur se composant de firmes de moins de 20 employés, peut-être effectuée par téléphone.

Il existe aussi d'autres avantages à aller directement aux établissements. Un de ces avantages qui se révèle d'une façon très distincte, est le fait que les employeurs fournissent une réponse plus objective sur l'occupation de leurs employés que le chef de famille.

2. ETENDUE DE L'ENQUÊTE

L'E.P.S. est une enquête qui recueille les données sur une période d'un an. Les unités répondantes sont choisies bien avant les entrevues. Chaque unité répondante est soumise seulement à une enquête au cours de l'année. A la fin de l'année, l'échantillonnage est complet.

L'unité déclarante est généralement une unité déclarante d'emploi (U.D.E.) utilisée dans l'enquête mensuelle sur l'emploi et la rémunération (E.M.E.R.). La désignation des diverses activités et groupes d'activités des unités déclarantes est fondée sur la classification des activités économiques de 1960. Une description de ces activités et de ces groupes figure dans le Manuel de la Classification Type des Industries, édition de 1960, numéro 12-501 au catalogue.

Quant au champ d'observation, tous les secteurs d'activités économiques sont couverts, sauf l'agriculture, la pêche et le piégeage, les services domestiques et l'élément non civil de l'administration publique et de la défense. Quant à la classification géographique, les unités déclarantes sont assignées à une province en fonction de la classification géographique type de 1972. Les données de Terre-Neuve, de l'Île-du-Prince-Édouard, de la Nouvelle-Écosse et du Nouveau-Brunswick sont regroupées dans quelques tableaux qui présentent les statistiques de la région de l'Atlantique. De même, les données du Manitoba, de la Saskatchewan et de l'Alberta sont regroupées pour donner les statistiques de la région des Prairies.

Les données sur la profession des salariés englobent tous les employés figurant sur les listes de paie des unités déclarantes durant la période de référence. Le concept de salarié utilisé par l'E.P.S. est conforme à celui de l'enquête mensuelle sur l'emploi et la rémunération; par conséquent, les propriétaires ou associés actifs d'une entreprise non-constituée en corporation et les directeurs inactifs ne sont pas inclus dans l'E.P.S. De ce fait, les employés à salaire et à traitement, qui travaillent à plein temps, à temps partiel ou occasionnellement contre rémunération, ou encore qui sont en congé payé durant la période de référence, ne sont pas comptés comme salariés.

3. PLAN DE SONDAGE

En peu de mots, l'enquête sur la profession des salariés est un plan de sondage stratifié qui s'effectue par téléphone ou par entrevue personnelle. L'unité de base dans l'échantillonnage est l'unité déclarante d'emploi (U.D.E.).

Il faut ajouter que le plan de sondage de l'E.P.S. a été dicté par deux autres enquêtes à la Division des méthodes d'enquêtes "entreprises". Ces deux enquêtes sont respectivement: les enquêtes mensuelles sur l'emploi et les feuilles de paie (ES-1 et ES-2) et l'enquête bi-mensuelle des postes vacants (E.P.V.).

L'ES-1 est un recensement postal des entreprises qui ont plus de 20 employés ou plus d'un établissement. La composition des unités déclarantes associées aux entreprises multiples coïncide plus ou moins avec la composition de la section de comptabilité de l'entreprise. Il y a environ 54,000 unités déclarantes dans l'ES-1.

L'ES-2 est une enquête postale des entreprises qui ne sont pas dans l'ES-1, c'est-à-dire les entreprises simples avec moins de vingt employés. L'univers ES-2 est composé d'environ 500,000 unités répondantes.

L'enquête sur les postes vacants est une enquête par sondage menée par la poste et par entrevues auprès d'employeurs qui représentent environ 90% de l'emploi au Canada (les secteurs exclus étant l'agriculture, le service domestique, le secteur militaire, la pêche et le piégeage). Chaque occasion (une occasion se répétant toutes les deux semaines) couvre approximativement 38,000 unités déclarantes de postes vacants.

Les enquêtes ont lieu deux fois par mois. Il faut ici dire que l'E.P.S. ne se sert que du secteur ES-2 de l'enquête des postes vacants. En effet, l'enquête des postes vacants dérive aussi de l'ES-1 et de l'ES-2.

L'enquête des postes vacants comporte deux phases: Une étape postale et une étape de l'entrevue. La partie de l'ES-2 dont on se sert dans l'E.P.S. est la phase de l'entrevue du secteur ES-2 de l'enquête des postes vacants. L'étape postale et l'étape par entrevue sont menées concurremment. L'échantillon de l'étape par entrevue est un sous-échantillon de l'échantillon de l'étape postale.

Nous allons maintenant décrire le plan de sondage utilisé par l'E.P.S. Pour fins d'échantillonnage, l'univers des U.D.E. est divisé en six secteurs, soit: (i) ES-1, comprenant les grandes entreprises qui emploient 20 personnes ou plus, (ii) ES-2, comprenant les entreprises employant 19 personnes ou moins, (iii) les établissements d'enseignement et autres institutions, (iv) le gouvernement fédéral, (v) les gouvernements provinciaux et (vi) les administrations locales.

L'univers des U.D.E. de chaque secteur est divisé en un certain nombre de strates (Classification Type des Industries - C.T.I.) selon la situation géographique, le groupe ou le genre d'activités économiques et la taille. Pour tous les secteurs, sauf le ES-2, la première étape est la suivante: chacune des strates (C.T.I.) est divisée, selon la situation économique, en deux sous-strates, que nous désignerons par sous-strate complète et sous-strate partielle. Pour simplifier les choses, disons que la sous-strate complète se comporte d'une partie des plus grandes U.D.E. de cette strate, et que le nombre total d'employés dans ces U.D.E. renferme 40 à 50% des salariés. La sous-strate partielle renferme les U.D.E. restantes. Les U.D.E. de la sous-strate complète sont échantillonnées à 100%, tandis que les U.D.E. de la sous-strate partielle le sont à un certain taux. Ce taux est proportionnel au nombre d'employés dans la strate en question. Il faut ici préciser que l'E.P.S. se sert d'un sondage à deux degrés. Comme l'a été dit, au premier degré du sondage, les U.D.E. de la sous-strate partielle sont échantillonnées. Si la collecte intégrale des données sur l'emploi d'une U.D.E. quelconque n'est pas possible, on passe au deuxième degré du sondage désigné sous le nom d'échantillonnage dans l'entreprise. Cette deuxième étape consiste en une collection systématique des données sur l'emploi selon un taux d'échantillonnage établi en fonction de la taille de l'U.D.E.

Pour fins de collecte, l'échantillon est réparti en quatre panneaux E.P.S. (qui correspondent plus ou moins aux saisons de l'année) couvrant toute l'année. A l'intérieur de chaque panneau E.P.S., la charge de travail est répartie parmi le personnel sur le terrain toutes les deux semaines, de façon à atteindre un total de six occasions par panneau E.P.S. Il est également à noter qu'à chaque occasion, l'échantillon est subdivisé en deux sous-échantillons redoublés ou sous-panneaux, en vue de l'estimation des variances. La formation de ces sous-panneaux est produite lors de la répartition initiale des U.D.E. aux panneaux.

the first of these is the fact that the
the second is the fact that the
the third is the fact that the
the fourth is the fact that the
the fifth is the fact that the
the sixth is the fact that the
the seventh is the fact that the
the eighth is the fact that the
the ninth is the fact that the
the tenth is the fact that the
the eleventh is the fact that the
the twelfth is the fact that the
the thirteenth is the fact that the
the fourteenth is the fact that the
the fifteenth is the fact that the
the sixteenth is the fact that the
the seventeenth is the fact that the
the eighteenth is the fact that the
the nineteenth is the fact that the
the twentieth is the fact that the
the twenty-first is the fact that the
the twenty-second is the fact that the
the twenty-third is the fact that the
the twenty-fourth is the fact that the
the twenty-fifth is the fact that the
the twenty-sixth is the fact that the
the twenty-seventh is the fact that the
the twenty-eighth is the fact that the
the twenty-ninth is the fact that the
the thirtieth is the fact that the
the thirty-first is the fact that the
the thirty-second is the fact that the
the thirty-third is the fact that the
the thirty-fourth is the fact that the
the thirty-fifth is the fact that the
the thirty-sixth is the fact that the
the thirty-seventh is the fact that the
the thirty-eighth is the fact that the
the thirty-ninth is the fact that the
the fortieth is the fact that the
the forty-first is the fact that the
the forty-second is the fact that the
the forty-third is the fact that the
the forty-fourth is the fact that the
the forty-fifth is the fact that the
the forty-sixth is the fact that the
the forty-seventh is the fact that the
the forty-eighth is the fact that the
the forty-ninth is the fact that the
the fiftieth is the fact that the
the fifty-first is the fact that the
the fifty-second is the fact that the
the fifty-third is the fact that the
the fifty-fourth is the fact that the
the fifty-fifth is the fact that the
the fifty-sixth is the fact that the
the fifty-seventh is the fact that the
the fifty-eighth is the fact that the
the fifty-ninth is the fact that the
the sixtieth is the fact that the
the sixty-first is the fact that the
the sixty-second is the fact that the
the sixty-third is the fact that the
the sixty-fourth is the fact that the
the sixty-fifth is the fact that the
the sixty-sixth is the fact that the
the sixty-seventh is the fact that the
the sixty-eighth is the fact that the
the sixty-ninth is the fact that the
the seventieth is the fact that the
the seventy-first is the fact that the
the seventy-second is the fact that the
the seventy-third is the fact that the
the seventy-fourth is the fact that the
the seventy-fifth is the fact that the
the seventy-sixth is the fact that the
the seventy-seventh is the fact that the
the seventy-eighth is the fact that the
the seventy-ninth is the fact that the
the eightieth is the fact that the
the eighty-first is the fact that the
the eighty-second is the fact that the
the eighty-third is the fact that the
the eighty-fourth is the fact that the
the eighty-fifth is the fact that the
the eighty-sixth is the fact that the
the eighty-seventh is the fact that the
the eighty-eighth is the fact that the
the eighty-ninth is the fact that the
the ninetieth is the fact that the
the ninety-first is the fact that the
the ninety-second is the fact that the
the ninety-third is the fact that the
the ninety-fourth is the fact that the
the ninety-fifth is the fact that the
the ninety-sixth is the fact that the
the ninety-seventh is the fact that the
the ninety-eighth is the fact that the
the ninety-ninth is the fact that the
the hundredth is the fact that the

On vient de décrire le plan de sondage des secteurs 1, 3, 4, 5 et 6. Pour ce qui est de l'ES-2, l'échantillon de ce dernier est fondamentalement le même que l'échantillon d'entrevue de l'enquête sur les postes vacants. Cet échantillon, lequel renferme environ 65,000 unités déclarantes par an, est restructuré conformément au plan général de l'E.P.S. exposé ci-dessus. Au départ, l'univers ES-2 a été stratifié en strates économiques (C.T.I.). A chacune de ces strates a été alloué un taux de sondage. L'échantillon de chacune de ces strates a été ainsi réparti sur M panneaux ES-2, M étant un multiple de 24. Le nombre de panneaux ES-2 présent en toutes occasions et par strate est 12. Toutes les fois qu'une nouvelle occasion se présente, un panneau ES-2 rentre dans l'échantillon, y demeure pendant 12 occasions et en ressort. Environ un douzième de la phase postale de l'ES-2 est entrevue personnellement par occasion. C'est de cette entrevue dont on se sert pour l'E.P.S.

4. ESTIMATION

L'estimateur pour l'E.P.S. dans une strate typique est:

$$\hat{M}_{..k} = x \frac{\sum_{h=1}^{L/2} \sum_{a=1}^2 \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{m_{hik}^a}{m_{hi}^a} M_{hi}^a}{\sum_{h=1}^{L/2} \sum_{a=1}^2 \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}^a}$$

ou

M_{hi}^a = nombre total d'employés dans la $i^{\text{ième}}$ U.D.E. du $a^{\text{ième}}$ réplikat dans la $h^{\text{ième}}$ strate C.T.I.,

m_{hik}^a = nombre d'employés échantillonnés qui sont dans la $k^{\text{ième}}$ classe d'occupation, dans la $i^{\text{ième}}$ U.D.E. du $a^{\text{ième}}$ réplikat dans la $h^{\text{ième}}$ strate C.T.I.,

m_{hi}^a = nombre d'employés échantillonnés dans la $i^{\text{ème}}$ U.D.E.
du $a^{\text{ème}}$ réplikat dans la $h^{\text{ème}}$ strate C.T.I.,

N_h = nombre total d'U.D.E. dans la $h^{\text{ème}}$ strate,

n_h = nombre d'U.D.E. échantillonnées dans la $h^{\text{ème}}$ strate,

h = strate C.T.I. peut aussi représenter un classement à
travers les panneaux et les réplikats des C.T.I.,

L = nombre de classements possibles,

X = valeur de base pour les employés appartenant à une
strate C.T.I. - ces valeurs sont dérivées
des sondages ES-1 et ES-2.

L'estimateur de variance pour l'estimateur donné ci-dessus est:

$$\begin{aligned} \text{var} (\hat{M}_{..k}) &= \left(\frac{X}{\hat{M}_{...}} \right)^2 \left[\sum_{h=1}^{L/2} \left(\frac{B_h - 4}{B_h} \right) \right. \\ &\quad \times \left. \left((\hat{M}_{h.k}^1 - \hat{M}_{h.k}^2) - \hat{R}_{ck} (\hat{M}_{h..}^1 - \hat{M}_{h..}^2) \right)^2 \right. \\ &\quad \left. + \sum_{h=1}^L \sum_{i=1}^{n_h} B_h M_{hi} (r_{hi} - 1) \frac{m_{hik}}{m_{hi}} \left(1 - \frac{m_{hik}}{m_{hi}} \right) \right], \end{aligned}$$

ou

$$\hat{M}_{...} = \sum_{h=1}^L \sum_{i=1}^{n_h} M_{hi}$$

= nombre total d'employés dans la C.T.I. mesuré
par l'E.P.S.,

$$\hat{M}_{h.k}^a = \sum_{i=1}^{n_h} B_h r_{hi} m_{hik}^a ; a = 1, 2$$

= nombre total d'employés qui sont dans la $k^{\text{ème}}$
classe d'occupation, dans le $a^{\text{ème}}$ réplikat
dans la $h^{\text{ème}}$ strate C.T.I.,

$$\hat{M}_{h..}^a = \sum_k \hat{M}_{h.k}^a ; a = 1, 2$$

= nombre total d'employés dans le $a^{\text{ième}}$ réplicat, estimé par l'E.P.S.,

$$r_{hi} = \frac{M_{hi}}{m_{hi}} ; i = 1, 2, \dots, n_h ; h=1, 2, \dots, L$$

= taux de sondage à l'intérieur de chaque U.D.E.,

$$B_h = \frac{N_h}{n_h}$$

= taux de sondage à l'intérieur de chaque sous-strate C.T.I.,

$$\hat{R}_{ck} = \frac{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{m_{hik}}{m_{hi}} M_{hi}}{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}}$$

= proportion estimée des employés dans la $k^{\text{ième}}$ classe d'occupation dans la strate C.T.I.

Il est à noter que l'estimateur de la variance fait apparaître les composantes des deux degrés de sondage: la variance inter prenant en considération les variations entre U.D.E. et la variance intra pour les variations entre employés sous-échantillonnés dans les U.D.E.

Le système de totalisation de l'E.P.S. peut fournir entre autre, des totalisations par secteur et par activité économique, et dans ces catégories, par province, au niveau des codes de professions, à 1, 2, 3 ou 4 chiffres. En général, les totalisations présentent les données d'une activité choisie ou d'un groupe de secteurs selon la classification des professions (C.C.D.P.) par province.

1. The first part of the paper discusses the importance of the study.

2. The second part of the paper discusses the methodology used in the study.

3. The third part of the paper discusses the results of the study.

4. The fourth part of the paper discusses the conclusions of the study.

5.

6. The fifth part of the paper discusses the implications of the study.

7. The sixth part of the paper discusses the limitations of the study.

8.

9. The seventh part of the paper discusses the future research.

10. The eighth part of the paper discusses the acknowledgments.

11. The ninth part of the paper discusses the references.

12. The tenth part of the paper discusses the appendices.

13. The eleventh part of the paper discusses the index.

14. The twelfth part of the paper discusses the glossary.

ABSTRACT

This article describes briefly the methodology of the Occupational Employment Survey, which has been conducted every second year since 1973. The article presents the scope of the survey, the sampling plan and the estimation procedure.

REFERENCES

- [1] Cochran, W.G., "Sampling Techniques", John Wiley and Sons, Inc., 1973.
- [2] Gower, D., "The Occupational Employment Survey Background Paper for Users", Division du travail, Statistique Canada, avril 1975.
- [3] Hidioglou, M.A., "OES Variance Estimation", Division des méthodes d'enquêtes "entreprises", Statistique Canada, mars 1977.
- [4] Koop, J.C., "On Theoretical Questions Underlying the Technique of Replicated or Interpenetrating Samples", American Statistical Association Proceedings, 1960.
- [5] Ostry, S. et Sunter, A.B., "Definitional and Design Aspects of the Canadian Job Vacancy Survey", Journal of the American Statistical Association, 65, 331, 1059-70, 1969.
- [6] Strauss, E.E. et Sunter, A.B., "The Canadian Occupational Employment Survey", Division des méthodes d'enquêtes "entreprises", Statistique Canada, 1973.
- [7] Strauss, E.E., "JVS/ES2 Relating Sample Specifications", Division des méthodes d'enquêtes "entreprises", Statistique Canada, octobre 1974.

SOME FACTORS AFFECTING NON-RESPONSE¹

R. Platek
Household Surveys Development Division

Non-response exists in any survey, but its magnitude depends upon the type of survey, the interviewers' ability to conduct an interview, and the respondents' motivation to respond to survey questions. This paper discusses non-response in relation to a number of household surveys and in particular the behaviour of non-response rates over time in a continuous survey such as the Canadian Labour Force Survey.

A profile of interviewers employed by Statistics Canada shows that the correlation between non-response and a number of interviewer characteristics is not significant. Respondents themselves, and their motivation, are the key elements in an interview process and therefore in respondent relations.

This article draws on the results of various studies conducted to investigate the effects of response burden, choice of respondent and response incentives to provide some insight into the characteristics of non-respondents.

1. INTRODUCTION

During the past several years, the number of surveys, as a means of collecting a variety of data, has greatly increased in most countries. Any survey, whatever its type and whatever the method of collecting data, will suffer from some non-response. Most practicing statisticians or data analysts recognize non-response as an important measure of quality of data since it affects the estimates by introducing both a possible bias and an increase in sampling variance. For example, since the sampling variance of the estimates is inversely proportional to the response rate, estimates based on a simple random sample with 80%

¹ Presented in substantially the present form as an invited paper at the 41st Session of the International Statistical Institute, New Delhi, 5-15 December, 1977.

response rate will have a sampling variance that is 12.5% higher than the variance of corresponding estimates with 90% response rate. In multi-stage clustered samples, the same relationship holds approximately but would affect mainly the final stage of sampling. The relationship between the bias and the size of non-response while perhaps more important is less obvious since it depends on both the magnitude of non-response and the characteristics of both respondents and non-respondents. A reduction of non-response in the field does not necessarily ensure a reduction in bias. In fact, it can be shown that if the procedures for reduction of non-response are not well thought out and appropriately executed, the bias may not be reduced and could even be increased. In most surveys, however, the elimination or reduction of non-response is very important and beneficial.

One way of dealing with non-response is through methods of imputation or adjustment of weights at the processing and estimation stage. While adjustments for non-response may be more or less effective in reducing bias, well designed data collection operations will keep non-response at an acceptable level and at reasonable cost, thus minimizing the necessity for the application of these methods for non-response.

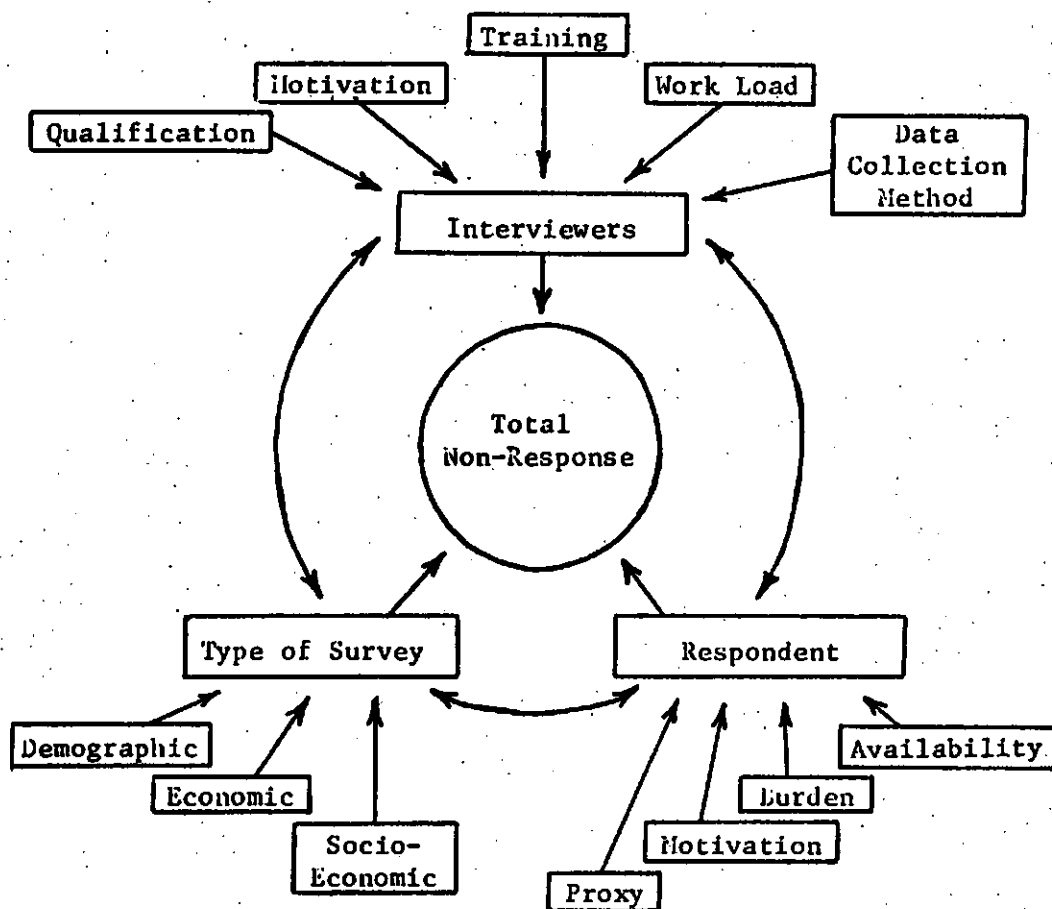
It is generally recognized that the problem of non-response is very complex since it is affected by many factors which, quite often, operate in different directions for various survey situations. It is, therefore, virtually impossible to think of one simple approach which would have a uniform application to every survey.

In this paper, non-response in household surveys will be discussed in relation to such factors as interviewers, respondents and content of surveys. In addition, some experimental studies are examined, providing some information on respondent burden, proxy and non-proxy reporting, profile of interviewers and profiles of non-respondents. Most of the results are drawn from the experiences of the Canadian Labour Force Survey (LFS) which is a monthly survey of about 55,000 households.

2. NON-RESPONSE

Non-response may be defined as a failure to obtain a usable report from a reporting unit, which legitimately falls into the sample in a particular survey.

The potential sources or causes of non-response are basically related to (i) content of surveys, (ii) data collection method and (iii) respondent. The main sources and other factors affecting non-response may be presented graphically as follows:



Total non-response may be broken down into a number of components, each of which has a different cause and requires a different treatment. For example, in a household survey, one can recognize the following components of non-response: a) Household temporarily absent, b) No one at home, c) Refusal, d) No interviewer available, e) Bad weather conditions, etc.

Non-response occurs due to operational difficulties, time and cost restraints, a lack of co-operation from respondents, the inability or unwillingness of interviewers to track down missing respondents, or for some other reason. The non-response rate measures the severity of non-response problems, and it is calculated as the percentage of non-respondent households out of all sampled households.

As an illustration, Tables I and II present some response and non-response rates pertaining to LFS and other surveys.

TABLE I

LFS Non-Response Rates by Component at the Canada Level

Month Year		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Overall													
	1973	7.3	7.2	6.8	7.9	7.0	8.4	15.1	10.9	6.5	5.7	5.2	6.6
A	1974	6.0	6.0	6.4	8.3	7.0	6.8	10.4	8.8	5.6	5.5	4.3	4.6
	1975	4.3	4.7	4.6	4.7	4.7	5.8	7.6	6.3	4.3	4.5	4.3	5.3
	1975	10.2	9.3	8.7	7.5	7.9	8.9	13.2	11.0	7.5	7.5	7.0	7.6
B	1976	7.9	7.7	7.3	8.8	9.2	8.6	11.7	9.2	5.8	5.9	6.0	5.3
	1977	6.1	5.9	5.8	4.5	5.8	5.7						
Temporarily Absent													
	1973	1.8	2.2	1.9	2.4	1.8	3.3	9.1	5.6	1.6	1.3	1.2	1.7
A	1974	1.7	1.8	1.9	2.0	1.5	2.0	6.1	4.7	2.0	1.7	1.0	1.4
	1975	1.4	1.6	1.6	1.2	1.2	2.2	4.2	3.0	1.1	0.9	0.7	1.2
	1975	2.0	2.2	2.1	1.5	1.6	2.8	7.1	5.0	1.7	1.4	1.2	1.4
B	1976	1.9	2.1	2.3	2.3	1.8	2.4	6.2	4.1	1.7	1.4	1.3	1.3
	1977	1.9	1.9	2.2	1.4	1.9	2.3						
No One at Home													
	1973	2.5	2.1	2.0	2.6	2.5	2.7	3.2	2.3	2.1	1.9	1.6	2.0
A	1974	1.5	1.7	1.8	2.8	1.9	1.8	1.7	1.7	1.4	1.7	1.4	1.2
	1975	1.0	0.9	1.0	1.2	1.1	1.3	1.2	1.2	1.1	1.4	1.2	1.6
	1975	2.7	2.2	1.8	1.5	1.9	1.8	2.2	2.1	1.8	1.7	1.7	1.8
B	1976	2.2	1.9	1.5	2.8	2.8	2.4	2.4	2.2	1.7	2.0	2.2	1.8
	1977	1.8	1.9	1.6	1.4	2.0	1.7						
Refusal													
	1973	1.7	1.9	1.9	2.0	2.0	1.9	1.9	2.3	2.1	2.0	1.9	1.7
A	1974	1.6	1.6	1.7	2.1	2.4	2.3	2.1	1.9	1.6	1.4	1.3	1.2
	1975	1.2	1.2	1.2	1.4	1.6	1.4	1.4	1.3	1.3	1.2	1.4	1.3
	1975	2.1	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.6	1.5	1.5	1.4
B	1976	1.5	1.6	1.6	1.9	2.5	2.4	2.0	1.9	1.6	1.6	1.7	1.6
	1977	1.6	1.5	1.4	1.3	1.3	1.3						

Cont'd

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

...the ... of ...
...the ... of ...
...the ... of ...
...the ... of ...

'Other'													
A	1973	1.3	1.0	1.0	0.9	0.7	0.5	0.9	0.7	0.7	0.5	0.5	1.2
	1974	1.2	1.0	0.9	1.4	1.2	0.7	0.5	0.5	0.6	0.7	0.6	0.8
	1975	0.7	1.0	0.8	0.9	0.8	0.9	0.8	0.8	0.8	1.0	1.0	1.2
B	1975	3.4	2.7	2.7	2.5	2.5	2.5	2.2	2.3	2.4	2.9	2.6	3.0
	1976	2.3	2.1	1.9	1.8	2.1	1.6	1.1	1.0	0.8	0.9	0.8	0.6
	1977	0.8	0.6	0.6	0.4	0.6	0.4						

A rates are those rates taken from the old Labour Force Survey.

B rates are those rates taken from the new Labour Force Survey.

A few trends in the behaviour of non-response rates in the LFS should be pointed out. The overall non-response rate always increases sharply during the months of July and August (often to over 10%). This is mostly due to respondents being Temporarily absent on vacation (5% to 9%). The size of non-response is mainly determined by "Temporarily absent" and "No one at home". The refusal rates have been fairly steady over a number of years with some downward trend recently and they appear to be slightly higher in summer months. One significant trend is that the Revised Labour Force Survey in 1975 had a relatively high non-response rate. This was probably due to the hiring of an almost completely new staff of interviewers along with heavy burdens on the field supervisors due to new procedures, new samples and very heavy training loads. As the survey settled down and supervisors were able to devote more attention to response rate, the non-response was reduced to more acceptable levels.

the first of these is the fact that the
 the second is the fact that the
 the third is the fact that the

the fourth is the fact that the

the fifth is the fact that the
 the sixth is the fact that the
 the seventh is the fact that the
 the eighth is the fact that the
 the ninth is the fact that the
 the tenth is the fact that the
 the eleventh is the fact that the
 the twelfth is the fact that the
 the thirteenth is the fact that the
 the fourteenth is the fact that the
 the fifteenth is the fact that the
 the sixteenth is the fact that the
 the seventeenth is the fact that the
 the eighteenth is the fact that the
 the nineteenth is the fact that the
 the twentieth is the fact that the
 the twenty-first is the fact that the
 the twenty-second is the fact that the
 the twenty-third is the fact that the
 the twenty-fourth is the fact that the
 the twenty-fifth is the fact that the
 the twenty-sixth is the fact that the
 the twenty-seventh is the fact that the
 the twenty-eighth is the fact that the
 the twenty-ninth is the fact that the
 the thirtieth is the fact that the
 the thirty-first is the fact that the
 the thirty-second is the fact that the
 the thirty-third is the fact that the
 the thirty-fourth is the fact that the
 the thirty-fifth is the fact that the
 the thirty-sixth is the fact that the
 the thirty-seventh is the fact that the
 the thirty-eighth is the fact that the
 the thirty-ninth is the fact that the
 the fortieth is the fact that the
 the forty-first is the fact that the
 the forty-second is the fact that the
 the forty-third is the fact that the
 the forty-fourth is the fact that the
 the forty-fifth is the fact that the
 the forty-sixth is the fact that the
 the forty-seventh is the fact that the
 the forty-eighth is the fact that the
 the forty-ninth is the fact that the
 the fiftieth is the fact that the
 the fifty-first is the fact that the
 the fifty-second is the fact that the
 the fifty-third is the fact that the
 the fifty-fourth is the fact that the
 the fifty-fifth is the fact that the
 the fifty-sixth is the fact that the
 the fifty-seventh is the fact that the
 the fifty-eighth is the fact that the
 the fifty-ninth is the fact that the
 the sixtieth is the fact that the
 the sixty-first is the fact that the
 the sixty-second is the fact that the
 the sixty-third is the fact that the
 the sixty-fourth is the fact that the
 the sixty-fifth is the fact that the
 the sixty-sixth is the fact that the
 the sixty-seventh is the fact that the
 the sixty-eighth is the fact that the
 the sixty-ninth is the fact that the
 the seventieth is the fact that the
 the seventy-first is the fact that the
 the seventy-second is the fact that the
 the seventy-third is the fact that the
 the seventy-fourth is the fact that the
 the seventy-fifth is the fact that the
 the seventy-sixth is the fact that the
 the seventy-seventh is the fact that the
 the seventy-eighth is the fact that the
 the seventy-ninth is the fact that the
 the eightieth is the fact that the
 the eighty-first is the fact that the
 the eighty-second is the fact that the
 the eighty-third is the fact that the
 the eighty-fourth is the fact that the
 the eighty-fifth is the fact that the
 the eighty-sixth is the fact that the
 the eighty-seventh is the fact that the
 the eighty-eighth is the fact that the
 the eighty-ninth is the fact that the
 the ninetieth is the fact that the
 the ninety-first is the fact that the
 the ninety-second is the fact that the
 the ninety-third is the fact that the
 the ninety-fourth is the fact that the
 the ninety-fifth is the fact that the
 the ninety-sixth is the fact that the
 the ninety-seventh is the fact that the
 the ninety-eighth is the fact that the
 the ninety-ninth is the fact that the
 the hundredth is the fact that the

TABLE II

Response and Non-Response by Type of Surveys

Name of Survey	Year	Collection Method	% Response	% Refusal	% Other
Family Expenditure	1967	Personal	69.0	20.6	10.4
"	1969	"	69.2	16.1	14.7
"	1971	"	81.2	8.3	10.5
"	1972	"	77.5	10.2	12.3
"	1976	"	75.3	14.0	10.7
Family Food Expenditure	1969	Self-enu-meration	65.3	11.1	23.6
	1974	"	67.7	10.3	22.0
	1976	"	71.9	8.5	19.6
Nutrition Canada	1972	Personal	50.0	3.0	22.0
				25.0*	
Consumer Finance	1972	"	80.0	7.9	12.1
	1973	"	89.5	8.7	1.8
	1974	"	81.3	8.7	10.0
	1975	"	87.8	8.6	4.0
	1976	"	80.1	8.3	11.6
Canadian Travel	1974	Personal	85.0	4.5	10.5
Housing Survey	1974	Personal	87.0	3.8	9.2
LFS	1971-75	"	92.7	1.7	5.6
RLFS	1975-77	"	92.2	1.7	6.1

*Refusal to come to clinic.

It can be seen from Table II that the non-response rates are very much much higher for the other household surveys than the corresponding rates for the Canadian Labour Force Survey. This is largely due to the "one shot" nature of the surveys, the sensitivity of the subject matter and to the length and complexity of the questionnaire. Other observations may be made from the table.

- (i) The refusal rates for the same type of surveys do not normally vary greatly from year to year; an exception would be the FE in 1971 when it was carried out as a supplement to the LFS.
- (ii) There is a variation in response between the types of surveys for the same method of data collection.
- (iii) Surveys conducted as supplements to the LFS have a much higher response, for example FE in 1971 and Consumer Finance in 1973 and 1975. However, the refusal rates remained unchanged.

- (iv) Self-enumeration in conjunction with a particular subject matter results in the lowest response rate (between 60% and 70% versus 70% to 90% by personal interview).

In discussing approaches to minimizing non-response, one can distinguish between two types. One type, such as "No one home" or "Temporarily absent" is in fact a "no contact" problem and is primarily operationally oriented. The other type is the true non-response problems, where contact has been made with the respondent but an acceptable response is not obtained.

The "no contact" type of problem is, of course, usually attacked with operational solutions. In a telephone or personal interview, the time and patterns of calling on the respondent are important. The size of assignment and the time allotted to data collection must be adequate. In a mail survey, ensuring correct addresses on the mailing list, efficient follow-up procedures, convenient materials are all essential. The size of non-response due to "No one at home" or "Temporarily absent" provides an important indication of the operational problems.

While the operational, logistical approach is very important, it is highly dependent on the conditions of each individual survey and is generally well controlled by a competent data collection agency. The paper will, therefore, concentrate mostly on the problems of refusals. It should be conceded at the outset that refusal rates are not always as straightforward as one might expect. An interviewer may prefer to record a refusal as a "No one at home" or a respondent may simply not answer the door as a means of refusing and thus being recorded as "No one at home". In a mail survey one is not always certain that the respondent received the questionnaire and if he has received it whether he simply neglected to mail it. Thus, the distinction between the true non-response and other causes is not easily established. In any event, regardless of how non-response is recorded, the problem seems to be to motivate the appropriate respondent to produce a valid response.

1. The first part of the paper is devoted to the study of the properties of the function $f(x)$ defined by the equation

$$f(x) = \int_0^x \frac{1}{1+t^2} dt.$$

It is shown that the function $f(x)$ is increasing and concave down on the interval $(-\infty, \infty)$. Moreover, the function $f(x)$ has a horizontal asymptote at $y = \frac{\pi}{2}$ as $x \rightarrow \pm\infty$.

2. In the second part of the paper, we consider the function $g(x)$ defined by the equation

$$g(x) = \int_0^x \frac{t}{1+t^2} dt.$$

It is shown that the function $g(x)$ is an odd function and that it has a horizontal asymptote at $y = \frac{\pi}{2}$ as $x \rightarrow \pm\infty$.

3. In the third part of the paper, we consider the function $h(x)$ defined by the equation

$$h(x) = \int_0^x \frac{1}{1+t^4} dt.$$

It is shown that the function $h(x)$ is an even function and that it has a horizontal asymptote at $y = \frac{\pi}{2}$ as $x \rightarrow \pm\infty$. Moreover, the function $h(x)$ is increasing and concave down on the interval $(-\infty, \infty)$.

4. In the fourth part of the paper, we consider the function $k(x)$ defined by the equation

$$k(x) = \int_0^x \frac{t^2}{1+t^4} dt.$$

It is shown that the function $k(x)$ is an odd function and that it has a horizontal asymptote at $y = \frac{\pi}{2}$ as $x \rightarrow \pm\infty$.

With respect to motivation, the respondent may be seen as being neutral towards the survey and only the influences which may motivate him either to respond or not to respond need be considered. Such factors as difficulty in understanding questions, use of respondent time, privacy, indifference, difficulties in recalling information, embarrassing or personal questions are all examples of motivation not to respond. On the other hand, examples of motivation to respond are an interest in the survey, willingness to help out, duty, understanding of the importance of survey results, etc.

The problem becomes, how to accentuate the positive motivation and reduce the negative motivation until the balance swings in favour of response. The key element is the respondent and anything which affects his ability and motivation to respond must be of interest and concern to a survey designer.

3. MOTIVATION OF RESPONDENT

It is a matter of common experience that in every day life when one asks a question, one normally receives an answer. What then, motivates a respondent not to respond to a survey? Invasion of privacy, respondent burden and general hostility or distrust are the three major reasons. In dealing with these problems, it is important to consider them from the respondents' point of view and not from some preconceived notions on the part of the sponsor or survey designer. Certain questionnaires may be perceived by the respondents as burdensome if they do not understand why they are being asked or how the survey is related to them. Under different conditions, however, the same questions may be perceived as very interesting and the respondent is motivated to participate in the survey. The general means by which respondent motivation can be sought may be divided into two parts, public relations and respondent relations.

Public relations are activities directed to the general public, and can take many forms. One long range objective is to create a climate

in the general public which will tend to motivate it in the direction of co-operation with surveys. The image of the survey organization as perceived by the respondent can be an important factor in his motivation to respond. The extent and manner of data dissemination can be used to impress upon the public the importance of the role of the statistical agency. The statistical agency must be alert to possibilities for favourable comment via the media and also to the necessity for timely and appropriate handling of criticism. A public relations approach, in addition to maintaining a favourable image, may be used to publicize specific activities and request co-operation in carrying them out. Ill conceived publicity may, however, fail to impress the respondents of the survey and may spark attacks on the survey, which may in fact have a negative effect. As a general rule, publicity campaigns are most suitable for population census operations where everyone is affected, and there is a need to motivate the population as a whole. Where the ratio of sample size to population is small, it is usually better to avoid direct publicity and to concentrate on respondent relations.

4. RESPONDENT RELATIONS

A working definition of respondent relations might be that it comprises any action directed toward the individual respondent which may affect his attitude and motivation with regard to the survey. As was previously stated, the problems of most concern are invasion of privacy, respondent burden, and general hostility or distrust. In regard to hostility or distrust, public relations and the agencies' image have already been mentioned. Identification of the interviewer and sponsor or agency conducting the survey is very important. Unwelcome callers have been known to use the pretext of a survey to establish contact or gain entry. It is important that the interviewer present official identification and that the survey materials convey an appearance of being official.

Introductory letters, examples of the uses of the data, and brochures describing the objectives and authority for the survey, are often excellent means of avoiding hostility and distrust.

Invasion of privacy is related to the content of the questionnaire although the reaction of different respondents is quite variable. Many procedures exist for minimizing the effect on the respondent and the specific procedure should be tailored to the specific situation. In some cases, it may be best to allow the respondent to reply in a completely anonymous fashion. This can be accomplished by self-enumeration with no identification whatsoever on the questionnaire. Quite often, though, it is essential to have some area code or sample designation for weighting and estimation purposes and in that event care must be taken that the respondent does not perceive this as a means of identifying his replies.

In addition to the assurance of privacy, some form of compensating the respondents for their time and effort have been practiced. A small gift is often used to encourage the respondent to co-operate.

5. RESPONSE INCENTIVE STUDY

A Response Incentive Experiment was carried out in the Labour Force Survey during 1975 and 1976 in order to determine the effectiveness of a "response incentive" on improving respondent relations and interviewer performance.

The response incentive used in the experiment was the "Canada Handbook", an annual Statistics Canada publication which is an attractive presentation in textual and pictorial format of economic, social and cultural developments in Canada. In order to measure the effectiveness of the "Canada Handbook" as a response incentive, interviewers across Canada were divided into two groups to form an experimental subsample and a control sample.

Interviewers in the experimental group distributed one copy of the "Canada Handbook" to each household entering the survey for the first time. They explained to the respondents that this publication was being given to them in appreciation of their co-operation during the

survey. When confronted with a refusal, interviewers still offered a copy of the "Canada Handbook" to the respondent with a brief explanation that these books were being distributed to every selected household. The control subsample, on the other hand, received no response incentive other than the usual introductory letter and presentation of the interviewer's identification card. Some conclusions drawn from the study are mentioned on the following page.

TABLE III

Refusal Rates (%) in the Control and Experimental Subsamples According to the Number of Months a Household Was in the Survey

No. of Months in the Survey	Self-Representing Units (SRU)		Non-Self Representing Units (NSRU)	
	Control	Experimental	Control	Experimental
1	2.07	2.01	1.12	1.01
2	1.55	1.48	1.11	1.01
3	1.75	1.33	1.48	1.37
4	2.05	1.76	1.73	1.47
5	2.43	1.81	1.88	1.63
6	2.47	2.05	1.97	1.72

Households which received the "Canada Handbook" had lower refusal rates than households which did not receive it (Table III). The observed differences in the refusal rates between the control and experimental groups were not significant with the exception of groups 4 and 5. But since all 6 groups had lower refusal rates in experimental group in both SRU and NSRU, the difference would be found significant by the sign test.

Qualitative information on the effectiveness of response incentives in terms of interviewers' acceptance and attitudes was obtained from an evaluation questionnaire which was completed by all the interviewers in the control and experimental subsamples at the conclusion of the

experiment. The results of this questionnaire showed that both interviewers and respondents reacted very favourably to the "Canada Handbook". The majority of interviewers indicated that they feel that response incentives are necessary and that they are helpful in establishing a good rapport with the respondent. This is probably more important than a slight decrease in refusal rates. The Response Incentives Experiment has also shown that there is a real need to provide the respondent with more information on the purposes of the survey data.

6. INTERVIEWER EFFECT ON RESPONDENT

In many household surveys, the interviewer is still the main, and often the only contact the respondent has with the survey. Therefore, the success of a survey is very much dependent upon the way the interviewer presents the survey to the respondent.

Until recently, most interviewers had not been properly trained in the public relations or motivational aspects of their job. Experience shows that respondents are persuaded to co-operate in a survey as much by the sincerity and capability of the interviewer as they are by the validity of the arguments he or she presents.

It is clear that interviewers must be properly motivated, and in fact continuously motivated, or they cannot possibly hope to motivate their respondents. Further, if interviewers are hired for their job on the basis of intelligence, general competence, availability and good health only, and there is no consideration given to their ability and willingness to get along with people, we can be reasonably well assured that the new interviewer will have a continuing history of non-response problems. This implies a thoroughly thought out and well-implemented program to train and retrain the interviewer in the skills of motivating respondents and making certain that the interviewer remain convinced of the importance of the survey and the validity of the arguments he or she delivers to respondents. To provide accurate information, however, the respondent

must know or have access to the information being sought. The choice of respondent is, therefore, important. In some cases, only the respondent himself can provide the required data. This is especially true if the subject is of such a personal or sensitive nature that proxy responses may introduce biases. In other cases, proxy response by any responsible member of the household is quite acceptable or inevitable.

7. CHOICE OF RESPONDENT

To obtain data from each individual (non-proxy) may be impossible within the time constraints of the survey or very expensive. Furthermore, non-proxy procedures may result in a large non-response, thus contributing to an increase in both sampling variability and non-response bias. Proxy reporting on the other hand may result in higher response rates and a possible reduction in respondent burden as well as in the cost of data collection.

A special study on non-proxy reporting using a Methods Test Panel (MTP) was conducted in 1974 in Canada. The LFS questionnaire was used and a comparison was made between the MTP study and the Canadian Labour Force Survey in relation to non-response. Non-response rates and proxy rates are described in Table IV.

TABLE IV

Non-Response Rates and Proxy Rates in MTP and LFS (same assignment area) as Percent. Average over Period of Study; Broken Down by Type of Non-Response and Type of Area (May to Nov. 1974)

	Self-Representing Units		Non-Self Representing Units		Both	
	MTP %	LFS %	MTP %	LFS %	MTP %	LFS %
Total Non-Response	11.6	6.9	9.3	3.9	11.0	5.9
Temporarily absent						
& No one Home	6.6	4.4	4.7	3.0	6.1	4.0
Refusal	4.4	2.0	4.5	0.7	4.4	1.5
Other	0.6	0.5	0.2	0.2	0.5	0.4
Avg. Proxy Rate	23.4	52.3	7.1	44.8	19.3	49.7

The experiment demonstrated that it was possible to reduce the proxy rate in the LFS from 52.3% in SRU's, 44.8% in NSRU and 49.7% overall to 23.4%, 7.1% and 19.3% respectively in MTP. However, the non-proxy requirement resulted in very much higher non-response rates than in the regular LFS. Both total non-response and refusal rates doubled in most areas and more than doubled in some areas.

The acceptance of suitable proxy respondents is, therefore, an effective means of lowering non-response and has the added advantage of usually providing a reduction in cost as well. But from the point of view of the respondent, he may still regard a survey as a burden in relation to his time and effort.

8. RESPONDENT BURDEN

Respondent burden has many facets such as length of questionnaire or interview, amount of detail required, sensitivity of the subject, etc.

In order to learn about the possible effect of respondent burden upon non-response, another study using the Methods Test Panel was conducted in order to measure the impact on the Labour Force Survey response rates of the supplementary questionnaires when the latter were attached to the LF sample. Respondent burden is evaluated in terms of changes in non-response rates in relation to the following type of questionnaires used with LFS: Consumer Finance (CF), Household Facilities and Equipment (HFE) and Job Mobility. The questionnaire for the Consumer Finance Survey dealt with such items as income and assets. Household Facilities and Equipment contains information on type of dwelling, heating equipment, home facilities such as refrigerators, radios, televisions, etc. The Job Mobility Survey dealt with possibility of mobility according to the availability of occupations.

A summary of the results of the experiments is shown in the following tables. The response burden is analysed by studying the trends over a 4-month period on the non-response and refusal rates.

4. The first of these is the fact that the
theoretical model of the firm is based on
the assumption that the firm is a profit
maximizing entity. This is a simplification
of reality, but it is a useful one. It
allows us to focus on the core of the
firm's decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

5. The second of these is the fact that
the firm is assumed to be a single
entity. This is also a simplification of
reality, but it is a useful one. It
allows us to focus on the firm's
internal decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

6. The third of these is the fact that
the firm is assumed to be a profit
maximizing entity. This is a simplification
of reality, but it is a useful one. It
allows us to focus on the core of the
firm's decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

7. The fourth of these is the fact that
the firm is assumed to be a single
entity. This is also a simplification of
reality, but it is a useful one. It
allows us to focus on the firm's
internal decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

8. The fifth of these is the fact that
the firm is assumed to be a profit
maximizing entity. This is a simplification
of reality, but it is a useful one. It
allows us to focus on the core of the
firm's decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

9. The sixth of these is the fact that
the firm is assumed to be a single
entity. This is also a simplification of
reality, but it is a useful one. It
allows us to focus on the firm's
internal decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

10. The seventh of these is the fact that
the firm is assumed to be a profit
maximizing entity. This is a simplification
of reality, but it is a useful one. It
allows us to focus on the core of the
firm's decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

11. The eighth of these is the fact that
the firm is assumed to be a single
entity. This is also a simplification of
reality, but it is a useful one. It
allows us to focus on the firm's
internal decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

12. The ninth of these is the fact that
the firm is assumed to be a profit
maximizing entity. This is a simplification
of reality, but it is a useful one. It
allows us to focus on the core of the
firm's decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

13. The tenth of these is the fact that
the firm is assumed to be a single
entity. This is also a simplification of
reality, but it is a useful one. It
allows us to focus on the firm's
internal decision-making process, and to
ignore the many other factors that
influence its behavior. This is a good
starting point, and it is one that we
can build upon as we develop our
model.

TABLE V(A)

LFS Non-Response Rates by Type of Burden

	Type of Burden	Month			
		April	May	June	July ¹
Total Non-Response Rate (%)	CF & HFE same month	9.35	9.99	14.38	27.46
	CF & HFE in 2 months (Consecutively)	10.58	11.28	12.13	17.19
Refusal Rate (%)	CF & HFE same month	3.25	4.41	4.31	5.03
	CF & HFE in 2 months	1.76	3.22	3.47	4.07
Avg. sample size per type of burden hhllds.		750	750	600	450

¹In July, another supplementary, Job Mobility Survey, was asked.

According to Table V(A), the tendency towards higher non-response and refusal rates occurs among households subjected to two long questionnaires in the same month as compared with households subjected to two long questionnaires one month apart. This tendency is more evident in later months among households exposed to the two types of questionnaires.

In July, two months after the main test, the rotation groups 2, 3 and 4 which had been already subjected to CF and HEF questionnaires, and remained still in the sample were given the Job Mobility questionnaire. The purpose of this test was to see the respondent reaction to yet another long supplementary questionnaire. The results are in Table V(B).

TABLE V(B)

LFS Response Rates by the Type of Burden

Rotation Group	Average Sample Size	Job Mobility in July with	
		CF & HFE Same Month	CF & HFE 2 Months
2	216	86.96	86.99
3	211	73.78	82.74
4	230	73.44	89.17
	657 (Total)	78.04 (Avg.)	86.39 (Avg.)

The source of burden (CF and HFE in the same months) resulted in significantly lower response rates compared with that part of the MTP sample receiving questionnaires in two successive months.

In addition to the findings for the LFS, some other results were noted in relation to the supplementary surveys themselves. Thus, (a) slightly lower response rates for CF supplementary resulted when it was combined with HFE (HFE and CF in the same months) than without HFE. It should be noted that CF is asked after HFE so that there is a possible effect due to the additional burden when the respondent has to answer both supplementaries in the same month; (b) those households which were burdened with two surveys in April were much less receptive to the July supplementary than those which were burdened with two surveys separately in April and May.

In April 1976, the Survey of Consumer Finance was carried out as a supplement to the LFS. The following table shows refusal rates by rotation group for a number of months. Each rotation group (RN) contains a panel of households which remain in the LFS for six consecutive months and then it is replaced by a new panel of households.

TABLE VI
LFS Refusal Rates by Rotation Groups (R.N.)
July 1975 to December 1976

Months	R.N. 1	R.N. 2	R.N. 3	R.N. 4	R.N. 5	R.N. 6
July '75	1.4	1.9	1.8	1.8	1.6	1.5
August	1.3	1.4	1.7	1.9	1.6	1.6
Sept.	1.5	1.4	1.6	1.8	1.7	1.6
Oct.	1.4	1.2	1.2	1.7	1.8	1.7
Nov.	1.5	1.4	1.2	1.5	1.5	1.7
Dec.	1.5	1.4	1.3	1.5	1.2	1.5
Jan. '76	1.7	1.5	1.3	1.7	1.3	1.4
Feb.	1.4	1.6	1.4	2.0	1.4	1.6
March	1.4	1.5	1.7	2.1	1.5	1.8
April *	1.8	1.8	2.0	2.2	1.7	2.1
May	2.6	2.6	2.9	2.6	1.8	2.5
June	2.4	2.4	2.8	3.1	1.5	1.1
July '76	1.3	2.4	3.0	2.9	1.4	1.2
August	1.1	1.4	3.0	3.0	1.5	1.3
Sept.	1.4	1.3	1.2	3.1	1.7	1.4
Oct.	1.8	1.6	1.2	1.7	1.8	1.6
Nov.	2.1	2.0	1.6	1.6	1.5	1.9
Dec.	2.1	1.9	1.6	1.6	1.3	1.5

* Consumer Finance Survey.

From the table, the following observations can be made:

- (a) Refusal rate for the households for the second month in the sample is smaller than that of the first month for all rotation groups, except for RN3 introduced in March, irrespective of its month of introduction.
- (b) Refusal rates increased for all rotation groups in April when the CF survey was introduced.
- (c) From April on, refusal rates increase with the length of time a particular rotation group remains in the sample. The rotation group which rotates out in April, has the smallest increase in the refusal rate. Evidently, the respondent is prepared to accept this additional burden provided it is for the last time.
- (d) Consumer Finance Survey seems to exert an additional burden on the LFS respondent which manifests itself in the increased refusal rates not only for one month at the time of the CF survey, but remains until a particular group rotates out of the CF survey.

9. INTERVIEWER PROFILE

In specialized and sensitive surveys, specialized interviewers are quite often required. In the majority of cases, however, the average interviewer with proper training can obtain very good results.

In this section, an attempt is made to present a profile of an interviewer employed for the LFS and other associated surveys. In addition, possible relationships between various characteristics and non-response rates are discussed.

Based on the record of 2,800 interviewers, the following information about the interviewers has been obtained.

- (a) The vast majority (close to 90%) of interviewers are females and only 13% are single.
- (b) Over 90% of the interviewers have at least 10 years of education; 27% have completed a post-secondary education, while 10% have a university degree.
- (c) Average age of interviewers is 45 years with the ages ranging from 18 years to 77 years.
- (d) Many interviewers (60 to 70%) have had previous experience in contacting the public.
- (e) 36% have been on the job for less than 1 year, 32% have worked for 1 to 2 years, while the remainder have worked from 3 to 10 years.
- (f) The turnover rate is about 30% per year.

Knowledge of interviewers' personal characteristics serves as a guide not only in hiring suitable replacements but also in determining the extent and type of training that may be required. It seems reasonable to assume that personal characteristics, in conjunction with such factors as the size of assignment, experience and turnover may determine the interviewers' performance. A simple regression analysis was carried out between non-response rates and interviewers' assignment. This was followed by a multiple linear regression with non-response as the dependent variable and size of assignment, years of experience, interviewers' age in years and years of schooling as four independent variables. The study was carried out separately in SRU and NSRU areas for some 500 interviewers in each type of area.

The following results were obtained:

- (a) Little or no significant regression coefficients were observed between non-response rates or any of its components and any of the four independent variables. Similarly, the multiple correlation coefficients were not significant.

- (b) A small but negative correlation between total non-response and size of assignment was observed for individual assignments but a large negative correlation was obtained when the assignment sizes were grouped into classes of ten household intervals. The decreasing trend in the non-response rates, as the assignment size increases, may be observed in Table VII, particularly in SRU areas. The same holds true for refusal rates in SRU areas, but the trend is reversed in NSRU areas as a result of relatively high refusal rates in 70-79 and 80-89 assignment size ranges.
- (c) It is interesting to note that the lowest refusal rates in SRU areas are associated with assignment sizes ranging between 50-59 households which happens to be the most frequent assignment. The same phenomenon does not seem to hold in NSRU areas (see Table VII).

One factor should be noted in connection with the above data, i.e. the role of supervisory staff. Each month supervisory staff carries out purposive allocation of interviewers' assignment size based on their empirical evaluation of interviewer capacity. This makes a statistical analysis of the data rather difficult.

It seems, however, that as far as the LFS interviewers are concerned, a balance has been reached between their characteristics and performance resulting in low non-response rates. Further marginal improvements could perhaps be achieved through better training and working conditions.

TABLE VII

Non-Response Rates and Refusal Rates by Size of Assignment
and Type of Area (Jan. 1977 LFS Data)

Size of Assignment	Self-Representing Units				Non-Self Representing Units			
	A	B	Average Rates		A	B	Average Rates	
			C (%)	D (%)			C (%)	D (%)
10 - 19	4	62	8.06	4.84	25	316	8.86	1.24
20 - 29	8	173	7.51	2.31	52	1,100	6.91	0.64
30 - 39	26	878	7.06	2.05	101	3,050	6.89	1.25
40 - 49	62	2,674	6.62	2.06	131	5,056	5.76	1.05
50 - 59	134	6,897	5.70	1.94	153	7,109	6.93	1.42
60 - 69	129	7,874	5.63	1.60	83	4,540	5.57	0.97
70 - 79	81	5,652	5.36	1.73	30	1,920	6.56	1.82
80 - 89	37	2,860	5.42	1.99	10	708	4.38	1.98
90 - 99	6	519	6.74	2.50	1	69	5.80	1.45
100 +	6	586	1.71	0.51	0	0	--	--
Total	493	28,175*	5.66	1.81	585	23,368*	6.33	1.24

* assignments of 1 to 9 households comprising 23 households in SRU and 29 households in NSRU omitted.

Legend: A = No. of Interviewers
B = No. of Households
C = Non-Response
D = Refusal

10. NON-RESPONSE PROFILE

Positive motivation of respondents is essential in maintaining non-response at an acceptable level. It may become considerably easier to change negative motivation to positive motivation if one can identify potential non-respondents before they become effective non-respondents. It is of interest therefore, to determine at least approximately their characteristics and background. To achieve this, a profile of households which responded at least once in a six month period in the Labour Force

Survey was studied. The differences in the characteristics according to the number of times households failed to respond may reveal characteristics which could be linked to non-response in such a way as to suggest procedures to reduce non-response.

The data presented in the following table is derived from the Six Month Data File which links together relevant survey data collected for a household over the six month period during which a household is in the sample. In Table VIII, some data on the patterns of non-response and some characteristics of non-respondents are presented.

TABLE VIII

Household Size and Unemployment Rate by Frequency and Type of Non-Response

Frequency of Non-Resp	Temp. Absent		No One Home		Refusal		Responding All Six Months	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
1	2.84	5.88	2.55	7.21	2.85	6.82		
2	2.37	7.34	2.31	10.78	3.14	6.90		
3	2.16	9.78	2.42	9.50	3.39	10.37		
4-5	<u>1.65</u>	<u>13.64</u>	<u>1.92</u>	<u>6.06</u>	<u>3.46</u>	<u>9.34</u>		
Average	2.62	6.46	2.44	7.92	3.23	8.39	3.39	7.92

(a) Average Household Size (b) Unemployment Rate

The following observations may be seen from the above table:

- i) Average household size, where contact was not made with the household, i.e. "No one at home" and "Temporarily absent" was smaller than those where contact was made, i.e. responding households and refusals.
- ii) In terms of average unemployment rate, the "No one at home" households are very much like the respondent households, while refusal households have higher, and "Temporarily absent" households have lower unemployment rate as compared to the responding households.

- iii) For both "Temporarily absent" and "No one at home" households, there is a decreasing trend noticed in the size of households as the frequency of non-response increases. For example, among "Temporarily absent", those households who did not respond for one month out of 6 monthly surveys (responded in 5 surveys) the household size is 2.84 and this size reduces to 1.65 for households which have not responded to 4 or 5 surveys out of 6. This implies that the hard core among "Temporarily absent" and "No one at home" are households comprised of only one or two persons. The trend for refusal households is, however, just the reverse.
- iv) The unemployment rate for "Temporarily absent" households shows an increasing trend as the frequency of non-response increases. This shows that long term "Temporarily absent" are likely to have higher unemployment rates.

A further study of the Six Month file demonstrated that the educational level of a potential non-respondent is about the same as that of the respondent.

11. CONCLUSIONS

There is evidence from many surveys and survey taking organizations, that the public is becoming more concerned about the frequency of surveys and their implications in terms of personal privacy. This attitude on the part of the public is reflected in an increased reluctance to co-operate in surveys. This is a serious matter since the respondent is the key element in data collection.

It has been noted in the paper that well established and controlled surveys such as the Canadian Labour Force Survey, even under the current conditions, are able to maintain an acceptable level of non-response. The fact that non-response is not at an alarming level is due to the efforts to maintain satisfactory respondent relations and a good reputation in terms of confidentiality, productivity and use

of the data. However, even in the Labour Force Survey, respondent burden as manifested in higher refusal rates can be observed if the supplementary surveys associated with the LFS are lengthy and of a sensitive nature.

Constant attention must be paid to the impact of sensitive questions on the respondents and to the extent the reporting burden is imposed. Furthermore, pilot testing of questions and of public reaction should be a rule rather than an exception in order to evaluate the acceptability of the questions and to determine the most suitable methodology of data collection. Training of interviewers, not only in the mechanics of locating respondents, administering questionnaires and recording replies, but also in the art of respondent relations, can be a major factor in obtaining complete and reliable data.

Finally, a strict control of the use of sample frames should be maintained to minimize overlapping of sample households in many surveys at the same time and to ensure that the households will not be sampled again before a specified period of time.

RESUME

La non-réponse est présente dans toutes les enquêtes mais son importance dépend du type de l'enquête, de l'habilité des interviewers à mener une entrevue et de la volonté des personnes interviewées à répondre aux questions de l'enquête. Cet article présente une discussion de la non-réponse en relation avec plusieurs enquêtes sur les ménages et, particulièrement, du comportement des taux de non-réponse dans le temps dans une enquête périodique telle que l'Enquête sur la Population Active du Canada.

Un profil du type d'interviewers employés par Statistique Canada montre que la corrélation entre la non-réponse et plusieurs caractéristiques des interviewers n'est pas significative. Ce sont les répondants eux-mêmes et la motivation dont ils font preuve qui sont les éléments clés d'une entrevue et par conséquent, des relations avec le répondant.

Cet article discute les résultats de plusieurs études qui ont été faites pour mieux comprendre les effets du fardeau qui est imposé au répondant, du choix du répondant, et de certaines formes d'incitation à la réponse, afin de mieux connaître les caractéristiques du non-répondant.

REFERENCES

- [1] Platek, R. and Singh, M.P. (1976), "Methodology of the Canadian Labour Force Survey", Statistics Canada, Household Surveys Development Division.
- [2] Singh, M.P. and Tessier, R. (1976), "Feasibility Test for Complete Non-Proxy Procedure Methodology and Analysis", Statistics Canada, Household Surveys Development Division.
- [3] Household Surveys Development Division (1975-76), Test on Respondent Burden.
- [4] Platek, R., Singh, M.P. and Tremblay, V. (1977), "Adjustments for Non-Response in Surveys", presented at the Symposium on Survey Sampling at the University of North Carolina.
- [5] Gower, A. (1977), "The Response Incentives Experiment in the Canadian Labour Force Survey", Statistics Canada, Household Surveys Development Division.

SURVEY OF CANADIAN RESIDENTS RETURNING BY LAND

J.H. Gough and P.D. Ghangurde
Household Surveys Development Division

The distribution of questionnaires to Canadian residents returning by land from the U.S. has been substantially modified, in an effort to improve sample yield at minimal additional cost. For each border crossing ("port") involved, a systematic sample of multi-day distribution stints has been selected. The sample selection method is described, the constraints which determined it are discussed, and some preliminary data on the method's effectiveness are presented.

1. INTRODUCTION

The target population of the system of international travel surveys conducted by Statistics Canada's International Travel Section (ITS) includes all persons entering Canada and clearing Canadian Customs. One segment of this population consists of Canadian residents returning by land, whatever the mode of transportation. The principal modes at the ports considered in this article are auto, bus and truck. A distinction is made between those parties which have spent at least one night outside Canada on the trip in question (known as "overnight" traffic), and those who have not ("same day"). Only the former are considered here; thus the target population of this particular survey is "(returning Canadian) residents, land, overnight". The overwhelming majority of these are returning from the U.S.

Until recently, most Canadian-U.S. land border ports have supposedly been following a one-day-in four questionnaire distribution scheme for overnight resident traffic. In theory, this consists of Canada Customs primary 'line officers' handing out a mail-back questionnaire to every travel party belonging to the target population on every fourth day throughout

the year, the days being chosen as a systematic random sample. If this scheme were reasonably well carried out, the resulting sample of questionnaires would be quite sufficient in size and ought not to be much affected by cyclical traffic patterns. Unfortunately, in spite of intensive liaison efforts, the scheme has been so irregularly followed that the sample is quite inadequate for the purposes for which it was intended. Since this distribution is only a minor part of the duties of Customs officers, over whom Statistics Canada has no authority, the method was difficult to control as regards either the timing or the quantity of distribution. In any event, response rates in the area of 20% must be taken into consideration.

As part of a continuing attempt to improve the ongoing international travel surveys at minimal cost, a new approach to field distribution has been developed for this survey. This consists of choosing two periods of several days each for continuous distribution, in each calendar quarter. Using estimates of traffic volume for the period, a trial of this method was conducted at the Ambassador Bridge at Windsor in the second quarter of 1976, the periods being designed to be five days long. As the results were reasonably encouraging, further trials were conducted at several ports during the remainder of the year, and eventually it was decided to use this method as an interim measure at all ports previously on the one-day-in-four plan. Hence, by the second quarter of 1977, 75 ports covering well over 98% of the eligible traffic will adopt a system of distributing questionnaires during two stints per quarter, each stint being expected to last from six to ten days, and successive stints beginning about 6½ weeks apart. The survey is expected to require distribution of about 350,000 questionnaires and to produce up to 70,000 responses, which can be processed under current operating procedures in ITS. Continuous liaison efforts will be employed to ensure the smooth operation of the survey and continued co-operation with Customs officers. It must, however, be stressed that the introduction of this method remains an interim measure only, an attempt to improve the quality of the sample as much as possible under the present conditions. This

method is in no way intended to replace or supersede development of a properly controlled survey, which has been proposed and is awaiting approval and funding.

2. USE OF CONTINUOUS STINTS

The principal reason for the adoption of stints consisting of several consecutive days was that the old method, being an on-and-off arrangement, required considerable monitoring to avoid having many days missed or incorrectly sampled. With over a hundred ports scattered across the country, and over ninety sample days per year, it was obviously impossible for ITS's liaison staff to check before each sample day. At the same time, Customs officers, with many more pressing duties, could hardly be expected to have this survey foremost in their minds every few days. It was felt that distribution on a continuous basis several times per quarter would give nearly as good a sample on a theoretical basis (the only potential loss being due to seasonal effects), and would make it possible to control distribution more closely by estimating traffic in advance, providing a suitable number of questionnaires and then simply indicating the starting date for each stint. Customs officers would then have only to begin at the right time and continue distribution until their supply of questionnaires was exhausted. With only two starts per quarter at each port instead of over twenty, the task of reminding port supervisors a day or two in advance of each start would become manageable. The idea was discussed with several Customs supervisors and liaison officers, and upon favourable reaction from personnel at Windsor, it was decided to try the method at the Ambassador Bridge.

At the same time, the question of the exact form of the stints, and of how to select a full-scale sample, had to be considered. At first it was hoped that each stint could be based on a fixed number of complete days rather than an uncontrolled time period, as this would allow the use of a better estimation procedure. In fact, when the use of continuous stints was first discussed, it was felt that a two-way stratification of time

and ports could be achieved. This was felt desirable in view of known variations in traffic patterns between ports, and of the complex seasonal patterns within ports. Since most provinces have a reasonably large number of ports, a two-way stratification within provinces was attempted. Details of the motivation and of the methodology developed for this procedure are given in a technical memorandum by P.D. Ghangurde [1]. Unfortunately, the various operational constraints make it impossible to avoid serious departures from such a design. The greatest problem in attempting to implement a two-way stratification was the impossibility of controlling stint length. True fixed-length stints had to be ruled out, as this would have necessitated over-supplying all ports with questionnaires (due to unknown traffic volumes) and instituting an elaborate system of counting and either returning or re-packaging the undistributed questionnaires for later use. The best that could be done was to estimate the eligible traffic and supply that number of questionnaires to be fully distributed. Thus from the outset, stint lengths would have been only approximated, to the accuracy of estimates of traffic volumes. This alone might not have had serious effects on the estimates, but with the first Windsor trial it became evident that distribution efficiency was a far more uncontrolled variable (see below). Not only would the sample in time not conform to the design, but since two stints per quarter were selected without replacement, stints could be adjacent or "nearly" adjacent; this would cause problems of overlap in the almost inevitable event of the earlier stint of a pair running on too long.

3. FIELD TRIALS

The first trial, at Windsor's Ambassador Bridge, was done in the second quarter of 1976. Using computations similar to those shown in section 5 below, a sampling fraction of 10% was felt appropriate. Two stints were chosen (from the originally proposed 2-way sample for that quarter), the total number of questionnaires was taken to be 10% of the previous year's eligible auto traffic for the quarter, and the two stints were assigned questionnaires in proportion to the traffic in the months in which they were to take place. The questionnaires were sent to the port several weeks ahead of time along with instructions to commence continuous distributions on the designated starting date. These instructions were followed up by a telephone call as a reminder a day or two before the start of the stint.

As might be expected, the actual coverage was far from 100%, and in spite of there being approximately the expected amount of traffic, the first two stints each lasted nearly three weeks. However, the plan as a whole was favourably received by Customs, and even these first results allowed an estimate of response rates and some further analysis which was impossible under the old scheme. Thus, it was decided to continue and expand the trials. For the third quarter, Windsor Bridge was kept on the trial and Lacolle (Que.) was added. As a result of increased liaison efforts, results improved sufficiently to allow a decision to proceed with six ports (one at least in each major region) for the final quarter of 1976, and to introduce the method on a full-scale basis as early as possible in 1977, providing only that liaison and operational problems could be handled with present resources. It was felt that this could, in fact be done without major problems, and that with a suitable system of follow-up and feed back to the ports on the part of ITS liaison staff, results could be maintained at a reasonably satisfactory level. Details of the operational and liaison steps have been compiled by J.G. Bailie [2]. Results of the 1976 trials and of the first quarter of 1977 are discussed in Section 6 below.

4. REVISED METHOD OF SAMPLE SELECTION

As the results of the first few trials were quite encouraging, particularly in terms of increased total response, it was decided to adopt the use of continuous stints for all major land ports beginning in 1977.

Of the approximately 120 land ports on the system, 75 were on the old sample plan. The remaining 45 or so very small ports have been and will remain on the 100% distribution plan (i.e. supposedly distributing questionnaires to all eligible parties). Since these ports account for a negligible fraction of the total traffic, and since it is well known that their distribution fraction is nowhere near 100%, they will make little or no difference to the approximations given below regarding sample size and questionnaire printing requirements. Because of the lack of

time for operations staff to obtain and package questionnaires, only eighteen of the busiest land ports used this system in the first quarter of 1977. For the second and subsequent quarters, all 75 ports on the old sample system have been converted. Details of the sample selection may be found in [3].

To minimize the impact of variable stint length and to conform as well as possible to the assumptions of the current methods, it is proposed that the present port stratification be maintained. All (75) ports are in the sample, but of these only 14 are "self-representing" in the sense that they are weighted separately; i.e., treated as strata of one port each. The remaining nine strata (or "factor groups" as they are called in ITS) consist of from 2 to 14 ports whose response and population counts are "pooled" before weighting. In most cases, these are the smaller ports; exceptions are the two Windsor ports taken together, and the three bridges at Niagara Falls. All other non self-representing strata have at least four ports.

It is desirable to have at least two stints per quarter in each port, both for variance estimation and for operational reasons. To avoid the potential problem of stints overlapping because of the lack of control on their length, a systematic sample is indicated; this has the further benefit of leaving a fairly long interval between stints, which may serve to reduce the perceived workload of those responsible for the distribution.

For strata of more than one port, it is similarly desirable that each port have two stints per quarter. Further, within any such stratum, it is preferable that all stints for the stratum be determined using a single systematic sample of starting points, as this will make the flow of documents as uniform as possible in time, while ensuring representativity as far as possible, since all estimation is done at the stratum level by quarter. To achieve these ends, the k ports in a stint are randomly ordered as p_1, p_2, \dots, p_k and a systematic sample of $8k$ stints is assigned as $p_1^{(1)}, \dots, p_k^{(1)}, p_1^{(2)}, \dots, p_k^{(2)}, p_1^{(3)}, \dots, p_k^{(8)}$. Thus each port has two stints per quarter and for any given port, the sampling interval between stints is as it would have been if the port were alone in its stratum.

The only problem in designing a systematic sample of stints is to meet the constraints of coverage and the problem of quarters of different lengths. Fortunately this is not so difficult to do. First of all, only starting points of stints are being selected; from there on, questionnaire quantities and distribution efficiency determine the time period covered, and responses received may be assumed a simple random sample from the traffic for that period.

The starting point of a stint is a day selected from the 365 (or 366) days of the year. In fact, it is the beginning of that day, but days are considered as discrete units for selection purposes. The only problem is that the number of days in the year is not evenly divisible by 8 ($365/8 = 45 \frac{5}{8}$), so that using 45 as the sampling interval leaves 5 days at the end of the year with zero probability of selection if the random start is taken from the first 45 days. If the sampling is circular systematic, with start R chosen from 1 to 365 and the other stints as $\{R + 45i\} \pmod{365}$, $i = 1, \dots, 7$, all days have equal probability of selection. However, there remains about a 5% probability of not getting the required two stints per quarter. Worse is the case of a stratum of $k > 1$ ports, where it is desired to draw a single circular systematic sample of $8k$ stints such that each port has two stints per quarter and within any port the sampling interval is as if $k = 1$. The worst case is B.C. with $k = 12$. If $365/96 \doteq 3.8$ is truncated to 3, the entire sample covers only about 10 months of the year, which is out of the question.

The problem can easily be solved by varying the sampling interval in circular systematic sampling according to a predetermined pattern. Then the truncation after selection is needed only in the sense that a stint will always start at the beginning of the day within which the starting point falls (see Appendix). It is easy to see that the selection of all starting days with equal probability is unaffected by this change.

In a few cases, however, (probability about 1%), the random start is such that the first quarter (with 90 days) has only one stint and either the third or the fourth (with 92 days) gets three. In such a case, the extra stint falls on the last day of the quarter. Thus, in the latter case, a start is moved from 31 December to 1 January; in the former case one is moved from 30 September to 1 October and consequently another from 30 or 31 December to 1 January. This is obviously not serious, and is made even less so by the fact that, in all cases, questionnaires from any stint handed out in the following quarter must be treated as coming from the latter quarter's sample, since they represent its population. This will clearly happen very often, even with stints beginning well before the end of a quarter. However, the systematic nature of the sample starts ensures stability of such patterns throughout the year. For purposes of variance estimation using the two stints as replicates, it is proposed that the two fractional (even-numbered) stints at the ends of the quarter be treated as one replicate and the mid-quarter stint as the other.

The current estimator is obtained by assuming simple random sampling within each quarter's traffic. For operational reasons this will not be changed at the moment, though sufficient data will be available from the questionnaire sample to permit a more refined estimator to be used.

5. EFFECTIVE SAMPLE SIZE

Because of the way the questionnaire distribution actually takes place in the field, effective sample size can be only roughly estimated using several assumptions and approximations. These involve the actual population size, ability of Customs officers to cover all traffic during stints, actual response rates of travellers to this type of questionnaire, and ITS's ability to process the resulting volume of forms.

The population consists of all returning travel parties of Canadian residents returning to Canada by land after an absence of one or more nights, excluding commercial trucks. In practice this means that the sampling unit is the vehicle in the case of auto traffic, although multiple parties may sometimes be detected. For bus traffic, an attempt

is made to identify the head of each party. This population in 1975 consisted of about 2,694,000 autos, roughly 270,000 bus parties (estimated from 549,000 overnight parties) and a few tens of thousands of "others", for a total of nearly 3 million. Applying the then estimated rate of increase of 8% for 1976 and a similar figure for 1977, we arrived at a total not far from 3,500,000. (In fact, the 1976 figure was well over 3,300,000, and current estimates for 1977 are in the area of 3,700,000.

Response rates estimated from the first six trial stints averaged 18.5%. If we expect national response rates to be similar, say in the range of 17-20%, a 100% distribution would yield from 600,000 to 700,000 responses. Since it is felt that ITS can (i) afford to print about 350,000 questionnaires and (ii) process at least 60,000 responses with current staff resources, a designed distribution rate of 10% is indicated as providing maximum affordable response of between 60,000 and 70,000 responses assuming that all questionnaires are distributed to eligible parties. In practice, there would be at least a small percentage of undistributed forms.

Now let us consider matters from the field point of view. A 10% rate of distribution implies, on average, 9 days' traffic per quarter or two stints of about $4\frac{1}{2}$ days each assuming 100% efficiency. Because of cyclic traffic patterns, it is desirable to have most stints cover at least a week; for liaison purposes, it is not desirable to let stints run beyond about 10 days. To keep stint length between say 6 and 10 days with $4\frac{1}{2}$ days worth of questionnaires implies a needed distribution efficiency of 45-75%, assuming of course that traffic volumes do not fluctuate too wildly. This range seems not too unreasonable. Considering that it is impossible to supply all booths in accurate proportion to their traffic flow, and that some parties will inevitably be missed, it seems unlikely that efficiency should frequently exceed 75%, giving too-short stints. On the other hand, if it slips much below 50%, the ports in the question must be reminded that one of the principal objects of this method is to reduce the length of time required for questionnaire distribution.

It may be noted that in practice, the chosen sampling fraction of 10% may be adjusted by quarter, if for example, the expected sample size for the busy third quarter would overwhelm available manpower resources. This has in fact been done for 1978; the third quarter sampling fraction will be only 8%. This flexibility extends up to the date of counting and packaging questionnaires for shipment to the ports.

6. RESULTS OF THE TRIALS

The first goal of this method was simply to increase distribution and thus response; the second, closely allied, was to make the sample more representative of the population. While it is felt that the second goal has been at least partly achieved, it is not possible to make any firm analysis in this area at the moment. However, it is possible to examine the numbers of responses and the perceived response rates in order to get some idea of the progress made toward the first goal.

Table 1 shows, for each stint of the 1976 trials, the estimated number of days' distribution, the number of questionnaires to be distributed, the number of responses and the response rate. The stint's length ranges from a low of five days (the nominal minimum) at Emerson to a high of 26 days at Lacolle. Of the 18 stints, all but four were completed within two weeks and can be considered acceptable; three of the other four were at Windsor where some improvement was noted in the later stints.

During the first six stints at Windsor, efficiency (estimated by dividing 5 by the actual number of days taken) ranged from 25% to 50%¹, the worst cases being the first two before certain modifications were made to the liaison procedure. At Lacolle, distribution efficiency during the first four stints ranged from 20% to 55%, and at the four ports added at the end of 1977, the range was from about 40% to "100%" (at Emerson). For the first quarter of 1976 only a few ports were below 40%

¹ We use 5 rather than 4 1/2 in the numerator to allow for the impossibility of distinguishing incomplete days in the denominator; there is further the problem that distribution may reasonably tail off over a day or two as some booths run out of questionnaires and do not replenish their stock from the few questionnaires remaining at other booths. However, "tails" much longer than this have occurred in some cases and are not considered in the same light.

(notably Fort Erie, which took 55 days for its first stint, though 80% of the returns come from the first ten days). In a few cases, efficiency appears to exceed 100%, due most likely to a few heavy traffic days as these stints all covered weekends. Similarly, some of the low "efficiencies" are due to low traffic; however, the overall average of 45-50% is not in dispute. The smaller ports in general may be expected to have somewhat higher rates of coverage due to having smaller crews, few lanes open, and less heavy peak traffic flow.

Response rates ranged from a low of 11.6% at Lacolle to a high of 32.9% at Lansdowne, with the majority of cases falling in the 13-22% range, only Lansdowne and St. Stephen approaching 30%. It may be that travellers through these ports are more receptive to our questionnaires. At any rate, looking at the average response rates, and assuming distribution to be more or less complete in all stints, we have a preliminary estimate of average response rates of 17-20% as used in the sample size calculations.

Table 1: Stint Length and Responses by Port: 1976 Trials

Port	Stint	Length in Days	Questionnaires Distributed	Responses	Response Rate (%)
Windsor Bridge (Ont.)	1	18	2,220	465	20.9
	2	20	2,380	367	15.4
	3	11	2,145	425	19.8
	4	17	3,135	655	20.9
	5	10	1,700	256	15.1
	6	12	1,800	250	13.9
Total		-	13,380	2,418	18.1
Lacolle (Que.)	1	14	12,550	2,221	17.7
	2	9	8,890	1,741	19.6
	3	26	2,200	358	16.3
	4	11	2,400	278	11.6
Total			26,040	4,598	17.7
St. Stephan (N.B.)	1	11	750	224	29.9
	2	11	150	44	29.3
Lansdowne (Ont.)	1	8	1,300	428	32.9
	2	8	1,100	301	27.4
Emerson (Man.)	1	8	1,300	280	21.5
	2	5	800	157	19.6
Pacific Highway (B.C.)	1	9	3,600	468	13.0
	2	13	2,800	504	18.0
1976 Totals		-	47,920	9,422	19.7

7. COMPARISON WITH THE OLD SYSTEM

It should be noted that the one-day-in-four system, if properly followed, implies distribution to approximately 25% of the quarterly traffic in each port; the only measure of response rates which can be calculated is on the assumption that this is the true distribution, but the interpretation of that measure bears as much on the known departures from this assumption as on the actual response rates.

The observed response rates from the trials, not far from 20%, may be compared with the apparent response rates (assuming full 25% distribution) for the same ports during the corresponding quarters of the previous year (auto only), shown in Table 2. These rates range from 0.5% at Windsor in the fourth quarter of 1975 to 7.7% at Emerson in the same quarter (the only value exceeding 3.3% among the 9 rates shown here). It may be added that these figures are typical of most other ports across the country, and have been for some years.

Assuming that true response rates were really in the 20% area in 1975 as well as 1976, we see that the distribution rates may range from as low as 2.5% to perhaps 35% at Emerson, with 5-15% the common range.

Thus, with the new method, we are supposedly distributing only two-fifths as many questionnaires as before, but receiving three or four times as many responses. This is the most convincing evidence of the effect of the scheme on Customs' performance.

Table 2: Apparent Response Rates Under Old System: 1975 (Auto only)

Port	Quarter	25% of Eligible Traffic	Responses	Apparent Response Rate (%)
Windsor (Bridge & Tunnel) (Ont.)	2	18,326	178	1.0
	3	19,596	353	1.8
	4	14,794	74	0.5
	Total	52,716	605	1.1
Lacolle (Que.)	3	40,406	1,331	3.3
	4	9,255	171	1.8
	Total	49,661	1,502	3.0
St. Stephen(N.B.)	4	2,383	31	1.3
Lansdowne(Ont.)	4	4,993	92	1.8
Emerson (Man.)	4	5,011	388	7.7
Pacific Hwy(B.C.)	4	18,029	144	0.8
	1975 Totals	132,793	2,762	2.1

8. IMPLEMENTATION IN EARLY 1977

Finally, in Table 3, we may examine a little further the results of implementing continuous stints at a larger number of ports during the first quarter of 1977. This gives the same general picture as Table 1.

In the case of Lacolle, it is known that up to a third of the questionnaires were not distributed during the second stint; it is suspected that something similar happened to the first stint as well. The extreme length of the Fort Erie stints has already been mentioned. With these two exceptions, both response rates and distribution efficiencies are generally in line with the results of the trials. Partial data available from the 75 ports for the second quarter of 1977 appear to support these conclusions.

Table 3: Stint Length and Responses by Port: First Quarter, 1977

Port	Stint	Length in Days	Questionnaires Distributed	Responses	Response Rate (%)
St. Stephen (N.B.)	1	13	250	37	14.8
	2	4	250	47	18.8
Woodstock Rd. (N.B.)	1	13	570	95	16.7
	2	3	285	47	16.5
Phillipsburg (Que.)	1	3	630	137	21.7
	2	13	1,040	253	24.3
Lacolle (Que.)	1	4	1,250	103	8.2
	2	16	2,450	186	7.6
Rock Island (Que.)	1	7	200	17	8.5
	2	9	400	85	21.3
Fort Erie (Ont.)	1	55+	2,300	357	15.5
	2	25	2,200	387	17.6
Lansdowne (Ont.)	1	13	800	179	22.4
	2	6	1,900	626	32.9
Niagara Rainbow (Ont.)	1	13	260	22	8.5
	2	17	260	80	30.8
Niagara Queenston (Ont.)	1	16	490	93	19.0
	2	14	820	187	22.8
Niagara Whirlpool (Ont.)	1	10	140	16	11.4
	2	15	400	27	6.8
Sarnia (Ont.)	1	7	470	136	28.9
	2	15	940	258	27.4
Sault Ste. Marie (Ont.)	1	10	170	27	15.9
	2	12	270	64	23.7
Windsor Bridge (Ont.)	1	22	2,560	606	23.7
	2	12	4,250	943	22.2
Windsor Tunnel (Ont.)	1	14	925	119	12.9
	2	12	1,550	261	16.8

Cont'd

Table 3 Cont'd

Port	Stint	Length in Days	Questionnaires Distributed	Responses	Response Rates (%)
Emerson (Man.)	1	10	680	175	25.7
	2	10	900	176	19.6
Aldergrove (B.C.)	1	5	160	30	18.8
	2	8	475	164	34.5
Huntingdon (B.C.)	1	5	450	129	28.7
	2	11	750	194	25.9
Pacific Highway (B.C.)	1	7	2,050	411	20.0
	2	13	4,200	1,003	23.9
Canada Totals	-	-	37,695	7,677	20.4

9. CONCLUSION

The system of questionnaire distribution during continuous stints is running smoothly as an interim measure. However, the methodology can be refined, and some modifications will undoubtedly be made in an attempt to further improve the results, as this system will have to stay in place for at least two or three more years.

ACKNOWLEDGEMENT

The survey described in this article was developed under the direction of Dr. M.P. Singh and Mr. M.W. Valiquette. The authors wish to acknowledge the major contribution of Mr. J.G. Bailie to the survey's development and successful operation. The authors also thank the referee for several helpful comments and suggestions.

RESUME

Afin d'augmenter la taille de l'échantillon à un coût additionnel minimal, des modifications importantes ont été apportées au système de distribution de questionnaires aux residents du Canada qui rentrent des Etats-Unis par voie terrestre.

On a choisi un échantillon systématique de périodes d'enquête, chacun comportant plusieurs jours consecutifs, à chaque poste frontière ("port") considéré dans cette enquête. Cet article décrit la méthode d'échantillonnage, les contraintes qui l'ont déterminée et certaines données provisoires portant sur son efficacité.

REFERENCES

- [1] Ghangurde, P.D., "Sample Design and Estimation for Auto Entry Survey", Technical Memorandum, March 1976.
- [2] Bailie, J.G., "Low Cost Distribution System: Operations Manual", International Travel Section, November 1976.
- [3] Gough, J.H., and Ghangurde, P.D., "Sample Selection for 1977 Questionnaire Distribution to Returning Canadians (Land 1+), Technical Memorandum, January 1977.

APPENDIX A

Formal Description of the Sampling Method

The sample selection method may be formalized using the following notation:

M : number of days in year

m : number of stints per port, per year

k : number of ports in stratum

R : a random start

$\lambda = km$: number of stints in stratum, per year

$\alpha = \left[\frac{M}{\lambda} \right]$: the greatest integer in $\frac{M}{\lambda}$

$f = M - \lambda\alpha$ so that f/λ is the fractional part of M/λ ; i.e. $M/\lambda = \alpha + f/\lambda$

t_i : sample points, $i = 0, 1, \dots, \lambda - 1$

The usual circular systematic sampling method is to choose R randomly from $\{1, \dots, M\}$ and to let the sample consist of

$$t_i = (R + i\alpha) \pmod{M}, \quad i = 0, 1, \dots, \lambda - 1.$$

If $f = 0$, circular systematic is equivalent to linear systematic sampling. However, in general

$$\begin{aligned} t_\lambda &= R + M - f \\ &\equiv t_0 - f \pmod{M}. \end{aligned}$$

Instead of taking $t_i = R + i \left[\frac{M}{\lambda} \right]$, we simply guarantee that t_i itself be an integer by letting $t_i = R + \left[i \frac{M}{\lambda} \right]$. Thus the sampling interval is varied according to a predetermined pattern, and

$$\begin{aligned} t_\lambda &= R + \left[\lambda \frac{M}{\lambda} \right] \\ &= R + M \\ &\equiv t_0 \pmod{M} \text{ as desired.} \end{aligned}$$

AN INVESTIGATION OF THE PROPERTIES OF RAKING RATIO ESTIMATORS: II WITH CLUSTER SAMPLING

H.R. Arora and G.J. Brackstone
Census Survey Methods Division

Results of an earlier paper on the use of raking ratio estimators are extended to the case of cluster sampling. An empirical study is discussed.

1. INTRODUCTION

In an earlier paper [1] the properties of raking ratio estimators up to four iterations were examined under the assumption of simple random sampling within each Weighting Area (WA). A description of the use of raking ratio estimators in the 1971 and 1976 Canadian Censuses of Population and Housing was given in [1]. In this paper we extend the results of [1] to the case where the sample is assumed to be a cluster sample within each WA. In the Census context the results in this paper are applicable to estimators of the number of persons with a particular personal characteristic since the Census sample was a cluster sample of persons, the clusters being households. The results in [1] were applicable only to household characteristics.

2. THE RAKING RATIO ESTIMATOR OF A DOMAIN TOTAL

Consider a population of A clusters of units. Suppose the units in the population are cross-classified into a two-dimensional matrix that is defined in terms of characteristics known for all units of the population. Let N_{hij} be the number of units from the h th cluster that fall in the (i,j) th cell of the matrix. Let Y_{hij} be the total of the variable Y for these N_{hij} units. Suppose a simple random sample, s , of ' a ' clusters is selected and all the units in the selected clusters are cross-classified according to the same matrix. Summation over a subscript will generally be indicated by replacing the subscript with a dot. However, summation over clusters will be denoted by dropping the subscript h (e.g., $n_{ij} = \sum_{hes} N_{hij}$ and $y_{ij} = \sum_{hes} Y_{hij}$).

The objective is to estimate the population total, Y , for the sample characteristic, y .

The estimator \hat{Y} of Y will be of the form $\hat{Y} = \sum_i \sum_j W_{ij} y_{ij}$ (2.1)

where $W_{ij} = f(\{n_{ij}, N_{ij}\})$ is a weight attached to each sample unit in the (i, j) th cell.

Let $W_{ij}^{(t)}$ denote the weight, W_{ij} , corresponding to the t^{th} iteration of the Raking Ratio Estimation Procedure (RREP). Then $W_{ij}^{(t)}$ is given by

$$\begin{aligned} W_{ij}^{(0)} &= \frac{N}{n} \text{ for all } i, j \\ W_{ij}^{(t)} &= W_{ij}^{(t-1)} \frac{N_{.j}}{\sum_u n_{uj} W_{uj}^{(t-1)}} \quad \text{if } t \text{ is even} \\ &= W_{ij}^{(t-1)} \frac{N_{i.}}{\sum_b n_{ib} W_{ib}^{(t-1)}} \quad \text{if } t \text{ is odd.} \end{aligned} \quad (2.2)$$

If t is even the weighted sample total in each column of the matrix is exactly equal to the known population total for the column, while if t is odd, this equality is exact for the rows of the matrix.

$$\text{Let } \hat{Y}^{(t)} = \sum_i \sum_j W_{ij}^{(t)} y_{ij}. \quad (2.3)$$

$$\text{We shall define } S^2(v_h) = \frac{1}{A-1} \sum_{h=1}^A (v_h - \bar{v})^2 \quad (\bar{v} = \frac{1}{A} \sum_{h=1}^A v_h)$$

$$S(v_h, w_h) = \frac{1}{A-1} \sum_{h=1}^A (v_h - \bar{v}) (w_h - \bar{w}), \text{ for the population,}$$

$$\text{and } s^2(v_h) = \frac{1}{a-1} \sum_{hes} (v_h - \bar{v})^2 \quad (\bar{v} = \frac{1}{a} \sum_{hes} v_h)$$

$$s(v_h, w_h) = \frac{1}{a-1} \sum_{hes} (v_h - \bar{v}) (w_h - \bar{w}) \text{ for the sample.}$$

$$\text{Also } \alpha = A^2 \left(\frac{1}{a} - \frac{1}{A} \right).$$

All other notation is as defined in Section 2 of [1].

3. BIAS AND VARIANCE OF THE NO-ITERATION ESTIMATOR

$$\hat{Y}(o) = \frac{N \sum_{hes} \sum_i \sum_j Y_{hij}}{\sum_{hes} \sum_i \sum_j N_{hij}} = \frac{N \sum_{hes} Y_{h..}}{\sum_{hes} N_{h..}}$$

$$E \left(\sum_{hes} \sum_i \sum_j Y_{hij} \right) = \frac{a}{A} \sum_{h=1}^A \sum_i \sum_j Y_{hij} = \frac{a}{A} \sum_i \sum_j Y_{ij} = \frac{a}{A} Y \quad \text{and}$$

$$E \left(\sum_{hes} \sum_i \sum_j N_{hij} \right) = \frac{a}{A} \sum_{h=1}^A \sum_i \sum_j N_{hij} = \frac{a}{A} \sum_i \sum_j N_{ij} = \frac{a}{A} N$$

If y, x are estimators of Y, X respectively with biases $= O(n^{-1})$, second order moments $= O(n^{-1})$, and higher order moments $= O(n^{-1})$, and $k = E(x) + O(n^{-1})$, then to order n^{-1}

$$\text{Bias} \left(k \frac{Y}{x} \right) \doteq \text{Bias} (y) - R \text{Bias} (x) - \frac{1}{E(x)} C(x, y-Rx) \quad (3.1)$$

$$\text{where } R = \frac{E(y)}{E(x)}.$$

$$\text{Using this result with } y = \frac{A}{a} \sum_{hes} \sum_i \sum_j Y_{hij}, \quad x = \frac{A}{a} \sum_{hes} \sum_i \sum_j N_{hij}$$

$$\text{and } k = N = E(x),$$

$$\begin{aligned}
 \text{Bias } (\hat{Y}^{(o)}) &\doteq - \frac{1}{N} C \left[\frac{A}{a} \sum_{h \in S} \sum_i \sum_j N_{hij}, \frac{A}{a} \sum_h \sum_i \sum_j (Y_{hij} - \frac{Y}{N} N_{hij}) \right] \\
 &= - \frac{1}{N} C \left[\frac{A}{a} \sum_{h \in S} N_{h..}, \frac{A}{a} \sum_{h \in S} (Y_{h..} - \frac{Y}{N} N_{h..}) \right] \\
 &= - \frac{1}{N} \alpha S (N_{h..}, Y_{h..} - \frac{Y}{N} N_{h..}) \quad (3.2)
 \end{aligned}$$

$$= -\alpha \frac{1}{A-1} \frac{1}{N} \sum_{h=1}^A (N_{h..} Y_{h..} - \frac{Y}{N} N_{h..}^2) \quad (3.3)$$

An estimator of $B(\hat{Y}^{(o)})$ can be obtained by substituting the sample covariance in (3.2) giving

$$b(\hat{Y}^{(o)}) = -\alpha \frac{1}{a-1} \frac{a}{An} \sum_{h \in S} (N_{h..} Y_{h..} - \frac{Y}{n} N_{h..}^2) \quad (a \geq 2). \quad (3.4)$$

We next derive the asymptotic variance of $\hat{Y}^{(o)}$.

$$\begin{aligned}
 v(\hat{Y}^{(o)}) &= v\left(\frac{N \sum_{h \in S} Y_{h..}}{\sum_{h \in S} N_{h..}} \right) \\
 &\doteq v\left[\frac{A}{a} \sum_{h \in S} (Y_{h..} - \frac{Y}{N} N_{h..}) \right] \\
 &= \alpha S^2 (Y_{h..} - \frac{Y}{N} N_{h..}) \quad (3.5)
 \end{aligned}$$

An estimator of $v(\hat{Y}^{(o)})$ can be obtained by substituting the sample variance in (3.5) giving

$$v(\hat{Y}^{(o)}) = \alpha \frac{1}{a-1} \sum_{h \in S} (Y_{h..} - \frac{Y}{n} N_{h..})^2 \quad (3.6)$$

4. BIAS AND VARIANCE OF THE ONE-ITERATION ESTIMATOR

$$\hat{Y}^{(1)} = \sum_i \sum_j N_{ij} \cdot \frac{Y_{ij}}{n_{i.}} = \sum_i N_{i.} \cdot \frac{Y_{i.}}{n_{i.}}, \quad (4.1)$$

$$E(Y_{i.}) = E\left(\sum_{h \in S} Y_{hi.}\right) = \frac{a}{A} \sum_{h=1}^A Y_{hi.} = \frac{a}{A} Y_{i.},$$

$$E(n_{i.}) = E\left(\sum_{h \in S} N_{hi.}\right) = \frac{a}{A} \sum_{h=1}^A N_{hi.} = \frac{a}{A} N_{i.}.$$

4.1 Asymptotic Bias of $\hat{Y}^{(1)}$

Using (3.1), we can write

$$\begin{aligned} \text{Bias } (\hat{Y}^{(1)}) &\doteq - \sum_i \frac{1}{E\left(\frac{A}{a} n_{i.}\right)} C\left(\frac{A}{a} n_{i.}, \frac{A}{a} (Y_{i.} - R_{i.} n_{i.})\right) \\ &= - \sum_i \frac{\alpha}{N_{i.}} S(N_{hi.}, Y_{hi.} - R_{i.} N_{hi.}) \end{aligned} \quad (4.2)$$

$$= - \frac{\alpha}{A-1} \sum_i \frac{1}{N_{i.}} \sum_{h=1}^A (N_{hi.} Y_{hi.} - R_{i.} N_{hi.}^2). \quad (4.3)$$

An estimator of $B(\hat{Y}^{(1)})$ can be obtained by substituting the sample covariance $s(N_{hi.}, Y_{hi.} - \frac{Y_{i.}}{n_{i.}} N_{hi.})$ in (4.2) giving

$$b(\hat{Y}^{(1)}) = - \frac{\alpha}{a-1} \sum_i \frac{a}{An_{i.}} \sum_{h \in S} (N_{hi.} Y_{hi.} - \frac{Y_{i.}}{n_{i.}} N_{hi.}^2). \quad (4.4)$$

4.2 Asymptotic Variance of $\hat{Y}^{(1)}$

$$\begin{aligned} v(\hat{Y}^{(1)}) &= v\left(\sum_i N_{i.} \cdot \frac{Y_{i.}}{n_{i.}}\right) \\ &\doteq v\left[\frac{A}{a} \sum_{h \in S} (Y_{h..} - \sum_i R_{i.} N_{hi.})\right] \\ &= \alpha S^2(Y_{h..} - \sum_i R_{i.} N_{hi.}). \end{aligned} \quad (4.5)$$

Thus

$$v(\hat{Y}^{(1)}) = \frac{\alpha}{A-1} \sum_{h=1}^A (Y_{h..} - \sum_i R_{i.} N_{hi.})^2. \quad (4.6)$$

An estimator of $V(\hat{Y}^{(1)})$ can be obtained by substituting the sample variance $s^2(Y_{h..} - \sum_i \frac{y_{i.}}{n_{i.}} N_{hi.})$ in (4.5) giving

$$v(\hat{Y}^{(1)}) = \frac{\alpha}{a-1} \sum_{hes} (Y_{h..} - \sum_i \frac{y_{i.}}{n_{i.}} N_{hi.})^2. \quad (4.7)$$

In the special case where clusters are of size one (i.e., s.r.s.), $N_{hij} = 0$ or 1 and it can be shown that (4.3) and (4.6) reduce respectively to zero and to the expression (5.2) in [1].

5. BIAS AND VARIANCE OF THE TWO-ITERATION ESTIMATOR

$$\hat{Y}^{(2)} = \sum_i \sum_j N_{.j} \frac{\hat{Y}_{ij}^{(1)}}{\hat{N}_{.j}^{(1)}} = \sum_j N_{.j} \frac{\hat{Y}_{.j}^{(1)}}{\hat{N}_{.j}^{(1)}}. \quad (5.1)$$

5.1 Asymptotic Bias of $\hat{Y}^{(2)}$

Using (3.1) we can write

$$B(\hat{Y}^{(2)}) = \sum_j B(\hat{Y}_{.j}^{(1)}) - \sum_j R_{.j}^{(1)} B(\hat{N}_{.j}^{(1)}) - \sum_j \frac{1}{\hat{N}_{.j}^{(1)}} C(\hat{N}_{.j}^{(1)}, \hat{Y}_{.j}^{(1)} - R_{.j}^{(1)} \hat{N}_{.j}^{(1)}).$$

By expanding the second and third terms we obtain

$$\begin{aligned} B(\hat{Y}^{(2)}) &= B(\hat{Y}^{(1)}) + \sum_i \sum_j R_{.j}^{(1)} \frac{1}{\hat{N}_{.j}^{(1)}} \alpha S(N_{hi.}, N_{hij} - \rho_{ij} N_{hi.}) \\ &- \sum_j \frac{1}{\hat{N}_{.j}^{(1)}} \alpha S(N_{h.j} - \sum_u \rho_{uj} N_{hu.}, Y_{h.j} - \sum_u \frac{y_{uj}}{N_{u.}} N_{hu.} \\ &- R_{.j}^{(1)} (N_{h.j} - \sum_u \rho_{uj} N_{hu.})). \end{aligned} \quad (5.2)$$

An estimator of $B(\hat{Y}^{(2)})$ can be obtained by substituting sample covariances in (5.2) giving

$$\begin{aligned} b(\hat{Y}^{(2)}) &= b(\hat{Y}^{(1)}) + \alpha \sum_i \sum_j \hat{R}_{.j}^{(1)} \frac{a}{An_{i.}} \frac{1}{a-1} \sum_{hes} N_{hi.} (N_{hij} - \hat{\rho}_{ij} N_{hi.}) \\ &- \alpha \sum_j \frac{1}{\hat{N}_{.j}^{(1)}} \frac{1}{a-1} \sum_{hes} (N_{h.j} - \sum_u \hat{\rho}_{uj} N_{hu.}) \{ Y_{h.j} - \sum_u \frac{y_{uj}}{n_{u.}} N_{hu.} \\ &- \hat{R}_{.j}^{(1)} (N_{h.j} - \sum_u \hat{\rho}_{uj} N_{hu.}) \}. \end{aligned} \quad (5.3)$$

5.2 Asymptotic Variance of $\hat{Y}^{(2)}$

By repeatedly linearizing the ratios in $\hat{Y}^{(2)}$ we obtain

$$\begin{aligned} v(\hat{Y}^{(2)}) &\doteq v \sum_j (\hat{Y}_{.j}^{(1)} - R_{.j}^{(1)} \hat{N}_{.j}^{(1)}) \\ &= \alpha S^2 \{ (Y_{h..} - \sum_i R_{i.} N_{hi.}) - \sum_j R_{.j}^{(1)} (N_{h.j} - \sum_i \rho_{ij} N_{hi.}) \}, \end{aligned} \quad (5.4)$$

and the estimator

$$v(\hat{Y}^{(2)}) = \frac{\alpha}{a-1} \sum_{hes} [(Y_{h..} - \sum_i \frac{y_{i.}}{n_{i.}} N_{hi.}) - \sum_j \hat{R}_{.j}^{(1)} (N_{h.j} - \sum_i \hat{\rho}_{ij} N_{hi.})]^2. \quad (5.5)$$

6. BIAS AND VARIANCE OF THE THREE-AND FOUR-ITERATION ESTIMATORS

By the procedure of repeatedly linearizing ratios similar to that used above, we obtain the following results.

$$\hat{Y}^{(3)} = \sum_i N_{i.} \frac{\hat{Y}_{i.}^{(2)}}{\hat{N}_{i.}^{(2)}}, \quad (6.1)$$

$$\begin{aligned} v(\hat{Y}^{(3)}) &= \alpha S^2 [Y_{h..} - \sum_i R_{i.} N_{hi.} - \sum_j R_{.j}^{(1)} (N_{h.j} - \sum_i \rho_{ij} N_{hi.}) \\ &\quad + \sum_i R_{i.}^{(2)} \sum_j \kappa_{ij}^{(1)} (N_{h.j} - \sum_u \rho_{uj} N_{hu.})], \end{aligned} \quad (6.2)$$

$$\begin{aligned} v(\hat{Y}^{(4)}) &= \alpha S^2 [(Y_{h..} - \sum_i R_{i.} N_{hi.}) - \sum_j R_{.j}^{(1)} (N_{h.j} - \sum_i \rho_{ij} N_{hi.}) \\ &\quad + \sum_i \sum_b R_{i.}^{(2)} \kappa_{ib}^{(1)} (N_{h.b} - \sum_u \rho_{ub} N_{hu.}) \\ &\quad - \sum_i \sum_j R_{.j}^{(3)} \rho_{ij}^{(2)} \sum_b \kappa_{ib}^{(1)} (N_{h.b} - \sum_u \rho_{ub} N_{hu.})]. \end{aligned} \quad (6.3)$$

Similar but longer expressions for $B(\hat{Y}^{(3)})$ and $B(\hat{Y}^{(4)})$ have also been derived.

Estimators of the above variances can again be obtained by substituting the sample variances in place of the population variances in (6.2) and (6.3). This procedure yields

$$v(\hat{Y}^{(3)}) = \frac{\alpha}{a-1} \sum_{hes} [(Y_{h..} - \sum_i \frac{Y_{i.}}{n_{i.}} N_{hi.}) - \sum_j \hat{R}_{.j}^{(1)} (N_{h.j} - \sum_i \hat{\rho}_{ij} N_{hi.}) + \sum_i \sum_j \hat{R}_{i.}^{(2)} \hat{\kappa}_{ij} (N_{h.j} - \sum_u \hat{\rho}_{uj} N_{hu.})]^2 \quad (6.4)$$

and

$$v(\hat{Y}^{(4)}) = \frac{\alpha}{a-1} \sum_{hes} [(Y_{h..} - \sum_i \frac{Y_{i.}}{n_{i.}} N_{hi.}) - \sum_j \hat{R}_{.j}^{(1)} (N_{h.j} - \sum_i \hat{\rho}_{ij} N_{hi.}) + \sum_i \sum_j \hat{R}_{i.}^{(2)} \hat{\kappa}_{ij} (N_{h.j} - \sum_u \hat{\rho}_{uj} N_{hu.}) - \sum_i \sum_j \sum_b \hat{R}_{.j}^{(3)} \hat{\rho}_{ij} \hat{\kappa}_{ib} (N_{h.b} - \sum_u \hat{\rho}_{ub} N_{hu.})]^2 \quad (6.5)$$

7. THE EMPIRICAL INVESTIGATION

To compare the variance and bias of raking ratio estimators for various iterations and for various characteristics, the formulae derived in the previous sections were applied to data from the 1974 Canadian Test Census. This Test Census utilized the sample procedure described in section 1 of [1]. The data used in the empirical investigations came from one Electoral District (ED) which contained 15 Weighting Areas (WA's). The WA's varied in size from a low of 628 households (213 in the sample) to a high of 1946 households (647 in the sample). The average size was 1262 households.

The initial cross-classification table or weighting matrix for persons is given in Appendix 1. The rows and columns of the initial table were collapsed if the following conditions were not met:

1. all $N_{i.}$ and $N_{.j} \geq 35$
2. all ratios $N_{i.}/n_{i.}$ and $N_{.j}/n_{.j} > 1$ but < 12
3. all $n_{i.}$ and $n_{.j} > 0$.

A collapsing strategy pre-defined the rows or columns to be collapsed when these conditions were not met. The collapsing procedure continued until all the above conditions were satisfied or until all steps in the collapsing strategy had been exhausted. The result was that the final weighting matrices would in general differ slightly from WA to WA. The largest weighting matrix had 28 rows and 26 columns while the smallest had 21 rows and 14 columns (compared to 34 rows and 30 columns in the initial matrix). Differences in collapsing between WA's explain some of the differences in variance reduction between WA's in the results presented below.

The population variables collected in the 1974 Test Census were the following:

100%	Sample (33 1/3%)
Relation to Head	School Attendance
Sex	Years of Schooling
Age	Post-Secondary Education
Marital Status	Academic Qualifications
Mother Tongue	Labour Force Status
	Address Five Years Ago

Since each category of each of the above variables, and each cell of each cross-tabulation is a potential characteristic (y) that could be examined (and all of these at any geographic level within the WA), some arbitrary selection of characteristics was essential. Altogether 30 categories were investigated and the results presented here represent a typical cross-section of these 30 categories.

Clearly our primary interest is in the sample variables. However, a selection of 100% variables has been considered for two reasons. First, in cross-tabulations of sample variables against 100% variables, estimates for 100% variables are published (though the variance of such estimates may be of little interest when the 100% value, and therefore the exact sampling deviation is available). Secondly, 100% variables can be used to some extent as examples of variables

with a relatively high correlation with the variables used to define the weighting matrix in order to see the effects of the raking ratio procedure for such variables. For the purposes of presentation we have broken down the categories considered into three classes.

- A. Categories defined by sample variables.
- B. Categories defined by 100% variables but not used for control in the weighting matrix.
- C. Categories defined by 100% variables and used for control in the weighting matrix.

The specific categories for which results are presented are the following:

- A1 Employed
- A2 Unemployed
- A3 Not in the Labour Force
- A4 Not Moved in Last 5 Years
- A5 Moved in Last 5 Years in Same Municipality
- A6 Highest Grade is 1 to 8
- A7 Highest Grade is 9 to 10
- A8 Attending School Full Time
- A9 Bachelor Degree or Higher

(Note that all the above categories were restricted to persons 15 and over in the 1974 Test Census)

- B1 Widowed
- B2 Mother Tongue German
- B3 Mother Tongue Italian

- C1 Single (Never Married)
- C2 Mother Tongue English
- C3 Mother Tongue French

In addition to the categories listed above, all of which are geographically at the WA level, we also examined certain categories at the EA level within WA 2. This WA contained three EA's. The categories considered at the EA level were A1, A4, A6, B2, C1. Note that the same weights, calculated at the WA level, are used in producing estimates at the EA level.

Tables A1 to C3 summarize the results. These tables list for each iteration ($i = 0, 1, 2, 3, 4$) estimates of the population totals ($\hat{Y}^{(i)}$), standard errors (SE_i), coefficients of variation (CV_i), and the "ratio of error" (RE_i) defined as the standard error of the i -iteration estimator expressed as a % of the standard error of the no-iteration estimator. These figures are given for a sample of 5 WA's arranged in size from largest to smallest. Corresponding figures are also given at the Electoral District level. These act as a summary of the WA data. A measure of $C(k-p)$ of the change in population estimates between iterations is also given. See [1] for the definition of $C(k-p)$.

Because of the high computational cost, the bias estimates were calculated only for one average-sized WA and only up to the second iteration. For the 15 categories considered it was found that the absolute value of the estimated bias as a percentage of the population estimate never exceeded .13%. This figure is small compared with the corresponding coefficient of variation figures. Because of this, and because of similar findings in the case of simple random sampling (see [1]) the decision not to pursue bias estimation any further was felt to be justified.

8. ANALYSIS OF RESULTS

The principal findings from an examination of the empirical results are listed below.

- a) As mentioned in the previous section the biases of the raking-ratio estimates appear to be negligible compared with their standard errors.

- b) For A-type categories there is, at the ED level, a reduction of at least 2.4 in the 'ratio of error' in moving from the 0th to the first iteration, and a further reduction of at least 1.1 in moving from the 1st to the second iteration. In general, the gains in efficiency seem to be much more evenly split between the first and second iterations than was the case with simple random sampling [1], where most of the gains occurred either at the first iteration or at the second iteration. An examination of the variables used in the rows and columns of the cross-classification tables (Appendix 1) reveals that some correlation between many of the A-type variables and both the rows and columns of the table is not unexpected. Indeed to some extent the variables defining the rows and columns are correlated with each other. This contrasts with the simple random sampling case (Appendix 1 in [1]) where there was a marked dichotomy between the housing variables that defined the columns and the head of household variables that defined the rows.

Beyond the second iteration, there is little further gain in efficiency and little change in the values of the estimates.

- c) Given that there is little gain in efficiency for A-type (ie. sample) variables beyond the second iteration, the justification for proceeding to four iterations is primarily in terms of improving the sample-population agreement for B and C-type (i.e. 100%) variables.
- d) A comparison of CV_0 and RE_4 for the individual A-type categories reveals a strong tendency for large gains in efficiency (i.e. low RE_4) to be associated with large categories (i.e. small CV_0). The exceptions to this tendency (e.g. A8) are generally explained by a strong association between a small sample category and the rows or columns defining the cross-classification table.

- e) As expected, the gains in efficiency for B-type categories depend heavily on the relationship of the category to the rows and columns of the cross-classification table and on the frequency of collapsing of rows and columns. For example consider category B1 (widowed). This category accounts for 7.9% of ever married persons and ever married persons in turn account for all persons in certain rows. Since no collapsing took place over Marital Status (except for age groups 65 or over), the gains are high. On the other hand for the smaller Mother Tongue categories (B2 and B3) the gains are only moderate. Although the proportion

$$\frac{\text{Italian Mother Tongue}}{\text{'Other' Mother Tongue}}$$

is higher than the proportion

$$\frac{\text{Widowed}}{\text{Ever Married}}$$

the gains for Italian are lower than for Widowed because the 'other' Mother Tongue category is almost always collapsed with English, French, or both, whereas the category 'ever married' is rarely collapsed.

- f) The gains in efficiency for C-type categories depend heavily upon the collapsing that took place within each WA. Where no collapsing of important rows or columns took place, variances at the WA level are clearly reduced to zero for some categories and certain iterations. The results for C-type categories highlight the need to choose carefully the collapsing criteria and strategy since these will have a profound effect on the variance, and on the sample-100% agreement, for certain C-type categories.

- g) Categories at the EA level (or any geographic level below the WA) can be regarded as small WA categories. As expected, the gains in efficiency at the EA level are found to be smaller than for the corresponding categories at the WA level. Findings a, c, d, e, f and g are similar to those in [1].

9. CONCLUSION

The results of the empirical study have shown some significant reductions in variance for sample variables through the use of the RREP.

For the sample categories, there is generally a reduction in variance at both the first and second iteration of the RREP. Thus, a level of efficiency is achieved that is greater than would have been achieved with a one-dimensional ratio estimation procedure using either the rows or the columns. Also, using the RREP, the marginal constraints for the rows are exactly satisfied, and those for the columns are nearly satisfied. This is an important consideration as far as the user is concerned. Inconsistencies between tabulations based on sample data and tabulations based on 100% data, although inevitable to some extent, often present a problem of reconciliation to the user and a problem of explanation for the survey-taker.

10. ACKNOWLEDGEMENT

The authors wish to acknowledge the valuable assistance of Ian MacMillan and Michael Bankier in writing computer programs and checking the results and Professor J.N.K. Rao for his advice and contributions to this study.

RESUME

Les resultats concernant l'usage des estimateurs de l'échantillon en formation qui ont été présentés dans un article précédent sont étendus ici au cas de l'échantillonnage par grappes. Une étude empirique est discutée.

REFERENCES

- [1] Arora, H.R. and Brackstone, G.J., (1977), "An Investigation of the Properties of Raking Ratio Estimators With Simple Random Sampling", Vol. 3, No. 1.

CATEGORIES AT WA LEVEL UNDER CLUSTER SAMPLING

A1

EMPLOYED

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.43 0.70 0.36 0.37 0.60 0.12

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	2,144.7	56.7	0.026	2,120.1	50.6	0.024	89.2	2,118.9	44.3	0.021	78.2	2,109.8	44.7	0.021	78.8	2,117.9	44.2	0.021	77.9
9	2,062.3	55.0	0.027	2,089.8	45.1	0.022	82.2	2,087.4	41.2	0.020	75.0	2,087.7	41.7	0.020	75.9	2,084.3	40.9	0.020	74.5
11	1,995.6	48.4	0.024	2,017.0	37.8	0.019	78.0	2,020.2	34.7	0.017	71.7	2,017.1	35.0	0.017	72.3	2,019.4	34.7	0.017	71.7
2	1,562.2	43.6	0.028	1,590.1	35.1	0.022	80.5	1,592.8	31.0	0.019	71.1	1,591.2	31.9	0.020	73.1	1,592.3	30.9	0.019	70.9
12	910.9	34.4	0.038	899.9	27.7	0.031	80.4	900.9	25.5	0.028	74.2	898.5	26.0	0.029	75.5	899.6	25.3	0.028	73.5
ED	24,464.4	180.9	0.007	24,511.7	152.4	0.006	84.3	24,485.7	137.7	0.006	76.1	24,478.0	139.0	0.006	76.8	24,480.2	137.3	0.006	75.9

A2

UNEMPLOYED

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
2.74 1.50 0.71 0.64 1.13 0.29

8	258.8	23.4	0.090	252.3	22.5	0.089	96.1	247.9	22.3	0.090	95.6	249.3	22.3	0.089	95.3	247.9	22.3	0.090	95.6
9	237.6	22.3	0.094	238.7	21.8	0.091	98.1	240.5	21.7	0.090	97.6	239.4	21.8	0.091	97.7	241.3	21.7	0.090	97.6
11	138.5	18.3	0.132	145.8	17.7	0.122	96.8	145.0	17.4	0.120	95.2	145.0	17.5	0.121	95.5	144.9	17.4	0.120	95.2
2	103.6	14.8	0.143	104.0	14.4	0.138	97.1	104.7	14.0	0.134	94.8	103.8	14.1	0.135	94.9	104.2	14.0	0.135	94.7
12	59.1	12.1	0.205	55.3	12.0	0.217	99.2	54.8	11.8	0.215	97.6	54.9	11.8	0.215	97.5	54.7	11.8	0.216	97.6
ED	2,234.5	71.1	0.032	2,223.3	68.9	0.031	96.8	2,219.7	68.0	0.031	95.7	2,223.0	68.1	0.031	95.7	2,221.0	68.0	0.031	95.7

A3

NOT IN THE LABOUR FORCE

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.69 1.02 0.39 0.45 0.89 0.15

8	1,841.3	54.7	0.030	1,826.8	50.2	0.028	91.8	1,834.3	42.9	0.023	78.4	1,839.5	43.2	0.023	78.9	1,835.5	42.8	0.023	78.1
9	1,468.3	50.1	0.034	1,494.4	43.9	0.029	87.5	1,497.7	40.5	0.027	80.8	1,496.1	39.9	0.027	79.7	1,500.0	40.3	0.027	80.4
11	1,131.9	41.8	0.037	1,139.1	35.9	0.031	85.8	1,143.8	33.1	0.029	79.3	1,139.1	33.0	0.029	78.9	1,144.6	33.1	0.029	79.2
2	878.7	37.6	0.043	892.5	34.3	0.038	91.9	898.5	30.4	0.034	81.0	891.6	30.9	0.035	82.2	899.5	30.3	0.034	80.7
12	588.5	28.2	0.048	578.9	25.7	0.044	91.1	578.3	23.7	0.041	84.2	580.7	23.8	0.041	84.4	579.7	23.5	0.040	83.3
ED	18,614.5	174.8	0.009	18,647.3	151.2	0.008	86.5	18,650.4	135.5	0.007	77.5	18,678.2	136.1	0.007	77.9	18,654.7	135.1	0.007	77.3

A4

NOT MOVED IN LAST 5 YEARS

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.70 1.32 0.32 0.33 1.12 0.17

8	2,109.0	61.3	0.039	2,064.3	70.8	0.034	87.1	2,074.4	65.1	0.031	80.1	2,071.7	64.8	0.031	79.7	2,071.7	64.6	0.031	79.5
9	2,190.2	85.4	0.039	2,214.9	69.4	0.031	81.2	2,231.3	64.1	0.029	75.0	2,224.0	64.4	0.029	75.4	2,232.9	64.1	0.029	75.1
11	1,795.1	78.2	0.044	1,812.5	66.5	0.037	85.1	1,842.9	61.7	0.033	78.9	1,837.9	62.1	0.034	79.4	1,842.2	61.6	0.033	78.8
2	772.2	59.9	0.078	787.0	54.4	0.069	90.8	793.9	51.1	0.064	85.4	789.1	51.2	0.065	85.5	794.3	50.9	0.064	85.1
12	671.3	51.4	0.077	649.2	45.6	0.070	88.7	659.8	43.1	0.065	83.8	659.4	43.3	0.066	84.2	660.4	43.0	0.065	83.7
ED	22,565.9	280.9	0.012	22,574.1	242.4	0.011	86.3	22,527.3	227.7	0.010	81.1	22,534.3	227.9	0.010	81.2	22,511.6	227.1	0.010	80.8

CATEGORIES AT WA LEVEL UNDER CLUSTER SAMPLING

A5

MOVED IN LAST 5 YEARS IN SAME MUNICIPALITY

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.25 1.16 0.44 0.31 1.07 0.22

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	1,121.4	61.7	0.055	1,113.7	60.4	0.054	97.9	1,113.2	59.1	0.053	95.7	1,112.2	58.9	0.053	95.5	1,114.5	59.0	0.053	95.5
9	840.8	57.7	0.069	852.1	56.0	0.066	97.0	840.3	53.6	0.064	92.9	842.3	53.9	0.064	93.4	838.8	53.6	0.064	92.9
11	790.0	56.8	0.072	797.3	56.0	0.070	98.6	787.1	54.5	0.069	96.0	784.4	54.6	0.070	96.1	785.9	54.4	0.069	95.8
2	710.1	51.7	0.073	725.7	50.7	0.070	98.0	728.0	50.5	0.069	97.7	725.2	50.4	0.070	97.5	727.5	50.4	0.069	97.5
12	210.0	28.2	0.134	205.0	27.2	0.133	96.2	201.7	26.6	0.132	94.3	200.4	26.6	0.133	94.4	201.0	26.6	0.132	94.1
ED	10,929.9	207.3	0.019	10,943.0	201.0	0.018	97.0	10,940.3	196.2	0.018	94.7	10,940.6	196.2	0.018	94.7	10,941.3	196.0	0.018	94.6

A6

HIGHEST GRADE IS 1 TO 8

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.92 1.61 0.75 0.71 1.17 0.18

8	1,228.5	49.2	0.040	1,222.3	44.7	0.037	90.9	1,235.9	41.6	0.034	84.6	1,230.1	41.9	0.034	85.2	1,236.4	41.7	0.034	84.7
9	609.2	37.2	0.061	625.4	35.3	0.056	94.9	629.2	33.8	0.054	90.7	630.3	33.7	0.054	90.6	630.4	33.7	0.054	90.6
11	427.4	32.9	0.077	436.2	29.6	0.068	90.2	440.5	28.0	0.064	85.3	439.3	28.1	0.064	85.4	440.6	28.0	0.063	85.1
2	378.7	32.1	0.085	377.2	30.1	0.080	93.9	379.7	27.9	0.074	87.0	378.4	27.9	0.074	87.0	380.8	27.9	0.073	86.9
12	298.7	26.5	0.089	286.3	23.7	0.083	89.6	293.6	21.3	0.073	80.5	292.4	21.6	0.074	81.7	294.7	21.3	0.072	80.3
ED	10,660.2	151.8	0.014	10,681.0	133.5	0.012	87.9	10,656.9	122.5	0.011	80.7	10,688.8	123.6	0.012	81.4	10,668.1	122.4	0.011	80.6

A7

HIGHEST GRADE IS 9 TO 10

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
2.12 0.89 0.30 0.30 0.85 0.10

8	913.2	40.5	0.044	898.7	39.7	0.044	98.1	897.6	38.7	0.043	95.5	899.6	39.0	0.043	96.2	898.4	38.7	0.043	95.5
9	813.3	38.6	0.048	832.2	38.1	0.046	98.6	827.7	36.6	0.044	94.6	826.4	36.7	0.044	95.0	827.1	36.6	0.044	94.6
11	680.9	37.2	0.055	694.7	34.9	0.050	93.9	706.3	34.4	0.049	92.4	705.5	34.3	0.049	92.3	706.6	34.3	0.049	92.4
2	473.4	32.3	0.068	484.2	31.0	0.064	95.9	490.1	30.2	0.062	93.5	487.9	30.5	0.062	94.3	490.2	30.3	0.062	93.6
12	446.6	27.9	0.062	434.6	26.9	0.062	96.6	424.6	26.5	0.063	95.3	426.4	26.4	0.062	94.7	424.8	26.5	0.062	95.0
ED	10,165.5	137.4	0.014	10,137.3	131.9	0.013	96.0	10,141.0	128.0	0.013	93.1	10,145.4	128.4	0.013	93.4	10,139.3	127.9	0.013	93.1

A8

ATTENDING SCHOOL FULL TIME

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
4.48 2.35 1.61 1.55 1.32 0.16

8	452.1	33.3	0.074	436.5	29.7	0.068	89.2	427.3	27.8	0.065	83.4	431.6	28.2	0.065	84.6	427.8	27.7	0.065	83.3
9	591.0	41.5	0.070	595.8	29.3	0.049	70.7	601.2	27.7	0.046	66.8	593.2	28.3	0.048	68.3	602.2	27.7	0.046	66.8
11	474.6	35.3	0.074	497.8	24.4	0.049	69.0	493.7	24.1	0.049	68.3	494.7	24.1	0.049	68.2	493.4	24.0	0.049	68.0
2	289.9	29.5	0.102	301.8	22.4	0.074	75.9	306.5	20.9	0.068	70.7	304.1	20.9	0.069	70.8	306.2	20.8	0.068	70.6
12	195.2	22.3	0.114	199.9	15.1	0.076	67.8	189.5	14.9	0.079	67.0	192.0	14.7	0.077	66.2	189.6	14.9	0.079	66.9
ED	5,583.0	121.3	0.022	5,590.9	92.6	0.017	76.3	5,606.8	88.2	0.016	72.7	5,613.7	89.0	0.016	73.4	5,604.5	88.1	0.016	72.6

CATEGORIES AT WA LEVEL UNDER CLUSTER SAMPLING

A9

BACHELOR DEGREE OR HIGHER

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.31 1.88 0.61 0.46 1.75 0.33

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	104.1	15.6	0.150	102.3	15.2	0.148	97.5	101.3	14.9	0.147	95.7	101.3	14.9	0.147	95.6	101.3	14.9	0.147	95.6
9	237.6	23.8	0.100	238.3	23.5	0.099	98.5	241.6	23.0	0.095	96.3	241.8	23.1	0.096	96.9	241.9	23.0	0.095	96.4
11	283.0	25.4	0.090	280.8	24.8	0.088	97.5	279.2	24.3	0.087	95.6	278.8	24.3	0.087	95.4	278.7	24.3	0.087	95.5
2	145.0	18.4	0.127	147.4	18.1	0.123	98.0	145.1	17.8	0.123	96.8	145.0	17.9	0.123	96.8	144.7	17.8	0.123	96.7
12	56.2	11.2	0.199	56.9	11.0	0.193	98.2	56.0	10.9	0.194	97.2	55.4	10.9	0.196	97.1	55.6	10.9	0.195	97.1
ED	1,898.3	68.4	0.036	1,903.8	66.8	0.035	97.6	1,896.8	65.5	0.035	95.7	1,897.1	65.5	0.035	95.8	1,897.3	65.4	0.034	95.6

B1

WIDOWED

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
6.40 3.79 2.58 1.64 2.45 1.44

8	431.3	30.5	0.071	468.3	23.4	0.050	76.6	459.3	18.3	0.040	60.0	467.0	17.3	0.037	56.7	464.9	16.7	0.036	54.9
9	262.0	23.7	0.090	248.5	17.7	0.071	74.7	251.8	15.9	0.063	67.4	249.1	15.2	0.061	64.1	250.8	14.9	0.059	62.8
11	129.7	16.3	0.126	139.8	12.1	0.087	74.2	139.4	11.5	0.082	70.4	138.6	10.5	0.076	64.6	139.2	10.6	0.076	64.9
2	106.5	15.9	0.149	101.1	13.1	0.129	82.3	96.8	11.0	0.114	69.2	99.3	10.9	0.109	68.3	99.7	10.6	0.107	67.0
12	41.4	10.3	0.248	47.0	8.1	0.172	79.0	51.8	6.3	0.123	61.9	48.5	6.1	0.126	59.7	50.1	6.0	0.120	58.8
ED	2,629.3	75.8	0.029	2,687.7	57.0	0.021	75.2	2,667.4	51.0	0.019	67.4	2,708.5	49.1	0.018	64.8	2,689.5	48.1	0.018	63.5

B2

MOTHER TONGUE GERMAN

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
10.41 3.18 1.88 0.65 2.83 1.63

8	8.9	5.5	0.616	10.9	5.2	0.473	94.1	11.3	5.2	0.459	93.9	10.8	5.2	0.480	93.9	10.9	5.2	0.473	93.9
9	30.5	9.8	0.323	33.4	9.2	0.276	93.8	34.4	9.2	0.267	93.7	34.4	9.2	0.268	93.6	34.3	9.2	0.268	93.7
11	61.9	17.4	0.280	66.7	14.9	0.224	86.0	68.1	14.8	0.217	85.0	65.9	14.7	0.223	84.8	66.2	14.7	0.222	84.8
2	44.4	14.9	0.336	44.9	12.7	0.283	85.4	45.1	12.6	0.280	84.7	44.3	12.6	0.284	84.5	44.5	12.6	0.283	84.6
12	20.7	6.3	0.304	20.1	6.2	0.309	98.8	21.2	6.2	0.290	97.9	21.1	6.1	0.291	97.8	21.1	6.1	0.291	97.7
ED	431.7	42.6	0.099	425.4	38.0	0.089	89.3	433.6	37.8	0.087	88.8	429.1	37.7	0.088	88.6	429.7	37.7	0.088	88.7

B3

MOTHER TONGUE ITALIAN

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
10.83 1.66 1.11 0.38 0.73 0.79

8	461.1	55.0	0.119	407.9	34.0	0.083	61.9	413.6	33.9	0.082	61.7	409.4	33.8	0.083	61.5	410.8	33.8	0.082	61.4
9	39.6	14.1	0.355	44.4	12.6	0.283	89.3	44.5	12.4	0.279	88.4	44.4	12.4	0.279	88.0	44.1	12.4	0.280	87.9
11	47.2	17.6	0.372	48.1	15.7	0.327	89.5	49.7	15.7	0.316	89.2	48.0	15.6	0.326	89.0	48.1	15.6	0.325	89.0
2	5.9	3.5	0.590	5.6	3.4	0.603	97.0	6.6	3.4	0.551	96.8	6.2	3.4	0.549	96.7	6.3	3.4	0.540	96.8
12	3.0	2.4	0.810	3.1	2.4	0.771	99.8	2.7	2.4	0.880	99.0	2.7	2.4	0.867	99.0	2.7	2.4	0.873	98.9
ED	1,148.7	89.3	0.078	1,059.8	65.0	0.061	72.8	1,067.1	64.4	0.060	72.1	1,059.2	64.3	0.061	72.0	1,059.8	64.2	0.061	72.0

CATEGORIES AT WA LEVEL UNDER CLUSTER SAMPLING

C1

SINGLE (NEVER MARRIED)

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.42 0.74 0.61 0.60 0.36 0.02

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	2,632.6	49.9	0.019	2,646.7	23.3	0.009	46.8	2,616.4	6.1	0.002	12.3	2,636.9	16.0	0.006	32.0	2,617.2	6.1	0.002	12.2
9	2,644.1	48.5	0.018	2,606.6	17.4	0.007	35.9	2,628.0	6.7	0.003	13.7	2,610.6	13.8	0.005	28.5	2,627.4	6.6	0.003	13.6
11	2,484.9	43.0	0.017	2,499.0	12.1	0.005	28.2	2,490.0	5.5	0.002	12.7	2,500.4	9.7	0.004	22.4	2,489.9	5.4	0.002	12.6
2	1,816.6	38.2	0.021	1,795.2	14.0	0.008	36.5	1,787.0	4.7	0.003	12.4	1,795.5	12.5	0.007	32.7	1,787.2	4.7	0.003	12.3
12	1,215.5	32.2	0.027	1,235.5	8.6	0.007	26.7	1,233.0	0.0	0.000	0.0	1,235.7	7.0	0.006	21.6	1,233.0	0.0	0.000	0.0
ED	23,046.0	163.5	0.005	33,073.4	57.2	0.002	35.0	33,015.4	22.5	0.001	13.7	33,034.8	45.4	0.001	27.8	33,018.2	22.3	0.001	13.6

C2

MOTHER TONGUE ENGLISH

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
2.14 0.22 0.20 0.04 0.06 0.17

8	3,417.9	89.2	0.026	3,511.9	17.4	0.005	19.5	3,505.0	18.4	0.005	20.6	3,511.7	17.1	0.005	19.2	3,509.6	17.1	0.005	19.2
9	4,212.9	77.3	0.018	4,125.7	19.5	0.005	25.2	4,133.4	20.0	0.005	25.8	4,126.1	19.2	0.005	24.9	4,128.0	19.3	0.005	25.0
11	4,247.6	64.0	0.015	4,201.9	18.1	0.004	28.2	4,191.2	18.8	0.004	29.4	4,200.1	17.9	0.004	27.9	4,198.8	17.9	0.004	27.9
2	3,156.9	56.2	0.018	3,188.2	15.2	0.005	27.0	3,187.3	15.4	0.005	27.4	3,188.2	15.1	0.005	26.8	3,187.1	15.0	0.005	26.7
12	1,960.8	52.3	0.027	2,009.3	15.5	0.008	29.6	2,008.7	15.3	0.008	29.3	2,011.0	15.2	0.008	29.1	2,010.8	15.2	0.008	29.0
ED	43,295.7	279.9	0.006	43,314.4	74.8	0.002	26.7	43,309.5	76.4	0.002	27.3	43,320.7	73.8	0.002	26.4	43,319.9	73.7	0.002	26.3

C3

MOTHER TONGUE FRENCH

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
3.81 0.35 0.33 0.05 0.10 0.29

8	1,377.3	80.3	0.058	1,362.0	0.0	0.000	0.0	1,362.5	4.9	0.004	6.1	1,362.0	0.0	0.000	0.0	1,362.3	0.7	0.001	0.8
9	913.9	74.7	0.082	984.7	9.6	0.010	12.9	976.3	11.2	0.011	15.0	985.0	9.7	0.010	12.9	983.4	9.7	0.010	13.0
11	480.5	59.3	0.123	519.8	10.5	0.020	17.7	524.4	11.3	0.022	19.1	521.0	10.5	0.020	17.7	521.4	10.5	0.020	17.7
2	479.3	52.1	0.109	446.2	12.5	0.028	24.1	444.7	12.9	0.029	24.7	444.7	12.5	0.028	23.9	445.0	12.5	0.028	24.0
12	414.0	52.1	0.126	365.2	9.1	0.025	17.6	366.3	9.3	0.025	17.8	364.0	8.9	0.024	17.1	364.2	8.9	0.024	17.0
ED	20,061.8	269.4	0.013	20,200.2	38.6	0.002	14.3	20,181.2	42.4	0.002	15.8	20,189.7	38.0	0.002	14.1	20,188.5	38.1	0.002	14.1

CATEGORIES AT EA LEVEL UNDER CLUSTER SAMPLING

A1

EMPLOYED

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.78 0.57 0.20 0.13 0.47 0.10

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
2	721.9	44.5	0.062	732.4	42.6	0.058	95.7	737.0	41.6	0.056	93.4	736.4	41.9	0.057	94.0	737.3	41.4	0.056	93.1
1	639.1	44.7	0.070	652.6	41.8	0.064	93.5	653.9	40.8	0.062	91.3	652.1	40.7	0.062	91.0	652.9	40.7	0.062	91.2
3	201.2	26.0	0.129	205.1	25.2	0.123	96.9	201.9	24.4	0.121	94.0	202.7	24.5	0.121	94.2	202.1	24.4	0.121	94.0

A4

NOT MOVED IN LAST 5 YEARS

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.89 0.96 0.57 0.66 0.44 0.14

2	281.1	38.7	0.138	285.1	38.0	0.133	98.2	287.9	37.2	0.129	96.3	286.0	37.3	0.130	96.4	287.8	37.2	0.129	96.3
1	443.8	49.7	0.112	452.7	46.7	0.103	94.0	457.0	44.7	0.098	90.0	454.7	44.6	0.098	89.6	457.7	44.4	0.097	89.4
3	47.3	18.6	0.392	49.2	18.3	0.372	98.6	49.0	18.2	0.371	97.8	48.4	18.2	0.376	97.8	48.7	18.2	0.373	97.7

A6

HIGHEST GRADE IS 1 TO 8

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.43 1.65 0.38 0.61 1.36 0.30

2	121.3	18.8	0.155	120.0	17.8	0.149	95.0	123.9	17.1	0.138	91.3	123.1	17.1	0.139	91.2	124.1	17.1	0.138	91.2
1	213.0	26.0	0.122	215.3	25.8	0.120	99.0	214.6	25.0	0.116	95.9	214.2	24.9	0.116	95.6	215.3	24.9	0.116	95.7
3	44.4	12.2	0.275	42.0	11.6	0.277	95.4	41.1	11.4	0.277	93.3	41.1	11.4	0.277	93.4	41.4	11.4	0.275	93.3

B2

MOTHER TONGUE GERMAN

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.60 2.99 2.61 0.46 1.52 2.23

2	29.6	12.5	0.424	29.6	11.1	0.375	88.5	30.4	11.0	0.361	87.6	29.5	11.0	0.371	87.4	29.7	11.0	0.370	87.5
1	14.8	8.2	0.552	15.3	7.8	0.506	95.1	14.7	7.7	0.523	94.4	14.8	7.7	0.519	94.3	14.9	7.7	0.517	94.3
3	0.0	0.0	-	0.0	0.0	-	-	0.0	0.0	-	-	0.0	0.0	-	-	0.0	0.0	-	-

C1

SINGLE (NEVER MARRIED)

C(0-1) C(1-2) C(2-3) C(3-4) C(1-3) C(2-4)
1.28 1.18 0.48 0.46 0.94 0.03

2	792.9	54.6	0.069	779.8	51.9	0.067	95.1	776.2	51.5	0.066	94.4	779.7	51.5	0.066	94.3	776.1	51.5	0.066	94.4
1	772.2	55.5	0.072	769.7	52.3	0.068	94.1	758.6	50.9	0.067	91.6	762.4	51.4	0.067	92.6	758.9	50.9	0.067	91.6
3	251.5	37.0	0.147	245.7	35.7	0.145	96.3	252.2	35.1	0.139	94.9	253.4	35.1	0.138	94.8	252.2	35.1	0.139	94.8

APPENDIX 1

Table PF1: Cross-classification Table For Population and Family Weight				FAMILY MEMBERS																		NON FAMILY MEMBERS																							
				Head in a Husband-Wife Family						Head In A One Parent Family			Spouse in a Husband-Wife Family						Children in the Families						Head of Household			Related Mem- ber of Hhld			Unrelated Mem- ber of Hhld														
				Without Children			With Children			Moth. Tongue	Eng	Fr	Oth	Moth. Tongue	Eng	Fr	Oth	Moth. Tongue	Eng	Fr	Oth	Moth. Tongue	Eng	Fr	Oth	Moth. Tongue	Eng	Fr	Oth	Moth. Tongue	Eng	Fr	Oth												
				Moth. Tongue			Moth. Tongue																											Moth. Tongue			Moth. Tongue			Moth. Tongue			Moth. Tongue		
				Eng	Fr	Oth	Eng	Fr	Oth																									Eng	Fr	Oth	Eng	Fr	Oth	Eng	Fr	Oth	Eng	Fr	Oth
Sex	Marital Status	Age	No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30												
M A L E	Never Married	0-4	1																																										
		5-9	2																																										
		10-14	3																																										
		15-19	4																																										
		20-24	5																																										
		25-34	6																																										
		35-44	7																																										
		45-54	8																																										
		55-64	9																																										
		65+	10																																										
	Ever Married	0-19	11																																										
		20-24	12																																										
		25-34	13																																										
		35-44	14																																										
		45-54	15																																										
		55-64	16																																										
		65+	17																																										
F E M A L E	Never Married	0-4	18																																										
		5-9	19																																										
		10-14	20																																										
		15-19	21																																										
		20-24	22																																										
		25-34	23																																										
		35-44	24																																										
		45-54	25																																										
		55-64	26																																										
		65+	27																																										
	Ever Married	0-19	28																																										
		20-24	29																																										
		25-34	30																																										
		35-44	31																																										
		45-54	32																																										
		55-64	33																																										

1 - Some children less than 6 years

2 - No children less than 6 years

LIST OF REFEREES/LISTE DES CRITIQUES

The Editorial Board wish to thank the following persons who have served as referees during the past year.

Le comité de rédaction désire remercier les personnes suivantes, qui ont bien voulu faire la critique des articles présentés au cours de l'année dernière.

G. Brackstone
N. Chinnappa
D. Dodds
P. Ghangurde
J-F. Gosselin
G.B. Gray
H. Hofmann
T.M. Jeays
M. Lawes
F. Mayda
M.S. Nargundkar
C. Patrick
D. Serrurier
K.P. Srinath
L. Swain
P.F. Timmons
J. Tourigny
V. Tremblay

DATE DUE
DATE DE RETOUR

MAR 18 1982

109

Julia

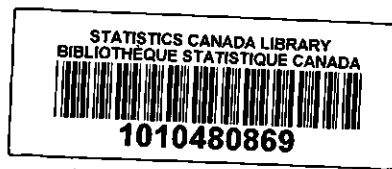
LOWE-MARTIN No. 1137

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

C O N T E N T S

Adjustment for Non-Response in Surveys R. PLATEK, M.P. SINGH and V. TREMBLAY	1
Some Estimators of Population Totals from Simple Random Samples Containing Large Units M.A. HIDIROGLOU and K.P. SRINATH	25
Redesign of the Sour Cherry, Peach and Grape Objective Yield Surveys in the Niagara Peninsula G. DAVIDSON	38
An Investigation of the Properties of Raking Ratio Estimators: I With Simple Random Sampling H.R. ARORA and G.J. BRACKSTONE	62
The Response Incentives Experiment in the Canadian Labour Force Survey A.R. GOWER	84
Estimation of Variance in Multi-Stage Sampling K.P. SRINATH and M.A. HIDIROGLOU	104
Yukon Labour Force Survey - Methodology of a Feasibility Study D. ROYCE	114



C.3

1976

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1976

Vol. 2

No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

C O N T E N T S

Double Frame Ontario Pilot Hog Surveys D. SERRURIER and J.E. PHILLIPS	138
An Approximation To The Inverse Moments Of The Hypergeometric Distribution M. HIDIROGLOU	171
The Methodology Of The 1971 Reverse Record Check J.-F. GOSSELIN	180
The Estimation Of Total Variance In The 1976 Census G.J. BRACKSTONE and C.J. HILL	195
1974 Survey Of Housing Units H. HOFMANN	209
Stratification Index: Methodology and Analysis G.B. GRAY	233
Rotation Group Biases In The Old And New Labour Force Survey R. TESSIER	256