Statistics Canada
Statistique Canada

SURVEY METHODOLOGY

TECHNIQUES D'ENQUÊTE

December - 1978 - Décembre

VOLUME 4

NUMBER 2 - NUMÉRO 2

A Journal produced by
Statistical Services Field
Statistics Canada

Publié par Le Secteur des
Services Statistiques,
Statistique Canada

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1978 Vol. 4 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

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1978 Vol. 4 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

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 de 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 1978 Vol. 4 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 libre de les faire paraître dans d'autres revues statistiques.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE VOL. 4 NO. 2

THE EVOLUTION OF A NATIONAL STATISTICAL AGENCY

J. Spear[1]

A chronological account of the development of Canada's central statistical agency is presented in this paper with emphasis on the importance of adapting the organization to the demands of the times.

To study the history of statistical progress in Canada since 1666 is to read of successes and failures as men of statistical vision attempted to respond to need as they saw it (see [1]). They sought to define appropriate mandates and to put organizations in place to carry them out. It was a long drawn out, difficult process.

Canada as a nation was born in 1867 as a largely rural society based on agriculture and other primary industry. In a little more than 100 years Canada has experienced very rapid social, industrial and political development. Paralleling this development has been the evolution of a supporting statistical system culminating in a centralized statistical agency as the hub of the system.

We have to thank the Deputy Minister of the Department of Agriculture, Dr. J.C. Taché, for appraising the statistical scene in Canada in 1864, and preparing a report which formed the base for the references to statistics in the British North America Act. One item in the BNA Act profoundly affected the future state of Canadian statistics. It named "statistics" as among matters under the sole control of the federal authority and provided the legal setting for a federal statistical agency.

---

[1] J. Spear, Institutions and Agriculture Statistics Field, Statistics Canada.

It recognized the impossibility of a federal system without statistics pertaining to the whole society, that is to say to those parts under provincial, as well as those under dominion jurisdiction. It appears that our constitutionalists intended that there should be provincial statistics, but that the two series, federal and provincial, should make one entity. Confederation greatly dispersed administrative authority; instead of a single government responsible for everything, there were now two governments operating in each province, each with its controls. Among Taché's recommendations in 1864 for the reform of the statistical process was a consulting board consisting of the heads of departments and agencies. They were to concern themselves with the best method of collecting, analyzing and arranging statistics connected with the concerns of their respective departments.

The idea of a board was approved and affiliation of Deputy Ministers as consulting auxiliaries was debated, and somehow the whole affiliation idea got lost in the process of setting up the new machinery of confederation. Notwithstanding the key item in Section 91 of the BNA Act, a practice of statistical decentralization was adopted, more through the practice of osmosis than by intent.

Rather surprisingly, from today's perspective, as early as 1871 there was evidence of pressure from the provinces for a central statistical system. The Registrar General for Ontario was very unhappy with the state of statistical collection and publishing in Canada, and recommended that there should be a statistical bureau in Ottawa to develop a uniform system of statistics covering all of Canada.

1879 was a key date for the Canadian statistical system. It was in this year that a new census and statistical Act was passed. This Act provided for the taking of a census in 1881, and "at the beginning of every 10th year thereafter". As well as changing some of the detail regarding censuses, it added a very significant clause, namely, Section 28, under

the heading of statistics. This provided that the Minister of Agriculture should from time to time make rules and regulations "for the purpose of collecting, abstracting, tabulating and publishing vital, agricultural, commercial, criminal and other statistics". It is also significant that the Act refers to "arrangements with the provinces and other public offices in order to obtain these statistics external to the census". The Act provided for "special investigations" and so for the very first time the requirement for national statistics as a totality was envisioned and written into the legislation. The Act, however, did not contain any directive as to the machinery to be used, nor set out the procedures by which the statistical arrangements with the provinces would be carried out. In fact, it took nearly another 40 years before an effective centralized statistical office was put in place. During these years, the provinces were left severely alone as far as their statistical progress was concerned. Dominion government statistics were departmentalized, and the policy of the government during this period was, despite Section 91 of the BNA Act, that statistics be produced through the administrative motions of government departments. If additional statistics were needed as a guide to policy then the relevant department produced them if it had the powers. If the department did not have the powers, then it sought them.

Many serious problems existed. For example, Canada's production figures at the time of confederation, and for many years after, were entirely inadequate. One of the main reasons was that at confederation the control of production had passed almost wholly to the provinces. Disorder ruled; 9 separate authorities were responsible for one issue which was in turn split into many parts for which a dozen different authorities were responsible within each province. With statistical chaos ruling between the provinces and with the census chronically out of date - the inevitable happened. Certain federal departments (Mines, Forestry, Agriculture) invaded the statistical field. As another example, under the BNA Act, agriculture was a field which was covered both by dominion and provincial jurisdictions but without central co-ordination statistical confusion

reigned. Most of the provinces had set up annual and even monthly crop and livestock reporting soon after the close of the century, but the figures were all at cross purposes, coinciding neither as to time, definition, nor general methods.

By 1901, the demand for a more adequate system of statistics had become more and more insistent. In this year a report by the Minister of Agriculture to Cabinet included the following, "The aim and intention of the several Acts is the establishment of a bureau of statistics which shall form part of the Department of Agriculture, and in which will be consolidated the general statistics of the country, the officers in charge of which shall have every facility necessary to enable them to obtain the needed statistics from the several departments of the federal government, the provincial governments, or by special statistical investigations. The revised statutes give the necessary legislative authority to enable the department to join the provincial authorities in making arrangements for the better collection of different kinds of statistics, without limiting the power of any department to enter upon provincial fields not worked by provincial organizations. By a combination of forces, the results would be more satisfactory than by any other system that would be originated by federal authorities. Instead of clashing statistics there could be statistics having a joint approval". Another five years went by before there was any action on these proposals.

In 1906, Sydney Fisher, the Minister of Agriculture, attempted to put an end to statistical fragmentation in Canada when he made the Census and Statistics Office permanent. The Act of 1906 supported this centralization in Section 19 and Section 23, but once again while the realization of the inadequacies was apparent, and legal authority was granted to permit their correction, implementation took a long time. The implementation stage appears to be a perennial government problem.

At long last, however, influences were at work which were to lead to radical changes. First the inadequacies of Canadian statistics had reached the state where they were handicapping governments at all levels, business and all users of statistics. Sir George Foster, then Minister of Trade and Commerce attended a Dominion Royal Commission in 1911 to take stock of the natural resources of the British Empire and to see if a greater exchange of products could be encouraged. One of the first things he discovered was the unsatisfactory state of statistics as the background for their investigations. The statistics they required were either non-existent or had been developed along very different lines in the countries of the empire. As a result of these inadequacies Sir George Foster decided that the Canadian statistical state of affairs had to be put in order.

Sir George Foster's first move in 1912 was to obtain the transfer of the Census and Statistics Office from the Department of Agriculture to the Department of Trade and Commerce where it would be under his immediate control. This explains why the central statistical office has been located in or linked with this Department ever since. The next step was to arrange for a thorough investigation of the existing statistical environment and product and Sir George Foster was responsible for setting up the inter-departmental Commission on the Official Statistics of Canada to carry this out. The commission included representatives from the Civil Service Commission, the Census and Statistics Office, the Department of Trade and Commerce, the Department of Labour and the Department of Customs. The representative from the Department of Labour was Robert Hamilton Coats, who was to become the chief architect of the Commission report and the developments which arose from it.

The Commission was required to report on "a comprehensive system of general statistics adequate to the necessities of the country in keeping with the demands of the time". The Commission was authorized to communicate with the various governments with a view to ascertaining "what branches of

statistical work are being conducted by the provinces and to what extent these may assist in or duplicate work being done by the dominion, looking to a system of co-operation". The Commission was also authorized to enquire into the statistical work being carried out in various departments. Coats describes the investigation thus, "As for existing conditions, a pilgrimage from department to department revealed them worse than had been suspected. From any general standpoint notwithstanding excellence in spots, imperfections of method, restrictions of outlook, lack of unity and co-ordination were found rampant". The report pointed particularly to "the lack of coherence and common purpose in the body of Canadian statistics as a whole" and concluded that "Each department or branch charged directly or indirectly with statistical investigation, has concerned itself primarily with the immediate purpose only in view. The effect statistically has been to inculcate routine and a neglect of opportunities".

Coats made three important recommendations resulting from the Commission's investigations. Firstly, that a central statistical office be created to organize, in co-operation with the several departments concerned, the statistical work undertaken by the Dominion Government. Secondly, he recommended consultation with provincial governments starting with an Interprovincial Conference on Statistics, and thirdly, that an Interdepartmental Statistical Committee be formed consisting of representatives selected from the statistical office, and from other departments engaged in the collection of statistics. He proposed that this Interdepartmental Committee be advisory and deliberative rather than executive with the following objectives:

- to prevent duplication and conflicting results;
- to better adapt the statistical material of one branch to the needs of another;
- to establish uniformity in definitions and methods;

- to ensure expansion and development along appropriate lines, including the suggesting of new work and its allocation to those branches best equipped to carry it out;
- to supervise the various statistical publications with a view to the proper distribution of statistical information.

The first step in carrying out the recommendations of the report of the Commission was taken on June 19, 1915 by the creation of the Office of the Dominion Statistician and Dr. Coats was appointed Dominion Statistician. The Commission report contained a chapter on statistical organization, and that chapter contained a key phrase repeated by Coats many times throughout his term of office. It was evidently at the core of his thinking.

> "The object of this organization should be to co-ordinate the statistics of Canada under a single comprehensive scheme and so to extend them that they meet the present needs of the country and follow the probable course of its development ... The object of such a reorganization should be primarily to constitute "a central thinking office" on the subject of statistics in Canada".

The translation of the Commission's suggestions into a series of implementation plans led to another major milestone in Canada's statistical progress: an Act establishing the Dominion Bureau of Statistics - the Statistics Act in 1918. Structurally the Act was a consolidation of previous statistical legislation of the dominion government. The Dominion Bureau of Statistics was charged with the general administration of the Act. Its duties were, "to collect, abstract, compile and publish statistical information relative to the commercial, industrial, social, economic and general activities and conditions of the people".

Particularly significant was the provision for collaboration with other departments. It was specified that where statistics originate as by-products of departmental administration for their own accounting purposes, they should, through consultation with the bureau, also conform to general

statistical needs. Thus, the right of enquiry possessed by the government for different purposes could be used to the best statistical advantage. This right of enquiry is conferred on departments having executive control in specific fields in order to exercise that control and upon the Dominion Bureau of Statistics for informational purposes. It was clearly set out at that time that where two sets of powers are exercised in parallel, they are to be organized in co-operation for statistical purposes. The right to collect all other statistics was invested in the Bureau.

In order to define the principle explicitly and to facilitate satisfactory interdepartmental arrangements, an Order-in-Council under the Act was passed on October 12, 1918. Extracts from this Order-in-Council are worth quoting:

(1) That all purely statistical investigations relative to the commercial, industrial, social, economic and general activities of the people shall be carried in the Dominion Bureau of Statistics.

(2) That with respect to such records of any department or branch of the Public Service as are of a statistical character, the Dominion Statistician shall confer with the head of such department or branch with a view to arranging that such records be collected, and compiled insofar as possible in conformity with the methods and organization established in the bureau, the object of such arrangement being the prevention of overlapping, the increase of comparability and the utilization of departmental organizations in the best manner for statistical ends.

(3) That after such conference, the Dominion Statistician shall, at an early a date as practicable, prepare a report on the statistical work of each department or branch of the Public Service, with a view to carrying out the above requirements, such report to be submitted to the Council for approval with a view to effecting a permanent arrangement for dealing with the statistics collected by the government, and

(4) To further promote efficiency and economy, all statistical compilations for the government be carried out insofar as practicable by mechanical appliances and for this purpose use be made of the machines installed in the Bureau of Statistics.

The Act also provided the machinery for provincial co-operation by a clause enabling the bureau to enter into arrangements for the collection and supplying of statistical data through provincial departments or offices.

During the years following the passage of the Statistics Act in 1918, statistical work in the federal government was transferred from various departments to the bureau by the authority of Orders-in-Council. It was also the period during which dominion-provincial co-operation was established to co-ordinate statistical work, for the primary purpose of ensuring unified practices and eliminating duplication. The first annual report of the Dominion Statistician contained this sentence - "In addition, there has been created what is frequently called a central "thinking office" in statistics, continuously in touch with the general conditions and the line of probable development". This was a clear statement of the overall criteria for accountability which the first Dominion Statistician set for the organization.

Once the early phases of the bureau organization were worked out, including the establishment of its eleven branches covering subject matter fields, the way was clear for the development of new statistical series and the expansion of existing ones. High unemployment, poverty and subsequent human suffering during the depression years brought about demands for revoluntionary changes in social services. In comparison with many other countries, Canada had made small progress in establishing a social security programme. The distribution of powers and jurisdictions under the British North America Act presented many difficulties, and these would not be investigated until the Rowell-Sirois Commission of 1935.

The Commission discovered that to bring about the social and economic reforms necessary, the redistribution of national income through the medium of such measures as old age pensions, family allowances and health insurance, was essential. The time was right for the dominion government to assume responsibility for problems of economic need arising out of unemployment and agricultural distress; it was prepared to accept this responsibility subject to a general revision of intergovernmental/ financial relationships. The Rowell-Sirois Commission was set up as the investigatory body and as a result of its work there was a prime need for improved statistics on finance. This led to a series of dominion-provincial conferences on the public finance statistics of the provinces and municipalities and a further strengthening of the Bureau's staff.

By 1939 it could be said that the broad framework of a unified and co-ordinated system of national statistics for Canada had been established. World War II brought an unprecedented demand for statistics. For example, the cost of living index became a key figure; employment statistics had to be expanded to meet the requirements of war departments; monthly payroll statistics were added to the bureau's employment series.

In 1942, Robert Hamilton Coats retired from the Office of Dominion Statistician. He was the dominant figure in Canadian statistics for the first half of this century. Coats was a centralist and throughout his career as the Dominion Statistician he worked towards the goal of creating a central statistical organization. His main objective was statistical objectivity and to separate the statistician from those with administrative or political interest in the figures. He worked hard and steadily towards the goal of removing statistical units from departments and placing them within the Dominion Bureau of Statistics so that they would not be influenced by departmental or political interests and pressures. Technical standards advanced under Coats, statistical objectivity became the essential by-word and professionalism increased in every area. In every way he increased the respect for, and the integrity and value of the statistician and the statistical process. The organization under his management responded to the needs of the times.

In 1943, an interdepartmental committee was set up by the Dominion Statistician to produce the reorganization which would enable the bureau to meet the post-war statistical needs. At this point in its growth two important developments occurred which were to have a profound influence on the progress of the bureau. One was the establishment of a central research and development staff, and the other was the establishment of a sampling organization to develop probability sampling. The function of the Research and Development Division was to integrate and analyze existing statistical data and to develop a new series of economic statistics. The National Accounts were the result. Not only would these estimates constitute a basic statistical background for financial and fiscal policy but the classification of these accounts into a separate statistical summary for various sections of the economy revealed the inadequacies in existing statistics, gaps which had to be filled and defects in the integration of statistical series. Statistical sampling opened new doors for securing statistical information and permitted the exploitation of many fields of information that had previously been unobtainable except in decennial or quinquennial censuses or not at all. The most important of the sample surveys was the labour force survey first taken in 1945. The wide coverage of this sample required the setting up of Regional Offices in Halifax, Montreal, Toronto, Winnipeg and Vancouver. Later, offices were added in Edmonton, St. John's, Newfoundland and in Ottawa. In 1948, the Statistics Act was amended to ensure legislative authority for the collection of statistics by means of sampling.

Major national social welfare programmes began to emerge in the fifties. Old age pensions on a universal scale were established in 1952. Unemployment insurance, health insurance, post-secondary educational facilities and welfare expenditures all received attention which resulted in increased growth and coverage of the bureau's data base.

Changes affected the bureau through the evolving needs of the user community, but changes also resulted from Federal Government investigations into its own activities. A prime example was the investigation undertaken by what was to be known as the "Glassco Commission".

In 1960, a Royal Commission on Government Organization was established under the chairmanship of J. Grant Glassco. The Commission's mandate was "to recommend the changes therein which they considered were the best to promote efficiency, economy and improved service in the dispatch of public business". The report which resulted in 1962 contained a recommendation for increased expenditures for statistical services and considered that no other conclusion was possible if quality was to be maintained and pressing needs adequately met. The report described economic and social statistics as "essential nutrients in the regular functioning of a complex society", and emphasized the need to pursue an "integrated statistical system for social statistics as well as for economic statistics". The Commission weighted the pros and cons of a centralized vs decentralized statistical system and ruled heavily in favour of a centralized and specialized statistical agency. Interestingly, the Commissioners recommended that the bureau audit the statistical programmes of all departments and agencies and report annually to Parliament on the state of government statistical services.

A recommendation of the Glassco Commission of key importance to the agency was implemented by Order-in-Council in January of 1965. The Dominion Bureau of Statistics was designated a "department" of the Federal Government and the Dominion Statistician was assigned the status and power of a Deputy Minister. The purpose of the recommendation was to "emphasize the independence of the Dominion Statistician because of the position of trust he holds with respect to those who are required by law to report confidential information to him". In addition, even though the Dominion Statistican would act as deputy for the Minister responsible for the Department of Trade and Commerce and continue an association with this department, the move had the advantage of making the Bureau an independent departmental entity, separate from the Department of Trade and Commerce.

The report of the Royal Commission on Government Organization (the Glassco Report) was studied by bureau officials during 1964 and a number of administrative improvements were put into effect as a result. There was a notable acceleration in the statistical needs of both federal and provincial government departments and agencies. The importance of statistics in the 1960's arose from a spectacular growth in technology, increasing professional expertise in internal and in user communities and an increasing attention given to a new phase of social statistics. By 1966 there were new demands placed on the bureau in the form of the need for broad national figures and information on regions and sub-provincial areas. These demands coincided with the planning and implementation of important and far-reaching government programmes with a great deal of emphasis placed on regional development. Increased attention was directed to the possible use of administrative statistics as a more economical method of obtaining information.

An important milestone took place in 1966 with the creation of a DBS Satellite Unit within the Department of Transport to deal with air transport statistics. This was a reversal of the traditional practice of physical centralization of statistics, but it was believed that the physical proximity to the Department of Transport would ensure its effective support for the work of the Satellite, and that the supervision by bureau personnel would promote statistical efficiency and consistency. The bureau continued to experience a period of rapid growth and in order to handle this more efficiently a major reorganization was effected in 1967. A Socio-Economic Statistics Branch was put into place to deal mainly with statistics derived from or related to households and individuals; the Economic Statistics Branch covered statistics derived from business establishments; and the Financial Statistics Branch dealt with corporations. The increased importance of automation was recognized by the creation of an Operations and Systems Development Branch responsible for data processing and computer programming.

In 1971, the Dominion Bureau of Statistics officially became known as Statistics Canada, as a result of a new Statistics Act which received royal assent on February 11, 1971. The new Statistics Act resulted from a basic review of the needs of users, the growing importance of the provinces and from the experience of the bureau with previous legislation. It reflected the needs of the times. The new Statistics Act significantly reinforced the authority of Statistics Canada as the co-authority in the national statistical system. The Act provided far more legislative authority by which Statistics Canada and the provincial statistical agencies could co-ordinate and integrate their activities. The Act also provided explicit legislative authority by which Statistics Canada had access to tax returns and confirmed its access to the administrative records of other federal government departments. The changes in the Act were an important step forward for the bureau and it is a great tribute to the foresight of those who drafted earlier versions of the Act and in particular to R.H. Coats who prepared the first Act in 1918, that the basic principles of the legislation remained untouched by a comprehensive review and revision over fifty years later.

In 1972, a new Chief Statistician of Canada was named. The Chief Statistician set up a study group to identify the critical challenges facing the agency in the future. The pace of change had escalated sharply and the agency had to be re-shaped if it was to respond to changing need and remain accountable for its performance and product. To meet the ever-increasing demand for official statistics, there was an expansion in the statistical activity in federal departments and in the provinces, but the brunt of the responsibility to meet the statistical demands of the times still fell on Canada's centralized statistical agency and increased and more complex statistical demands had to be supported by changing the statistical environment. Technology had increased the capability of users to retrieve and manipulate data. Users had become more sophisticated and so the inter-action and data linkage between producers and users demanded greater attention and data

consistency. At this point in time the agency's 1973-74 budget was $73 million, almost 2 1/2 times greater than the expenditures of five years earlier. The work force consisted of 5,000 people with a core of 680 statisticians and economists.

The study group charged with re-shaping the agency and ensuring its accountability identified three critical challenges relevant at this point in time:

(1) making statistics more usable and useful;
(2) upgrading the nation's overall statistical capability over the long haul; and
(3) maintaining public support.

The process of change was escalating in another direction, however, and in the mid seventies Canada in common with most industrialized nations began to experience greatly reduced economic growth which resulted in increasing inflation and growing unemployment.

In 1975, a new Chief Statistician was charged with responsibility for the agency and he foresaw the inevitability of drastic change affecting the organization as the government moved towards a policy of fiscal restraint and zero growth in order to cope with its economic ills. In an important policy statement in July of 1976, he outlined his view of the future of the organization preparing it for an external environment which would impact on every aspect of its activities. He described a future in which the statistical system would be more visibly associated with an information industry. Statistics Canada would increasingly be viewed as only one node in the larger statistical system, albeit a dominant one, in which there would be numerous data bases connected by a common data base management system. Such a system would embody quality control with special emphasis on the production of clean microdata bases - the data capital of the system. Integration would become absolutely essential to this informational system and must be designed to be extremely adaptable to meet the diverse needs of users. This adaptability would be obtained as a result of the increased emphasis on the analytical function of the statistical system. In such an

environment, an essential requirement would be efficient control/co-ordination mechanisms. The Chief Statistician viewed it as a prime responsibility that the bureau would take the initiatives in fostering such a statistical informational system.

The changing expertise and awareness of users also demanded that the Bureau become more user oriented. The Chief Statistician designated that the program of collection would in future be related to "spheres of observation" - households, institutions, non-farm businesses and farms ... Such a conception was expected to solve many of the difficulties experienced in integrating information embodied in different surveys, as the spheres of observation would help define the primary level of integration.

Internal to the Bureau many changes would be necessary to prepare the organization to cope with this vision of future need and respond to it. The key requirements were described as follows:

- Reorganization into spheres of observation to permit economies of scale, to give impetus to integration and to make optimum use of specialization and professional skills.
- The development of the program control function concerned principally with establishing policy, setting priorities, allocating funds and evaluating performance.
- The fostering of the content and analysis function to support users' needs, define their requirements and provide expert consultants to their user communities.
- The recognition of the operations function concerned with survey design, survey operations and the generation of a clean data base as a professional activity and given its proper place of importance in the system.

In his policy paper, the Chief Statistician warned that the Bureau should prepare itself for a new statistical leadership and co-ordination role - a role which must be continually changed and modified in response to

changes in the environment in which it must operate. The Chief Statistician believed that the mid 80's would find more and more active participants in the national statistical system and that the new organization and functional separation of activities he proposed would accomplish two major objectives:

1) Create an organization adaptable to change,
2) Serve the reality of the new environment.

Or in other words meld the programs and structure of the agency so that they could form "a comprehensive system of statistics adequate to the necessities of the country in keeping with the demands of the times" to use the 1912 criteria for the Bureau's conception.

In January of 1978, a document "Statistics Canada - The Medium Term" was distributed to all main users in the user community throughout Canada. The document contained the bureau's mandate statement and a clear description of its main strategic thrusts based on the Chief Statistician's appraisal of future needs.

The most fundamental change recognized was the likelihood of zero or negative growth in statistical budgets resulting from the programme of fiscal restraint in the government. This would of course result in an intensive re-examination of programmes and priorities because of the necessity to fund new endeavours at the expense of existing programmes.

The document described environmental changes to which the agency must adapt:

- rising public concern with privacy and confidentiality
- heightened resentment of response burden
- concern about the cost of government
- criticism of the data published by Statistics Canada.

As a result of this examination of external concerns the strategic thrusts of the bureau over the next five years were identified as:

a) improved service to users,
b) reduction in response burden,
c) enhanced efficiency,
d) statistical leadership and co-ordination.

What was suggested was a gradual movement away from areas where others are able to assume the statistical responsibility or where respondent costs are high in favour of more national responsibilities and greater reliance on analysis and uses of administrative data.

The January 1978 declaration of mandate contains the following paragraph which captures succinctly the framework within which those who manage this organization and strive for statistical excellence are working.

> "The mandate, as thus set out, differs little from that of the original legislation of 1918 which first brought into being a centralized statistical agency in Canada. It is broad and not suitable as a basis for prescribing specifically what should be done at any one time. This, however, should be regarded as an advantage rather than as a drawback. Those who first drafted the mandate recognized that a generalized statement of the responsibilities assigned to Statistics Canada would give it the necessary flexibility to change, in accordance with the needs of the times, its conception of what those responsibilities mean, the relative importance to be attached to each one of them, and the means for carrying them out".

The environment within which Canada's central statistical agency must operate is an ever changing one. In the seventies and looking into the eighties, the rate of change will continually escalate. It is important to remember that worthwhile institutions have lives of their

own that continue long after those who pass through them have gone. But they only thrive if those who work for them appreciate the continuing role of the institution, understand its mandate and work to protect its integrity even as they respond, on a daily and monthly basis to the changing demands of the time. R.H. Coats, the first Dominion Statistician, was one of many such individuals. When he was appointed he described his view of the organization:

> "The object of this organization should be to co-ordinate the statistics of Canada under a single comprehensive scheme and so to extend them that they meet the present needs of the country and follow the probable course of its development ..."

Some thirty years later in 1946 his view had not changed.

> "The statistical objective is to get a good body of statistics on each and every public interest, and at the same time see that these dovetail and provide a good conspectus of the whole: there are the rooms and there is the house. An edifice of this kind is never done building".

RESUME

Cet article présente les grandes étapes de l'histoire de l'organisme statistique central du Canada; l'accent est mis sur l'adaptation de l'appareil aux exigences de l'époque.

REFERENCE

[1] Spear, J., "Historical Milestones in Canadian Statistics", November 1975, unpublished paper.

NON-RESPONSE AND IMPUTATION

R. Platek and G.B. Gray[1]

The problems of dealing with non-response at various stages of survey planning are discussed with implications for the mean square error, practicality and possible advantages and disadvantages. Conceptual issues of editing and imputation are also considered with regard to complexity and levels of imputation. The methods of imputation include weighting, duplication, and substitution of historical records. The paper includes some methodology on the bias and variance.

1. INTRODUCTION

The reliability of survey estimates is governed by many factors, one of which is the effect of missing and inconsistent or incomplete data. Any survey, whatever its nature, suffers from some non-response or responses which fail data edit procedures. The question that should be answered is "what should we do with this kind of incompleteness in the data"? One can argue, of course, that if the magnitude of deficient data is less than one percent, one might not worry about it at all. But in practice, the size of non-response is more like 10%, 15% or more, depending on the subject matter.

To disregard the effect of non-response of such size may lead to survey results of unacceptable quality and it will definitely mean that population totals could not be estimated since they would be based on partial data only. On the other hand, the reliability of averages and proportions will be affected less than that of totals by non-response and one can also argue, with some justification, that in general, the effect of non-response on national estimates will be smaller than for some sub-national levels. Nevertheless, the elimination and the reduction of the effect of non-response and invalid responses is very

[1] R. Platek, and G.B. Gray, Household Surveys Development Division, Statistics Canada.

important and it should be undertaken at various stages of survey design as well as in the field. Despite these efforts, however, some non-response and deficiencies will remain in the data and, in practically all surveys, some form of adjustment or imputation for non-response will have to be considered.

Imputation may be defined as the assignment of data to empty fields (including total non-response) or a replacement of invalid data following certain rules. There is no known unbiased method of imputing unless several assumptions are made regarding non-respondents and respondents. There is, however, some evidence that certain methods may be more efficient than others.

2. DEALING WITH NON-RESPONSE

(i) Survey Planning and Development

At the planning stage, an awareness of the effect of non-response on the Mean Square Error of survey data will undoubtedly lead to a survey design with as little non-response as possible. Consequently, one of the important factors in planning a survey is a decision on the tolerance level of non-response and an experienced survey designer can estimate fairly accurately the level of response for a particular survey that can be expected under various survey conditions. It can be argued that for some surveys when only national estimates are required and when the characteristics of non-respondents are not strikingly different from those of respondents, a non-response rate (20-30%) may be tolerated even though this will result in an increase in sampling and perhaps in response variance. The same arguments can be applied to surveys whose objective is to provide some notion about trends and proportions. However, for surveys whose estimates must be precise and are required at various sub-national levels, the non-response rate should be kept as low as 5% or less and pockets of large non-response in local areas should also be avoided.

The survey cost is another item which will affect many factors in survey development including non-response. It is important to balance the other factors against the cost so as to achieve a non-response rate sufficiently low to serve the goals of the survey. It should also be realized that within reasonable limits, it is sometimes better to accept a somewhat smaller sample than originally planned and to transfer the resources to appropriate data collection, follow-up and estimation procedures. This would be particularly advantageous if the survey designer suspects large differences between respondents and non-respondents in their characteristics.

Apart from intuition and experience which undoubtedly play an important part in survey planning and development, one can identify a number of factors which are important in the design of surveys. These factors can be classified into three groups:

Group I
- a) sample size
- b) stratification
- c) degree of clustering
- d) sample allocation
- e) method of selection

Group II
- a) sampling frame
- b) method of interviewing
- c) selection, training and control of staff
- d) length of questionnaire and wording
- e) sensitivity of questions
- f) type of area in which the survey is taken
- g) feasibility of call-backs and the number of them
- h) publicity

Group III
- a) edit and imputation
- b) estimation
- c) variance estimation and other data analysis.

All these operations certainly affect the Mean Square Error to a varying degree. It is true that in practice we often lack enough data on the effect of most of the factors. However, since not all these factors are of equal importance, an examination of the more important components of the Mean Square Error would be very helpful. Let us suppose that the Mean Square Error can be decomposed into the following components:

$$MSE = V_S + V_R + V_{CR} + (B_S + B_R)^2$$

where

V_S = sampling variance
V_R = response variance
V_{CR} = correlated response variance
B_S = sampling bias
B_R = response bias.

Sampling variance (V_S) and sampling bias (B_S) are affected by all the factors in Group I, by estimation procedures and also by the size of non-response. The larger the size of non-response, the greater the effect it has on sampling variance and bias. For example, since the sampling variance of the estimates is inversely proportional to the response rate in the case of a simple random sample, estimates based on such a sample with 80% 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 it affects 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. In considering non-response it has to be taken in account that a reduction of non-response in the field does not necessarily ensure a reduction in bias. In fact, if the procedures for the reduction of non-response are not well thought out and appropriately executed, the bias may not be reduced and could even be increased.

In some surveys, survey conditions may affect the sampling variance and sampling bias. For example, the wording of the questionnaire and/or the training of the interviewers may operate in such a manner that legitimate extreme values are eliminated. A low sampling variance but a high sampling bias may result. The artificially low sampling variance may occur because the variance between units of the expected responses without extreme values will be lower than the variance between the true values with the extreme values. The extreme values on opposite sides of the mean value will not necessarily balance so that a high sampling bias could result. Consequently, the survey conditions may affect sampling variance and sampling bias.

Non-sampling components of Mean Square Error (V_R, V_{CR}, B_R^2) which also include non-response are affected to a varying degree by all the factors in Group II. In addition, the Mean Square Error is also affected by some factors in Group I. For example, clustering may affect the correlated variance in much the same way as it affects sampling variance since households in clusters may produce higher correlations in response errors than households further apart. Since the estimate is a function of the observed values, which in turn are subject to non-sampling error, and since each distinct estimation procedure involves a different function, then the non-sampling variance will also be affected by the estimation procedure.

(ii) Data Collection Stage

Non-response can be reduced by persistent efforts of interviewers and by motivation of non-respondents to become respondents. The persistent efforts are usually in the form of repeated attempts to contact a respondent and to gather information about him or her. There is a point beyond which it is impossible to attempt further callbacks, either because the survey is to be completed by a certain date or because there are not sufficient

funds. In the case of telephone interviewing, the cost is only that of repeated telephone attempts and with mail surveys that of subsequent reminders. However, in the case of personal interviews where, for reasons of cost, the sample must usually be clustered to minimize travelling time and distance between successive calls, repeated callbacks often result in a greater distance between households and the cost per unit may become unacceptably high, without any reductions in the variance.

Further, if the probability of non-response were the same for each unit, the non-respondents become a random subsample of the full sample and there would be no non-response bias in the estimate (apart from a ratio estimate bias) when the data are further weighted by the inverse of the response rate. A slight ratio estimate bias may result because of the variation in the respondent sample. Since, in the majority of cases the probability of non-response is not known, every effort should be made to minimize the size of non-response. However, even if we did know the probability of non-response, there may still be response bias in the estimate based on the subsample just as would be present if there was no non-response.

Another major component of non-response is that of refusals and these can only be prevented, in many instances, by motivating them to respond. However, it is possible that those respondents who were initially reluctant to respond may commit larger response errors than those who were willing to co-operate so that while we have reduced the imputation error ${}^{NR}\varepsilon_i$, we may have increased the response error ${}^{R}\varepsilon_i$ (Platek, Singh and Tremblay [7]). Just to convert every refusal into a respondent may therefore lead to a false sense of security with respect to the validity of the responses. A well-trained interviewer will certainly succeed in motivating more refusals to respond and in obtaining more reliable responses than a poorly-trained interviewer.

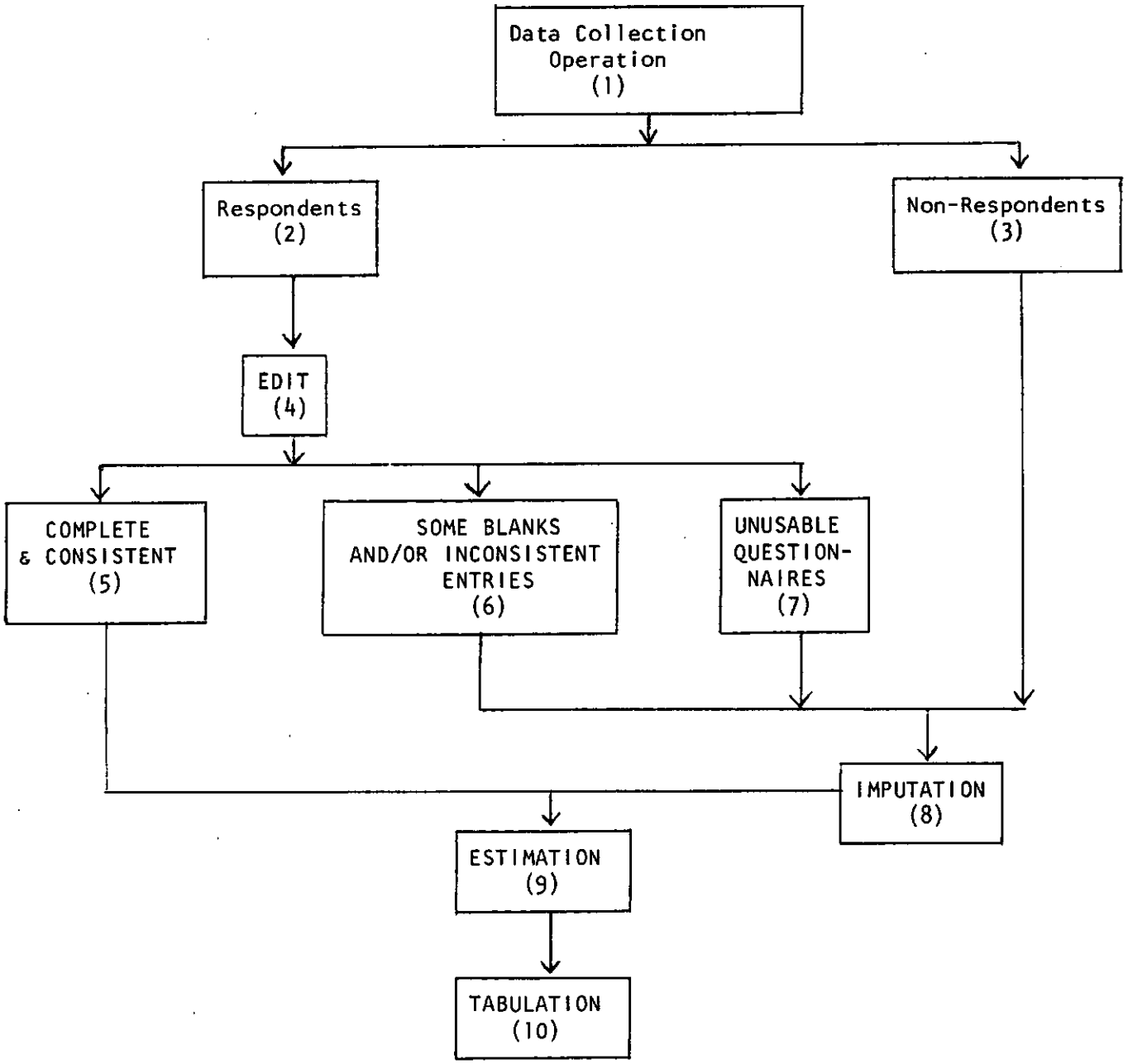
One method of dealing with non-response at the data collection stage is to substitute other previously unselected units in the field; for example, a next-door neighbour. Unfortunately, this would lead to a sampling bias. While any unit may be selected with known probability according to the sample design, substitution of other previously unselected respondents to replace unco-operative respondents in some uncontrolled manner, or even in a controlled manner, will alter the inclusion probabilities to such an extent that they cannot even be calculated. While a sampling bias of unknown magnitude would exist (since the selection probabilities are unknown for several reasons), the sampling variance may be reduced because of an increase in the effective sample size but there would probably be no reduction in the response error or the non-response bias. Even if the inclusion probabilities could be calculated, the non-response bias would remain since the unco-operative units essentially have no chance of inclusion.

In addition to callbacks or substitution of units in the field, interviewers may apply (i) double sampling (selecting a subsample of non-respondents and making an intensive effort to obtain responses from these units, or (ii) Politz scheme (considering "best time to call" as one of the weighting groups). These schemes are also expensive and must be carefully planned if they are to be used to tackle non-response.

3. TYPES OF RESPONSES AND CONCEPTUAL ISSUES OF IMPUTATION

As the information flows from data collection to tabulation, the various types of responses can be identified and are presented as follows in Chart I.

Chart 1: Flow Chart Pertaining to Each Sampled Unit



This is, of course, a highly simplified diagram of the process and it is produced only for the purpose of the discussion of this paper.

Looking at Chart 1, two of the three groups following the edit stage require some action prior to estimation. These are the unusable questionnaires and the questionnaires containing some blanks and/or inconsistent entries. The unusable questionnaires can be classified as total non-response or they can be associated with the respondent households with some blank or inconsistent entries. There remain two groups that require some attention. The first group consists of blank and/or inconsistent responses, the second group consists of non-respondents. Non-respondents (at least in household surveys as opposed to the census) are usually weighted up in some manner. The deficient questionnaires, on the other hand, fall into two categories such as inconsistent entries or illegitimate blanks.

The inconsistent entries can be either logical impossibilities or they can be plausible but highly unlikely. It seems natural that if the entries are logical impossibilities and they can be detected as such, they ought to be adjusted even though they may not affect the data to any great extent. The adjustment would save a great deal of embarrassment on the part of subject matter analysts associated with the published reports.

In the case of plausible but highly unlikely entries, one is faced with a difficult choice between remaining with observations in an unnatural distribution or removing the extreme values of the distribution which may actually represent the real life situation. Ideally, one ought to opt for one or the other choice on the basis of experience with error mechanisms and the nature of the substantive distribution based on the knowledge of subject matter. In any case, one has to be able to identify the problem cases, i.e., one has to have suitable edit rules whenever one encounters impossible or highly unlikely events and a method of dealing with them.

There is a fundamental distinction between editing and imputation. Let us consider the set of all possible code combinations on a

questionnaire. Editing can be defined as the division of this set into two mutually exclusive subsets: those combinations which are judged acceptable and those which are unacceptable, the latter including questionnaires with invalid blanks and inconsistent entries. Thus, editing is basically a diagnosis and operationally it must be defined by a set of rules. Imputation, on the other hand, is more in the nature of a treatment of data, although the two clearly interact.

As far as editing is concerned, the detection of logically impossible entries and invalid blanks presents no conceptual problems and with respect to the detection of inconsistencies, there are a number of options available. For example, one can compare pairs of fields and decide that the two are inconsistent and hence, one of them has to be changed. One can continue this procedure by comparing some other pairs of fields (or three fields at a time). Having detected a particular inconsistency, one may either impute immediately one of the fields involved to make these fields consistent with one another, or else complete the entire edit process before imputation begins. However, by looking at two or three fields at a time, one does not take into account all the possibilities. For example, if one makes all combinations of, say two or three fields consistent with one another, it does not mean that the whole record will be consistent. A system which has been developed in Statistics Canada is based on the approach that identifies all inconsistencies before any corrective action is taken. Then, in the face of all known inconsistencies between the fields of the given record, together with all the logical impossibilities and invalid blanks, one decides which field or set of fields, if corrected, would remove all the inconsistencies in the whole record.

Having determined which fields are going to be changed, the next step is, of course, to carry out imputation for them. The simplest situation occurs when there is only one possible value which can be imputed for that field in such a way that after the imputation the record will be consistent. Sometimes, there may be more than one value which would

make the record consistent. If this is the case, one would choose a particular value which is more predominant in the field or more plausible. A good example of this kind can be found in the Labour Force Survey where in the fall to spring months, for 15 and 16 year old persons, if there is no Labour Force characteristic entered, one imputes that they are "attending school", although it is not at all impossible they do not attend school. So long as the proportion of such cases is sufficiently small, the effect of this will be a slight increase in bias. At the same time, there will be some reduction in variance and the added advantage of the reduction of complexity in imputing.

In other situations where one could reasonably impute a whole range of values, one needs some other criteria. One possible criterion would be to minimize the mean square error of the resulting estimates. The question, however, arises, the mean square error of which estimates? With the continuously increasing demand for micro data tabulated in a number of different and unforeseen ways one really does not know which mean square error one ought to minimize. Furthermore, one would not know all the kinds of aggregates to which a particular record may contribute in different kinds of tabulations. Consequently, one would like to use some other criterion which would produce the most appropriate entry for a field in a particular record in relation to the other information in the record. In other words, how can one best predict the value of one field on the basis of knowing the other fields on the record. A good example of this kind of imputation is the use of previous month's data in the Labour Force Survey: for a particular person, one could hardly find a better imputed value, particularly in those cases where demographic characteristics change slowly. If one does not have information based on the past, the best imputed value may be the result of some sort of regression equation. For some household surveys, however, the application of regression is somewhat restricted due to the qualitative nature of variables. Consequently, one may adopt as a reasonable criterion that

the distribution after imputation should remain as close as possible to the distribution prior to imputation with respect to marginal distributions or preferably, if it is possible, with respect to joint distributions of all the variables to be imputed.

In most cases, to impute for non-response at the micro level as opposed to some aggregate level is mainly justified because of the lack of advanced knowledge as to the kind of aggregates that will be produced from the micro data file. However, in some situations where one knows one can limit the tabulation requirements in advance, imputing at the individual level may not always be necessary. This notably applies to surveys based on area samples where the primary sampling units are not likely to be split up in any subsequent tabulations. In this case, one can hardly do better in terms of the mean square error of any of the possible aggregates that one will produce, but to impute the average of that primary sampling unit[1].

4. PROCESSING AND ESTIMATION

One of the most common procedures for accounting for non-response at the processing and estimation stage is that of the design-dependent balancing area, in which the weights are further inflated by the inverse of the response rate. In a balancing area b, an estimate of the characteristic total is given by $\hat{X}_b = \sum_{i=1}^{n_b} x_i/\Pi_i$, where x_i is the response, Π_i = the inclusion probability, and n_b is the sample size in the balancing areas. If only m_b units respond, then the weight Π_i^{-1} is further inflated by the inverse of the response rate, m_b/n_b, i.e. by the factor n_b/m_b.

The balancing areas should, preferably, be determined at the planning stage rather than at the processing stage and they could be individual strata, groups of strata, a province, primary sampling unit, or a

---

[1] While the weight adjustment at the PSU level is justified for complete non-response, it would be inappropriate for either partial non-response or fields whose entries have been rejected on account of editing. If one carried out weighting at the individual field level, one could not properly cross-tabulate the data since records would have more than one weight.

cluster. The choice of balancing area is quite crucial since the non-response rates and the bias may differ from area to area.

An important methodological problem, for example, is to determine an optimum or in some way appropriate size of balancing area where "appropriate" refers to a proper size to ensure a sufficient response rate in order to prevent excessive weights and at the same time ensure the advantages due to the measures of homogeneity to help reduce the non-response bias. It can readily be shown that weight inflation of all the records in a balancing area to compensate for non-respondents is equivalent to the substitution of the mean values of all the weighted respondents to each non-respondent in the area. If a characteristic has a high measure of homogeneity (increasing with decreasing size of area), then weighting (or substitution of mean value) in small areas vs. large areas would tend to result in mean values that are more similar to the actual characteristic value of the non-respondent than would be the case in larger areas. Thus, the non-response bias would tend to be lower in the case of small balancing areas than in the case of large balancing units. What about the variance? As balancing areas become smaller, the weight inflation becomes more unstable as the variation in response rates becomes more unstable among many small balancing areas as opposed to a few large balancing areas and the instability of the weight would tend to increase the variance. Clearly, there is some trade-off on the size of the balancing area between small areas to take advantage of the measure of homogeneity and large areas to ensure stability in the weight adjustments. The possible extreme values of the sizes of balancing areas are the whole sample at the upper end and a size of '1' at the lower end. However, in the latter case, one is faced with the problem of what should be done if that unit fails to respond. Instead of substitution of the mean value, one would have to resort to regression analysis or superpopulation models (an entirely different approach to substitution) or else employ historic values.

The choice of the size of balancing area depends not only on the measure of homogeneity but also on the sample design, the sample size and the response rate. Surveys with low response rates would require larger balancing units than those with high response rates. One could utilize small balancing areas and adopt some procedure of collapsing them until the response rate reaches some respectable level (not too much below the overall response rate) so as to minimize the instability of the weight. The collapsing of balancing areas however adds a complex dimension to the variance estimation since one would have to consider the probabilities of collapsing 1, 2, 3, 4, etc., balancing areas and the choice of 1, 2, 3 or 4 balancing areas. While such a procedure is undertaken in LFS, the need to collapse is infrequent enough not to warrant special treatment for variance estimation purposes. Consequently, if any variance calculations or analysis other than mere averages or totals are contemplated, balancing areas that are expected to be stable without much collapsing should be incorporated into the sample design. That is, the response rates should be sufficiently large with high probability to avoid the necessity of collapsing balancing areas if variance estimation is contemplated. This would discourage one from using small areas to balance for non-response.

Instead of weighting by the inverse of the response rate in a balancing area, one could duplicate a sufficient number of units among the m_b respondents to bring the apparent sample size up to the original level of n_b units. However, it can be shown that an additional variance component occurs over that incurred when simple weighting is applied and in the case of srs, the sampling variance is considered alone, the increase would be up to about 12%, depending upon the response rate (see Hansen et al [3]). The main advantage of duplication vs. weighting lies in ensuring that integral rather than fractional weights are applied to each record. In certain types of published data, e.g., the number of persons with some characteristic, integral weights would tend to avoid rounding errors when sub-classifying data. When one estimates means or proportions or certain types of quantitative totals such as gross national product, the use of integers rather than fractional weights are of no advantage.

Apart from the comments in the above paragraph, the methodological problems concerned with weighting in balancing areas also apply to duplication in balancing areas.

Another important method of imputation for non-response is one of substitution of historical (previous month's data) or external source data (administrative files, other surveys, Census data). Once the substitution of historical or external source data has been undertaken to the extent possible for non-respondents, the weighting or duplication may be affected within balancing areas. In the case of weighting, one would inflate the weight Π_i^{-1} by the factor $n_b/(m_b+m_b')$, where m_b respondents were obtained as before and for m_b' of the (n_b-m_b) non-respondents, historical records were substituted for the missing data. In such a method of imputation, the sampling variance is reduced from that which occurs in the weighting scheme since we have increased the effective sample size from m_b to somewhere between m_b and (m_b+m_b') units. The increased sample, including those records imputed from historical records will not be as good as m_b+m_b' since historical or external source data are not as good as current response information unless there has been no change in the characteristics of the units for which substitution of historical data was undertaken.

Alternatively, one may wish to duplicate respondents; i.e., take a sample from respondents equal in size to the number of non-respondents and apply a weight of 2 instead of inflate the weight for all the respondents. However, one may wish to avoid duplication of those non-respondents for which substitution of historical information had been undertaken but one would rather subsample $n_b - (m_b+m_b')$, say m_b^* units from the m_b respondents to duplicate in order to bring the apparent sample size from (m_b+m_b') to n_b units in balancing area b. The estimated total for balancing area b would be

$$\hat{X}_b = \sum_{i=1}^{m_b} w_i\, x_i/\Pi_i + \sum_{j=m_b+1}^{m_b+m_b'} x_j'/\Pi_j, \qquad (4.1)$$

where x_j' is the imputed value for unit j and w_i = 1 or 2 (2 for a random subsample of $n_b - m_n - m_b'$ units among the m_b respondents). The expected value of $\hat{X}_b$ over all possible ways of duplicating is

$$\hat{X}_b^* = (n_b - m_b')/m_b \sum_{i=1}^{m_b} x_i/\Pi_i + \sum_{j=m_{b+1}}^{m_b+m_b'} x_j'/\Pi_j. \qquad (4.2)$$

Consequently, $V(\hat{X}_b) = V(\hat{X}_b^*)$ + (additional component of variance as a result of subsampling among the respondents). $\hat{X}_b^*$ is not the same as the estimate $n_b/(m_b+m_b')[\sum_{i=1}^{m_b} x_i/\Pi_i + \sum_{j=m_{b+1}}^{m_b+m_b'} x_j'/\Pi_j]$ and the variance of $\hat{X}_b^*$ is also different from that of the estimate where the weight inflation of $n_b/(m_b+m_b')$ is applied (see appendix).

The estimation procedures dealt with above include weighting or duplicating in design-dependent balancing areas. If historical or external source data are available for some of the non-respondents, these could be employed for imputation purposes prior to weighting or duplication in balancing areas. Instead of balancing areas, one could utilize weighting classes for imputation purposes and these are discussed in the next paragraph.

Weighting classes are distinguished from balancing areas in that they generally comprise characteristics of ultimate units (e.g., dwelling types, special income groups, etc.) as opposed to geographic areas, though one could conceivably group areas according to some distinct characteristics that are not related to the sample design. Usually, one defines weighting classes as well as balancing areas prior to the survey gathering procedure and makes adjustments through collapsing if the response rates are unacceptably low or the sample too small to employ any type of adjustment of the weights. In some imputation procedures, however, weighting classes are defined after the survey data have been gathered where factor analysis or other analytical tools are employed

to determine the most efficient set of weighting classes. After the weighting classes have been determined, the estimation procedures are essentially identical to those used in balancing areas. The biases and the variances (at least in terms of individual and joint inclusion probabilities of the ultimate units and other parameters not related to the sample design) are identical. However, upon further expansion of the variance to take into account the particular sample design, the variances of the estimates pertaining to balancing areas and weighting classes will be quite different. In order to utilize weighting classes for imputation purposes some knowledge about the non-respondents (such as income class, size of household dwelling type) must be available. In practice, when such information is not available, the procedure cannot be used.

The estimation formula for the methods of imputation discussed here may be written as below.

$$\hat{X} = \sum_{b} \hat{X}_b \quad \text{estimates the total of some characteristic,}$$

where b is either the balancing area or weighting class. The estimate for a given balancing area or weighting class is in turn given by:

$$\hat{X}_b = \sum_{i=1}^{m_b} w_i x_i/\Pi_i + \sum_{j=m_b+1}^{m_b+m_b'} w_j x_j'/\Pi_j, \text{ where } \Pi_i \text{ or } \Pi_j \text{ is the} \quad (4.3)$$

inclusion probability and m_b is the number of units that responded out of n_b units in balancing area b.

x_i = response value for unit $i = X_i$ (true value) + $^{R}\varepsilon_i$ (response error) (4.4)

x_j' = historical value for unit j (if available), given that unit j failed to respond. Among the $(n_b - m_b)$ non-responding units m_b' possess historical records in balancing area b.

$$x_j' = X_j \text{ (true value)} + {}^{R}\varepsilon_j' \text{ (response error of historical value, relative to } X_j) \quad (4.5)$$

w_i and w_j are weights, appropriate to the imputation method and the weights are listed in Table 1.

Table 1: Imputation Method in Balancing Area/Weighting Class

		w_i	w_j
(a)/(c)	Weighting by Inverse of Response Rate m_b/n_b	n_b/m_b	0
(b)/(d)	Duplication of a random subsample of (n_b-m_b) units from m_b respondents	2 for (n_b-m_b) units & 1 for $(2m_b-n_b)$ units	0 0
(e1)	Substitution of Historical Records for m_b' of (n_b-m_b) non-respondents, followed by weighting	$n_b/(m_n+m_b')$	$n_b/(m_b+m_b')$
(e2)	Substitution as in (iii), followed by duplication of respondents only	2 for $(n_b-m_b-m_b')$ units & 1 for $(2m_b+m_b'-n_b)$ units	1

(c) and (d) refer to weighting classes while (a) and (b) refer to balancing areas.

In the case of duplication, we have assumed the response rate m_b/n_b to be at least 0.5. If it is exactly 0.5, then duplication and weighting would yield identical estimates. Let us suppose that $n_b/m_b = W_b+d_b$, where W_b is an integer and d_b, a fraction in the range $0 \leq d_b < 1$, then m_b would be partitioned into m_{b1} units, subsampled at random requiring a weight of W_b and $m_{b2} = (m_b-m_{b1})$ units, requiring a weight of (W_b+1). Thus, $n_b = W_b m_b+d_b m_b = W_b m_{b1}+(W_b+1)\ m_{b2} = W_b m_b+m_{b2}$. Hence, $m_{b2} = d_b m_b$ and $m_{b1} = (1-d_b)\ m_b$. Consequently, a random subsample of $d_b m_b$ respondents would be assigned a weight of (W_b+1) and the remaining $(1-d_b)\ m_b$ respondents assigned a weight of W_b. If $W_b = 1$, then $n_b = m_b+d_b m_b$ or $d_b m_b = (n_b-m_b)$ units would require a weight of 2, as indicated in Table 1. Whatever the value of W_b, the expected value of the estimates

by method (b) or (d) over all possible subsamples of $d_b m_b$ respondents which would be assigned a weight of (W_b+1) instead of W_b is just the estimate by the weight inflation as of method (a) or (c).

In the case of method (e2), use of historical or external source data, followed by duplication of respondents, one would most likely confine the duplication only to a subsample from the m_b respondents rather than from the (m_b+m_b') units that either responded or utilized historical records. In such a case, the conditional expected value over all possible random subsamples of units assigned for duplication, given the sample, is not the estimate by method (e1) but rather an estimate with $w_i = (n_b-m_b')/m_b$ for the m_b respondents and $w_j = 1$ for the m_b' non-respondents with available historical records.

5. BIAS OF ESTIMATES

The bias of $\hat{X}_b$ according to imputation procedure may be readily obtained simply by finding $E\,\hat{X}_b$. Since $\hat{X}_b$ is a ratio estimate with the responding sample m_b a variable and similarly for (m_b+m_b'), a ratio estimate bias exists in addition to the response and non-response biases but we have neglected this in Table 2 where the biases are given for the estimates which are defined in Table 1. In the table, α_i is the probability of unit i responding while $\bar{\alpha}_b$ is the expected response rate in balancing area b and may be written as $\bar{\alpha}_b = E_b^* \alpha_i$. ${}^{R}B_i$ denotes the response bias pertinent to unit i while ${}^{R}B_i'$ denotes the bias of the historical value, relative to the true value X_i. α_i' is the probability of unit i possessing historical data and finally, Cov $\delta_i\delta_i'$ is the covariance between the event of responding or not responding ($\delta_i = 1$ or 0) and the existence or non-existence of historical data ($\delta_i' = 1$ or 0).

It will be noted in Table 2 that the bias is identical for weighting and duplicating and the reason for this is that, as pointed out before, the expected value of the estimate using duplication for imputation purposes over all possible subsamples of units to be duplicated is just the estimate using the weight inflation. The overall expected value of the two estimates is consequently the same.

The bias under method (e1) may be readily compared with the bias under methods (a) to (d). The non-response bias under method (e1) is given by $(\bar{\alpha}_b + \bar{\alpha}_b'')^{-1}$ En_b Cov_b^* $(\alpha_i + \alpha_i'', X_i/\Pi_i)$, which will reduce to the non-response bias according to methods (a) to (d) when $\alpha_i'' = 0$ and $\bar{\alpha}_b'' = 0$. As the combined probabilities $\alpha_i + \alpha_i''$ approach one, the population covariance between $\alpha_i + \alpha_i''$ and X_i/Π_i or Cov_b^* $(\alpha_i + \alpha_i, X_i/\Pi_i)$ approaches zero. In fact, if $\alpha_i + \alpha_i''$ were equal for all i, the covariance would be zero and there would be no non-response bias. The same holds for methods (a) to (d) if α_i's were all equal. The non-response bias under method (e1) would be expected to be lower than under methods (a) to (d) because of an anticipated decrease in the population covariance. Depending upon the extent of the availability of historical records, $\alpha_i + \alpha_i''$ would exceed α_i and would most likely have a smaller population variance. If Cov_b^* $(\alpha_i, X_i/\Pi_i) = r_{\alpha_i, X_i/\Pi_i} \sqrt{V_b^*(\alpha_i)\, V_b^*(X_i/\Pi_i)}$ and if Cov_b^* $(\alpha_i+\alpha_i', X_i/\Pi_i) = r_{\alpha_i+\alpha_i', X_i/\Pi_i} \sqrt{V_b^*(\alpha_i+\alpha_i')} \cdot \sqrt{V_b^*(X_i/\Pi_i)}$, then Cov_b^* $(\alpha_i+\alpha_i', X_i/\Pi_i)$ would most likely be smaller than Cov_b^* $(\alpha_i, X_i/\Pi_i)$ because one would expect $(\alpha_i + \alpha_i')$'s to be closer to one and presumably less variable among the units than α_i's alone, implying that V_b^* $(\alpha_i + \alpha_i') < V_b^*$ (α_i). A further decrease in the non-response bias would occur under methods (e1) than under methods (a) to (d) because of the larger denominator $(\bar{\alpha}_b + \bar{\alpha}_b'')$ pertaining to method (e1) compared with $\bar{\alpha}_b$ in the denominator of the bias pertaining to methods (a) to (d).

A lower non-response bias may be partially offset by a larger response bias pertaining to method (e1). If $^{R}B_i$'s were about the same magnitude as $^{R}B_i$'s on an average, then the response biases would be about the same but one would expect $^{R}B_i'$'s to be slightly larger than $^{R}B_i$'s since historical data would not be as close to the truth as current responses.

Table 2: Bias of Estimate, According to Imputation Procedure[1]

Method	Bias of Estimate	
Weighting (a)/(c) & Duplicating (b)/(d)	$\bar{\alpha}_b^{-1} En_b \, Cov_b^* (\alpha_i, X_i/\Pi_i)$	... non-response bias
	$+ \bar{\alpha}_b^{-1} \sum_{i=1}^{N_b} \alpha_i \, {}^R B_i$	... response bias
(e1) Substitution of Historical Records, then weighting	$(\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} En_b \, Cov_b^*(\alpha_i, X_i/\Pi_i)$	... non-response bias contributed by the use of weight inflation of respondents
	$+ (\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} En_b \, Cov_b^*(\alpha_i'', X_i/\Pi_i)$	... non-response bias, contributed by substitution of historical records
	$+ (\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} \sum_{i=1}^{N_b} (\alpha_i \, {}^R B_i + \alpha_i'' \, {}^R B_i')$	... response bias, contributed respectively by respondents and by non-respondents with historical data
(e2) Substitution of Historical Records, then duplication or weighting of respondents only	$\bar{\alpha}_b^{-1} En_b (1-\bar{\alpha}_b'') Cov_b^*(\alpha_i, X_i/\Pi_i)$	... non-response bias contributed by duplication
	$+ \bar{\alpha}_b^{-1} En_b (1-\bar{\alpha}_b'') \sum_{i=1}^{N_b} \alpha_i \, {}^R B_i$ ...	response bias contributed by respondents
	$+ En_b \, Cov_b^*(\alpha_i'', X_i/\Pi_i)$ ...	non-response bias, contributed by substitution of historical records
	$+ \sum_{i=1}^{N_b} \alpha_i'' \, {}^R B_i'$ ...	response bias, contributed by imputation by historical records

In (e1) and (e2), $\alpha_i'' = (1-\alpha_i)\alpha_i' - Cov \, \delta_i \, \delta_i'$ and $\bar{\alpha}_b'' = E_b^* \alpha_i'' = (1-\bar{\alpha}_b) \, \bar{\alpha}_b' - Cov_b^* \alpha_i \alpha_i' - E_b^* Cov \delta_i \delta_i'$

[1] Bias derived for method (e1) in Appendix 1.

6. VARIANCE OF ESTIMATES

The variance of $\hat{X}_b$ as defined for method (e1) is partially derived in Appendix 2 by regarding $\hat{X}_b$ as a combined product and ratio expression and employing Tayler series expansions. The same holds for $\mathrm{Cov}(\hat{X}_b, \hat{X}_c)$,

$$V(\hat{X}_b) \doteq \{X_b + (\bar{\alpha}_b + \bar{\alpha}_b'')^{-1}[En_b\ \mathrm{Cov}_b^*(\alpha_i + \alpha_i'', X_i/\Pi_i) + \sum_{i=1}^{N_b}(\alpha_i\ {}^{R}B_i + \alpha_i''\ {}^{R}B_i')]\}^2$$

$$\times \{V_s[\frac{n_b}{En_b} + \frac{\sum_{i=1}^{n_b}\Pi_i^{-1}[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]}{\sum_{i=1}^{N_b}[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]}$$

$$- \frac{\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')}{En_b(\bar{\alpha}_b+\bar{\alpha}_b'')}] \qquad \ldots \text{ sampling variance (s referring to a specific sample)}$$

$$+ E_s\ V[\frac{\sum_{i=1}^{n_b}\Pi_i^{-1}[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i')}{\sum_{i=1}^{N_b}[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]}$$

$$- \frac{\sum_{i=1}^{n_b}(\delta_i+\delta_i'')}{En_b(\bar{\alpha}_b+\bar{\alpha}_b'')}]|s \quad \ldots\ldots \text{ non-sampling variance } \} \qquad (6.1)$$

where expanded forms of non-sampling variances and covariances may be obtained in Appendix 2. Similarly $\mathrm{Cov}(\hat{X}_b, \hat{X}_c)$ may be expressed. In the above formula, $\delta_i'' = (1-\delta_i)\delta_i'$ and $\alpha_i'' = E\delta_i''$.

In the case of methods (a) and (c), formula 6.1 also holds with all α_i'', δ_i'' and $\bar{\alpha}_b''$ equal to zero.

For methods (b) and (d); viz., duplication of a subsample of units at random to boost the sample from m_b to n_b units in balancing area b, formula 6.1 yields one component of variance. There is an additional component, arising from the variation in the choice of subsampled units to be duplicated.

The additional variance component is given by:

$$E_s \, E\Big[\sum_{i=1}^{n_b} \left(w_i - \frac{n_b}{m_b}\right) \delta_i (X_i + {}^{R}\varepsilon_i) \, \Pi_i^{-1} \Big]^2 \Big| s, \qquad (6.2)$$

where s is a given sample of n_b units and the second E is taken over all possible responses and non-responses within a particular sample. For a given response rate m_b/n_b, $E\, w_i = n_b/m_b$ for all respondents in balancing area b and the response rate is assumed to be at least 0.5 so that $w_i = 1$ or 2. In the case of srswor, assuming m_b and n_b both constant, Hansen et al [3] showed that the additional variance component caused by duplication instead of weighting is as much as 12% for a response rate of about 3/4. Similar results occur when ppswor is undertaken. However, when m_b and n_b both vary, further studies on the expansion of 6.2 must be carried out.

It is difficult to compare the variance of $\hat{X}_b$ under method (e2) with that under method (a) or (c) from formula 6.1 without substitution of numerical values. Intuitively, one would expect the variance under method (e1) to be lower than that under method (a), the extent of the decrease depending upon the size of the non-response utilizing historical records and the correlation between historical and current information. The variances need to be explored, perhaps upon rewriting 6.1 in terms of average parametric values of α_i, ${}^{R}\sigma_i^2$, α_i'', etc. in the balancing area.

7. CONCLUSION

The conceptual issues cover the difficulty of non-response and pros and cons of different methods of dealing with them. Empirical data will be needed to obtain the parameters in the formulae stated in this paper for comparison purposes. An important fact to be noted is the additional variance component that occurs in duplication as opposed to weighting when a given response rate occurs in a given sample. The effect of duplication must be further studied as sample size and response rates both vary.

Much of the methodological development of the bias and the variance of estimates under different imputation procedures depends upon the knowledge of response probabilities which are rarely known in real life. Some estimates of response probabilities can be obtained from longitudinal studies of response profiles in the case of continuous surveys; otherwise, special experimental studies of non-respondents outside the sample used for publication purposes may be needed to obtain approximate estimates of response probabilities.

It is very important to note that, under the usual imputation procedures of duplication or weighting, there is non-response bias only if the response probabilities vary among the units and if there exists a correlation between response probabilities and the characteristic values of the units. Response bias, however, will occur whether or not we have full response.

8. ACKNOWLEDGMENT

The authors sincerely appreciate the comments and suggestions of the referee, editor, and A. Ashraf, Senior Methodologist, Household Surveys Development Division.

RESUME

L'article analyse les problèmes posés par les mesures applicables, à diverses étapes de la planification d'une enquête, pour contrer la non-réponse, les répercussions de ces mesures sur l'écart-type moyen, ainsi que l'utilité pratique, les avantages et les inconvénients de ces mesures. Il examine aussi certaines questions théoriques touchant la complexité et les niveaux d'imputation. Il existe diverses méthodes d'imputation: par pondération, par reproduction et par substitution d'enregistrements. L'article traite aussi de certaines questions méthodologiques concernant le biais et la variance.

REFERENCES

[1] Fellegi, I.P. and Holt, D., "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association (1976), Vol. 71, pp. 17-35.

[2] Ghangurde, P.D. and Mulvihill, J., "Non-Response and Imputation in Longitudinal Estimation in LFS", Household Surveys Development Staff, Statistics Canada Report (February 1978).

[3] Hansen, M.H., Hurwitz, W.N. and Madow, W.G., "Sample Survey Methods and Theory", Vol. II, Theory, pp. 139-141, John Wiley and Sons, Inc. (1953).

[4] Nordbotten, S., "The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means for Improving the Quality of Statistics", Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Belgrade 41, (September 1965), pp. 417-441.

[5] Platek R., "Imputation for Household Surveys in Statistics Canada", report prepared for European Statisticians' Conference held in Geneva, March 1978.

[6] Platek, R., "Some Factors affecting Non-Response", Survey Methodology (Statistical Services Field, Statistics Canada), Vol. 3, No. 2 (December 1977), pp. 191-214.

[7] Platek, R., Singh, M.P. and Tremblay, V., "Adjustment for Non-Response in Surveys", Survey Methodology (Statistical Services Field, Statistics Canada), Vol. 3, No. 1 (June 1977), pp. 1-24.

[8] Platek, R. and Gray, G.B., "Imputation Methodology", Household Surveys Development Division, technical paper describing biases and variances of estimates, using different methods of imputation.

[9] Szameitat, K. and Zindler, H.J., "The Reduction of Errors in Statistics by Automatic Corrections", Bulletin of the International Statistical Institute, Proceedings of 35th Session, Belgrade 41, (September 1965), pp. 395-417.

APPENDIX 1

Bias of Estimate in Balancing Area/Weighting Class

Consider the estimate $\hat{X}_b$ as defined by 4.3 in general, and in particular for case (e1) as of Table 1, viz., substitution of historical records for m_b' of $(n_b - m_b)$ non-respondents, followed by weighting.

$$\text{Then} \qquad \hat{X}_b = \frac{n_b}{m_b + m_b'} \left[\sum_{i=1}^{m_b} x_i/\Pi_i + \sum_{j=m_b+1}^{m_b+m_b'} x_j'/\Pi_j \right] \qquad \text{A1.1}$$

To derive the bias of $\hat{X}_b$, let us define δ_i as 1 or 0 according as unit i responds or not and δ_i' = 1 or 0 according as historical records are available and used for imputation or not. Then $m_b = \sum_{i=1}^{n_b} \delta_i$ and $m_b' = \sum_{i=1}^{n_b} (1-\delta_i)\,\delta_i'$. In the case of methods (a) to (d) all $\delta_i' = 0$ and consequently $m_b' = 0$.

$$\text{Then} \qquad \hat{X}_b = \frac{n_b}{\sum_{i=1}^{n_b} [(\delta_i + (1-\delta_i)\delta_i']} \sum_{i=1}^{n_b} \Pi_i^{-1} [\delta_i (X_i + {}^{R}\varepsilon_i) + (1-\delta_i)\delta_i' (X_i + {}^{R}\varepsilon_i')] \qquad \text{A1.2}$$

in which x_i and x_j' defined by 4.4 and 4.5 respectively have been substituted.

To determine the bias, one needs only to derive $E\hat{X}_b$ as of A1.2. We shall neglect the ratio estimate bias and also the covariance between n_b and the ratio with $\sum_{i=1}^{n_b}$ in the numerator and denominator, a covariance which may exist when the sample size, n_b, is a variable.

$$\text{Then} \quad E\hat{X}_b \triangleq \frac{En_b \sum_{i=1}^{N_b} [\alpha_i (X_i + {}^R B_i) + (\alpha_i' - \alpha_i \alpha_i' - \text{Cov}\ \delta_i \delta_i')(X_i + {}^R B_i')]}{\sum_{i=1}^{N_b} \Pi_i [\alpha_i + (1-\alpha_i)\alpha_i' - \text{Cov}\ \delta_i \delta_i']} \qquad \text{A1.3}$$

noting that $E^R \varepsilon_i = {}^R B_i$ and $E^R \varepsilon_i' = {}^R B_i'$.

We have not assumed independence between δ_i and δ_i' since the presence of historical record may be related to the tendency to respond or not to respond. Hence, $E(1-\delta_i)\ \delta_i' = (1-\alpha_i)\alpha_i' - \text{Cov}\ \delta_i \delta_i'$.

Further simplification of A1.3 is possible by utilizing "average" parameters such as, for example, $E_b^* T_i = \sum_{i=1}^{N_b} (\Pi_i / En_b)\ T_i = \bar{T}_b$, whatever T_i may be. Other expressions such as $\text{Cov}_b^* (T_i U_i) = E_b^* T_i U_i - E_b^* T_i E_b^* U_i$ and $V_b^*(T_i) = \text{Cov}_b^*(T_i, T_i)$ may be derived. E_b^* is a weighted average of individual parameter values, using Π_i / En_b as the weights, noting that $\sum_{i=1}^{N_b} \Pi_i = En_b$.

$$\text{Thus,} \quad \sum_{i=1}^{N_b} \alpha_i X_i = \sum_{i=1}^{N_b} \Pi_i \alpha_i\ X_i / \Pi_i = En_b\ E_b^* \alpha_i\ X_i / \Pi_i$$

$$= En_b [E_b^*\ \alpha_i\ E_b^*\ X_i/\Pi_i + \text{Cov}_b^*(\alpha_i,\ X_i/\Pi_i)]$$

$$= En_b [\bar{\alpha}_b (En_b)^{-1}\ X_b + \text{Cov}_b^*(\alpha_i,\ X_i/\Pi_i)]$$

$$= \bar{\alpha}_b\ X_b + En_b\ \text{Cov}_b^*(\alpha_i,\ X_i/\Pi_i) \qquad \text{A1.4}$$

Now let $(1-\alpha_i)\alpha_i' - \mathrm{Cov}\delta_i\delta_i' = \alpha_i''$ A1.4a

$$E_b^* \alpha_i'' = (1-\bar{\alpha}_b)\bar{\alpha}_b' - \mathrm{Cov}_b^*\alpha_i, \alpha_i' - E_b^* \mathrm{Cov}\delta_i\delta_i' = \alpha_b'' \text{ say} \qquad \text{A1.5}$$

In a similar manner as undertaken in A1.4,

$$\sum_{i=1}^{N_b} \alpha_i'' X_i = \bar{\alpha}_b'' X_b + En_b \mathrm{Cov}_b^*(\alpha_i'', X_i/\Pi_i), \text{ where } E_b^*\alpha_i'' = \bar{\alpha}_b''.$$

Then $\hat{EX}_b \doteq En_b[(\bar{\alpha}_b+\bar{\alpha}_b'') X_b + En_b \mathrm{Cov}_b^*(\alpha_i+\alpha_i'', X_i/\Pi_i)$

$$+ \sum_{i=1}^{N_b} (\alpha_i R_{B_i} + \alpha_i'' R_{B_i}')]/En_b(\bar{\alpha}_b+\bar{\alpha}_b''), \text{ where the bias equals}$$

$$(\bar{\alpha}_b+\bar{\alpha}_b'')^{-1}[En_b \mathrm{Cov}_b^*(\alpha_i+\alpha_i'', X_i/\Pi_i) + \sum_{i=1}^{N_b} (\alpha_i R_{B_i} + \alpha_i'' R_{B_i}')] \qquad \text{A1.6}$$

The bias under imputation methods (a) to (d) immediately follow by putting $\alpha_i'' = 0$ and $\bar{\alpha}_b'' = 0$ in A1.6.

The bias of $\hat{X}_b$ under method (e2) may be similarly derived except that the covariance between δ_i and δ_j'' has been omitted in order to simplify the formula.

APPENDIX 2

Summary of Development of Variance of Estimate $\hat{X}$

$$\hat{X} = \sum_b \hat{X}_b, \text{ where } b = \text{balancing area or weighting class}$$

$$V(\hat{X}) = \sum_b V(\hat{X}_b) + \sum_{b \neq c} Cov(\hat{X}_b, \hat{X}_c), \text{ and hence, the need to derive}$$

$V(\hat{X}_b)$ and $Cov(\hat{X}_b, \hat{X}_c)$. A covariance may exist between the estimates based on different balancing areas or weighting classes, depending upon the definition of balancing areas or weighting classes as well as the sample design.

We will deal with method (e1), where historical information is substituted for non-responses, whenever available or appropriate and then weighting or duplication of records to boost the sample up to the required level. We will deal with the weighting first, where $\hat{X}_b$ is defined as in 4.3 or A1.1 or A1.2 (the most convenient form for the development of the bias and variance).

$\hat{X}_b$ may be regarded as a complex expression of the form $n_b y_b / z_b$ with n_b, y_b and z_b all variables. In some sample designs, n_b remains constant but it need not be in the general developments.

Then, by the use of Taylor series expansion,

$$Cov(n_b y_b/z_b, n_c y_c/z_c) \doteq (En_b\ Ey_b/Ez_b)(En_c\ Ey_c/Ez_c)$$

$$[\text{Rel Cov } n_b,n_c + \text{Rel Cov } y_b,y_c + \text{Rel Cov } z_b,z_c$$

$$+ (\text{Rel Cov } n_b,y_c + \text{Rel Cov } n_c,y_b) - (\text{Rel Cov } n_b,z_c + \text{Rel Cov } n_c,z_b)$$

$$- (\text{Rel Cov } y_b, z_c + \text{Rel Cov } y_c, z_b)] \qquad \text{A2.1}$$

and Rel Var $(n_b y_b/z_b)$ immediately follows by putting c=b.

To derive $V(\hat{X}_b)$ and $Cov(\hat{X}_b, \hat{X}_c)$, we require the following expressions. Here $(1-\delta_i)\delta_i$ as defined in A1.2 will be abbreviated by δ_i'' so that $E\delta_i''/i = \alpha_i''$ as defined in A1.4a.

En_b cannot be further simplified than $\sum_{i=1}^{N_b} \Pi_i$

$$V(n_b) = \sum_{i \neq j}^{N_b} \Pi_{ij} - En_b(En_b - 1)$$

Ey_b is given by [ ] in A1.6 and $Ez_b = En_b(\bar{\alpha}_b + \bar{\alpha}_b'')$, with $\bar{\alpha}_b''$ given in A1.5

Additional expressions involve variances and covariances, which are stated below without proof but have been developed by Platek and Gray [8].

$$V \sum_{i=1}^{n_b} \Pi_i^{-1}[\delta_i(X_i + {}^R\varepsilon_i) + \delta_i''(X_i + {}^R\varepsilon_i')]$$

$$= V_s\{ \sum_{i=1}^{n_b} \Pi_i^{-1}[\alpha_i(X_i + {}^RB_i) + \alpha_i''(X_i + {}^RB_i')]\}$$

$$+ E_s\{ V \sum_{i=1}^{n_b} \Pi_i^{-1}[\delta_i(X_i + {}^R\varepsilon_i) + \delta_i''(X_i + {}^R\varepsilon_i')] | s \ , \qquad \text{A2.2}$$

where s means a specific sample of n_b units.

The second line, viz., the non-sampling variance component is given by:

$$\sum_{i=1}^{N_b} V[\delta_i(X_i + {}^R\varepsilon_i) + \delta_i''(X_i + {}^R\varepsilon_i')]\Pi_i^{-1}$$

$$+ \sum_{i \neq j}^{N_b} \Pi_{ij}\Pi_i^{-1}\Pi_j^{-1} \ Cov[\delta_i(X_i + {}^R\varepsilon_i) + \delta_i''(X_i + {}^R\varepsilon_i'), \delta_j(X_j + {}^R\varepsilon_j) + \delta_i''(X_j + {}^R\varepsilon_j')].$$

Here, $V[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i')] = \alpha_i\ {}^{R}\sigma_i^2 + \alpha_i''\ {}^{R}\sigma_i'^2$

$$+ 2(\alpha_i\alpha_i''+ \mathrm{Cov}\delta_i\delta_i'')r_{2ii''}\ {}^{R}\sigma_i\ {}^{R}\sigma_i'+V[\delta_i(X_i+{}^{R}B_i)+\delta_i''(X_i+{}^{R}B_i')]$$

and $\mathrm{Cov}[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i'),\delta_j(X_j+{}^{R}\varepsilon_j)+\delta_j''(X_j+{}^{R}\varepsilon_j')]$

$$= (\alpha_i\alpha_j+\mathrm{Cov}\delta_i\delta_j)\ r_{2ij}\ {}^{R}\sigma_i\ {}^{R}\sigma_j+(\alpha_i\alpha_j''+\mathrm{Cov}\delta_i\delta_j'')\ r_{2ij}''\ {}^{R}\sigma_i\ {}^{R}\sigma_j'$$

$$+ (\alpha_i''\alpha_j+\mathrm{Cov}\delta_i''\delta_j)\ r_{2ji}''\ {}^{R}\sigma_i'\ {}^{R}\sigma_j+(\alpha_i''\alpha_j''+\mathrm{Cov}\delta_i''\delta_j'')\ r_{2ij}'\ {}^{R}\sigma_i'\ {}^{R}\sigma_j'$$

$$+ \mathrm{Cov}[\delta_i(X_i+{}^{R}B_i)+\delta_i''(X_i+{}^{R}B_i'),\delta_j(X_j+{}^{R}B_j)+\delta_j''(X_j+{}^{R}B_j')].$$

In the above, $\mathrm{Cov}\ {}^{R}\varepsilon_i\ {}^{R}\varepsilon_j = r_{2ij}\ {}^{R}\sigma_i\ {}^{R}\sigma_j$ = covariance between current responses of pairs of units, $\mathrm{Cov}\ {}^{R}\varepsilon_i\ {}^{R}\varepsilon_j' = r_{2ij}''\ {}^{R}\sigma_i\ {}^{R}\sigma_j'$ = covariance between current and historical responses (applicable also for j=i), and $\mathrm{Cov}\ {}^{R}\varepsilon_i'\ {}^{R}\varepsilon_j' = r_{2ij}'\ {}^{R}\sigma_i'\ {}^{R}\sigma_j'$ = covariance between historical responses.

If we replace α_i by $\alpha_i^2+V(\delta_i)$ and similarly for α_i'', the above variances and covariances would be symmetrically described.

$$V\Big[\sum_{i=1}^{n_b}(\delta_i+\delta_i'')\Big]$$

$$= V_s\Big[\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')\Big]$$

$$+ E_s\ V\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')\,|s$$

$$= V_s\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'') + \sum_{i=1}^{N_b}\Pi_i[V(\delta_i+\delta_i'')]$$

$$+ \sum_{i\neq j}^{N_b}\Pi_{ij}[\mathrm{Cov}(\delta_i+\delta_i''),(\delta_j+\delta_j'')]\ . \qquad \text{A2.4}$$

$$\text{Cov} \sum_{i=1}^{n_b} (\delta_i+\delta_i''),\ \sum_{i=1}^{n_b} \Pi_i^{-1}\,[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i')]$$

$$= \text{Cov}\ \{ \sum_{i=1}^{n_b} (\alpha_i+\alpha_i''),\ \sum_{i=1}^{n_b} \Pi_i^{-1}\,[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]\}$$

$$+ E_s\ \text{Cov}\{ \sum_{i=1}^{n_b} (\delta_i+\delta_i''),\ \sum_{i=1}^{n_b} \Pi_i^{-1}\,[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i')]\}\,|s \qquad \text{A2.5}$$

The second line, viz., the non-sampling covariance is given by:

$$\sum_{i=1}^{N_b} \text{Cov}(\delta_i+\delta_i''),\ [\delta_i(X_i+{}^{R}B_i)+\delta_i''(X_i+{}^{R}B_i')]/i$$

$$+ \sum_{i\neq j}^{N_b} \Pi_{ij}\,\Pi^{-1},\ \text{Cov}(\delta_i+\delta_i'')\cdot[\delta_j(X_j+{}^{R}B_j)+\delta_j''(X_j+{}^{R}B_j')]\,|i,j \qquad \text{A2.6}$$

For the covariance expressions involving balancing areas b and c, V_s is replaced by Cov_s, $\sum_{i=1}^{N_b}$ does not exist and $\sum_{i\neq j}^{N_b}$ is replaced by $\sum_{i=1}^{N_b}\sum_{j=1}^{N_c}$.

$$\text{Cov}\ \{ n_b,\ \sum_{i=1}^{n_b} \Pi_i^{-1}\,[\delta_i(X_i+{}^{R}B_i)+\delta_i''(X_i+{}^{R}B_i')]\}$$

$$= \text{Cov}_s\ \{ n_b,\ \sum_{i=1}^{n_b} \Pi_i^{-1}\,[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]\} \qquad \text{A2.7}$$

and finally, $\text{Cov}\ n_b,\ \sum_{i=1}^{n_b} (\delta_i+\delta_i'') = \text{Cov}_s[n_b,\ \sum_{i=1}^{n_b} (\alpha_i+\alpha_i'')]$ A2.8

Now $V(n_b y_b/z_b)$ may be written approximately as:

$$\frac{(En_b)^2(Ey_b)^2}{(Ez_b)^2} \; V\left(\frac{n_b}{En_b} + \frac{y_b}{Ey_b} - \frac{z_b}{E_b}\right)$$

Likewise, $Cov(n_b y_b/z_b,\ n_c y_c/z_c)$ can be approximately written as:

$$\frac{En_b\ Ey_b}{Ez_b} \quad \frac{En_c Ey_c}{Ez_c}\ Cov\ \left(\frac{n_b}{En_b} + \frac{y_b}{Ey_b} - \frac{z_b}{Ez_b}\right),\ \left(\frac{n_c}{En_c} + \frac{y_c}{Ey_c} - \frac{z_c}{Ez_c}\right)$$

and by partial substitution of the formulae derived we may obtain $V(\hat{X}_b)$ as stated in 6.1 and $Cov\ (\hat{X}_b, \hat{X}_c)$.

For $V(\hat{X}_b)$ under methods (a) to (d), one simply puts all δ_i'', α_i'', $\bar{\alpha}_b''$ equal to zero. All $Cov\ \delta_i\delta_i''$, $Cov\ \delta_i\delta_j'' = 0$ for methods (a) to (d).

THE APPLICATION OF A SYSTEMATIC METHOD OF AUTOMATIC EDIT AND IMPUTATION TO THE 1976 CANADIAN CENSUS OF POPULATION AND HOUSING[1]

C.J. Hill[2]

I.P. Fellegi and D. Holt proposed a systematic approach to automatic edit and imputation. An implementation of this proposal was a Generalized Edit and Imputation System by the Hot-Deck Approach, that was utilized in the edit and imputation of the 1976 Canadian Census of Population and Housing. This paper discusses that application, evaluating the strengths and weaknesses of the methodology with some empirical evidence. The system will be considered in relation to the general issues of the edit and imputation of survey data. Some directions for future developments will also be considered.

1. INTRODUCTION

This paper is a discussion of the application of a Systematic Method of Automatic Edit and Imputation originally developed by I.P. Fellegi and D. Holt [1] to the 1976 Canadian Census of Population and Housing. The implementation of this methodology as a computer system within Statistics Canada is the system known as 'CAN-EDIT'. This was described by Graves [2]. The Can-Edit system, in turn, became a component of the "Census Edit and Imputation Processing System" which included several other custom-built modules [3]. Some of these modules handled certain special edit and imputation problems. Others operated in conjunction with the CAN-EDIT system and addressed methodological issues not covered by Fellegi and Holt. Some discussion of the methodology of these modules is included here in that they were essential to the application of the Fellegi-Holt method.

---

[1] Adapted from a paper presented at the Annual Meeting of the American Statistical Association, August 14-17, 1978, San Diego, California, U.S.A.

[2] C.J. Hill, Census Survey Methods Division, Statistics Canada.

The census is a multi-purpose survey consisting of both population and housing questions. The housing questions in 1976 were primarily concerned with identifying the type and tenure of the dwelling. The population questions were divided into two parts, a basic set of questions asked of all persons, and a set of sample questions asked of persons 15 years of age and over in 1/3 of all private households, and all collective dwellings. The basic questions were demographic questions on age, sex and marital status, a question on relationship to head, and one on mother tongue. The sample questions were on education, labour force status and mobility status. The 'CAN-EDIT' system was used in the edit and imputation of most of the variables. The only variables not handled by this system were mother-tongue and mobility status.

This paper presents the rationale for the edit and imputation of the Census and a brief non-technical description of the methodology in sections 2 and 3. An evaluation of the method is then given in section 4,with a final section suggesting directions for further work on the development of edit and imputation methodologies arising from the experience of the application to the 1976 Canadian Census.

2. THE RATIONALE FOR THE EDIT AND IMPUTATION OF THE CENSUS DATA

The terms 'edit and imputation' (E&I) as used here in reference to the Census are twin aspects of a single operation. 'Edit' refers to the detection of an error, 'imputation' to the correction of an error. Edit can be considered separately from imputation in that it may be used to initiate a corrective action involving a return to an earlier state in the processing. Editing may also be undertaken merely to flag erroneous records. Imputation as the correction of an error is taken to mean any modification of the data that produces a record that will pass the edits, other than by reference back to the source of the data to elicit a 'true' response. This operation of edit and imputation is undertaken with the intention of minimizing the errors in the data at the micro level.

The reason for imputing, rather than making a correction attempting to obtain a 'true' value, is that after a certain stage in the operation it becomes costly, if not impossible, to retrace one's steps. The choice at this stage is either to edit and impute the data or to publish data that include unspecified or erroneous information.

Among others, the following three important reasons influenced the undertaking of edit and imputation in the 1976 Census.

(1) To obtain the required estimates, adjustments must be made for errors at either the macro or the micro level. Correction (by edit and imputation) at the micro level can make maximum use of the available information and in principle achieve the best estimate.

(2) Subsequent operations in the Census, for example, the formation of families would be much more complicated, if not impossible, with incomplete and inconsistent data. In certain cases, the number of invalid records would increase considerably.

(3) Consistent official estimates are essential as a service to the users both outside and within Statistics Canada. Few users will wish to take responsibility for adjusting the estimates, and difficulties may arise as a result of differing unofficial estimates.

3. THE METHODOLOGY AND ITS IMPLEMENTATION

3.1 The Methodology Objectives

Fellegi and Holt state three objectives for the methodology underlying the edit and imputation system.

(1) As much as possible of the original data should be retained by changing the minimum number of fields in a given dirty record in order to produce a clean record.

(2) The data after imputation should retain, as far as possible, the distributional properties of the clean records.

(3) The imputation action should arise directly out of the edit rules.

These objectives are clearly aimed at ensuring data quality; their validity will be discussed below in the section on evaluation. The third objective is a practical consideration as it serves to greatly simplify the operation of defining imputation.

3.2 The Implementation of These Objectives

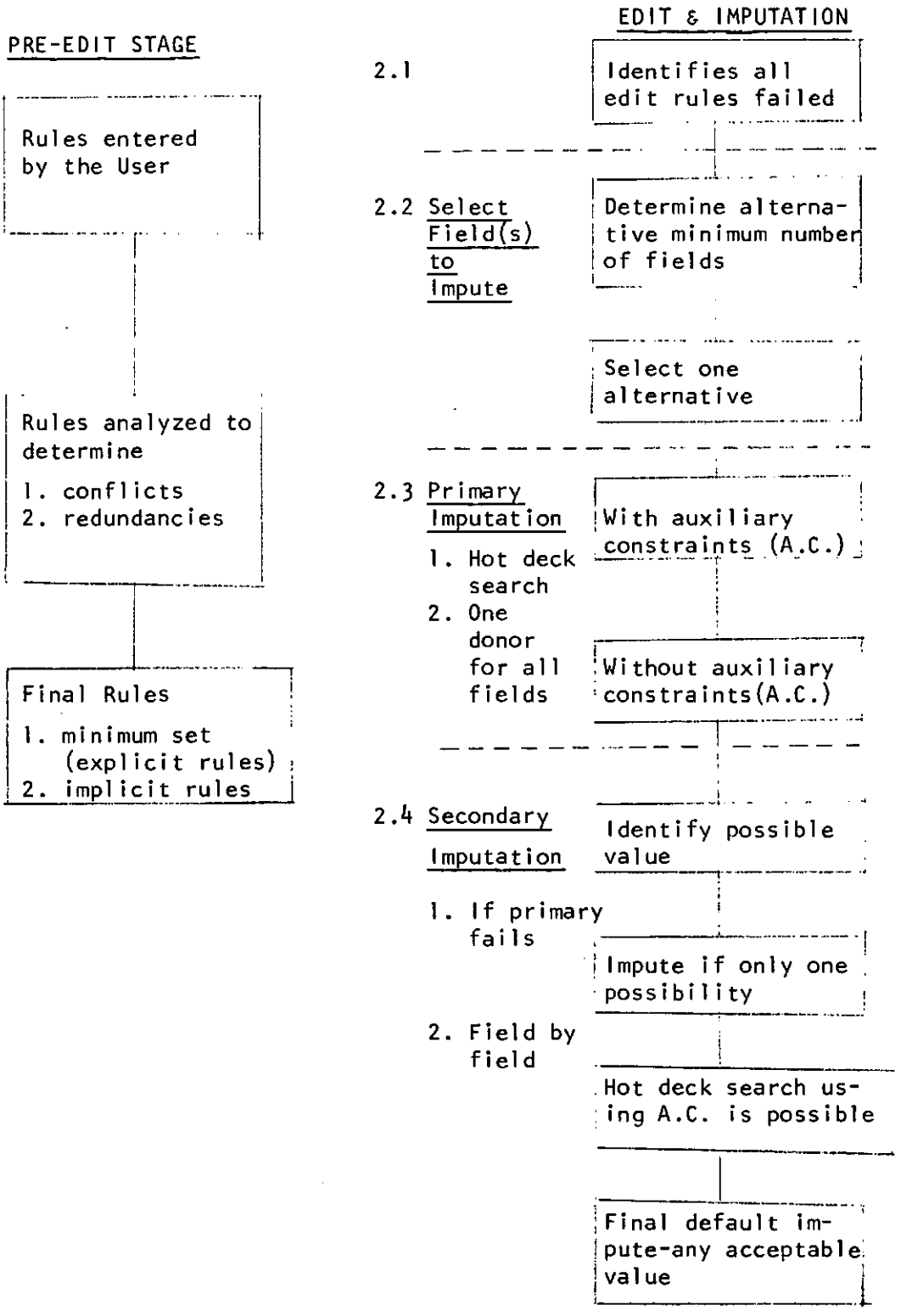
The initial attempt at the implementation of the methodology was by a system that consisted of two basic sub-systems:

(1) A system to analyze the edit rules.
(2) The edit and imputation system that operates on the data.

These operations are shown in Diagram 1.

Diagram 1

A Flowchart of 'CAN-EDIT' Processes to be Undertaken for Each Stratum*



* Stratification and Auxiliary Constraints are explained in Section 3.3.

(1) The System to Analyze the Edit Rules

The first stage in the edit and imputation operation is the analysis of the edit rules. This stage consists of the following steps:

The edits are written in a conflict form. They may be either within-person edits or between-person edits.

> An example of a within-person edit is:
> 'It is a conflict if the third person in the household is married and is less than 15 years of age'.

> An example of a between-person edit is:
> 'It is a conflict if the sixth person in the household is the parent of the head of the household and male and the ninth person in the household is the parent of the head of the household and male'.

It is important to note that one concept requires many edit rules. If, for example, an edit is required to exclude the possibility that the head of household has two parents of the same sex, edit rules have to be written between all possible pairs of persons. This essential feature creates some limitations to the system that will be discussed in a later section.

The edit rules are then analyzed and the output defines:

i) Any inconsistencies or conflicts between the rules.
ii) Any redundancies in the rules.

Once inconsistencies are removed, the final output is:

i) A minimum set of edit rules (explicit rules).
ii) A set of implied edit rules, that are generated from the minimum set.

These two sets combined comprise the complete set.

(2) The Edit and Imputation System

The analysis of the rules having been completed, the edit and imputation of the data can be undertaken. This operation divides into four stages:

(2.1) The edit that defines which rules have failed for each record.

(2.2) The selection of fields to impute. This has two parts:

i) the identification of which field(s) represent(s) the minimum number of field(s) that need to be changed to ensure a clean record,

ii) the selection at random from among alternatives if there is more than one minimal set. The information that existed in the fields selected for imputation is now ignored and will in no way influence the imputation action.

There are two stages of imputation, known as primary and secondary imputation.

(2.3) Primary imputation is a method by which one donor record gives a 'dirty record' all the values necessary to complete the imputation. To do this the donor must match the 'dirty record' for those fields that will not be changed, and are linked by an edit rule to the fields to be imputed. These conditions ensure that a new record is clean. (Refinements on this principle will be discussed below). A donor record is found by selecting at random an acceptable record from a file of about 2,000 records. This is a form of the method of imputation known as 'hot deck' imputation. If no acceptable record is found, the search continues by the method of secondary imputation.

(2.4) Secondary imputation is a method of field-by-field hot deck imputation. In this method certain matching conditions may be applied during the search for a donor. However, the crucial condition for accepting a donor is not a perfect match which has already proved impossible, but rather that the new record will pass the edit rules involving fields left unchanged or previously imputed. Once a field is imputed, it is incorporated into the record for the search to continue so as to impute the next field.

One important discovery that was made during the development testing of 'CAN-EDIT' is that for primary imputation only the minimum set of rules is required, whereas secondary imputation needed the complete set of rules. Failure to use the complete set could result in creating a situation in which a partially imputed record could become impossible to complete.

3.3 Modifications and Enhancements Consistent With the Original Methodology

As a result of experience in attempting to apply the system, various modifications and enhancements were introduced. Some of these were consistent with the methodology, four of which are considered here. Section 3.4 will consider two modifications that conflicted with the original objectives. Two are important refinements to the principles of imputation within the 'CAN-EDIT' system. These are (1) 'Auxiliary Constraints' and (2) 'Data Dependent Decoupling'. The other two are elements of the 'Census System' that address methodological problems not covered by Fellegi and Holt. These are (3) the Stratification Sub-System and (4) the choice between single or multiple unit editing.

(1) Auxiliary Constraints

Auxiliary constraints are fields used in matching during the search for a donor record irrespective of whether or not they are required as a matching condition to ensure a clean record. They are used in both primary and secondary imputation. Fields used as auxiliary constraints will normally be those highly correlated with the fields to be imputed. This enhancement was suggested in the paper by Fellegi and Holt.

In primary imputation, they have to be used as a complete set or not at all. The system was designed this way because there is no very obvious algorithm for relaxing constraints when the entire record is imputed simultaneously. In effect, therefore, primary imputation has two levels of matching, the optimum matching conditions that include auxiliary constraints and a degraded option matching on the necessary fields only.

In secondary imputation, with field by field imputation, one can attempt to match on as many fields as specified and take the best match.

(2) Data Dependent Decoupling

During a test of an early version of 'CAN-EDIT' excessive matching conditions forced a large number of records to have to go to secondary imputation. An analysis of the problem indicated that the matching conditions in the search for a donor were too restrictive.

In the original version, a match was made with every field linked to the fields to be imputed by edit rules. However, because two fields are linked by edit rules, it does not necessarily mean that the value in the field to remain unchanged restricts the acceptable values in the field to be imputed. An example of this is in the field "relationship to head", with reference to the previously mentioned rule preventing two parents of the head with the same sex. Clearly, if there is a person in the household coded head's parent and male, this places a restriction

upon imputing the code parent to another male. If on the other hand there is no such person, there need no longer be this restriction.

(3) The Stratification System

The function of the stratification system was to partition the data into subsets that (1) shared a common set of edit rules and (2) manifested a degree of homogeneity beyond that of sharing edit rules. Edit and Imputation is then undertaken independently within each stratum.

The control variables[1], document type and collective dwelling type were used for this purpose, for the 100% data, together with a variable defined in terms of the mix of persons in the household. Age, sex and collective dwelling type were used to stratify the sample data.

A full appreciation of the nature of stratification needs to be considered in conjunction with the question of single and multiple unit editing, since one of the dimensions of stratification for multiple unit editing was the number of persons in the household.

(4) Single or Multiple Unit Editing

In a sense, the Census represented three if not four surveys rolled into one and part of the complexity of attempting to edit it lies in this multiple nature. The dwelling data stands alone and presented only minor problems. The difficulty lies in the interrelationship between person, family and household data. At the start of the editing operations the number of persons (the low level unit) in households (the high level unit) has been frozen. There is, of course, variation in household size.

---

[1] The operation prior to edit and imputation determined whether a household was a private or a collective dwelling, occupied or unoccupied and whether or not it was in the sample. It also ensured that all collective dwellings had an identified type, e.g. hospital, orphanage, hotel. This information was frozen as the control variables document type and collective dwelling type.

The family at this stage has yet to be defined. There is now a choice between treating the person or the household as the editable unit.

This problem, which was not addressed by Fellegi and Holt, represented a major practical issue when integrating 'CAN-EDIT' into the 'Census System'. The methodology is based on a Cartesian data space which in a specific case, i.e. a household of a certain size, has a fixed number of dimensions. It was not possible to have sets of edit rules that addressed spaces of different dimensions, because each rule spans all dimensions of the space. Therefore, if there are to be edit rules between persons each size of household requires a unique set of edit rules.

Single unit editing is the method of editing in which the person is the editable unit. This means there can be no edit rules between persons.

Multiple unit editing is a method of editing in which the household is the editable unit. This method allows edit rules between persons. However, this is achieved at certain cost.

i) The data have to be stratified by size of household.
ii) The potential size of the editable unit becomes very large.
iii) There is a cut-off point beyond which it is totally unrealistic to take multiple unit editing which means there must be single unit editing for residual persons in large households.

In 1976, multiple unit editing was used for editing the 100% data in private households principally because of the need to establish clean family data. Single unit editing was used to edit most of the persons in collective dwellings, the 13th person onwards in very large households, and all sample data.

3.4 Modifications and Enhancements Inconsistent with the Original Methodology

In developing the Census system, two features were included that conflicted with the original objective, set out by Fellegi and Holt, of changing the minimum number of data fields. These two features were both systems external to the 'CAN-EDIT' system but utilized a specific property of that system to achieve their effect. They were : (1) a derive system used prior to edit and imputation and (2) a hierarchical edit and imputation structure. The Fellegi-Holt methodology specified that the amount of change in the observed data should be minimized. By implication all fields are equal candidates for change. The 'CAN-EDIT' system for very good reasons recognized that there were control variables fixed prior to editing and that the system should include the possibility of distinguishing between 'Imputable' and 'Non-Imputable' fields.

The Derive System: This piece of software is a semi-generalized system that creates an environment within which additional variables may be derived for the edit and imputation operation.

i) To combine two or more fields into one.
ii) To derive a variable for stratification.
iii) To create class values of a variable.
iv) As a means of forcing an imputation action.

It is this last function that is important to consider here as it conflicts with original objectives. The derived variable was frozen as an non-imputable variable. This meant that where an edit involved this field and other fields, some of the other fields were forced to change. This was used to force a specific imputation outcome. In general, this meant changing more than the minimum number of fields. This is explained in detail in section 4.3.3.

Hierarchical Editing: Hierarchical editing is a system of editing in which one set of fields is edited, imputed and frozen before another set of fields is edited, and in which there exists at least one edit rule linking the two sets. If there are no rules linking the two sets, the order is irrelevant. If, however, there are linking rules, freezing some fields in an earlier hierarchy may force more than the minimum change in the record as a whole. The principle of minimum change only applies to a single hierarchy.

In 1976, there were two main hierarchies: one for the 100% data and one for the sample data. This structure clearly only had implications for the sample questionnaire, primarily in relation to the age question. Age was frozen in the first hierarchy and may have been inconsistent with the data on education, labour force status and mobility status. In practice, such inconsistencies were rare and the effect on the data was negligible. An additional minor hierarchy was used for questions within filters in the sample data.

4. AN EVALUATION OF THE EDIT AND IMPUTATION METHODOLOGY

4.1 Introduction

The method may be evaluated as an instrument in allowing the successful edit and imputation of the data and objectively by an external evaluation against a source of true data. A project is underway to achieve the latter. The findings of this project will be reported in a census publication [4]. The discussion here, however, is a consideration of the system as an instrument for producing a clean data base.

4.2 The Evaluation of the Method as an Instrument for the Edit And Imputation of the Data

The following points will be considered in evaluating the generalized system as a means of achieving a successful edit and imputation operation.

(1) The methodological scope of the system, i.e. the range of types of variable and edit conditions the system is designed to handle.
(2) Finiteness, i.e. the practical limits to which the system conforms.
(3) The appropriateness of the three objectives outlined by Fellegi and Holt.

4.2.1 The Scope of the Method

In their paper, Fellegi and Holt write "At the beginning, let us restrict ourselves to records containing only qualitative (coded) data, i.e. data which are not subject to a meaningful metric".

In developing a generalized edit and imputation system, it was necessary to limit the scope of the types of data that it could handle. As indicated by Fellegi and Holt, the methodology addressed itself primarily to qualitative data.

Quantitative fields can, of course, be treated as if they were qualitative variables and therefore be handled in the same system. There are, however, two important objections to doing this:

(1) The loss of information in throwing away the metric.
(2) The potentially vast number of edit rules that may be generated in attempting to treat arithmetic rules as logical rules between categories.

Despite these objections, the system was applied in the Census to records that contained a mixture of quantitative and qualitative data. This was justified insofar as the variables were predominantly qualitative and the edits applied to the quantitative variables were of a limited nature. However, as the Census was attempting to edit variables outside the scope for which the editing system was designed, the results were not totally satisfactory.

The only quantitative variable in the 100% data was date of birth or, by implication, age.

Date of birth was defined by 3 variables: decade, year, and month of birth, this last being more correctly the two periods January to May, June to December. Each of these taken separately could be used as a qualitative variable and indeed was so treated. There were two main problems:

(1) A crucial age barrier occurs at age 15. The sample questions were only to be answered by persons at or over this age. Also certain conditions were only allowable at or above this age, e.g. Head of household or Married. The problem was that after edit and imputation there were more than the expected numbers of certain groups of persons close to the 15 year age boundary, in particular widowed or divorced persons. The only consolation was that the problem was greatly reduced when compared with the 1971 data.

(2) It was impossible to write edits to ensure reasonable age spacing between parents and children. The number of edits required to ensure a 15 year minimum difference was very large as this would have required an edit rule for each individual age difference. The decision was therefore:

   i) to limit such edits to age differences between the Head and Spouse and their children, (the main group of edits this excluded was edits between the Head and his parents);

   ii) to use only decade of birth in the edits;

   iii) to ensure that at least one parent was born in an earlier decade than all the children. (It is theoretically possible for a step-parent to be younger than an adult child).

The application of these rules removed some, but not all of the erroneous data. A successful solution to this problem awaits the development of a methodology that can be implemented as a system that will not only edit and impute quantitative data but quantitative data in combination with complex qualitative data.

4.2.2 Finiteness

The population of Canada is 23 million. The number of households is 7 million. The complete data space representing households has very many more cells than the total number of households. For households of size 'n', this space contains approximately $(2000)^n$ cells. The number of edit rules required to partition this space is also potentially very large. A particular between-person edit condition that could apply between most persons in the household, in almost all positions, would have generated 100 million edit rules. A tabulation of the data indicated that in fact there were only 1700 persons in Canada who could potentially fail these rules.

The total number of edit rules is a function of household size and the set of edit conditions to be applied. A realistic utilization of computer resources set a limit of 2048 upon the total number of edit rules. This limit was implemented by restricting multiple unit editing to households of 12 or less, or the first 12 persons in large households, and by excluding certain types of conditions from the set of edit rules. A special 'clean-up' programme was used to edit and impute these residual problems.

There are also data limitations in trying to push the method too far. The imputation was by a hot-deck method. In attempting to edit and impute large households, the system came up against the data limit that the number of available records for the hot-deck had become very small. With very large households a point is reached at which the operation is very costly, the number of records is very small and the

quality of the imputation is much reduced by the small hot-deck size. The finite limitations of the system are probably a minor constraint upon the effectiveness of the method given the finite nature of the data.

4.2.3 The Methodological Basis

Editing is an essentially very straightforward operation and is passive in relation to the final data. The only problem presented by editing is to ensure that the edit rules are clean and consistent. The issue to be discussed here is the methodological basis of the imputation action. The three criteria set out by Fellegi and Holt were outlined above in the description of the methodology and will now be assessed.

4.2.3.1 Changing the Fewest Possible Items of Data

The principle of changing the fewest possible data items (fields) is considered by Fellegi and Holt to be of overwhelming importance. This position is more than justified as a reaction against the enthusiastic over-correction of data that has been known to occur. Their formulation, however, is a specific case of a general principle that data modification should be kept to a minimum. The problem is that the number of fields is a somewhat arbitrary count. The number of fields covering the same information may be modified by changes in the questionnaire or in its data capture. A simple, easily defined concept may be reliably captured by one question, whereas a number of questions may be used to define a single potentially ambiguous concept. On the other hand, one cannot pretend to start counting concepts as if they had the same concrete existence as a question.

This problem is implicitly recognized by Fellegi and Holt in the suggestion they made that weights could be attached to fields in relation to their reliability. This suggestion was not implemented for use in the system applied to the 1976 Census. However, careful analysis is required before any alternatives are introduced.

Alternative formulations of the principle of minimum change may be considered.

(1) Changing the fewest possible data items.
(2) Changing a weighted minimum number of data items.
(3) Moving the minimum distance in some conceptual space.

The first of these formulations is given by Fellegi and Holt and the second one is an alternative they suggest. The justification for using the second alternative may, however, relate to the conceptual intentions of the questionnaire rather than the reliability of each field. This may be illustrated with reference to the questions on education.

One education question asks for the respondent's highest school grade, three other questions ask for the respondent's post-secondary education and qualifications. By 'post-secondary' the Census had intended to refer to education of an advanced nature requiring a certain minimum schooling as an entrance requirement. Unfortunately, a surprisingly high proportion of respondents interpreted this as any education obtained after leaving school. Typically, the respondents making this error were giving two wrong answers consistent with each other but in conflict with the highest grade that was too low for entry into post-secondary education. In this case the minimum change was causing the highest grade to be incorrectly up-graded. It was finally decided that the best strategy was to modify certain rules to avoid the risk of serious distortion of the highest grade response by imputation.

4.2.3.2 Imputation Rules Derived from Corresponding Edit Rules

Among the subject-matter-oriented benefits of the system listed by Fellegi and Holt are:

(1) "Given the availability of a generalized edit and imputation system, subject-matter experts can readily implement a variety of experimental edit specifications whose impact can therefore be evaluated without extra effort involving systems development. This is particularly important given the generally heuristic nature of edit specifications".

The main problem with the data in both these two examples is that they are cases of infrequent errors on common conditions being mis-allocated to infrequent conditions.

(3) One type of error that created special problems was erroneous responses associated with common-law relationships. The intention of the Census was that consensual unions should be treated the same way as legal unions, hence allowing the identification of families. However, the frequent response pattern in these cases was to give the legal marital status, i.e. 'not married', together with the de facto relationship to head, either spouse or common-law spouse.

A typical patterns of response was:

Person 1.	Head of Household	Divorced
Person 2.	Spouse of Head	Single

in such a case the minimum change of data fields was to change the relationship to head of person 2 rather than the marital status of both persons. Problems of this nature were identified during the test Census. It was decided that the best strategy was to force the data using an uneditable derived variable. This was given a value 'Spouse Confirmed' whenever cases such as the above occurred. Then the responses were forced into the pattern:

Person 1.	Head of Household	Married
Person 2.	Spouse of Head	Married

There remained a residual problem as to how to edit children of the common-law partner in these cases. Certain distortions in the data were considered too critical to be left uncorrected. Additional strategies for correction were therefore adopted, either prior to the application of the Fellegi-Holt methodology as with common-law spouses or in certain cases as a clean-up afterwards. Evaluation is currently being undertaken to assess the correctness of the actions taken during the entire edit and imputation.

These particular problems which may be remedied by systematic corrections must, however, be weighed against the advantages of the method. There are very many rules to which the data should conform, each failed by a small number of records. Separate imputation rules for each of these would have required a much more complicated system.

The first of the two benefits, 'the parametric approach' referred to above must also be weighed against the loss of flexibility in specifying the imputation. However, for these edits even a very imperfect imputation action would have had a negligible impact on the final data.

The system created a framework within which alternative edit specifications could be reviewed, evaluated and modified very easily. It required a certain amount of work on the part of subject matter personnel to familiarize themselves with the system and its language. Once this had been achieved, however, considerable progress could be made in understanding the problems in the data and refining the edits.

One incident illustrated the flexibility of the system. Tabulations were run on the data at stages during the production. A tabulation indicated that a rule had been omitted from one particular set of rules. The erroneous condition detected was a rare condition that had not occurred in the test data, but was a condition that would never the less cause difficulties in the subsequent family formation programme. This omission was corrected within 48 hours. The system naturally cannot ensure that the user has included a complete set of edits, but it can ensure that the existing set is clean and consistent. It took much longer to make corrections to tailor made programmes with always the risk that a correction introduced a new error.

4.2.3.3 Retaining the Distributional Properties of the Clean Data

In the absence of any additional information, retaining the distributional properties of the clean data is the most appropriate strategy to take during imputation. The effectiveness of the system to achieve this was increased by the use of auxiliary constraints, that is fields used as matching criteria in the hot-deck search by reason of their correlation with the field to be imputed irrespective of any links by edit rules. There were, however, situations in which the dirty records were clearly drawn from a distribution very different from that of the clean records. These situations are equally true for any sub-sets of the population defined by other fields in the records. The inadequacy of the imputation as reflected in the final data in this case is a function of the difference between the two distributions and the proportion of dirty records.

There were two main reasons for this type of problem arising:

(1) Certain sub-groups of the population have difficulty selecting the correct response and are therefore more likely to fail to respond;

(2) Many questions include a 'null' or 'none' category. No device has yet been invented to prevent the relatively high non-response from persons who fall into this group.

This problem is illustrated by Table 1. This tabulates Labour Force Status defined from the unedited data. Clearly, there is a tendency for non-response to increase as the proportion of persons not in the Labour Force increases. This suggests that there is a tendency for non-respondents to be drawn more heavily from the non-participating population. It is possible to control imputation with respect to the variables in the Census, but not for any relationship beyond these.

An evaluation of this problem is currently being undertaken. Some consideration has also been given to possible enhancements to the methodology to adjust for this differential non-response. However, in order to utilize such enhancements, external information is needed to estimate the differential non-response rates with respect to the target variable.

Labour Force Status Identified from the Unedited Census Weighted Data

	Status Defined				Not Defined		Total
	In the Labour Force		Not in the Labour Force		Labour Force Status		
	Count	%	Count	%	Count	%	
BOTH SEXES							
15-19	346,243	42.39	424,653	51.99	45,872	5.72	816,768
20-24	534,746	71.99	175,551	23.63	32,554	4.38	742,851
25-54	2,065,994	68.46	851,709	28.22	100,203	3.32	3,017,906
55-64	340,744	50.57	304,740	45.23	28,304	4.20	673,788
65+	74,373	9.46	675,876	85.93	36,298	4.61	786,547
Total	3,362,100	55.68	2,432,529	40.29	243,231	4.03	6,037,860
MALES							
15-19	194,651	46.50	200,104	47.80	23,892	5.71	418,647
20-24	300,829	30.33	56,673	15.13	16,979	4.53	374,841
25-54	1,320,626	86.85	152,289	10.01	47,745	3.14	1,520,660
55-64	231,330	70.99	82,003	25.16	12,533	3.85	325,866
65+	52,347	15.73	264,988	79.61	15,541	4.67	332,876
Total	2,099,783	70.64	756,057	25.43	116,690	3.93	2,972,530
FEMALES							
15-19	151,592	38.08	224,549	56.50	21,980	5.52	398,121
20-24	233,917	63.50	118,878	32.27	15,575	4.23	368,370
25-54	745,368	49.78	699,420	46.71	52,458	3.50	1,497,246
55-64	109,414	31.45	222,737	64.02	15,771	4.53	347,922
65+	22,026	4.86	410,888	90.57	20,757	4.58	453,671
Total	1,262,317	41.18	1,676,472	54.69	126,541	4.13	3,065,330

5. CONCLUSIONS

The edit and imputation system developed from the methodology outlined by Fellegi and Holt was designed to be a generalized system. The major motive behind the development, however, was the needs of the Census as manifested in problems experienced during the edit and imputation of the 1971 Census. It was an attempt to bring order to a complex and potentially chaotic operation.

The system was very successful in achieving this objective. The edited data were available relatively earlier than the 1971 data. There has been no need for post-edit fixes. The residual problems in the data in general seem less serious than those found in 1971. There is a great deal more knowledge about data problems and means of correcting them.

This system has in fact allowed a much more critical analysis of the data and made it possible to identify problem areas such as systematic response error and non-response bias. Future work can be concentrated on a better handling of these problems within a controlled structure.

The following four issues are some of the key issues that need to be or are currently being addressed:

(1) A means for handling systematic errors that can be integrated with the existing system needs to be found.

(2) Alternatives to the principle of changing the minimum number of fields need to be investigated. Such alternatives may prove of limited value compared with the handling of systematic errors.

(3) Strategies for the handling of non-response to adjust for the differences between the responding and non-responding population should be considered.

(4) An experimental system for arithmetic edit and imputation is already being developed. The integration into this system of means of handling both quantitative and qualitative variables is among the possible long term plans.

Errors cannot be avoided no matter how carefully the survey is designed. The appropriateness of the edit and imputation strategy lies in its ability to recover the 'true' values. To achieve this there is a need for more empirical evidence concerning the nature of errors in the data.

RESUME

À partir de la méthode systématique de vérification et d'imputation proposée par I.P. Fellegi et D. Holt, on a mis au point un système général de vérification et d'imputation par la méthode du hot-deck et on l'a appliqué aux données du recensement de la population et du logement de 1976. Le présent article étudie cette application de la méthode Fellegi-Holt et évalue les points forts et faibles de la méthodologie à partir de certains exemples empiriques. La présentation du système est faite dans un contexte plus vaste, celui des grands problèmes posés par la vérification et l'imputation des données d'enquête. L'auteur énumère aussi quelques avenues de développement possibles.

REFERENCES

[1] Fellegi, I.P. and Holt, D., "A Systematic Approach to Automatic Edit and Imputation". Journal of the American Statistical Association, March 1976. Volume 71, Number 353.

[2] Graves, R.B., "Can-Edit, A Generalized Edit and Imputation System In A Data Base Environment". A report to the working party on electronic data processing, Conference of European Statisticians. (CES/WP.9/142). Feb. 1976.

[3] Turner, M.J., "The Use of Data Base Technology in Large Systems". A report to the working party on electronic data processing, Conference of European Statisticians, April 1977.

[4] Quality of Date, '76 Census, Series I: Sources of Error - Effect of Edit and Imputation Procedures on the Quality of Data from the 1976 Census of Population and Housing. Catalogue No. 99-843.

LARGE SCALE IMPUTATION OF SURVEY DATA [1]

M.J. Colledge, J.H. Johnson, R. Pare, and I.G. Sande [2]

Owners of small businesses complain about the quantity of forms they are required to complete and tend to blame the collectors of statistics. Administrative data are an alternative source but do not usually include all the information required by the survey takers.

The "Tax Data Imputation System" makes use of tax data collected from a large number of businesses by Revenue Canada and data obtained by sample survey for a small subset of these businesses. Survey data is imputed (estimated) for all the businesses not actually surveyed using a "hot-deck" technique, with adjustments made to ensure certain edit rules are satisfied. The results of a simulation study suggest that this procedure has reasonable statistical properties. Estimators (of means or totals) are unbiased with variances of comparable size to the corresponding ratio estimators.

1. INTRODUCTION

The demand for statistical information to aid government and management decision making has been increasing for many years. In the past, Statistics Canada was able to cope with this situation by expanding the scope and number of their surveys. Recently, such expansion has become inhibited as a result of two factors. Firstly, there is an increasing sensitivity to complaints from respondents about the burden of completing questionnaires. Secondly, current fiscal policies prevent growth in manpower. There is no indication that either of these factors is likely to be shortlived. Thus, in order to cater for an increased demand for information without raising costs or response burden, Statistics Canada is committed to making the best possible use of existing data, including data collected by other agencies for administrative purposes. One particular manifestation of this policy was the decision

---

[1] Adapted from a paper presented at the Annual Meeting of the American Statistical Association, August 14-17, 1978, San Diego, California, U.S.A.

[2] M.J. Colledge, J.H. Johnson, R. Pare, and I.G. Sande, Business Survey Methods Division , Statistics Canada.

to use financial data from Revenue Canada to supplement two annual surveys of businesses for the 1975 reference year. This paper deals with the systems which evolved as a result.

The Census of Construction (COC) is concerned with about 80,000 businesses in Canada whose primary activity is construction. The COC had been a census, but a decision was made to reduce the response burden of smaller businesses. For the 1975 reference year, only businesses with gross business income (GBI) of at least $5,000 were considered in scope. These were divided into two groups: "small" businesses having a GBI of less than $500,000 and "large" businesses. The latter group were the subject of a census operation; all large businesses were mailed a questionnaire asking for a comprehensive set of data. Small business information was derived from two sources: from Revenue Canada and from a mailout as follows. A sample of businesses stratified by GBI, was selected from Revenue Canada tax files, the largest business being selected with certainty. Basic financial data was transcribed for these businesses. For a subsample, secondary (more detailed) financial data was obtained from the tax return. The size of this subsample was limited by the costs of the additional transcription. A second subsample, designed to overlap the first to some extent, was selected and mailed a survey questionnnaire requesting only non-financial data. The size of the second subsample was limited by the need to reduce response burden and costs. Thus in comparison with a full census, the COC response burden was reduced by sampling and reducing the number and type of questions asked.

Arrangements for the Motor Carrier Freight Survey (MCF) were along the same general lines. The significant differences were that the universe of about 25,000 was divided into "small" and "large" by a GBI threshold of $100,000, no subsample of secondary financial data was obtained and the survey questionnaire requested a full range of information (not just non-financial).

The decision to utilize administrative tax data for the COC and MCF came quite abruptly and in advance of experience, existing software, data or feasibility study. The short time scale combined with a restricted budget dictated certain constraints on the design. Firstly, program development and testing had to be substantially achievable before any real data were available. Secondly, the programs had to be robust and easily modifiable in order to allow adjustment for unexpected characteristics of the data. Thirdly, the programs had to interface with existing systems associated with the surveys, in particular, the tabulation systems which had been developed for census operations in previous years. Thus the following design decisions were made:

i) data from tax and survey sources would be combined at the micro level, i.e. level of individual businesses;

ii) a complete set of data (all financial and non-financial items) would be imputed at micro level for all businesses using a "hot-deck" technique with constraints to ensure that imputation was consistent with prescribed edit rules;

iii) the data would be inflated to universe level by replication to allow tabulation by existing systems which had not been developed to handle weights;

iv) programs would be modular and readily adaptable to new or modified imputation and edit rules.

The following sections of this paper elaborate upon the design features and describe the systems implementation which processed 1975 data for the COC and MCF. An evaluation of the procedures is given in section 5.

2. OVERVIEW

The central feature of the system is the imputation procedure, discussed in detail in sections 3 and 4. The purpose of this section is to outline the environment within which the procedure operates by describing the complete system. The scale of processing is illustrated by reference to figures for the small business portion of the COC universe.

A system flow chart is shown in figure 1.

MERGE The first module labelled MERGE brings together data records from tax and survey sources. The input data files have been individually cleaned and edited. The output is a set of records, one per business, each of which contains a basic tax data segment and may (or may not) contain secondary tax data or survey data segments. The existing segments may have sporadic missing entries in various fields, also, some entries may be inconsistent with one another.

CHECKIN The essential purpose of the second module, CHECKIN, is to prepare data for imputation by screening out unusable or unwanted data. The module reformats the records, strips off irrelevant fields, identifies out of scope or duplicate records, checks entries against a set of prescribed edit rules, blanks out inconsistent entries and identifies all missing fields. Any record which is out of scope or a duplicate or contains insufficient useful data is flagged ("dropped"); the remainder are subject to processing by the next module, IMPUTE.

Columns 1 and 2 of figure 2 illustrate the results of processing COC data. Some 9106 of the 50,538 merged records were declared out of scope (by being in the wrong industry or too large, for example). Of the remainder,462 were dropped leaving 40970 "good" records.

IMPUTE This is the major processing module. Its function is to impute all missing fields on every record. For the COC data, 884 records contained all segments, 3963 records required imputation of just the secondary financial segment, 2186 records required imputation of just the survey segment and 33937 records required both (see figure 2, column 3). In addition, some entries in existing segments were missing.

CHECKOUT Although, in principle, imputation is constrained by the edit rules, in practice inconsistent values may be imputed due to shortcomings in specification or programming. Furthermore, imputation may fail in the sense that no suitable value for a field can be located. Thus,the function of CHECKOUT is to check the records against the same prescribed set of rules as were applied to the data at input, and to identify and "drop" records containing inconsistent or missing entries.

From columns 2 and 3 of figure 2, it can be deduced that 194 COC records were inconsistent or incomplete and had to be dropped.

INFLATE The function of the last processing module in the system is to raise the sample of good records to the population level and thereby generate an output file which can be tabulated by the census tabulation system. Inflation is achieved by replicating each record according to its weight after "correction". All records entering the system carry a weight which is the inverse of the probability with which the record entered the basic tax sample. Three types of correction are applied prior to replication :

i) Duplication correction. Some businesses are represented by more than one record.

ii) Out of scope correction. There are instances where the tax data information suggests the business is in scope, whereas the survey data indicates it is not. The survey data is assumed to be more reliable. In order to allow for possible inclusion of out of scope records containing tax data only, a correction factor is applied based on data from businesses for which tax and survey information is obtained.

iii) Dropped record correction. Records for some in scope businesses are dropped because of inadequate or inconsistent data.

Only the last type of correction is relevant in the imputation context. It implies that the imputation procedure need not be 100% successful for every record as a correction can be made.

Figure 2 indicates that after weight correction and inflation, a file of 78,563 small Construction businesses was obtained.

Imputed data is clearly identified on all files and the sponsor has access to the intermediate files to check on the reasonableness of the imputation. Some auditing and tabulation functions are also provided. The final output file has to be written in a format which can be accepted by a tabulation system which predates the imputation system and so special identifiers do not appear on this file.

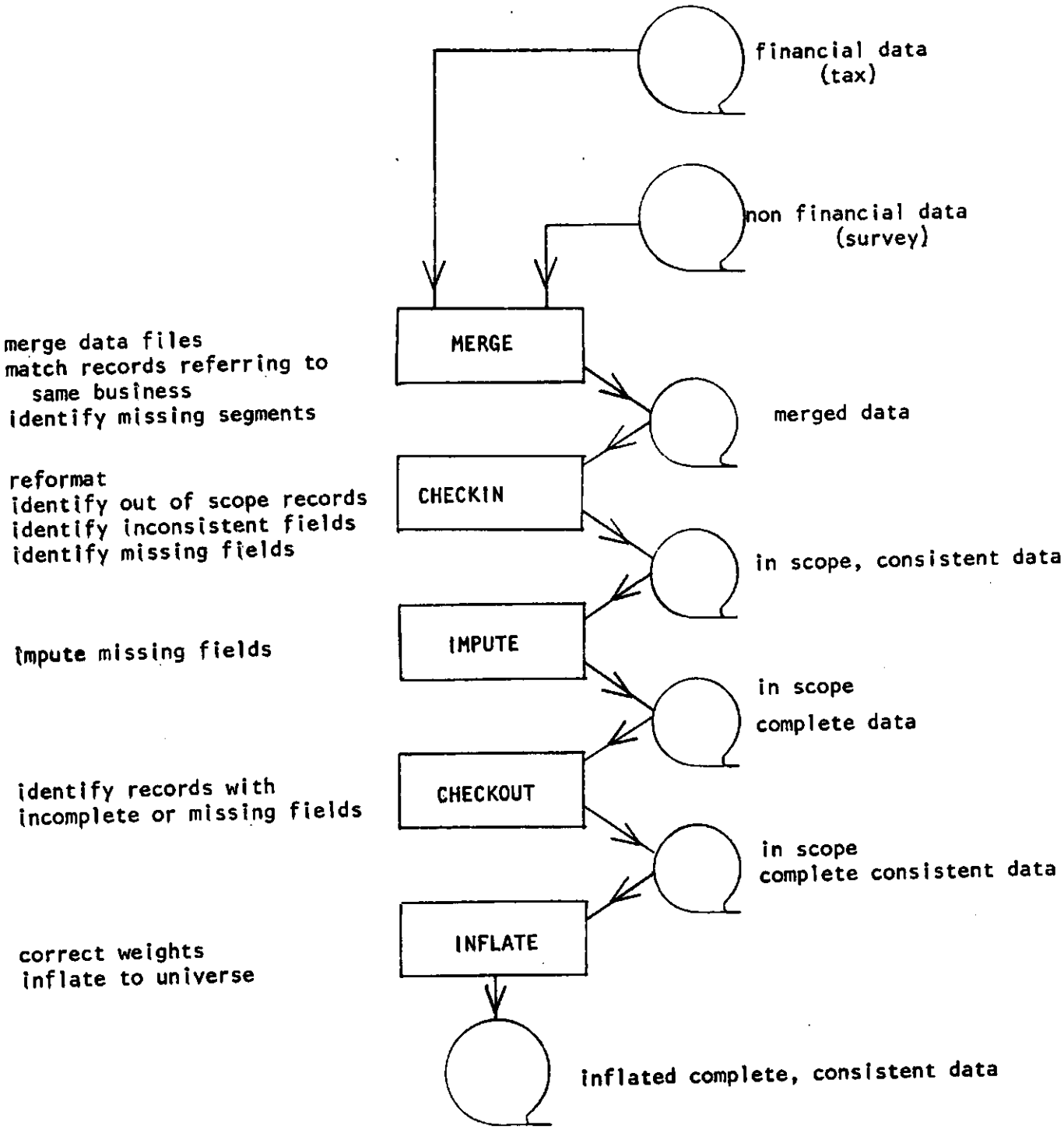


Figure 1. System Flow Chart

Figure 2. Summary of Results of Processing Census of Construction, 1975 Reference Year.

				1 At input to System	2 After Checkin	3 At output from System	4 Blown up to Universe
Out of Scope				9106	9106	9106	-
In Scope	Data not Good			0	462	656	-
	Segments Present						
	Tax		Survey				
	Basic	Secondary					
Good	XXX			34,181	33,937	0	-
	XXX	XXX		2,316	2,186	0	-
	XXX		XXX	4,027	3,963	0	-
	XXX	XXX	XXX	908	884	40,776	-
Total Good				41,432	40,970	40,776	78,563
Total				50,538			

3. IMPUTATION METHODOLOGY

For purposes of imputation, the record for each business can be considered as consisting of four types of segment:

i) Key fields. These consist of fields used for classification or matching and are collected or derived from the tax return. The actual fields used were the standard industrial classification (SIC), province, salaries and wages indicator (SWI, set to 1 or 0 according as there is any indication that salaries or wages were paid or not), gross business income (GBI), net business income (NBI). If any of these fields were missing, the record was not used in the imputation.

ii) Basic financial data collected from the tax return, e.g. depreciation, purchases, closing inventory. An attempt is made to collect this data for all businesses sampled, but the information available with the return may be insufficient or unclear. Thus some or all of these fields may be missing, i.e. the segment may be incomplete. If not, all fields are present and the segment is complete.

iii) Secondary financial data, collected from tax returns for a subsample of records. These detailed financial data, e.g. balance sheet, detailed expense breakdowns, were collected only for the Census of Construction; but, potentially, one or more such subsamples might exist. This segment may be either complete (all fields present), incomplete (some fields present) or missing (no fields present, as in the case of records not in the subsample).

iv) Survey data, collected for a subsample of records. This segment may be complete, incomplete or missing. In addition, there are a variety of control fields and flags.

The imputation problem is to complete the incomplete segments and to supply the missing segments.

A possible imputation procedure would be to model the missing fields in terms of those that are present. If the number of fields were very large (as it is here) and the constraints (or edit rules) on the fields were at all complex, structuring the model would be very difficult. One would have to evaluate several models to determine the best fit and this would have to be done after the data had been collected and edited. As a result, a great deal of time would be spent experimenting with the data just when one could least affort it - when the publication deadlines were approaching and a great deal of processing had yet to be done.

Thus, modelling the data did not seem a very attractive option and a type of hot-deck technique was devised. In this procedure, a record requiring imputation (candidate) is matched with a complete record (donor). The donor supplies the missing fields, possibly with some adjustment so that the edit rules are satisfied. This procedure produces realistic looking data and can be expected to preserve the underlying distributions, whereas modelling tends to produce smoothed data and distorts distributions. Another advantage is that the imputation can be set up and ready to run before the data collection is finished.

The hot-deck requires a reasonable supply of complete records, but in fact there are few records with all segments complete. If one attempted to impute for all missing fields in a single pass, the same donors would be used excessively, no use would be made of records with partial information, and the matches would be poor. In addition, a matching procedure appropriate for one segment may not be appropriate for another. Therefore, the imputation is broken up into several phases, each corresponding to a segment or sub-segment.

Phase 1. Candidates are records with Segment A incomplete (but not missing).
Donors are records with Segment A complete. At the end of Phase 1, all records have Segment A complete or missing.

Phase 2. Candidates are records with Segment A missing. Donors are records with Segment A complete (including records which were Phase 1 candidates). At the end of Phase 2, all records have Segment A complete.

Phase 3. Candidates are records with Segment B incomplete (but not missing). Donors are records with Segment B complete. At the end of Phase 3, all records have Segment B complete or missing. Those with Segment B complete are eligible as donors in Phase 4.

In order to match candidates with donors, the file of all records is stratified by Province (or Region), SIC and SWI. The collection of potential donors (i.e. the hot-deck) as well as the collection of candidates are identified for the particular phase. Within the stratum, the records are ordered by GBI. A sequence of records from a stratum might be represented like this:

GBI: \$25K \$26K \$27K \$28K \$29K

...CCDCCDCCD$_{-5}$CCD$_{-4}$CCD$_{-3}$CCCD$_{-2}$CCCD$_{-1}$C$_o$CD$_1$CCD$_2$CCCD$_3$CCD$_4$D$_5$CCD ....

The C's are candidates and the D's are donors (other records not involved in this phase are not represented). In order to impute for C_o, only the nearest 5 potential donors on "either side" of C_o are considered, a total of 10 possible donors which are all about the same size (in terms of GBI) as the candidate. The number 5 is quite arbitrary - it could as well be 3 or 10, or the two sides could be of different lengths, but the imputation seems relatively insensitive to this parameter. From the "nearest" 10 donors, that one is chosen which minimizes a distance function DIST (C,D). DIST can be quite a complex function, but the basic structure used was

$$DIST\ (C,D) = |\log EXP_C - \log EXP_D|$$

where EXP = GBI - NBI = total expenses, and the subscripts C and D denote values from the candidate and donor records respectively. EXP was used because many of the fields to be imputed are detailed expense breakdowns or correlated with expenses.

Note that GBI and NBI are key fields, so that DIST is always determined. DIST may also depend on other key fields, or fields which have already been imputed in an earlier phase, or even meta-data. In particular, the distance function may be modified to spread donor usage, e.g.

$$DIST\ (C,D) = |\log EXP_C - \log EXP_D| \quad (1 + p.n_d)$$

where n_D = number of times the potential donor D has already been used as an actual donor in the phase,

and p = the proportional penalty for each usage (e.g. .02). The size of p depends on the amount of imputation to be done and the degree of concern over having one donor used much more frequently than another.

After a suitable donor has been identified, the candidate's missing fields are supplied from the corresponding fields in the donor record. Some adjustment or transformation may be necessary to ensure that the constraints (edits) are satisfied. For example, three fields, X,Y and Z may have to satisfy $X + Y \leq Z$ with X, Y and Z all non-negative. The donors's values for these fields are X_D, Y_D and Z_D while the candidate has X and Y missing and the value Z_C in the Z field. If the values X_D and Y_D are simply written into the corresponding candidate fields, we may find that $X_D + Y_D > Z_C$, which violates the edit. Therefore, it is better to prorate X_D and Y_D to ensure that the edit holds:

$$X_C = (X_D/Z_D)\ Z_C; \quad Y_C = (Y_D/Z_D)\ Z_C \quad .$$

In other words, the proportions X_D/Z_D and Y_D/Z_D are transferred to the candidate. A common example is

$$FUEL_C = (FUEL_D/EXP_D)\ EXP_C$$

where FUEL is the amount spent on fuel and EXP is the total expenses. This imputation estimates that the candidate spent the same proportion of his total expenses on fuel as did the donor.

The transformation needed to impute a field may be more complex if the field is involved in several edits. For example, the four fields W, X, Y, Z, may have to satisfy $X + Y \leq Z$ and $X \leq W$, where all fields are non-negative. The donor's values for these fields are W_D, X_D, Y_D, Z_D. The candidate has W_C, X and Y missing, and Z_C. An appropriate imputation (but not necessarily the only one) is

$$X_C = \text{Min}\ \ W_C,\ (X_D/Z_D)\ Z_C$$

$$Y_C = (Y_D/Z_D)\ Z_C.$$

When the edit rules are even more complex a decision table may be required, where the form of imputation depends on which set of conditions holds. In desperate situations, a table of default values may be used.

If a field is not involved in any edits, it may be prorated using a correlated variable in the case of a numeric field. Categorical data may simply be copied from donor to candidate.

The imputation specifications are written separately for each field - no generalized transformation is used. They are written in such a way as to produce consistent data and this involves not only accommodating constraints, but also ensuring that constraints are not violated due to roundoff error.

4. IMPLEMENTATION

The systems design was based on the following premises:

a) The breakdown into phases each of which is functionally the same. except in detail, suggested a general system which would be tailored separately for each phase.

b) To simplify data-set control, the output produced from a phase would have the same record description as the input and all records would be carried forward. Each phase would identify its donors and candidates, perform imputation, and copy all other data as is.

c) Instrumentation of the system would mostly be done offline by analysis of a log file describing imputation "events", and by investigation of the output of each phase.

d) Fields would either have a value or be missing. If missing, any value which it might have had would be ignored for imputation purposes.

e) Fields would be identified as missing only at beginning of processing. Once imputed to a value, the field stays imputed. Thus, inconsistencies must be removed at the beginning and never introduced by imputation.

f) The control language should be quite flexible to allow unusual imputation rules, but should still be quite readable since it would be the final specification of side effects in unusual situations.

g) One donor only would be used in each phase.

The effect of these considerations on the design was to simplify the systems development and operation of the system while retaining flexibility in the details of imputation. This would facilitate final turning without holding up production more than necessary.

Consideration a) resulted in the general phase structure shown in Fig.3. Basically four modules are involved along with three utility sorts:

i) CNVT is responsible for identifying that subset of the file that is to be involved in imputation. For each donor or candidate it writes out an "Imputation Control Segment" (ICS) which contains match fields for donor assignment as well as space for indicating the donor actually assigned.

ii) NEBR performs the assignment of donor to candidate on the basis of match fields. The ICS file has been stratified by sorting on a KEY. A local search is performed in a large circular buffer (about 2000 segments) and the best match according to some measure is selected.

iii) MERG combines a copy of the appropriate donor record to each ICS record.

iv) IMPT then performs consistent imputation using the donors assigned.

Consistent imputation (for linear edits) was aided by a routine that kept track of the current upper and lower bounds for each field, determined by the edits and the fields already assigned. For each field to be imputed, assignment would be done if the value were in range, and the ranges of the remaining unassigned fields would then be adjusted appropriately. The routine caused the actual assignment to be made and a log entry to be written.

Where it could be applied, this approach simplified the work enormously. Unfortunately, it could not be made universally applicable without in effect solving an integer programme at each field assignment. Nonetheless, the edit rules which occurred were predominantly positivity restrictions and simple sums. Some conditional edits could be handled by selectively activating edits. Others were handled by taking great care with the imputation rules. However, the potential for an inconsistent imputation still remained.

Flexibility (consideration (f))was ensured by allowing the control language to be a number of inclusions into the general programmes which could then be compiled to produce executable modules. The environment of each inclusion is carefully documented and service routines are provided for certain common functions.

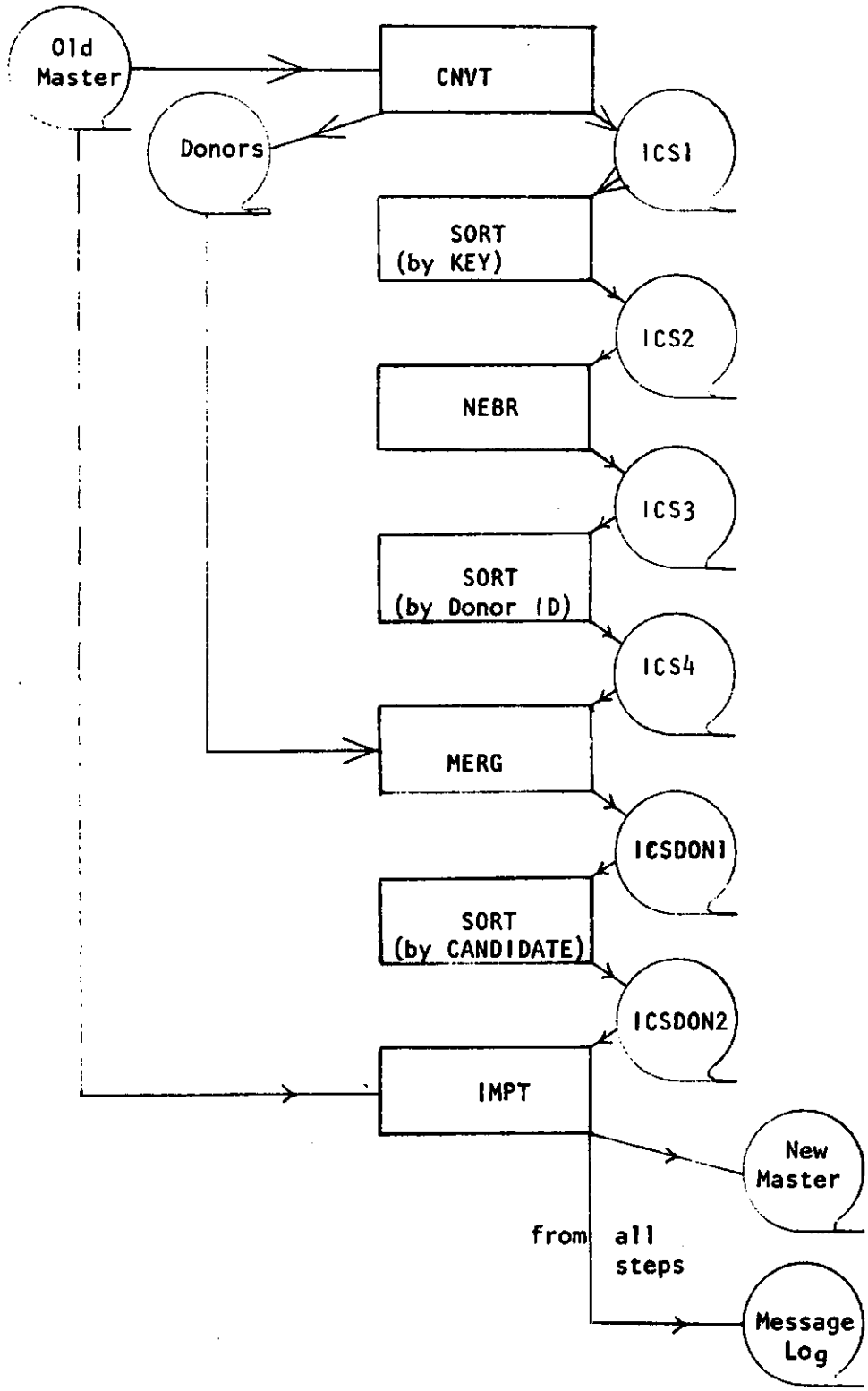


Figure 3. General Phase Structure

5. EVALUATION

The imputation procedure described in section 3 will produce estimates of the population totals (or means), but some assessment of the quality of these estimates, in terms of bias and variation, is required. One would like to know how the quality of the estimate varies with (a) the sampling bias, (b) the population size, (c) the sampling rate, (d) the correlation or relationship between the imputed variable and the auxiliary variable used for prorating, (e) the size of the window used to determine the number of eligible donors, (f) the complexity of the edits, (g) the distance function, and (h) the control of donor usage. One would also like to compare the "imputation" estimate with some natural competitors, such as the usual sampling (expansion) estimate and the ratio estimate.

A small simulation study has been done to examine the effects of sampling bias (in a nominally simple random sample) and sampling rate for a population of fixed size.

A population of 1000 units was created, each consisting of five variables corresponding to GBI, NBI and the "expense items": "salaries", "depreciation" and "purchases". GBI and NBI were the auxiliary variables. All quantities except NBI are non-negative and, in addition, we have the edit rule.

$$\text{Salaries} + \text{depreciation} + \text{purchases} \leq \text{GBI} - \text{NBI} = \text{EXP}.$$

We omit the gory details, but the distribution of the non-negative variables is skewed towards zero.

Sampling was either unbiased or biased. Biased samples were created by ordering the population on GBI and (a) selecting 25% of the sample from below the median GBI and 75% of the sample from above the median GBI (bias up), or (b) reversing the percentages in (a)(bias down).

The sampling fractions were 10%, 20% and 50%.

For each sampling bias and sampling rate, twenty-five independent samples were selected from the same population. For each sample, a new file was created for the population in which GBI and NBI were retained for all records and salaries, depreciation and purchases were included for the sampled records only. Salaries, depreciation and purchases were then imputed for the non-sampled records, using the sampled records as the hot-deck and prorating on EXP. For each replicate, the imputation, sampling and ratio estimates of the population means were calculated. These could then be compared with the known population values.

Table I gives the mean over 25 replicates divided by the population mean for each type of estimate, bias condition, sampling rate and variable. The t statistic, evaluating the "significance" of the difference between the population mean and the average value of the 25 estimates is given in parenthesis. The population correlation between the imputed variable and the prorating variable is given in parenthesis in the first column. For the unbiased case, all types of estimates do quite well, except that the ratio estimate begins to show bias at a 50% sampling rate. For the biased cases, the imputation estimate clearly does better than the ratio estimate. The sampling estimate does very badly as one would expect.

Table II gives the coefficient of variation of the estimates in the form of the standard deviation calculated for the 25 replicates divided by the population mean. For the unbiased case, the coefficients of variation are about the same for the imputation and ratio estimates, while that of the sampling estimate is much larger. This is also true for the upward biased case. In the downward biased case, the position is less clear and the estimates appear to be roughly equivalent; but if one considers the root mean square error divided by the population mean, the bias dominates and the imputation estimate is clearly superior.

The implication of Table II is that in order to estimate the variance of an imputation estimate (in a "real" situation where replicates are not available) one may formally use the estimate of the variance of the corresponding ratio estimate as a reasonable approximation.

It will be noticed in Table I that the correlations between the imputed and prorating variables are quite high, higher than one might expect in "real" data. We would expect the difference between the imputation and the ratio estimate to become less pronounced as the correlation decreased; but no systematic work has been done to investigate this.

When the correlations are high, the size of the window appears to have no effect on the quality of the imputation estimate.

We have some evidence to suggest that when the correlations are low and the sampling rates are very low, all estimates are bad.

Table 1 : Relative Bias of Estimates

Mean over 25 replicates / Population mean (t 24)

Variable ρ	Sampling Fraction	UNBIASED Imputation	UNBIASED Sampling	UNBIASED Ratio	BIASED UP Imputation	BIASED UP Sampling	BIASED UP Ratio	BIASED DOWN Imputation	BIASED DOWN Sampling	BIASED DOWN Ratio
	10%	.996	1.004	.998	.995	1.329	1.026	1.005	.690	.961
Salaries		(-.6)	(.2)	(-.2)	(-1.1)	(16.6)	(4.2)	(.6)	(-23.6)	(-4.8)
(.95)										
	20%	.998	1.000	.999	.997	1.330	1.029	.999	.673	.947
		(-.4)	(0.0)	(-.2)	(-.7)	(32.0)	(8.9)	(-.1)	(-34.7)	(-6.9)
	50%	1.000	1.000	.996	.996	1.338	1.030	.996	.670	.944
		(-.1)	(-.02)	(-3.4)	(-2.1)	(55.7)	(21.8)	(-1.2)	(-115.4)	(23.9)
	10%	1.004	1.003	1.000	1.004	1.254	.993	.993	.746	1.041
		(.5)	(.2)	(.1)	(.7)	(30.2)	(-9.0)	(-.8)	(-25.5)	(4.9)
Depreciation										
(.89)	20%	1.001	1.000	1.001	1.004	1.251	.968	1.002	.754	1.056
		(.2)	(0.0)	(.2)	(.8)	(43.8)	(-9.0)	(.3)	(-47.4)	(7.0)
	50%	1.000	.993	1.003	1.005	1.258	.969	1.004	.751	1.059
		(-.2)	(-1.4)	(2.0)	(2.0)	(88.1)	(-17.2)	(1.1)	(-80.7)	(23.8)
	10%	.993	1.008	1.002	1.000	1.342	1.036	1.004	.681	.946
Purchases		(-.6)	(.3)	(.2)	(0.0)	(12.9)	(2.8)	(.2)	(-17.3)	(-3.2)
(.82)	20%	.999	1.000	.999	.993	1.341	1.038	.979	.659	.921
		(-.1)	(0.0)	(-.1)	(-.9)	(23.4)	(5.1)	(-1.4)	(-25.9)	(-6.2)
	50%	.999	.996	.992	.995	1.343	1.034	.994	.658	.927
		(-.3)	(-.6)	(-2.5)	(1.8)	(45.1)	(12.7)	(-.8)	(-72.4)	(-12.7)

Table II : Coefficients of Variation of Estimates

Standard Deviation of 25 Replicates/Population mean.

Variable	Sampling Fraction	UNBIASED Imputation	UNBIASED Sampling	UNBIASED Ratio	BIASED UP Imputation	BIASED UP Sampling	BIASED UP Ratio	BIASED DOWN Imputation	BIASED DOWN Sampling	BIASED DOWN Ratio
Salaries	10%	.039	.106	.039	.024	.099	.031	.043	.066	.041
	20%	.021	.081	.022	.023	.052	.017	.040	.047	.038
	50%	.010	.028	.006	.011	.030	.007	.018	.014	.012
Depreciation	10%	.039	.078	.038	.024	.042	.031	.043	.050	.041
	20%	.021	.059	.021	.024	.029	.018	.040	.026	.040
	50%	.010	.025	.007	.012	.015	.009	.019	.015	.012
Purchases	10%	.066	.127	.065	.055	.132	.065	.101	.092	.084
	20%	.039	.091	.042	.039	.073	.037	.073	.066	.065
	50%	.021	.036	.016	.014	.038	.013	.041	.024	.029

6. CONCLUSION

Planning for the 1975 imputation system started in April 1976 and the final output data were delivered in August 1977. Most of the delays were due to problems with data collection and survey processing. Publications based partly on the imputed data have been released.

For 1976 data, the imputation system and methodology were refined and at least one survey, the Census of Construction, should run on virtually the same system with 1977 data.

Large-scale imputation appears to be a useful new weapon in the arsenal; but more evaluation should precede more widespread use. At the moment, assessment of its feasibility in any situation is based more on hunches than facts. Unfortunately, thorough and systematic evaluation promises to be a lengthy process and the best we can hope for are piecemeal results.

RESUME

Les petits entrepreneurs se plaignent de la quantité de formules qu'il leur faut remplir et ont tendance à accuser les responsables de la collecte des statistiques. Les dossiers administratifs constituent une autre source possible, mais il y manque souvent des renseignements essentiels aux enquêteurs.

Le système d'imputation à l'aide des données fiscales a recours aux données fiscales recueillies par Revenu Canada auprès d'un grand nombre d'entreprises et aux données obtenues par sondage auprès d'un petit sous-ensemble de ces entreprises. Les données sur les entreprises qui ne font pas partie de l'échantillon du sondage sont imputées (estimées) par la méthode du hot-deck, certaines corrections étant apportées pour assurer le respect de diverses règles de validation. Les résultats d'une simulation semblent indiquer que cette méthode possède des propriétés statistiques raisonnables. Les estimateurs (des moyennes ou des totaux) sont sans biais, et leurs variances présentent des grandeurs comparables à celles des variances des estimateurs obtenus par la méthode du quotient.

SOME METHODS FOR UPDATING SAMPLE SURVEY FRAMES[1] AND THEIR EFFECTS ON ESTIMATION

J.D. Drew, G.H. Choudhry, and G.B. Gray[2]

Frames designed for continuous surveys are sometimes used for ad hoc surveys which require selection of sampling units separate from those selected for the continuous survey. This paper presents an unbiased extension of Keyfitz's (1951) sample updating method to the case where a portion of the frame has been reserved for surveys other than the main continuous survey. A simple although biased alternative is presented.

The scope under Platek and Singh's (1975) design strategy for an area based continuous survey requiring updating is then expanded to encompass rotation of first stage units, establishment of a separate special survey sub-frame, and procedures to prevent re-selection of ultimate sampling units.

The methods are evaluated in a Monte Carlo study using Census data to simulate the design for the Canadian Labour Force Survey.

1. INTRODUCTION

Sample surveys frequently incorporate designs utilizing unequal probabilities of selection of units within strata. Since many characteristics are highly correlated with the relative sizes of the units, estimates based on such designs are in general more efficient than estimates based on designs where the sizes of the units are ignored. In continuous surveys, the sizes of the sampling units may change over time because of births and deaths of ultimate sampling units (e.g., construction or demolition of dwellings in the case of household surveys). An even rate of growth among the sampling units results in a decrease in the correlation between the characteristics being measured from the survey and the size measures, and consequently results in less efficient estimates than in the initial period.

---

[1] Adapted from a paper presented at the Annual Meeting of the American Statistical Association, August 14-17, 1978, San Diego, California, U.S.A.

[2] J.D. Drew, G.H. Choudhry, and G.B. Gray, Household Surveys Development Division, Statistics Canada.

In the case of sample designs based on area frames, a solution to the problem of out of data relative sizes lies in their periodic check by regularly scheduled field counts, followed by a revision of the selection probabilities, and finally a necessary change in the sample to reflect the new probabilities. Keyfitz [4] presented a method whereby revised selection probabilities could be incorporated while maximizing the probability of retaining the originally sampled unit in a stratum. More recently, Kish and Scott [5] adapted Keyfitz's procedure to other cases, in particular, where units are shifted from one stratum to another. The chief drawback of the above methods is that they can be applied only to sample designs in which one unit is selected per stratum. This implies that unbiased variance estimates cannot be obtained.

Rao, Hartley, and Cochran [7] devised a sampling procedure referred to as the random group method in which unbiased estimates and their variances can be obtained while selecting one unit per random group. As suggested by Platek and Singh [6], the Keyfitz update procedure may be applied to each random group.

In Section (2), we present an unbiased extension of Keyfitz's [4] sample updating procedure to the case where one first stage unit (fsu) is selected per stratum with unequal probability but where a portion of the fsu's, excluding the selected one, is reserved exclusively for special survey use. The units are reserved by applying some known probability mechanism, and at the time of sample update, the continuous survey is restricted to the non-reserved portion of the frame. The method incorporates "Working Probabilities" following an approach similar to that used by Fellegi [1] in his PPSWOR selection procedure.

In Section (3), we extend the study of update strategy to a rotating sample in which the random group method is applied. After selecting one unit with pps in each random group for the continuous survey, a specified portion of the remaining units within each group is reserved

with SRSWOR for special surveys. For the particular rotation scheme under consideration, it is shown that when units are reserved in the above manner, the probabilities of selection for the continuous survey remain unaffected prior to update. The unbiased updating procedure in Section (2) is adapted to accommodate the rotation scheme. As an alternative, a biased updating procedure, which approximates Working Probabilities by the revised probabilities of selection, is considered.

In Section (4), the reserved units from each random group within a stratum are merged together to form a special survey frame. Hartley and Rao's [3] randomized pps systematic method is employed to select samples from the special survey frame and an estimation procedure for special surveys is described.

In Section (5), we report the results of a Monte Carlo study based on the random group design. This design is used by the Canadian Labour Force Survey in self representing areas.

2. SAMPLE UPDATE WHEN A PORTION OF THE FRAME IS RESERVED: (NON-ROTATING CASE)

Consider a stratum which has N first stage sampling units. A size measure X_i is associated with the ith unit in the stratum; i=1,2,...,N. One unit from the stratum is selected for a continuous survey with pps where p_i, the probability of selecting unit i for the continuous survey is given by

$$p_i = X_i / \sum_{i=1}^{N} X_i ; \qquad i=1,2,...,N.$$

We assume that there is no rotation of fsu's for the continuous survey. Following the initial selection of one unit for the continuous survey, some of the remaining fsu's are reserved for use by special surveys, by some unknown probability mechanism. At the time of sample updating, the continuous survey is restricted to the non-reserved portion of the frame.

Let s denote a set of n units reserved for special surveys, and let S by any such set (note that S is a function whereas s is a realization), then Pr(s) is the probability of reserving the set s of units in any order. Let C denote the continuous survey. We have

$$\Pr(s) = \sum_{j \notin s} \Pr(j \text{ selected for C}).\ \Pr(s \mid j \text{ selected for C})$$

$$= \sum_{j \notin s} P_j.\Pr(s \mid j \text{ selected for C}). \qquad (2.1)$$

The only restriction placed on methods of reserving units is that the computation of Pr(s) should be practical.

At the time of update, revised size measures X_i' are obtained for each unit i=1,2,...,N. We require that the new probabilities of selection for the continuous survey C should be:

$$p_i' = \frac{X_i'}{\sum_{i=1}^{N} X_i'} \qquad i=1,2,\ldots,N. \qquad (2.2)$$

Note that the revised selection probabilities for the continuous survey are constrained by the non-selection of the reserved units. We therefore define, "Working Probabilities" $p_i(2)$, i=1,2,...,N, such that the overall probability of selecting unit i when averaged over all possible reserved sets of n out of (N-1) units excluding unit i should equal p_i', i.e.,

$$\sum_{s}' \Pr(s) \left(\frac{p_i(2)}{1 - \sum_{j \in s} p_j(2)}\right) = p_i' \qquad i=1,2,\ldots,N\ , \qquad (2.3)$$

where Σ_s' denotes the sum over all possible unordered n-tuples from (N-1) units, excluding unit i, and Pr(s) is defined by expression (2.1). Therefore, from (2.3) we have:

$$p_i(2) = \frac{p_i'}{\Sigma_s' \frac{Pr(s)}{1 - \Sigma_{j\varepsilon s} p_j(2)}} \qquad i=1,2,\ldots,N. \tag{2.4}$$

The solution for $p_i(2)$'s can be obtained iteratively by using p_i as initial values. Note that as N and n increase combinatorial difficulties quickly arise since N $\binom{N-1}{n}$ summations are involved for each iteration. The post-update conditional probability of selecting unit i, given the set s of reserved units, is:

$$\Pi'_{i|s} = \frac{p_i(2)}{1 - \Sigma_{j\varepsilon s} p_j(2)} . \tag{2.5}$$

The posterior probability for the continuous survey to contain the ith unit as the selected one given that the set of s of units was reserved is

$$\Pi_{i|s} = \frac{Pr\ (i \text{ selected for C and the set s of unit reserved})}{Pr(s)}$$

$$= \frac{p_i.(Pr(s|i \text{ selected for C}))}{Pr(s)} . \tag{2.6}$$

We now perform Keyfitz's type update based on (N-n) available units by comparing $\Pi'_{i|s}$ with $\Pi_{i|s}$ for $i \notin s$. In order to revise the conditional probabilities $\Pi_{i|s}$ to $\Pi'_{i|s}$, we undertake the Keyfitz updating procedure.

Define conditionally increasing and decreasing sets of units I and D, such that

$$i \in I \qquad \text{if } \Pi'_{i|s} \geq \Pi_{i|s}$$

and $i \in D$ otherwise.

If $i \in I$ retain the unit. If $i \in D$ retain the unit with probability $\Pi'_{i|s}/\Pi_{i|s}$ and if rejected, as it would be with probability $(1-\Pi'_{i|s}/\Pi_{i|s})$, select one unit from the set I with probability

$$\frac{\Pi'_{i|s} - \Pi_{i|s}}{\sum_{i \in I} (\Pi'_{i|s} - \Pi_{i|s})} \qquad \text{for } i \in I.$$

Then $P_{i|s}$, the conditional probability of selecting unit i under Keyfitz's procedure given the set s of reserved units, will be:

$$P_{i|s} = \Pi_{i|s} \left(\frac{\Pi'_{i|s}}{\Pi_{i|s}}\right) = \Pi'_{i|s} \qquad \text{for } i \in D$$

$$P_{i|s} = \Pi_{i|s} + \sum_{j \in D} \Pi_{j|s} \left(1 - \frac{\Pi'_{j|s}}{\Pi_{j|s}}\right) \left(\frac{\Pi'_{i|s} - \Pi_{i|s}}{\sum_{i \in I} (\Pi'_{i|s} - \Pi_{i|s})}\right)$$

$$= \Pi_{i|s} + \Pi'_{i|s} - \Pi_{i|s} = \Pi'_{i|s}. \qquad \text{for } i \in I$$

Therefore, at update the ith unit is selected with conditional probability $\Pi'_{i|s}$. Averaging over all possible reserves of n out of (N-1) units, excluding unit i, we obtain the overall average probability, P_i for unit i to be selected following update, as:

$$P_i = \sum_s' \Pr(s)(\Pi'_{i|s})$$

$$= p'_i \text{ by (2.3 and 2.5)} \qquad i=1, \ldots, N.$$

Therefore the updating scheme is unbiased. Since only one unit is selected per stratum for the continuous survey, the variance is a function of the probabilities of selection of units and is unaffected by the reserving of units.

3. SAMPLE UPDATING WHEN A PORTION OF THE FRAME IS RESERVED: (ROTATING CASE)

The results of the preceding section are applied to the Platek and Singh strategy [6] for a continuous, area-based sample requiring updating. The scope under this strategy is expanded to the case where the continuous survey incorporates rotation of fsu's. Here, only self weighting designs are considered for the continuous survey, so that when a portion of the frame has been reserved, it is required that the reserving mechanism does not affect probabilities of selection of units for the continuous survey as the sample rotates.

For simplicity we have considered as a model a two-stage random group design with pps selection of fsu's (clusters), systematic selection of ultimate sampling units (dwellings) and sample rotation within and between fsu's: this design is used by the Canadian Labour Force Survey in large cities. The results can be generalized for designs with more than two stages of selection.

As before, we have N units within a stratum (random group) and a size measure X_i associated with each unit i=1, 2, ..., N. We wish to sample within the stratum at the rate 1/R. Then we define cluster inverse sampling ratios as integers:

$$R_i \geq 1 \qquad i=1, 2, \ldots, N$$

$$\text{such that} \quad \sum_{i=1}^{N} |R_i - R.\, p_i| \quad \text{is minimized} \qquad (3.1)$$

$$\text{and} \quad \sum_i R_i = R$$

It should be noted that inverse sampling ratios in the form of integers are more convenient than non-integers for implementation in the field and for sample rotation.

Define R unique ordered samples within each random group as

$$j|R_i \qquad j=R_i, R_i-1, \ldots 2, 1; \qquad i=1, 2, \ldots, N$$

consisting of a sampled cluster i to be systematically sub-sampled at the rate $1/R_i$ for j successive occasions before rotation of fsu's occurs. That is, we have the following set of R ordered samples

$$R_1|R_1,\ (R_1-1)|R_1,\ \ldots,\ R_N|R_N,\ \ldots,\ 1|R_N.$$

Initially one of the above samples is selected by generating a random number r, $1 \leq r \leq R$. Suppose the selected sample is $j|R_i$, where $\sum_{k=0}^{i-1} R_k < r \leq \sum_{k=1}^{i} R_k$ for some $i \varepsilon \{1, 2, \ldots, N\}$, and $j = \sum_{k=1}^{i} R_k - r + 1$; R_o is defined to be zero. Then another random number r_i, $1 \leq r_i \leq R_i$ is generated and the systematic samples determined by the random starts r_i, $(r_i+1) \bmod R_i$, ..., $(r_i+j-1) \bmod R_i$ are respectively associated with the samples $j|R_i$, $(j-1)|R_i$, ..., $1|R_i$[1]. After each pre-specified constant interval of time, rotation takes place into the next sample on the list. At the time of rotation into the next cluster, i.e. cluster

[1] $R_i \bmod R_i$ is taken equal to R_i instead of 0. This convention will be adopted throughout in this paper.

$i^* = (i+1) \bmod N$, with sample $R_{i^*}|R_{i^*}$; a random number r_{i^*}; $1 \leq r_{i^*} \leq R_{i^*}$ is generated and the systematic samples determinded by the starts r_{i^*}, $(r_{i^*}+1) \bmod R_{i^*}$, ... $(r_{i^*} + R_{i^*} - 1) \bmod R_{i^*}$ are associated with the samples $R_{i^*}|R_{i^*}$, $(R_{i^*}-1)|R_{i^*}$ ..., $1|R_{i^*}$ respectively, and so on. In practice, random numbers r_i, i=1, 2, ..., N are all generated at the time of initial introduction of the sample and the rotation schedule is created in terms of the actual systematic samples or starts.

Following this rotation scheme, the probability of selecting cluster i at any point in time is given by:

$$Pr(i \varepsilon C) = R_i/R \doteq p_i \quad .$$

Given that cluster i is selected, the probability of each start being in the sample at any point in time is given by $1/R_i$, so that the overall probability of selecting each start is $(1/R_i)(R_i/R)$ or $1/R$. Consequently, since the design is self weighting, if y_{ik} is the characteristic total for start k in cluster i, then $R\, y_{ik}$ is an unbiased estimator of the group total y.

Now consider what happens to probabilities of selection when reserves are made from the frame, adopting the rule that if the unit that would have rotated is reserved, rotation will take place into the next unreserved unit. For simplicity we consider the case of one reserved unit. Since the probability of selecting a cluster at any point in time is given by R_i/R, we can assume with no loss of generality that at time t=0 cluster i is in the continuous survey, and that at time $t \varepsilon (0,1)$, one cluster, say $k \neq i$ is reserved with probability $p^*_{k|i}$. Then at t=1, the occasion of next rotation of the sample, the probability for cluster i to be in the sample for the continuous survey C, i.e. $Pr(i \varepsilon C | t=1)$ is given by:

$$\Pr(i\in C \mid t=1) = \Pr(i\in C \mid t=0).\ \Pr(\text{cluster } i \text{ will not rotate out at } t=1)$$

$$+ \Pr(i-1\in C \mid t=0)\cdot\Pr(\text{cluster } i-1 \text{ will rotate out at } t=1)\cdot \Pr(\text{cluster } i \text{ not reserved})$$

$$+ \Pr(i-2 \mid C\ t=0)\cdot \Pr(\text{cluster } i-2 \text{ will rotate out at } t=1)\cdot \Pr(\text{cluster } i-1 \text{ is reserved})$$

$$= \frac{R_i}{R}\left(1-\frac{1}{R_i}\right) + \frac{R_{i-1}}{R}\,\frac{1}{R_{i-1}}\left(1-p^*_{i\mid i-1}\right) + \frac{R_{i-2}}{R}\,\frac{1}{R_{i-2}}\,p^*_{i-1\mid i-2}$$

$$= \frac{R_i-1}{R} + \frac{1}{R}\left(1-p^*_{i\mid i-1}\right) + \frac{1}{R}\,p^*_{i-1\mid i-2}\ . \qquad (3.2)$$

Now (3.2) equals R_i/R if and only if $p^*_{i\mid i-1} = p^*_{i-1\mid i-2}$ for all i. This condition holds non-uniquely if one cluster is reserved with equal probability, excluding the unit selected for the continuous survey. The posterior probability for unit i to be in continuous survey C given that unit j was reserved is given by:

$$\Pi_{i\mid j} = \frac{\Pr(i\in C,\ j \text{ reserved})}{\Pr(j \text{ reserved})}$$

$$= \frac{p_i\,\frac{1}{N-1}}{\sum_{i\neq j} p_i\,\frac{1}{N-1}} = \frac{p_i}{1-p_j}\ . \qquad (3.3)$$

Thus, the expression for $\Pi_{i|j}$ is simplified if one unit is reserved with equal probability.

In general, it can be shown that when n out of N-1 clusters are reserved with equal probability excluding the continuous survey selection, the probabilities of selection for the continuous survey are preserved, and the expression for the posterior probability $\Pi_{i|s}$ simplifies to:

$$\Pi_{i|s} = \frac{p_i}{1 - \sum_{j \in s} p_j} \quad . \tag{3.4}$$

However, for the same reason that we have chosen a pps sampling scheme for the continuous survey, such a design in most instances would be advantageous for the special survey. Thus, instead of selecting one or more units specifically for a particular special survey with equal probability excluding the selection for the continuous survey, rather, our strategy will be to reserve a portion of the frame, say one-third, following the above mechanism for reserving fsu's and then to select units for the special survey from within the reserved portion following a pps scheme.

If reserves are made in the above manner, there will be no bias of selection for the continuous survey prior to update. In the remainder of this section, we show how the general method described in Section (2) can be adapted to the particular rotation scheme under consideration to achieve desired post-update probabilities while preventing overlaps of dwellings between the pre- and post-update samples.

Under this method of reserving fsu's, (2.1) and (2.3) reduce respectively to:

$$Pr(s) = \left(1 - \sum_{i \in s} p_i\right) \frac{1}{\binom{N-1}{n}} \tag{3.5}$$

$$\text{and} \quad \sum_{s}{}' \left(1 - \sum_{i\varepsilon s} p_i\right) \frac{1}{\binom{N-1}{n}} \left(\frac{p_i(2)}{1 - \sum_{i\varepsilon s} p_i(2)}\right) = p_i' \qquad (3.6)$$

$$i=1, 2, \ldots, N \quad ,$$

where p_i' are defined in (2.2).

By applying Keyfitz's sample updating procedure using conditional probabilities as described in Section (2), a cluster $i\varepsilon s$ could be selected for the continuous survey with conditional probability $\Pi'_{i|s}$ given by:

$$\Pi'_{i|s} = \frac{p_i(2)}{1 - \sum_{j\varepsilon s} p_j(2)}$$

so that when averaged over all possible reserves, the probability of selecting cluster i becomes p_i'. However, having retained a cluster in this fashion at update, it would be desirable to remain in the cluster only long enough so that sampling can be restricted to unused dwellings. This suggests a mapping (see Appendix A) from the possible pre-update samples into the possible post-update samples, such that following the rotation scheme, no overlap of dwellings would occur, and the required post-update probabilities would be achieved.

The cluster isr's based on new sizes will be defined as before, with R_i' replacing R_i and p_i' replacing p_i, N in expression (3.1).

Since we will be using a one to one mapping from the possible pre-update samples into the possible post-update samples to perform Keyfitz's type sample update as described in Appendix A, and there could be only $(R - \sum_{j\varepsilon s} R_j)$ possible pre-update samples, we define post-update cluster isr's as integers $R_{i|s}(2) \geq 1$ for $i\notin s$

$$\left.\begin{array}{l} \text{such that } \sum_{i \notin s} \left(R_{i|s}(2) - (R - \sum_{j \in s} R_j).\ \Pi'_{i|s}\right) \\ \text{is minimized and that} \\ \qquad \sum_{i \notin s} R_{i|s}(2) = R - \sum_{j \in s} R_j. \end{array}\right\} \qquad (3.7)$$

Thus in this fashion cluster $i \notin s$ will be selected with conditional probability

$\dfrac{R_{i|s}(2)}{R - \sum_{j \in s} R_j}$ instead of $\Pi'_{i|s}$. Note that this computational procedure is only subject to error in rounding to integer sizes. In expression (3.6), to calculate working probabilities $p_i(2)$, p'_i was taken as

$\dfrac{X'_i}{\sum_i X'_i}$ instead of R'_i/R so that the effect due to rounding to integers is not introduced twice.

Since we will be sampling at the rate $R_{i|s}(2)$ instead of R'_i in the selected cluster i, we will apply a compensating weight equal to the ratio $\dfrac{R_{i|s}(2)}{R'_i}$ at the estimation stage. As before, if y_{ij} is the characteristic total for the selected sample k in cluster i, then

$R\left(\dfrac{R_{i|s}(2)}{R'_i}\right)\ y_{ik}$ is an estimator for the stratum total, whose only bias is due to rounding to integers.

Due to the complexity involved in computing "Working Probabilities" and practical limitations of this method, a simple although biased alternative is presented here. It was observed empirically that, when $n/(N-1) \leq 1/3$

$$p_i(2) \doteq p_i' \qquad i=1, 2, \ldots, N$$

so that we now define the conditional probability of selecting unit i for the continuous survey C, given that the set s of units was reserved, as

$$\Pi^*_{i|s} = \frac{p_i'}{1 - \sum_{j \in s} p_j'}$$

and we define the isr's $R'_{i|s} \geq 1$ for $i \notin s$ by replacing $R_{i|s}(2)$ by $R'_{i|s}$ and $\Pi'_{i|s}$ by $\Pi^*_{i|s}$ in (3.7).

Then $R\left(\frac{R'_{i|s}}{R'_i}\right) y_{ik}$ is the estimator for the stratum total, and the mapping of pre-update samples into post-update samples is identical to the previous case.

It should be noted that if the number of post-update samples could be chosen as $R - \sum_{i \in s} R'_i$ instead of $R - \sum_{i \in s} R_i$, then the weights $\frac{R_{i|s}(2)}{R'_i}$ would in general be close to one, and the departure from a self-weighting design would be minimized. However, the mapping procedure for the case where the number of pre-update and post-update samples are not equal, becomes very complicated. Moreover, under this mapping, the probability of retaining the currently selected cluster will not be maximized as under Keyfitz's method.

4. STRATEGY FOR USE OF SPECIAL SURVEY FRAME

Within a stratum, the reserved units (clusters) from each random group are merged to form the special survey frame. Before presenting the methodology for the special survey frame, it should be pointed out that if it were not necessary to provide a capacity for updating the frame and the sample, surveys other than the continuous survey could also use the frame, avoiding overlap with the continuous survey by merely spacing their selections at some interval from those for the continuous survey. However, at the time of update, whether via Keyfitz's method or an independent selection, the continuous survey selection could change resulting in conflict with samples selected for special surveys. On the other hand, if the special survey is restricted to the same cluster in which the continuous survey selection happens to be, this may operationally link the continuous and special surveys to a degree that is detrimental to both. For instance, the special survey would be tied into the continuous survey's lead times for introduction of sampling units, while on the other hand, sporadic special survey use of the frame would have a disruptive effect on sample maintenance operations for the continuous survey.

Since the sample size may vary for different special surveys, a randomized pps systematic design [3] is proposed as this method is flexible with regard to the number of units selected [2]. Successive special surveys would, to the degree possible, utilize common fsu's to minimize listing costs; however, when the frame is updated, a completely independent selection wuld be carried out within the special survey frame, avoiding overlap at the dwelling level by means of the re-order mechanism described in Appendix (A).

Suppose that for each random group g, we select n_g clusters with SRS from the (N_g-1) available clusters excluding the continuous survey selection, where g=1, 2, ..., G. Thus within a sub-unit $n = \sum_{g=1}^{G} n_g$ out of $N = \sum_{g=1}^{G} N_g$ clusters are reserved for the special survey frame.

Since the continuous survey is more likely to be in larger clusters, the overall probability of a cluster being reserved for the special survey frame decreases as the size of the cluster increases. An unbiased design which takes this into account is likely to be less efficient than a biased design which assumes that the probability of cluster i to be in the special survey frame is equal to n/N for all i. Under the latter assumption, for an overall sampling rate of $1/R_o$ from the sub-unit, let $1/W_o$ be the equivalent sampling rate from the special survey frame. Then

$$\frac{n}{N}\,(1/W_o) = 1/R_o$$

or

$$W_o = \frac{n}{N}\,R_o .$$

Define $W_o' = [\frac{n}{N}\,R_o]$.

A compensating weight, ω, to offset the effect of rounding will be applied at the estimation, where

$$\omega = \frac{W_o'}{W_o} = \frac{W_o'}{\frac{n}{N}\,R_o} \quad .$$

Then inverse sampling rates for clusters in the special survey frame are defined as integers $W_i \geq 1$ for $i \in s$ such that

$$\sum_{i\in s} W_i = W_o' \quad \text{and} \quad \sum_{i\in s} \left(W_i - W_o' \left(\frac{X_i}{\sum_{i\in s} X_i}\right)\right)$$

is minimized, which partitions the special survey frame into W_o' systematic samples. Selection of M of these samples for a special survey corresponds to an M/R_o sampling rate from the entire frame.

Let y_m = response from mth selected sample.

Then $y = \sum_{m=1}^{M} y_m$ = total response from the sample.

Two estimators for the population total are considered:

$$\hat{y}_1 = \omega R_o y/M$$

$$= (\frac{N}{n}) W_o' y/M, \qquad (4.1)$$

and $$\hat{y}_2 = \frac{(\frac{X}{N})}{(\frac{X_s}{n})} \omega R_o y/M$$

$$= (\frac{X}{X_s}) W_o' y/M, \qquad (4.2)$$

where $X = \sum_{i=1}^{N} X_i$, $X_s = \sum_{i \varepsilon s} X_i$.

The ratio adjustment $\frac{(\frac{X}{N})}{(\frac{X_s}{n})}$ in $\hat{y}_2$ compensates for discrepancies in the size of the special survey frame relative to an n/N sub-sample from the frame, introduced as a result of sampling variability as well as the bias due to the assumption of simple random sampling for reserving units from the entire sub-unit.

It was observed in the Monte Carlo studies that $\hat{y}_2$ performed consistently better than $\hat{y}_1$, therefore the estimator considered for the special survey frame in Section (5) is $\hat{y}_2$.

5. MONTE CARLO STUDY

a) Description

The Canadian Labour Force Survey follows a multi-stage stratified sample design [6]. In the self-representing areas consisting of large cities and metropolitan areas, accounting for over 2/3 of the country, a two-stage stratified sample design is employed. The strata consist of sub-units whose populations vary from 6,000 to 25,000 while fsu's (clusters) consist of city block faces, and ultimate sampling units consist of dwellings.

To evaluate the gains in reliability of data as a result of updating procedures, and the suitability of the procedure suggested for special surveys, a Monte Carlo study was carried out for seven Labour Force sub-units (strata) with varying growth rates between 1966 and 1971 Censuses.

For the Census Enumeration Areas (EA's) comprising these sub-units, 1971 Census data was obtained at the individual level for the 1/3 sample of households which received a detailed census questionnaire. For the purpose of the study, institutions such as hospitals, and old age homes were excluded. For the most part, 1971 EA's were chosen to represent LFS clusters. However, in order that the distribution of cluster sizes within sub-units closely approximated the known distribution of cluster sizes by province and type of area for the LFS design, some of the larger EA's were sub-divided to form two or more clusters. The new size measures were obtained from the household counts pertaining to the 1/3 sample, while the corresponding old size measures were obtained by taking 1/3 of the dwelling counts for 1966 EA's and utilizing conversion tables from 1971 to 1966 EA's.

In this study we have considered estimation of the following six characteristics:

i) Population,

ii) Number of Households,

iii) Number of Persons Employed,
iv) Number of Persons Unemployed,
v) Number of Persons Not in Labour Force,
vi) Total Income.

Five different methods were simulated 1,000 times independently within each sub-unit. A method is defined as a selection scheme associated with an estimation procedure. The methods are described below.

Method 1 - Random group method using new size measures with complete frame available for the continuous survey.

Method 2 - Following select-on as in Method 1, a one-third portion from each random group was reserved with equal probability excluding the cluster selected for the continuous survey and the reserved clusters from each random group were merged together to form the special survey frame. Within the special survey frame the design and estimation procedure described in Section 4 were followed.

Method 3 - Same as Method 1, but using old size measures.

Method 4 - Following selection by Method 3, one-third portion from each random group was reserved, and the sample was updated utilizing the "Working Probability" scheme described in Section 3.

Method 5 - Same as Method 4, except the sample was updated via the "revised probability" scheme described in Section 3.

Let Y_h = the characteristic total for sub-unit h based on the 1971 Census; (h=1, 2, ..., 7),

and $y_{hr}^{(m)}$ = the estimate of Y_h from the rth replication using method m; (r=1, 2, ..., 1,000; m=1, 2, ..., 5).

Then the average value of 1,000 estimates for method m, sub-unit h is given by:

$$\bar{y}_h^{(m)} = \frac{1}{1,000} \sum_{r=1}^{1,000} y_{hr}^{(m)} .$$

Combining all the 7 sub-units, the population total Y is given by:

$$Y = \sum_{h=1}^{7} Y_h,$$

and similarly combining the estimates for all sub-units, we have:

$$y_r^{(m)} = \sum_{h=1}^{7} y_{hr}^{(m)}$$

and

$$\bar{y}^{(m)} = \sum_{h=1}^{7} \bar{y}_h^{(m)}$$

$$= \frac{1}{1,000} \sum_{r=1}^{1,000} y_r^{(m)}$$

Define the discrepancy of method m, $D^{(m)}$, to be the deviation of the average of 1,000 estimates, using method m, from the population total y, viz.

$$D^{(m)} = \bar{y}^{(m)} - y,$$

and % relative discrepancy by:

$$RD^{(m)} = 100(\bar{y}^{(m)} - y)/y.$$

The estimate of standard deviation of $y_{hr}^{(m)}$ is:

$$\hat{S.D.}(y_{hr}^{(m)}) = [\frac{1}{1,000} \sum_{r=1}^{1,000} (y_{hr}^{(m)} - \bar{y}_h^{(m)})^2]^{1/2}$$

Therefore, the estimate of the standard deviation of $y_r^{(m)}$ is

$$\hat{S.D.}(y_r^{(m)}) = (\sum_{h=1}^{7} [\hat{S.D.}(y_{hr}^{(m)})]^2)^{1/2} ,$$

and the estimate of the standard deviation of $\bar{y}^{(m)}$ is

$$\hat{S.D.}(\bar{y}^{(m)}) = \hat{S.D.}(y_r^{(m)})/(1,000)^{1/2} .$$

The estimated % coefficient of variation is then given as:

$$\hat{C.V.}(\bar{y}^{(m)}) = 100\ \hat{S.D.}(\bar{y}^{(m)})/\bar{y}^{(m)}$$

Within sub-unit h, define the efficiency of method m relative to method 1 as:

$$EFF_h\ (m \text{ vs } 1) = 100\ (MSE)_h^{(1)}/(MSE)_h^{(m)}$$

where

$$(MSE)_h^{(m)} = \frac{1}{1,000} \sum_{r=1}^{1,000} (y_{hr}^{(m)} - y_h)^2 .$$

Finally, define the overall efficiency for method m relative to method 1 as:

$$EFF(m \text{ vs } 1) = 100\ (MSE)^{(1)}/(MSE)^{(m)}$$

where

$$MSE^{(m)} = [\hat{S.D.}(y_r^{(m)})]^2 + (\sum_{h=1}^{7} D_h^{(m)})^2 .$$

b) Analysis of Results

Although the primary purpose of the study was to evaluate the two updating schemes (i.e. methods 4 & 5) and the performance of the proposed special survey frame, it was also possible to study the gains resulting from updating the sample when the entire frame is available. Let us briefly then examine these gains.

It can be observed from Tables (5.1) and (5.2) that with the exception of the characteristic unemployed, which is not very highly correlated with size measures, efficiencies tend to decrease (hence gains tend to increase) with decreasing correlation between the old and new size measures. Whereas, one might expect that in practice the greater the growth rate, the lower this correlation would be, sub-units 83112 and 95135 do not confirm these expectations. Even for areas of fairly moderate overall growth, substantial gains in simple survey estimates can result from updating as demonstrated by sub-unit 51201. However, due to the efficiency of techniques commonly utilized in estimation procedures for large scale surveys such as post-stratification by age-sex categories, the gains in precision for final survey estimates are likely to be smaller. It would be of interest to investigate this aspect further.

Table 5.1: Correlations[1] and % Growth[2]

	sub-unit						
	33102	83112	95135	51201	80114	53120	51110
correlation	.87	.79	.78	.65	.63	.51	.48
% growth	5.83	54.00	17.41	11.06	18.37	39.16	39.02

Table 5.2: Efficiency of Method 3 vs Method 1

characteristic	sub-unit						
	33102	83112	95135	51201	86114	53120	51110
population	87.8	27.4	25.3	30.0	48.1	23.8	8.6
households	33.6	6.6	4.3	5.1	3.0	4.0	1.8
employed	78.3	37.3	58.6	39.0	29.9	24.6	13.5
unemployed	82.1	85.4	86.4	99.3	78.3	79.3	88.3
not in LFS	87.2	57.7	43.1	50.7	89.4	55.4	31.7
income	93.3	42.1	46.2	35.4	26.5	26.5	10.8

1 correlation between old and new size measures

2 % growth for the period between 1966 and 1971 Censuses.

The performances of updating methods (4 and 5) and of the special survey frame relative to method 1 can be seen from an analysis of Tables 5.3 and 5.4.

From an efficiency point of view (Table 5.3) when one-third of the frame has been reserved, there is little difference between updating methods 4 and 5. Efficiencies under both methods are lowest for characteristics unemployed and not in labour force (91-93%). This small loss in efficiency for method 4 is most likely attributable to rounding to integer sizes, and to the departure from the self-weighting design, since otherwise, as noted in section (1), the variance under methods 1 and 4 should be identical. It seems plausible to attribute the loss in efficiency under method 5 to the same causes.

Table 5.3: Overall Efficiencies

Characteristic	Method			
	1	2	4	5
population	100	103.9	98.6	98.1
households	100	107.8	102.0	100.7
employed	100	101.1	101.5	100.4
unemployed	100	95.1	91.1	92.4
not in LFS	100	96.7	91.8	93.2
income	100	103.2	101.4	99.9

For remaining characteristics, efficiencies are in the range 98-102%. The efficiency of the special survey frame drops to 95% for unemployed and 96.7% for not in LF, but for other characteristics, ranges from 101-108%. The efficiencies do not appear to be appreciably affected by the procedure of reserving a portion of the frame, and then drawing the sample from the reserved portion as opposed to drawing the sample from the whole frame. This phenomenon seems to be attributable to both the design within the special survey frame and the proposed ratio estimator (4.2).

Table 5.4: % Relative Discrepancies/
Estimated % Coefficient of Variation

Characteristic	Population value	Method 1	2	4	5
population	49,389	.17 .1485	- .12 .1458	.00 .1497	.11 .1500
households	14,264	.07 .0512	.01 .0493	.01 .0507	.02 .0510
employed	19,951	.30 .1731	- .45 .1719	- .05 .1721	.08 .1730
unemployed	1,615	.35 .7391	- .22 .7578	.70 .7739	.22 .7687
not in LFS	12,288	- .10 .2414	.30 .2454	.52 .2515	.53 .2495
income ($1000's)	250,547	.08 .0972	- .02 .0957	- .06 .0965	- .03 .0972

From Table (5.4), it can be observed that the % relative discrepancies are low in all cases. Comparing the % RD for the theoretically unbiased methods (1 and 4) with those of the other methods, suggests that the bias under methods 2 and 5 is not serious. It should be noted that while significant t-statistics at 95% level were obtained for the characteristic employed under method 2 and not in Labour Force for both methods 4 and 5, these biases appear nevertheless of no practical significance, being less than 1% of the population value. Also, it is worth noting that although we have not presented discrepancies for individual sub-units, these were calculated, and it was observed that no methods either under-estimated or over-estimated a characteristic for all sub-units.

In conclusion, we feel that Tables 5.3 and 5.4 demonstrate the overall suitability of the strategy we have presented, from the perspective of both the continuous survey and special surveys. We conjecture that under circumstances similar to those in the study, the two updating schemes will perform equally well, so method 5 should be preferred on the grounds of computational simplicity.

RESUME

Les bases conçues pour des enquêtes permanentes servent parfois à effectuer des enquêtes spéciales qui nécessitent un échantillon distinct de celui de l'enquête permanente. Cet article présente une méthode sans biais de mise à jour d'une base de sondage, qui prolonge celle de Keyfitz (1951) en l'appliquant au cas où une partie de la base a été réservée à des enquêtes autres que l'enquête permanente. Une autre méthode, simple mais biaisée, est aussi exposée.

Les auteurs élargissent ensuite la portée de la technique de Platek et Singh (1975) sur la conception d'un échantillon permanent à partir d'une base aréolaire nécessitant des mises à jour, en incorporant à cette technique le renouvellement des unités d'échantillonnage de premier degré, l'établissement d'une base réservée aux enquêtes spéciales et des procédures visant à éviter de tirer deux fois la même unité finale.

Pour évaluer les méthodes proposées, les auteurs appliquent la méthode de Monte Carlo à des données du recensement, en simulant le plan de sondage de l'EPA.

REFERENCES

[1] Drew, J.D., "Sample Update - A Mapping Procedure for Cases Where the number of Pre- and Post-Update Sample Are Not Equal", Internal Statistics Canada Technical Memorandum, Household Surveys Development Division (1978).

[2] Fellegi, I.P., "Sampling With Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples", Journal of the American Statistical Association, Vol. 58 (1963), pp. 183-201.

[3] Gray, G.B., "On Increasing the Sample Size (number of psu's)", Internal Statistics Canada Technical Memorandum, Household Surveys Development Division (1973).

[4] Hartley, H.O. and Rao, J.N.K., "Sampling With Unequal Probabilities and Without Replacement", Annals of Mathematical Statistics (1962), Vol. 33, pp. 350-374.

[5] Keyfitz, N., "Sampling With Probabilities Proportional to Size: Adjustment for Changes in the Probabilities", Journal of the American Statistical Association, Vol. 46 (1951), pp. 105-109.

[6] Kish, L. and Scott, A., "Retaining Units After Changing Strata And Probabilities", Journal of the American Statistical Association, Vol. 66 (1971), pp. 461-470.

[7] Platek, R. and Singh, M.P., "A Strategy for Updating Continuous Surveys", Survey Methodology (Statistical Services, Statistics Canada), Vol. 1, No. 1 (June 1975), pp. 16-26.

[8] Rao, J.N.K., Hartley, H.O. and Cochran, W.G., "On a Simple Procedure of Unequal Probability Sampling Without Replacement", Journal of the Royal Statistical Society, Series B, Vol. 27 (1962), pp. 482-491.

[9] Statistics Canada (Household Surveys Development Division), "Methodology of the Canadian Labour Force Survey (1976)", Catalogue 71-526 occasional (published October 1977), pp. 33-38.

APPENDIX (A)

Operational Aspects of Sample Update Using Keyfitz's Procedure

Consider a stratum having N units, with inverse sampling ratios R_i; i=1, 2, ..., N; defined according to (3.1), and with the rotation scheme as described in Section 3 (page 8).

At some point in time, revised household counts are obtained, and revised inverse sampling ratios R_i'; i=1, 2, ..., N; are defined as before so that $\sum_{i=1}^{N} R_i' = R$. Then the R unique ordered samples based on the revised sizes are:

$$R_1'|R_1',\ (R_1'-1)|R_1',\ \ldots,\ R_N'|R_N',\ \ldots,\ 1|R_N' .$$

Thus, at the time of the next sample rotation, the probabilities of selection of clusters must be adjusted so that they are proportional to their revised isr's. Since we have the same number of post-update samples as the number of pre-update samples, a simple one-to-one mapping of pre-update samples into post-update samples can be defined such that:

i) Keyfitz's criteria of adjusting probabilities are satisfied.

ii) The post-update samples can be restricted to previously unselected dwellings, for which, if the same cluster is retained, a necessary but not sufficient condition is that

$$x_i/R_i \geq x_i'/R_i', \tag{A.1}$$

where $x_i|R_i$ is the sample that would have resulted had there been no update and $x_i'|R_i'$ is the post-update sample. A further condition relates to the choice of the post-update start and is discussed later.

Such a mapping (non-unique) can be carried out as follows:

a) If $i \in D$, i.e. $R_i' < R_i$, then the samples $R_i|R_i$, $(R_i-1)|R_i$, ..., $(R_i-R_i'+1)|R_i$ are mapped respectively into the samples $R_i'|R_i'$, $(R_i'-1)|R_i'$, ..., $1|R_i'$ and the samples $(R_i-R_i')|R_i$, $(R_i-R_i'-1)|R_i$, ..., $1|R_i$ are temporarily left unmapped.

b) If $i \in I$, i.e. $R_i' \geq R_i$, then the samples $R_i|R_i$, $(R_i-1)|R_i$, ..., $1|R_i$ are mapped respectively into the samples $R_i|R_i'$, $(R_i-1)|R_i'$, ..., $1|R_i'$, leaving the samples $R_i'|R_i'$, $(R_i'-1)|R_i'$, ..., $(R_i+1)|R_i'$ as available samples.

c) Since $\sum_{i \in D} (R_i-R_i') = \sum_{i \in I} (R_i'-R_i) = f$, say, the unmapped pre-update samples in the decreasing clusters can be mapped in a one-to-one fashion into the available post-update samples in the increasing clusters. There are f! possible mappings. Ideally, we might choose that mapping which maximizes the time interval (i.e. number of rotation periods) before any post-update sample rotates back into its corresponding pre-update cluster and begin re-using dwellings. However, evaluating all f! mappings will not always be practical, so we suggest the following procedure:

Let $D = \{i_1', i_2', \ldots, i_d'\}$ define the set of decreasing clusters ordered by increasing serial numbers, and $v = \{v_1, v_2, \ldots, v_d\}$ be the corresponding changes in their number of samples. Define $I = \{i_1'', i_2'', \ldots, i_e''\}$ and $w = \{w_1, w_2, \ldots w_e\}$ analogously for the set of increasing clusters.

For each $\ell = 1, 2, \ldots, d$, the procedure described below determines a mapping beginning with the decreasing cluster i_ℓ'. The minimum time interval in which a post-update sample will rotate back into its corresponding pre-update cluster and begin re-using dwellings is also obtained for each mapping. If a_ℓ is the minimum time interval

for mapping ℓ, then the mapping ℓ^* for which $a_{\ell}^* = \max\{ a_1, a_2, \ldots, a_d\}$ is chosen. For a given ℓ, the mapping is defined as follows:

Find the first cluster $k_1 \epsilon I$ with $i''_{k_1} > i'_{\ell}$; that is, the first increasing cluster which will rotate into the sample after cluster i'_{ℓ}. There are v_{ℓ} unmapped samples in the decreasing cluster i'_{ℓ} - map all of these samples in the increasing cluster i''_{k_1}, $i''_{(k_1+1) \bmod e}$, ... exhausting w_{k_1} available samples in the increasing cluster i''_{k_1} before proceeding to $i''_{(k_1+1) \bmod e}$ and similarly for $i''_{(k_1+1) \bmod e}$, $i''_{(k_1+2) \bmod e}$,...using as many of the increasing clusters as required. After mapping the v_{ℓ} samples from decreasing cluster i'_{ℓ} into increasing clusters i''_{k_1}, $i''_{(k_1+1) \bmod e}$, ..., the corresponding counts of available samples i.e. w_{k_1}, $w_{(k_1+1) \bmod e}$, ... are adjusted. Next, take the decreasing cluster $i'_{(\ell+1) \bmod d}$ and find the first cluster $k_2 \epsilon I$ with $i''_{k2} > i'_{(\ell+1) \bmod d}$ and as before map all the $v_{(\ell+1) \bmod d}$ unmapped samples in the decreasing cluster $i'_{(\ell+1) \bmod d}$ into the available samples in the increasing clusters i''_{k_2}, $i_{(k_2+1) \bmod e}$, ... Repeat this process for clusters $i'_{(\ell+2) \bmod d}$, $i'_{(\ell+3) \bmod d}$, ..., $i'_{(\ell+d-1) \bmod d}$.

The following example for the case where we have 4 clusters with old and new isr's as given in Table (A.1) illustrates the procedure.

Table (A.1)

Cluster No.	Old isr	New isr
1	4	2
2	3	4
3	2	4
4	3	2
	12	12

The set of decreasing clusters D = {1,4} and the corresponding changes in isr's, i.e. V = {-2, -1}, and similarly for the set of increasing cluster I = {2,3}, W = {1,2}. Fig. (1) below shows the mapping of pre-update samples into the post-update samples.

Mapping of Pre-Update Samples Into the Post-Update Samples

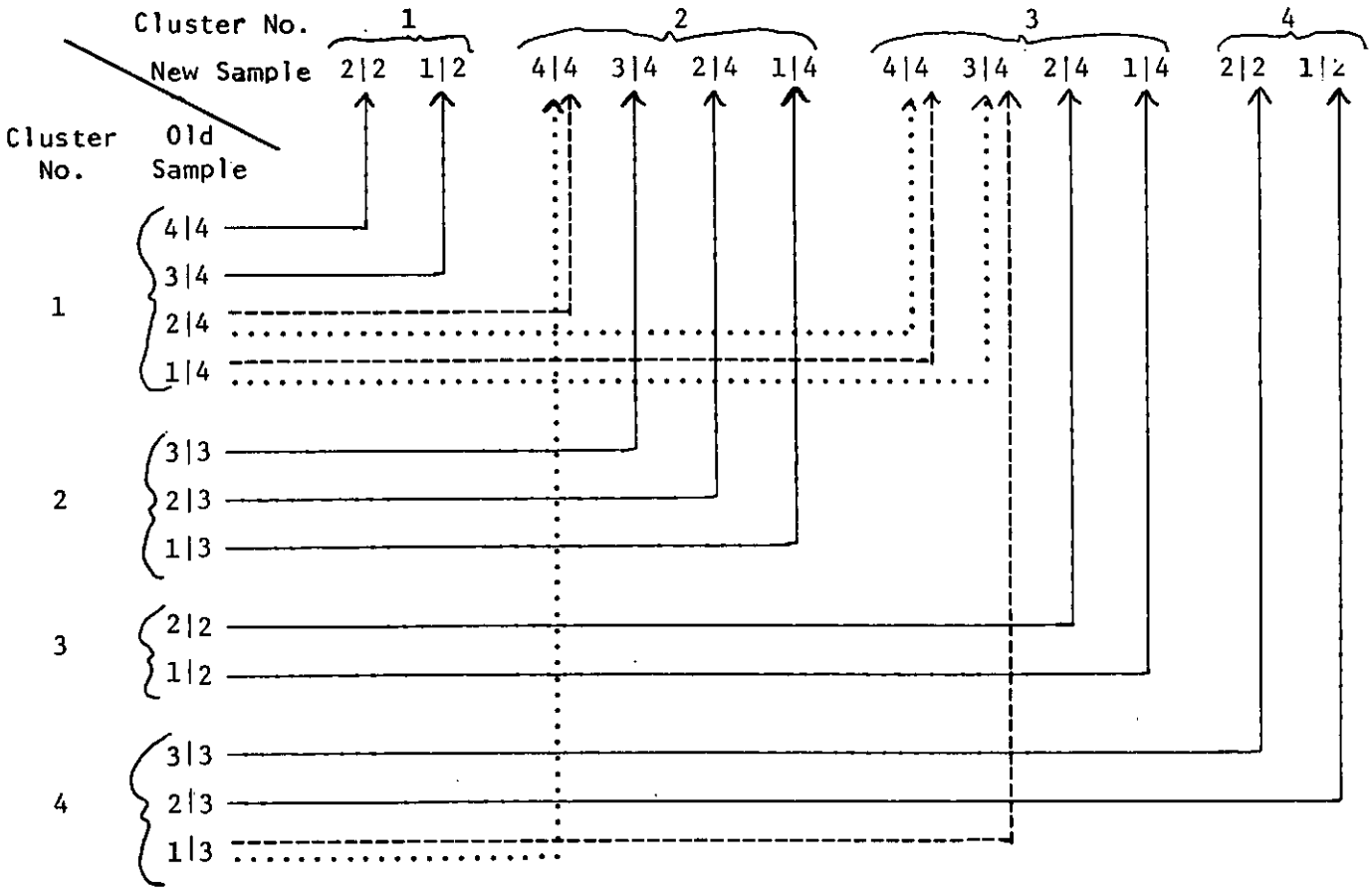


Fig. (1)

The solid lines correspond to the pre-update samples being mapped into the post-update samples in the same cluster, i.e. the cases where old selected cluster is retained. The unmapped pre-update samples in the decreasing clusters can be mapped into the post-update available samples in the increasing clusters starting from the decreasing cluster I

(broken lines) or starting from the decreasing cluster 4 (dotted lines). The minimum time interval for the re-selection of dwellings for the mapping indicated by broken lines is 3 and for the mapping indicated by dotted lines this time interval is 5. In the former mapping (broken lines) the minimum time interval corresponds to the pre-update sample $1|3$ in cluster 4 being mapped into the post-update sample $3|4$ in cluster 3, in which case following use of the samples $3|4$, $2|4$, $1|4$ in cluster 3, re-selection of dwellings in the pre-update cluster, 4, would occur with sample $2|2$. In the latter mapping (dotted lines) time interval corresponds to the pre-update sample $1|4$ in cluster 1 being mapped into the post-update sample $3|4$ in cluster 3. Thus, the mapping indicated by dotted lines will be used.

Clearly under the above mapping scheme:

i) The clusters are selected with probability proportional to their revised isr's as required.

ii) Each post-update sample is equally likely so that under the rotation scheme these probabilities will be preserved.

iii) Keyfitz's conditions on rejection and retention of clusters hold, and

iv) The condition necessary to avoid re-selection of dwellings also holds.

Having identified the post-update sample in the preceding mapping process, it remains to determine post-update random starts. The following 3 contigencies arise:

i) At the time of update the old cluster is rejected and a <u>new</u> cluster i is selected. Then a random start r_i', $1 \leq r_i' \leq R_i'$ is chosen, and if the sample to be introduced is $j|R_i'$, then the systematic samples determined by the starts r_i', $(r_i'+1) \bmod R_i'$, ..., $(r_i'+j-1) \bmod R_i'$ are associated with the samples $j|R_i'$, $(j-1)|R_i'$, ..., $1|R_i'$ respectively.

ii) The previously selected cluster i is retained and $R_i' = R_i$. In this case, the sequence of rotation within i remains unchanged.

iii) The previously selected cluster is retained and $R_i' \neq R_i$. In this case, we require a mapping of the old starts into the new starts such that the overall probability for each new start equals $1/R_i'$, and such that the number of dwellings to be used under the post-update starts never exceeds the number of dwellings used prior to update. The first condition ensures unbiased selection at the start level, while the second condition allows us to re-order the dwellings, as described later, such that no dwelling re-selections occur.

Let $Pr(s \rightarrow s')$ denote the probability that the pre-update start $s(s=1,2, \ldots, R_i)$ will be mapped into the post-update start $s'(s'=1,2, \ldots, R_i')$. Thus we need to determine an $R_i \times R_i'$ matrix P so that $Pr(s \rightarrow s')$ is given by $P_{ss'}$, where

$$\sum_{s'=1}^{R_i'} P_{ss'} = 1 \qquad \text{for all } s$$

$$\sum_{s=1}^{R_i} \frac{1}{R_i} P_{ss'} = \frac{1}{R_i'} \qquad \text{for all } s',$$

and the condition necessary to prevent re-selection of dwellings also holds. This can be achieved by determining an $R_i \times R_i'$ matrix A such that

$$\sum_{s'=1}^{R_i'} a_{ss'} = R_i' \qquad \text{for all } s \qquad \text{(A.2)}$$

$$\sum_{s=1}^{R_i} a_{ss'} = R_i \qquad \text{for all } s', \tag{A.3}$$

and assigning the maximum possible values to the elements of the matrix A in the order $a_{11}, a_{12}, \ldots, a_{1R_i'}, a_{21}, \ldots, a_{R_i 1}, a_{R_i 2}, \ldots, a_{R_i R_i'}$ subject to the constraints (A.2) and (A.3). Then the $Pr(s \to s')$ is simply given by $a_{ss'}/R_i'$ i.e. the matrix P will be defined as

$$P = \frac{1}{R_i'} A \tag{A.4}$$

The probabilities $P_{ss'}$ ($s=1,2, \ldots, R_i$, $s' = 1,2, \ldots, R_i'$) defined by (A.4) will always map the old start with largest permissible probability into the smallest new start at each step beginning with old start 1, then old start 2, and so on up to old start R_i.

The matrix A which defines the mapping for the case $R_i = 6$ and $R_i' = 7$ is given in Table (A.2).

Table (A.2)

Matrix A to Obtain the Probability for Post-update Start Given the Pre-update Start

	Post-update Start						
Pre-update Start	1	2	3	4	5	6	7
1	6	1	0	0	0	0	0
2	0	5	2	0	0	0	0
3	0	0	4	3	0	0	0
4	0	0	0	3	4	0	0
5	0	0	0	0	2	5	0
6	0	0	0	0	0	1	6

From the previous table, we find $Pr(1\rightarrow1) = \frac{6}{7}$, $Pr(1\rightarrow2) = \frac{1}{7}$ etc. It can be easily checked that if the mapping for the case $R_i = 6$, $R_i' = 7$ is given by the above matrix A, then the mapping for the case $R_i = 7$ and $R_i' = 6$ will be given by A^T where A^T is the transpose of matrix A, and this is true in general.

It can be readily verified that the mapping of pre-update starts to post-update starts combined with the earlier mapping of pre- to post-update samples, ensure that the number of dwellings to be used following update in retained clusters is less than or equal to the number unused prior to update. All that is required is to re-order the dwellings so that previously selected dwellings all appear under post-update starts that will not be used.

Before considering the re-ordering, it should be noted that in all cases for future clusters rotating into the sample following update, a random start r_i', $1 \leq r_i' \leq R_i'$ is chosen and a rotation schedule comprising a sequence of systematic samples is determined in the same manner as prior to update.

Re-ordering of Dwellings

The cluster isr, R_i, and the number of dwellings N_{it} in cluster i at time t determine the number of dwellings that will be selected under each start in the cluster. If $b_{it} = [\frac{N_{it}}{R_i}]$ and $Q_{it} = N_{it} - R_i.b_{it}$,

then the first Q_{it} starts have $b_{it}+1$ dwellings and the remaining ones have b_{it} dwellings. A schema or incomplete matrix is defined by N_{it} and R_i, as illustrated on the following page, for the case $N_{it} = 16$, $R_i = 6$.

starts	1	2	3	4	5	6
dwellings	X	X	X	X	X	X
	X	X	X	X	X	X
	X	X	X	X		

Fig. (2)

Ordinarily the dwellings are loaded row-wise into this schema, viz.

starts	1	2	3	4	5	6
dwellings	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16		

Fig. (3)

so that the dwellings 1, 7, and 13 would be selected with start 1, etc. New dwellings are added in a row-wise fashion, expanding the size of the matrix. If the isr is changed to R_i' at update with a post-update start of r_i', then the reorder would work as follows.

The dwellings under the unused starts are listed column-wise from left to right from the above schema, say there are L_i such dwellings. A random number ℓ_i; $1 \leq \ell_i \leq L_i$, is determined. Then in the order ℓ_i, (ℓ_i+1) mod L_i, ..., (ℓ_i+L_i-1) mod L_i, the unused dwellings are loaded column-wise into the schema under new isr beginning with the column r_i' and proceeding to the first column of the schema after the end of the last column is reached. Taking the remaining starts in the order in which they were used, dwellings are similarly loaded starting from the position following the last unused dwelling.

To illustrate, consider that at t=1, cluster i with $R_i = 6$, $r_i = 1$ was selected with the sample 6|6, and that $N_{i1} = 16$. At t=4, the sample is updated, so that $r_i^* = 4$, where r_i^* is the start that would have resulted

had there been no update. Say we have $R_i' = 7$, then the required mappings specify respectively that (i) the post-update sample should be 3|7, and (ii) the post-update start should be $r_i' = 4$ with probability 4/7 and $r_i' = 5$ with probability 3/7. Say we have $r_i' = 4$. From Fig. (3), the dwellings under the old unused starts (i.e., starts 4, 5, and 6) are {4, 10, 16, 5, 11, 6, 12}. Say $\ell_i = 3$, then the following re-order would result.

new starts	1	2	3	4	5	6	7
dwellings	7	8	9	16	11	12	10
	13	14	15	5	6	4	1
	2	3					

Fig. (5)

After using starts 4, 5 and 6, rotation would take place into the next cluster.

It should be noted that if r_i' had been chosen as a random integer between 1 and R_i', then we could have had $r_i' = 1$ in which case under the post-update starts 1, 2, 3 a total of 8 dwellings are to be selected whereas $L_i = 7$; that is a dwelling re-selection would have occurred.

It can be demonstrated with the above example that the re-order procedure is slightly biased for selection at the dwelling level. Given the pre-update sample 3|6, the unused starts can be {1, 2, 3}, {2, 3, 4}, {3, 4, 5}, {4, 5, 6}, {5, 6, 1}, or {6, 1, 2}, with equal probability where r_i^* is the first start in each case. For $N_{i1} = N_{i4} = 16$, the dwellings under each of these starts are all determined. The mapping of starts at update takes: $r_i^* = 1$ to $r_i' = 1$ with probability 6/7 and to $r_i' = 2$ with probability 1/7, after which in each case 3 dwellings out of the 9 dwellings under pre-update starts {1, 2, 3} will be selected with equal probability; $r_i^* = 2$ to $r_i' = 2$ with probability 5/7 after which 3 out of 9 dwellings are selected with equal probability, and $r_i^* = 2$ to $r_i' = 3$ with probability

2/7 after which 2 out of the 9 dwellings are selected with equal probability, etc. The overall probabilities at time t=4 are {.14484, .14749, .14749, .13955, .13690, .13690} for dwellings under pre-update starts {1, 2, ..., 6} respectively; whereas under the new isr of 7, the post update probabilities of dwellings should each equal 1/7 ≐ .14286. Given the choice between the inherent risks of respondent burden resulting from dwelling re-selections, and the slight selection bias at the dwelling level due to re-ordering, the latter has been deemed preferable.

ALTERNATIVE ESTIMATORS IN PPS SAMPLING

M.P. Singh[1]

Some estimators alternative to the usual PPS estimator are suggested in this paper for situations where the size measure used for PPS sampling is not correlated with the study variable and where data are available on another supplementary variable (size measure). Properties of these estimators are studied under super-population models and also empirically.

1. INTRODUCTION

It is well known that selection with probability proportional to size (PPS) generally improves the efficiency of the estimate of the population total for the characteristic under study provided the auxiliary variable (x) used as size measure is highly positively correlated with the study variable. Usually, therefore, in large scale multipurpose surveys where data are collected on several characteristics on a continuous basis, PPS sampling is used. The size measure (x) chosen for PPS selection in such surveys is such that it is highly correlated with the most important variable(s) under study at the time of designing the survey. However, as the time passes the initial size measure used to determine the initial selection probabilities becomes more and more out of date resulting in loss of correlation and hence the loss in efficiency of the survey estimates. In order to prevent such decline in efficiency quite often more up to date data on new size measure (z) are collected. Such data may be used either for reselection (updating) of the sample or for improving the estimation procedure. Use of new size measures in updating the sample has been discussed earlier for different sampling schemes by Keyfitz [4], Fellegi [3], Kish and Scott [5], Platek and Singh [6] and

---

[1] M.P. Singh, Household Surveys Development Division, Statistics Canada.

Drew, Choudhry and Gray [2]. In this paper, data on new size measures have been used at the estimation stage and the properties of the estimators which were introduced earlier by Singh [8] are studied.

Such estimators may also be used in the context of multi-purpose survey for those characteristics (y) that are not correlated with the size measure chosen for PPS sampling. Rao [7] has suggested an estimator alternative to the usual PPS estimator for such situations. The estimators suggested in this paper are compared with Rao's estimator and the usual PPS with replacement estimator under super-population models followed by an empirical study.

2. ALTERNATIVE ESTIMATOR

For a sample of size n selected with replacement with PPS of x, the usual unbiased estimator of the total $Y = \sum_{1}^{N} y_i$ is

$$\hat{Y}_p = \frac{1}{n} \sum_{1}^{n} \frac{y_i}{p_i} \tag{2.1}$$

with variance

$$V_p = \frac{1}{n} \sum_{1}^{N} \frac{y_i^2}{p_i} - \frac{Y^2}{n} \tag{2.2}$$

where $p_i = x_i/X$ and N is the number of units in the population, $X = \sum_{1}^{N} x_i$.

An unbiased estimator of Y in equal probability sampling (SRS) is

$$\hat{Y}_S = \frac{N}{n} \sum_{1}^{n} y_i \tag{2.3}$$

with variance

$$V_S = \frac{1}{n} \left(N \sum_{1}^{N} y_i^2 - Y^2 \right) \tag{2.4}$$

If y is uncorrelated with x then V_s would be smaller than V_p (Cochran, [1]). On this consideration, Rao [7] suggested an estimator alternative to $\hat{Y}_p$ for situations where y and x are unrelated even if the sample is selected with PPS. Rao's estimator entails 'undoing' of the PPS weights and is obtained by replacing x_i by 1 in the expression for $\hat{Y}_p$. Thus Rao's estimator is

$$\hat{Y}_o = \frac{N}{n} \sum_{1}^{n} y_i \tag{2.5}$$

and has variance

$$V_o = \frac{N^2}{n} [\sum_{1}^{N} y_i^2 \, p_i - (\sum_{1}^{N} y_i \, p_i)^2]. \tag{2.6}$$

Note that although $\hat{Y}_S$ and $\hat{Y}_o$ have the same form, their variances V_S and V_o are different due to difference in selection procedures.

Using the same reasoning, that is, whenever y and x are highly positively correlated substantial gains are achieved in using $\hat{Y}_p$ with PPS in contrast to $\hat{Y}_S$ with SRS, we consider an alternative estimator

$$\hat{Y}_{p'} = \frac{1}{n} \sum_{1}^{n} \frac{y_i}{p_i'} \tag{2.7}$$

where

$$p_i' = \frac{z_i}{Z}, \quad Z = \sum_{1}^{N} z_i.$$

Note that this estimator assumes the knowledge of an additional size measure z which is highly positively correlated with y.

The estimator $\hat{Y}_{p'}$, like $\hat{Y}_o$ is biased and their biases respectively are

$$B_{p'} = \sum_{1}^{N} y_i \left(\frac{p_i}{p_i'} - 1\right) \tag{2.8}$$

and

$$B_o = \sum_1^N y_i \, (Np_i - 1). \tag{2.9}$$

Variance of $\hat{Y}_{p'}$ is obtained by simply replacing y_i in (2.2) by $y_i p_i / p_i'$. Thus

$$V_{p'} = \frac{1}{n} \left[\sum_1^N \frac{y_i^2 \, p_i}{p_i'^2} - \left(\sum_1^N \frac{y_i \, p_i}{p_i'} \right)^2 \right]. \tag{2.10}$$

In the following section we compare these estimators under super-population models and then two other estimators are suggested in section 4 for similar situations and compared among themselves.

3. COMPARISON UNDER SUPER-POPULATION MODEL

The super-population model Δ_1 often used when y is highly positively correlated with z is

$$y_i = z_i + \eta_i \qquad i = 1, 2, \ldots, N \tag{3.1}$$

where

$$\varepsilon_1 \, (\eta_i | z_i) = 0, \; \varepsilon_1 \, (\eta_i^2 | z_i) = a \, z_i^g,$$

and

$$\varepsilon_1 \, (\eta_i \eta_j | z_i, z_j) = 0, \; a > o, \; g \geq o.$$

The smybol ε_1 denotes the average overall finite populations that can be drawn from the super-population.

Under this model Δ_1

$$\varepsilon_1 (B_{p'}) = \beta \Sigma z_i \left(\frac{p_i}{p_i} - 1\right)$$

$$= 0 \quad \text{for any } p_i, \text{ by substituting } z_i = p_i' Z.$$

Thus, $\hat{Y}_{p'}$ is unbiased under the model. However, in general, $\hat{Y}_{p'}$, like $\hat{Y}_o$, is biased and the bias does not depend on the sample size. Thus, neither estimator is consistent.

The expected variances of $\hat{Y}_o$ and $\hat{Y}_{p'}$ under the model Δ_1 are

$$\varepsilon_1(V_o) = V_o^* = \beta^2 V(\hat{Z}_o) + \frac{N^2 a Z^g}{n} \sum_1^N p_i'^g p_i(1-p_i) \tag{3.2}$$

and $$\varepsilon_1(V_{p'}) = V_{p'}^* = \frac{a Z^g}{n} \sum_1^N p_i'^{g-2} p_i(1-p_i) \tag{3.3}$$

where $$\hat{Z}_o = \frac{N}{n} \sum_1^N z_i \quad \text{and} \tag{3.4}$$

$$V(\hat{Z}_o) = \frac{N^2}{n} \left[\sum_1^N z_i^2 p_i - \left(\sum_1^N z_i p_i\right)^2\right]. \tag{3.5}$$

Further, in developing the estimators $\hat{Y}_o$ and $\hat{Y}_{p'}$, the underlying assumption is that y and x are unrelated. The super-population model Δ often used for this situation (Rao, [7]) for comparison of estimators is $y_i = \mu + e_i$, where $\varepsilon(e_i|x_i) = 0$, $\varepsilon(e_i^2|x_i) = b$, $b > 0$ and $\varepsilon(e_i e_j|x_i,x_j)=0$ and ε is defined for Δ like ε_1. Since Rao [7] has shown that $\varepsilon(V_o) < \varepsilon(V_p)$, it is enough to compare the average variances of $\hat{Y}_o$ and $\hat{Y}_p$, under the model Δ_1. In order to facilitate this comparison, we shall use the following model Δ_2 for the characteristic x, similar to the model Δ for y.

Δ_2: $\quad p_i = m + e_i' \qquad i = 1, 2, \ldots, N$

$$\varepsilon_2(e_i'|z_i) = 0,\ \varepsilon_2(e_i'^2|z_i) = a',\ a'>o \tag{3.6}$$

$$\varepsilon_2(e_i'e_i'|z_i,z_j) = 0$$

where ε_2 is defined as ε_1.

Thus, the expected variances are

$$\varepsilon_2(V_o^*) = V_o^{**} = \beta^2\varepsilon_2\, V(\hat{Z}_o) + \frac{N^2\, aZ^g}{n}\,(m-m^2-a')\sum_1^N p_i'^g \qquad \text{and} \tag{3.7}$$

$$\varepsilon_2(V_{p'}^*) = V_{p'}^{**} = \frac{aZ^g}{n}\,(m-m^2-a')\sum_1^N p_i'^{g-2}\ . \tag{3.8}$$

Therefore from (3.7) and (3.8)

$$V_o^{**} - V_{p'}^{**} = \beta^2\varepsilon_2\, V(\hat{Z}_o) + \frac{aZ^g}{n}\,(m-m^2-a')\sum_1^N p_i'^g\,(N^2 - \frac{1}{p_i'^2}). \tag{3.9}$$

Since $\quad \varepsilon_2(p_i) = m,\ \varepsilon_2(p_i^2) = m^2 + a'$ $\tag{3.10}$

and $\quad \varepsilon_2(p_i) > \varepsilon_2(p_i^2)$ because $p_i^2 > p_i$ for all possible values

except 1 or 0,

we have that $(m-m^2-a') \geq 0$. $\tag{3.11}$

Also, $\sum_1^N p_i'^2 \geq \frac{1}{N}$, with equality with all $p_i' = 1/N$. $\tag{3.12}$

For $g = 2$, the second term in (3.9) becomes $\frac{aZ^2}{n}\,(m-m^2-a')\sum_1^N (N^2p_i'^2-1) \geq 0$ because of the inequalities in the expressions (3.11) and (3.12). Therefore, for $g = 2$, in the model Δ in (3.1), the suggested estimator $\hat{Y}_{p'}$ performs better than Rao's estimator $\hat{Y}_o$.

The conditions for the choice of $\hat{Y}_p$, over $\hat{Y}_o$ for other values of g are quite complex to interpret in practice. However, as seen from the empirical study, considerable gains would be achieved in using the suggested estimator for situations where y and z are highly correlated and the coefficient of variation for x is relatively higher than that of z.

4. RATIO ESTIMATION

Two estimators of Z, namely $\hat{Z}_p$ and $\hat{Z}_o$, similar to $\hat{Y}_p$ in (2.1) and $\hat{Y}_o$ in (2.5) may be obtained using data on the new size measure z. These estimators are used to construct ratio estimators $\hat{Y}_{RP}$ and $\hat{Y}_{RO}$ for PPS with replacement sampling. Thus $\hat{Y}_{RP}$ is

$$\hat{Y}_{RP} = \frac{\hat{Y}_p}{\hat{Z}_p} \quad Z \tag{4.1}$$

where $\hat{Y}_p$ is defined in (2.1) and $\hat{Z}_p = n^{-1} \sum_1^n \frac{z_i}{p_i}$, $p_i = \frac{x_i}{X}$.

$\hat{Y}_{RP}$ has usual ratio estimation bias and variance which are approximated by

$$B_{RP} = Z^{-1} \left[RV(\hat{Z}_p) - R\ \mathrm{Cov}(\hat{Y}_p, \hat{Z}_p)\right] \tag{4.2}$$

and

$$V_{RP} = V(\hat{Y}_p) + R^2\ V(\hat{Z}_p) - 2R\ \mathrm{Cov}\ (\hat{Y}_p, \hat{Z}_p) \tag{4.3}$$

where $R = Y/Z$, $V(\hat{Y}_p) = V_p$ in (2.2),

$$V(\hat{Z}_p) = \frac{1}{n} \sum_1^N \frac{z_i^2}{p_i} - \frac{Z^2}{n},$$

and

$$\mathrm{Cov}(\hat{Y}_p, \hat{Z}_p) = \frac{1}{n} \sum_1^N \frac{y_i z_i}{p_i} - \frac{YZ}{n}.$$

It is of interest to note that $B_{p'}$ in (2.8) for PPS with replacement sampling may be approximated by (see Appendix)

$$B_{p'} = Z^{-1} \left[R\, V\left(\frac{Z_i}{p_i}\right) - \mathrm{Cov}\left(\frac{y_i}{p_i}, \frac{Z_i}{p_i}\right)\right] = n\, B_{RP} \tag{4.4}$$

and $V_{p'}$ in (2.10) may be approximated by V_{RP} in (4.3). Therefore, $\hat{Y}_{RP}$ may be prefered over $\hat{Y}_{p'}$ on account of having less bias.

An alternative ratio estimator for situations when y and x are unrelated is

$$\hat{Y}_{RO} = \frac{\hat{Y}_0}{\hat{Z}_0}\, Z \tag{4.5}$$

where $\hat{Y}_0$ and $\hat{Z}_0$ are as defined in (2.5) and (3.4) respectively.

$\hat{Y}_{RO}$ like $\hat{Y}_{RP}$ is biased but it will contain additional terms in the bias due to the fact that $\hat{Y}_0$ and $\hat{Z}_0$ themselves are biased estimates of Y and Z respectively. Approximate bias and variance of $\hat{Y}_{RO}$ may be written as (see Appendix)

$$B(\hat{Y}_{RO}) = B_0 + RB_0^* + Z^{-1}(RB_0^{*2} - B_0B_0^*) + Z^{-1}\left[R\, V(\hat{Z}_0) - \mathrm{Cov}(\hat{Y}_0\hat{Z}_0)\right] \tag{4.6}$$

and

$$V(\hat{Y}_{RO}) = V(\hat{Y}_0) + R^2\, V(\hat{Z}_0) - 2R\, \mathrm{Cov}(\hat{Y}_0, \hat{Z}_0), \tag{4.7}$$

where B_0, $V(\hat{Y}_0)$, $V(\hat{Z}_0)$ are as defined in (2.9) and(2.6) and (3.5) respectively.

Further,

$$B_0^* = B(\hat{Z}_0) = \sum_1^N z_i(Np_i - 1) \tag{4.8}$$

$$\mathrm{Cov}(\hat{Y}_0, \hat{Z}_0) = \frac{N^2}{n}\left[\sum_1^N y_i z_i p_i - \left(\sum_1^N y_i p_i\right)\left(\sum_1^N z_i p_i\right)\right]. \tag{4.9}$$

For comparing $\hat{Y}_{RP}$ and $\hat{Y}_{RO}$, we obtain their expected variances under the model Δ_1 (ie., assuming that y and z are highly correlated). We find that

$$\varepsilon_1 \text{ Cov}(\hat{Y}_p, \hat{Z}_p) = \beta \left(\frac{1}{n} \sum_1^N \frac{z_i^2}{p_i} - \frac{Z^2}{n}\right)$$

$$= \beta \, V(\hat{Z}_p) \tag{4.10}$$

and

$$\varepsilon_1 \text{ Cov}(\hat{Y}_0, \hat{Z}_0) = \beta \frac{N^2}{n} \left[\sum_1^N z_i^2 \, p_i - (\Sigma \, z_i p_i)^2\right]$$

$$= \beta \, V(\hat{Z}_0). \tag{4.11}$$

Both (4.10) and (4.11) are obtained by substituting (3.1) and noting that $En_i | i = 0$.

Thus, from (4.3) and (4.7), we have under model Δ_1

$$\varepsilon_1 \, V(\hat{Y}_{RP}) = V(\hat{Y}_p) + V(\hat{Z}_p)(R^2 - 2R\beta) \tag{4.12}$$

and

$$\varepsilon_1 \, V(\hat{Y}_{RO}) = V(\hat{Y}_0) + V(\hat{Z}_0)(R^2 - 2R\beta) \tag{4.13}$$

Further, if $\beta = R$ and

$$V(\hat{Y}_p) = V(\hat{Z}_p), \; V(\hat{Y}_0) = V(\hat{Z}_0) \tag{4.14}$$

then,

$$\varepsilon_1 \, V(\hat{Y}_{RP}) = V(\hat{Y}_p)(1 - R^2) \tag{4.15}$$

and

$$\varepsilon_1 \, V(\hat{Y}_{RO}) = V(\hat{Y}_0)(1 - R^2), \tag{4.16}$$

which shows that under the condition (4.14)

$$\varepsilon_1 \, V(\hat{Y}_{RO}) < \varepsilon_1 \, V(\hat{Y}_{RP})$$

since $V(\hat{Y}_0) < V(\hat{Y}_p)$ under the model Δ (Rao [7]).

However, in general, that is if (4.14) is not satisifed, then,

$$\varepsilon_1 \, V(Y_{RO}) \lesseqgtr \varepsilon_1 \, V(\hat{Y}_{Rp})$$

depending on

$$\{V(\hat{Y}_p) - V(\hat{Y}_0)\} + (R^2 - 2R\,\beta)\;\; V(\hat{Z}_p) - V(\hat{Z}_0)\} \lesseqgtr 0 .$$

Note that this comparison does not depend on the value of g.

Further, from (4.16), it is observed that $\hat{Y}_{RO}$ is more efficient than $\hat{Y}_0$ under usual conditions of ratio estimation. As both $\hat{Y}_{RO}$ and $\hat{Y}_{RP}$ are biased, the choice between them may be made on the basis of their biases as well. These estimators may be made unbiased or almost unbiased following usual techniques of bias reduction. In the following section, examples are given in which efficiency of $\hat{Y}_0$ and $\hat{Y}_p$, are compared.

5. NUMERICAL EXAMPLES

We have constructed 5 sets of data using two digit random numbers and each set is treated as a stratum. In each stratum, N = 20 random numbers are first drawn (designated as x) and then independently another 20 numbers are drawn (designated as y) so that y and x are unrelated. Further, the corresponding values of z are obtained by selecting 20 single digit random numbers and adding them to the numbers designated as y in order that y and z are highly correlated. Relative efficiencies of the estimates $\hat{Y}_0$ and $\hat{Y}_p$, are defined as:

$$e_{p0} = \frac{V_p}{mse(\hat{Y}_0)} \quad , \quad e_{pp'} = \frac{V_p}{mse(\hat{Y}_{p'})}$$

and

$$e_{0p'} = \frac{mse(\hat{Y}_0)}{mse(\hat{Y}_{p'})} \quad .$$

The following Table gives the bias and the relative efficency of the estimators. The correlation coefficients (δyx, δyz and δxz) and the coefficient of variations C_x, C_y and C_z are also given. The sample size in each stratum is assumed to be 2.

Table: Relative Bias and Efficiency of Alternative Estimators

	Stratum				
	1	2	3	4	5
δyx	0.092	0.007	0.012	0.069	0.070
δyz	0.998	0.995	0.997	0.998	0.998
δxz	0.099	-0.031	0.008	0.074	0.069
Cx	61	72	60	58	84
Cy	60	39	40	65	51
Cz	55	36	38	60	49
ΣY_i	1,034	1,160	1,178	983	1,063
B_0	35.4	2.2	3.5	25.5	32.3
$B_{p'}$	-44.8	6.3	-6.5	-103.7	-19.1
e_{p0}	765	3,446	1,184	342	2,530
$e_{0p'}$	1,194	4,581	8,732	455	5,824
$e_{pp'}$	9,137	157,895	103,480	1,572	147,393

Although Rao's estimate is highly efficient compared to the usual PPS estimator (e_{p0}), substantial gains are further achieved by utilizing information on z in the suggested estimator ($e_{0p'}$) for all the 5 strata.

The correlation patterns in the 5 strata are the same, that is, δ_{yx} and δ_{xz} are around zero and δ_{yz} is around 0.99 but stratum 2, 3 and 5 show considerably higher gains than those in stratum 1 and 4. This may be explained by the relative magnitude of coefficient of variation in these strata. In strata 1 and 4, C_x, C_y and C_z are approximately equal, but for strata 2, 3 and 5, we have $C_y = C_z = C_{x/2}$, which implies that the alternative estimators will perform much better if the model is satisfied and in addition if C_x is relatively higher than C_y and C_z. Bias in both the estimators seems to be usually small relative to the population total being estimated.

ACKNOWLEDGEMENT

The author wishes to thank the referee and Dr. M. Hidiroglou for their comments.

RESUME

On suggère dans cet article que certains estimateurs pourraient remplacer l'estimateur habituel basé sur l'échantillonnage avec probabilité proportionnelle à la taille dans le cas où la mesure de taille utilisée dans l'échantillonnage avec probabilité proportionnelle à la taille n'est pas corrélée avec la variable étudiée et où l'on dispose de données sur une autre variable supplémentaire (mesure de taille). On étudie les propriétés de ces estimateurs dans le contexte des modèles basés sur une population infinie, ainsi qu'empiriquement.

REFERENCES

[1] Cochran, W.G. (1963): "Sampling Techniques", John Wiley & Sons, Inc., New York.

[2] Drew, D. Choudhry, H., and Gray, G.B., "Some Methods For Updating Sample Survey Frames and Their Effects on Estimation", presented at the Annual Meeting of the American Statistical Association, August 14-17, 1978, San Diego, California, U.S.A.

[3] Fellegi, I.P., "Changing the Probabilities of Selection When Two Units Are Selected With PPS Without Replacement", proceedings of Social Statistics Section, American Statistical Association, 1966, 434-442.

[4] Keyfitz, N., "Sampling With Probability Proportional to Size - Adjustment for Changes in Size", Journal of the American Statistical Association 58, 1961, 183-201.

[5] Kish, L. and Scott, "Retaining Units After Changing Strata and Probabilities", Journal of the American Statistical Association 66, 1971, 461-470.

[6] Platek, R. and Singh, M.P., "A Strategy for Updating Continuous Surveys", Metrika, 1978, 25, 1-7.

[7] Rao, J.N.K., "Alternative Estimators in PPS Sampling for Multiple Characteristics", Sankhya, Series A, 28, 47-60.

[8] Singh, M.P., "Alternative Estimators in PPS Sampling", Abstract, Bulletin of the Institute of Mathematical Statistics, January 1975, 4.1, 29-30.

APPENDIX

Approximate expressions for bias and variance are derived here using Taylor's series expansion and considering terms of second order only, as is usually the case with ratio estimation.

(1) Bias of $\hat{Y}_{p'}$ in (4.4).

$$B(\hat{Y}_{p'}) = B_{p'} = E(\hat{Y}_p - Y) = E\left(\frac{Z}{n}\sum_1^n \frac{y_i}{z_i} - Y\right)$$

$$= E\left[\frac{Z}{n}\sum_1^n \frac{y_i/p_i}{z_i/p_i} - Y\right]$$

$$= YE[(1+e_{1i})(1+e_{2i}) - 1]$$

where $e_{1i} = (y_i/p_i - Y)/Y$ and $e_{2i} = (z_i/p_i - Z)Z$. Thus, assuming $|e_{2i}|<1$, under usual approximation

$$B_{p'} = Y[E(e_{1i}) - E(e_{2i}) + E(e_{2i}^2) - E(e_{1i}e_{2i})] \quad . \tag{A.1}$$

For PPS with replacement sampling, we have

$$E(e_{1i}) = E(e_{2i}) = 0$$

$$E(e_{2i}^2 = Z^{-2}\, E(z_i/p_i - Z)^2 = Z^{-2}\, V(z_i/p_i)$$

$$E(e_{1i}e_{2i}) = (YZ)^{-1}\, Cov(y_i/p_i, z_i/p_i).$$

Thus, under usual approximation $B_{p'}$ for PPS with replacement sampling is

$$B_{p'} = Z^{-1}\,[R\,V(z_i/p_i) - \mathrm{Cov}(y_i/p_i, z_i/p_i)], \text{ where } R = Y/Z. \tag{A.2}$$

Similarly, it is easy to show that $V_{p'} \doteq \frac{Y^2}{n} E(e_{1i}-e_{2i})^2 = V_{RP}$ in (4.3).

(2) Bias of $\hat{Y}_{RO}$ in (4.6):

$$B(\hat{Y}_{RO}) = E[\frac{\hat{Y}_0}{\hat{Z}_0} Z - Y]$$

$$= E[\frac{n^{-1}\sum_1^n Ny_i}{n^{-1}\sum_1^n Nz_i} Z - Y]$$

$$= YE[(1+e_3)(1+e_4)^{-1} - 1]$$

where $e_3 = (\hat{Y}_0-Y)/Y$, $e_4 = (\hat{Z}_0-Z)/Z$. Again assuming $|e_4|<1$, $B(\hat{Y}_{RO})$ is approximated by

$$B(\hat{Y}_{RO}) = Y[E(e_3)-E(e_4)+E(e_4^2) - E(e_3e_4)] \ . \tag{A.3}$$

Expressions for the expections involved in (A.3) are computed below for cases of PPS with replacement scheme.

$$YE(e_3) = E(n^{-1}\sum_1^n Ny_i - Y)$$

$$= \sum_1^N y_i(Np_i-1) = B(\hat{Y}_0) = B_0 \tag{A.4}$$

$$YE(e_4) = \sum_1^n z_i(Np_i-1) = B(\hat{Z}_0) = B_0^* \tag{A.5}$$

$$Z^2E(e_4^2) = E[n^{-1}\sum_1^n Nz_i - ENz_i + ENz_i - Z]^2$$

$$= E[n^{-1} \sum_{1}^{n} Nz_i - ENz_i + B_0^*]^2$$

$$= E(n^{-1} \sum_{1}^{n} Nz_i)^2 - (ENz_i)^2 + B_0^{*2}$$

$$= \frac{N^2}{n} [\sum_{1}^{N} z_i^2 p_i - (\sum_{1}^{N} z_i p_i)^2] + B_0^{*2}$$

$$= V(\hat{Z}_0) + B_0^{*2} \quad . \tag{A.6}$$

Similarly, $Y^2 E(e_3^2) = V(\hat{Y}_0) + B_0^2$. (A.7)

$$YZ\ E(e_3 e_4) = E(n^{-1} \sum_{1}^{n} Ny_i - Y)(n^{-1} \sum_{1}^{n} Nz_i - Z)$$

$$= E(n^{-1} \sum_{1}^{n} Ny_i - E(Ny_i) + B_0)[n^{-1} \sum_{1}^{n} Nz_i - E(Nz_i) + B_0^*]$$

$$= E[n^{-1} \sum_{1}^{n} Ny_i - E(Ny_i)][n^{-1} \sum_{1}^{n} Nz_i - E(Nz_i)] + B_o B_o^*$$

$$= n^{-2}[nE(N^2 y_i z_i) - nE(Ny_i)E(Nz_i)] + B_0 B_0^*$$

$$= \frac{N^2}{n} [\sum_{1}^{N} y_i z_i p_i - (\sum_{1}^{N} y_i p_i)(\sum_{1}^{N} z_i p_i)] + B_0 B_0^*$$

$$= Cov(\hat{Y}_0, \hat{Z}_0) + B_0 B_0^* \quad . \tag{A.8}$$

Using A.4, A.5, A.6 and A.8 bias for $\hat{Y}_{RO}$ in A.3 is

$$B(\hat{Y}_{RO}) = B_0 + RB_0^* + Z^{-1}(RB_0^{*2} - B_0B_0^*) + Z^{-1}[RV(\hat{Z}_0) - Cov(\hat{Y}_0\hat{Z}_0)].$$

Similarly, approximate expression for the mean square error of $\hat{Y}_{RO}$ is

$$M(\hat{Y}_{RO}) = Y^2[E(e_3^2) + E(e_4^2) - 2E(e_3e_4)].$$

Again using A.6, A.7 and A.8, and ignoring the bias terms, approximate expression for the variance of $\hat{Y}_{RO}$ is

$$V(\hat{Y}_{RO}) = V(\hat{Y}_0) + R^2V(\hat{Z}_0)\ 2R\ Cov(\hat{Y}_0, \hat{Z}_0)$$

where $R = Y/Z$.

PUBLICATION ANNOUNCEMENT

A Compendium of Methods of Error Evaluation in Censuses and Surveys

Statistics Canada has recently issued a publication entitled 'A Compendium of Methods of Error Evaluation in Censuses and Surveys'. Recognizing that a primary obligation of any survey-taking agency is to provide users of its data with information on the quality of the data it produces and disseminates, this publication attempts to catalogue a variety of methods useful in the assessment of errors in censuses and surveys.

The general approach taken in the publication is to present a section on each of seven principal types or sources of error: coverage, response and measurement, non-response, coding, data capture, edit and imputation, sampling and estimation. Within each section, the type or source of error is described, methods of controlling such errors through design or during implementation are briefly discussed, and then specific methods of measuring these errors are presented. References to applications of each method are included.

Readers interested in purchasing a copy of this publication (70 cents) should contact:

Publications Distribution,
Statistics Canada,
Ottawa, Ontario.
K1A 0T6

mentioning Catalogue number 13-564E (English) or 13-564F (French).

Avis de publication

Répertoire de méthodes d'évaluation des erreurs dans les recensements et les enquêtes

Statistique Canada a récemment fait paraître une publication intitulée "Répertoire de méthodes d'évaluation des erreurs dans les recensements et les enquêtes". Puisque l'une des obligations de tout organisme d'enquête est de fournir aux utilisateurs de l'information sur la qualité des données produites et diffusées, cette publication tente de cataloguer une variété de méthodes utiles permettant d'évaluer les erreurs dans les recensements et les enquêtes.

Nous avons choisi de présenter dans cette publication une section sur chacun de sept principaux types ou sources d'erreur: couverture, réponse et mesure, non-réponse, codage, saisie des données, vérification et imputation, échantillonnage et estimation. Chacune de ces sections vise à définir les méthodes connues de contrôler ces erreurs lors de l'élaboration et la mise en oeuvre, et à présenter des méthodes précises d'évaluation de ces erreurs. On fait mention de certaines applications pratiques des méthodes en question.

Les lecteurs intéressés à l'achat d'une copie de cette publication (70 cents) doivent contacter:

Distribution des publications,
Statistique Canada,
Ottawa, Ontario.
K1A 0T6.

Prière d'indiquer le numéro de catalogue 13-564F (Français) ou 13-564 E (Anglais).

The Editorial Board wish to thank Mr. D.A. Worton, Assistant Chief Statistician, Marketing Services Field, for his continuing support of the publication of this Journal.

Thanks are also due to Miss D.E. Weisenberg, for her patient typing of the Journal, as well as for the execution of numerous other duties associated with its production.

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

R.G. Carter
G.B. Gray
M.A. Hidiroglou
C.J. Hill
H.G. Hofmann
T.M. Jeays
M.C. Lawes
M.S. Nargundkar
H.A. Puderer
M. Rahman
K.P. Srinath
P.F. Timmons
T.J. Tomberlin

Le comité de rédaction désire remercier M. D.A. Worton, Statisticien en chef adjoint, Secteur des services de diffusion, pour son appui constant de la publication de cette revue.

On remercie également Mlle D.E. Weisenberg, qui a dactylographié soigneusement la revue, et qui s'est acquittée de nombreuses autres tâches associées à sa publication.

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

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.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1978 Vol. 4 No. 1

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