Statistics Canada Statistique Canada

SURVEY METHODOLOGY

1983
Volume 9
Number 1

A Journal produced by
Statistics Canada

Canada

SURVEY METHODOLOGY

1983 Vol. 9 No. 1

A Journal produced by Statistics Canada.

CONTENTS

8-3200-501
Reference No.
Z - 079

ISSN: 0714-0045

SURVEY METHODOLOGY

1983 Vol. 9 No. 1

A Journal produced by Statistics 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 Boad or the Department.

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, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.

THE REDESIGN OF A SURVEY TO MEASURE COMMODITY ORIGIN AND DESTINATION MOVEMENTS BY THE FOR-HIRE TRUCKING INDUSTRY IN CANADA[1]

Robert Lussier and Steven Mozes[2]

This paper firstly provides an overview of the For-hire Trucking Survey background and of the steps that were involved in the revision that led to its re-design. It secondly describes the general direction of the methodology of the re-designed survey which is being implemented for reference year 1981.

1. INTRODUCTION

The For-hire Trucking Survey was initiated by Statistics Canada in 1971 to measure commodity origin and destination movements by the for-hire trucking industry in Canada. For the purpose of the survey, this industry was defined as the sum of trucking establishments engaged in transportation of freight for compensation. The survey was a probability sample survey of shipments recorded on the shipping documents retained by Canada's for-hire trucking firms. Since 1971, the demand for more reliable and more detailed information has been increasing steadily. This increased demand can be attributed to a number of factors such as the dramatic growth of trucking activity since the early fifties; the increased sophistication of users of transportation statistics; the growing interest in the subject of economic regulation versus deregulation, and finally the increased market share of trucking, at the expense of other modes of transport, within the overall freight transportation market. In the late 1970's, Statistics Canada in cooperation with major users embarked on a complete revision of the survey.

---

[1] This is a revised version of an invited paper presented at the Joint Statistical Meetings of the American Statistical Association, the Biometric Society, ENAR and WNAR, and the Institute of Mathematical Statistics, Cincinnati, August 16-19, 1982.

[2] Robert Lussier, Business Survey Methods Division and Steven Mozes, Transportation and Communications Division, both of Statistics Canada.

It is the intention of this paper to serve two purposes. Firstly, it provides an overview of the background of the survey and a description of the steps in the revision process, and secondly, it describes the methodology of the redesigned survey which is being implemented for the reference year 1981. It should be noted that the details of the methodology of some phases have not yet been finalized; this paper will, however, include descriptions of the general direction of the incomplete phases.

2. BACKGROUND

2.1 Brief overview of the Canadian for-hire trucking industry

The Canadian for-hire trucking industry is characterized by a very large number of small operators and by a high degree of heterogeneity as manifested by the variety of commodities carried, size of operators, and area of operations.

Small carriers defined as carriers earning less than $100,000 represent 88% of the industry, measured in terms of numbers; however, they represent only 20% of the industry when measured in terms of operating revenues. The existence of this large number of operators, their volatility and their relative insignificance in terms of revenues lead to the decision to exclude them from the survey population.

Trucking firms are involved in the transportation of widely differing commodities, requiring different kind of equipment and operating practices. The various carrier types (e.g. general freight, bulk petroleum, household goods movers, etc.) differ from each other not only in terms of the commodities they carry but also in terms of shipment size.

Heterogeneity can also be illustrated by describing the area of operation. Some trucking firms operate locally only, others intraprovincially, and some of the larger carriers, in each province as well as internationally.

The combination of these factors have implications on the survey design especially on stratification.

2.2 History of Canadian truck origin and destination surveys

(a) The Motor Transport Traffic Survey (1957-1963)

The first attempt to measure truck traffic in Canada was made in 1957 with the introduction of the Motor Transport Traffic Survey (MTTS). This survey was a sample survey of motor vehicles engaged in freight transportation. The survey frame was a list of registered motor vehicles. It originated from the motor vehicle registration files maintained by the provincial or territorial governments. This frame was stratified by type of operation and gross vehicle weight.

The sample size was approximately 10% of all registered vehicles. The sample was selected in four quarterly segments with approximately one fourth of the total sample seleted each quarter. Each quarterly sample was spread over three survey weeks with one third of the sample being used for a seven day period per month.

As the survey was conducted on a vehicle basis, no information was requested regarding the detailed origin and destination of the commodities carried. It was a truck origin and destination survey; commodities had secondary importance. Data relating to the vehicle such as the description of the vehicle, miles travelled, fuel consumed and the operating cost associated with the vehicle were also collected.

The survey was in operation from 1957 to 1963 inclusive. It was discontinued in 1964 because of changes in vehicle licensing systems, structural changes in the industry, and most importantly because of a very se-rious deterioration in response rates.

(b) The For-hire Trucking Survey (1969-1979)

Initial work on a survey to measure domestic intercity origin and destination traffic movements of goods by the total Canadian for-hire trucking industry began in 1969. At that time, a study of various methods of collecting commodity origin and destination statistics was carried out. The study results showed that a sample survey of the carriers' administrative records, namely their shipping documents, was a viable approach to the collection of the required data.

In 1970, a pilot survey was conducted to assess the effectiveness of the survey approach. The pilot survey involved the examination of the shipping documents of 187 for-hire trucking firms throughout the country.

The favourable response to the pilot survey and the availability of origin, destination, commodity, weight and revenue information on the shipping documents indicated that the survey approach was feasible.

For-hire trucking surveys were therefore conducted for reference years 1970 and 1971 with the above-mentioned objective. For reference year 1972, the objective was modified to restrict the survey to Canadian domiciled for-hire carriers earning $100,000 or more annually from inter-city trucking. For reference year 1973, an updated and better-defined frame of regulated motor carriers was used and a more effective sampling procedure was developed. Since reference year 1973, the survey has been conducted and the results published on an annual basis by the Transportation and Communications Division of Statistics Canada [1] [5].

3. REVISION OF THE FOR-HIRE TRUCKING SURVEY

The revision process consisted of two main phases. Firstly, a critical review of the existing survey was initiated. Secondly, on the basis of the recommendations made during the review process, a complete survey redesign was

undertaken.

3.1 Survey review

(a) Reasons for the review

In early 1978, Statistics Canada initiated a review of the For-hire Trucking Survey for the following reasons. First, it has been the policy of the Transportation and Communications Division of Statistics Canada to conduct a periodic review of each of the ongoing surveys. The For-hire Trucking Survey had not been reviewed since 1973. Secondly, the current and anticipated future needs of users for increased details for commodity origin and destination statistics could not be satisfied within the constraints of the survey. Thirdly, experience gained during the undertaking of the For-hire Trucking Surveys and other related surveys provided additional information upon which the frame, the stratification variables and the imputation techniques could be improved. Fourthly, some developments in the trucking industry, such as the availability of origin and destination information in machine readable format lead to the belief that computer tapes could be utilized to increase the data base and reduce reporting burden at the same time.

In addition, increased sophistication of users required the development of improved data dissemination procedures while changes in data processing technology had made the present data processing system not only obsolete from a technical point of view but cost inefficient as well.

(b) Phases of the review

The survey review was originally organized into two parts, namely Phase I and Phase II.

The objective of Phase I was to outline recommendations which concentrated on improvements to the survey within its existing framework using only limited additional resources. The recommendations had to focus and indeed did focus

on a redefinition of the survey population, improvements in stratification variables and an increase in the sample size of shipping documents. The recommendations were presented in a report [2].

The objectives of Phase II were to assess the Phase I recommendations from a user point of view, to present various cost and implementation alternatives for the accepted recommendations and to complete further survey analyses. Phase II reformulated some of the Phase I recommendations and added additional recommendations which aimed at a smaller population of firms better stratified into more homogeneous groups. In addition to recommending the implemention of these recommendations, four alternatives for increasing the sample size were considered namely, the status quo; an increase of 50% to the sample of shipping documents; an increase of 100% to the sample of shipping documents; and finally an increase of 25% to the sample of shipping documents together with the processing of available carrier data tapes for presumably 40 or so firms. Based on an assessment of the advantages and costs of each of these alternatives, the latter one was approved in principle because it offered the potential for substantial sample size increases with a minimum of cost and data collection burden. The recommendations and the supporting details were tabled in a report [3].

A preliminary assessment of the impact of the recommendations revealed that further work was necessary especially to determine the full costs for the use of carrier waybill tapes. Therefore a Phase III was added tothe survey review process. In general, its terms of reference were to conduct the investigations required to formulate and recommend general specifications for a revised survey. The investigations had to follow the recommendations of the Phase II review.

In June 1980, Phase III proposed that the survey be redesigned to accept four types of input namely, tapes from selected respondents; transcriptions from samples of shipping documents drawn from each Document Storage Location Point (D.S.L.P.) having over 1.5 million intercity domestic revenue annually; transcriptions from samples of shipping documents drawn from a sample of D.S.L.P.'s

having between $350,000 and $1.5 million intercity domestic revenue annually; and macro information from D.S.L.P.'s with annual intercity domestic revenue between $100,000 and $350,000. The decision to collect macro information from the smaller carriers was based on that fact that these firms do not keep the documentation needed for sampling purposes.

3.2 Survey redesign

(a) Objective of the redesign

After the completion of Phase III of the survey review process, it was decided to carry out a complete redesign of all aspects of the survey. The objective of the redesign was to provide more reliable and more detailed commodity origin and destination statistics relating to the Canadian for-hire trucking industry. It was expected that both the reliability and the amount of regional and commodity detail available could be increased when compared with the "old survey".

(b) Constraints on the redesign

The main constraints imposed on the redesign were: that the survey population exclude some types of for-hire trucking firms namely, own account household goods carriers and oil field carriers; that the stratification be improved to be more in line with the economic structure of the industry; that three types of input be accepted, namely, tapes from selected respondents, transcriptions from samples of shipping documents drawn from D.S.L.P.'s of a sample of firms having more than $350,000 intercity domestic revenue annually and macro-information from a sample of firms having annual intercity domestic revenue between $100,000 and $350,000; and finally that the redesigned survey be implemented for the reference year 1981, data collection starting in the spring of 1982.

4. POPULATION AND FRAME

The population of the survey covers all shipments made during the reference year by those trucking firms which are defined as in scope for the survey. A shipment is defined as a quantity of merchandise transported by the carrier from one person or organization to another person or organization. The in-scope firms include those which earn more than $100,000 annually from intercity freight transportation, whose main activity is trucking and who are Canadian domiciled. Excluded from this population are the shipments of certain types of specialized carriers such as the oilfield carriers and own account household good movers.

However, this ideal population is not accessible. As a substitute, firms are used as natural clusters of shipments for the first-stage sampling units of the design.

For this reason, the frame consists of a list of all firms which have domestic intercity revenue over $100,000. Firms may further be segregated into D.S.L.P's. This is the case for those firms whose shipping documents are not stored at a central location. The frame is derived from an annual census survey of for-hire trucking conducted by Statistics Canada, the Motor Carriers-Freight and Household Goods Movers Survey[3].

5. ULTIMATE SAMPLING UNIT

The survey accepts three different inputs, namely tapes from selected carriers, transcribed information from sampled shipping documents and finally macro information from carriers earning between $100,000 and $350,000 annually.

---

[3] The Motor Carriers-Freight and Household Goods Movers Survey of Statistics Canada is an annual census survey of trucking establishments. Its objective is to obtain establishment-oriented input-output data such as revenues, expenses, balance sheet information and equipment operated.

The tapes contain information relating to individual shipments, the characteristics of which are the same as those which are recorded on shipping documents. Therefore, the ultimate sampling unit for those firms which either provide tapes or whose shipping documents are sampled is the shipment. For the carriers in the $100,000 to $350,000 range, macro information is obtained as these firms do not usually keep the necessary documentation relating to shipments. For these carriers, the ultimate sampling unit is the firm.

6. INFORMATION COLLECTED

The principal characteristics needed from each shipment sampled from carriers earning more than 350,000 intercity domestic revenue annually are the true origin and the final destination; the description of the commodity(ies) carried; the weight and the unit of weight; the transportation revenue earned and the interlined shipment information. Interlining occurs when a consignment is moved by a carrier to an intermediate point and then moved by another carrier to another point. The interlined shipment information is used to eliminate duplications.

The secondary characteristics needed are the date of shipment; the quantity of commodity and the unit of measurement (e.g. 5 board feet, 20 gallons, 15 sacks); some information regarding the shipment weight transcribed (e.g. minimum weight, convenient weight used for calculating revenue); the rate charged and the rate condition codes (e.g. a code indicating where rate is minimum, per 100 lb., per hour) and the revenue condition codes. (e.g. a code indicating where exact transportation revenue is not available, where the shipment is out-of-scope).

The macro information collected from the smaller carriers describe the average or typical shipments in terms of originating province, destination province, commodity, average revenue, average weight and number of shipments.

7. ADMINISTRATIVE RESTRICTIONS

The amount of resources available for data collection and processing and the goal to reduce the burden imposed on the respondent put a limit on the number of firms selected and on the number of shipments selected and transcribed.

7.1 Maximum number of firms in the sample

The population of the 1981 For-hire Trucking Survey[4] consists of 2,711 firms of which 1,288 earn more than $350,000 annually while 1,423 earn between $100,000 and $350,000 annually.

As data collection is very expensive due to the very high cost associated with travelling to remote areas, efforts are being made to limit the number of D.S.L.P.'s selected in the sample from those carriers earning over $350,000 annually. The limit is set at 875 D.S.L.P.'s per year, which has been the historical number during the last ten years of the old survey.

7.2 Maximum total number of transcriptions

The second administrative restriction relates to the total number of transcriptions. The present budget allocation allows a maximum of 418,000 transcriptions. This number may vary from year to year depending on negotiations between Statistics Canada and users who are also cofinancers of the survey.

7.3 Maximum number of transcriptions per firm

There is also an administrative restriction which relates to the maximum number of transcriptions per firm. There is an implicit limit imposed on the number of days the data collection team can spend at any particular location, so that the respondents are not burdened by the presence of the Statistics Canada regional operations personnel.

---

[4] 1981 For-hire Trucking Survey means the survey conducted in 1982 for reference year 1981.

8. STRATIFICATION AND SAMPLE ALLOCATION

Using the results of the previous year's Motor Carriers-Freight and Household Goods Movers Survey, the firms are stratified according to their in-scope transportation revenue, type of operation and area of operation. These variables were chosen because they characterize the heterogeneity of the industry. The in-scope transportation revenue indicates if the firm is a Class 1, a Class 2 or a Class 3 firm i.e. if the firm earned $2.7 million or more, between $350,000 and $2.7 million or between $100,000 and $349,999 dollars of revenue respectively from Canadian intercity non-armoured and non-household goods freight transport. The type of operation characterizes the firms as specializing in general freight small shipments, general freight large shipments, automobiles, liquid petroleum, dump trucking, forest products, building materials, dry bulk and/or refrigerated liquids, heavy machinery, refrigerated solids, explosive and/or other dangerous goods, agricultural products, animals and van lines. The general freight small shipment carriers are general freight carriers for which the average revenue per shipment is less than $85.00; the general freight large shipment carriers are the rest of the general freight carriers. The area of operation indicates the specific Canadian province, Yukon or Northwest Territories, or that combination in which the which the firm operates. For example, an area of operation could be New Brunswick, meaning that the firm operates in New Brunswick only. Another example would be Atlantic which means that the firm operates in 2 or more of Newfoundland, Prince Edward Island, Nova Scotia and New Brunswick but nowhere else in Canada. There are 20 of these areas of operation.

The dollar cut-offs used in the stratification by revenue and by type (i.e. $85, $350,000 and $2.7 million) are flexible and may vary in the years to come depending on the changes occurring in the population.

The above stratification creates 840 strata of which 355 were non-empty in the 1981 For-hire Trucking Survey.

Once the frame is stratified, subject matter officers may identify take-all

firms i.e. firms that they want to be included in the sample with probability one. Next, a methodologist determines the number of firms to be selected among the non take-all firms in the stratum. To do so, he goes through several steps from which the take-all firms are excluded.

First, a computer program calculates the initial number of firms to be selected in each stratum to meet a target coefficient of variation of the estimate of in-scope revenue in the stratum. This target coefficient of variation is the coefficient that one would like to obtain if the estimate were calculated using the reported total revenue from a sample of firms selected using simple random sampling from a population of firms for which the distribution of the in-scope revenue is the same as the distribution of the in-scope revenue of the previous year's Motor Carriers-Freight and Household Goods Movers Survey. The formula is:

$$ {}_1n_h = \frac{N_h^2 S_h^2}{N_h S_h^2 + Y_h^2 (C.V._h)^2} $$

where ${}_1n_h$: initial number of firms to be selected among the non take-all firms in stratum h;

N_h : number of non take-all firms in stratum h;

Y_h : total in-scope revenue of the non take-all firms in stratum h;

S_h^2 : variance of the in-scope revenue of the non take-all firms in stratum h; and

$C.V._h$: target coefficient of variation in stratum h (the value used is the same for all strata of a given class but may vary from class to class).

Secondly, the initial sample sizes are revised to ensure that a minimum number of firms is selected from each stratum i.e.

$$_2n_h = \min \{\max (m, {}_1n_h), N_h\}$$

where $_2n_h$: revised initial number of firms to be selected among the non take-all firms in stratum h; and

m : minimum number of firms to be selected in stratum h if possible[5].

Then the revised initial sample sizes are summed over the strata to get a total revised initial sample size.

Next, the sample sizes are again reviewed to ensure that the sample size in a given stratum is greater or equal to the sample size that one would have obtained if he had distributed the total revised initial sample size of a class across the strata of the class proportionally to the square root of the number of firms in each stratum i.e.

$$_3n_h = \max \left\{ \frac{\sqrt{N_h}}{\sum_h \sqrt{N_h}} \sum_h {}_2n_h, {}_2n_h \right\}$$

where the summation is done over all strata of the same class than stratum h.

Finally, the survey manager may subjectively adjust the sample sizes to $_4n_h$. The above sample allocation method has been retained because it is an algorithm which has given satisfactory results during the testing phase as well as has made use of the only variable that was available for all firms namely the in-scope revenue of the firms. Nevertheless, it should be realised that the in-scope revenues of the firms are not collected directly in the For-hire Trucking Survey but revenues from a sample of shipments are collected. Therefore the above method of sample allocation ignores completely the second stage of sampling.

---

[5] For reference year 1981, this minimum was set to 3 for all strata.

9. FIRST STAGE SAMPLE DESIGN

The first stage consists of selecting in each stratum a number of firms corresponding to the number of firms ${}_4n_h$, determined at the sample allocation stage.

All firms earning $2.7 million or more of in-scope transportation revenue were made take-all i.e. were selected with probability one in the 1981 For-hire Trucking Survey. The reason for this approach is that these firms are known to be heterogeneous with respect to the principal statistics to be estimated and are known to be contributing a large proportion of the revenue figures to be estimated.

The sample of firms is finally converted to a sample of D.S.L.P.'s by including in the latter sample all D.S.L.P.'s of the selected firms.

10. SECOND STAGE SAMPLE DESIGN

The second stage of the sample design for D.S.L.P's of Class 1 and Class 2 firms consists of selecting a systematic sample of shipments from the files of each selected D.S.L.P. This selection is done by Statistics Canada Regional Operations Division interviewers at the D.S.L.P. The sampling intervals used are different depending on the number of shipments carried by the firms. They are generally obtained from a table provided to the interviewers. This table gives various file size ranges with their corresponding sampling interval. However, the sampling intervals may be pre-determined for any given firm by Statistics Canada Head Office staff. This is especially the case of multi-D.S.L.P. firms because the interviewer at a given D.S.L.P. may not know how many shipments were carried by the firm as a whole. This is also the case for firms having special characteristics, such as firms carrying dangerous goods, and others for which the survey manager may want a larger data base. In subsequent years, this may also be the case for firms contributing to domains where the reliability of the estimates in the previous year was less or more

than what was desired.

For D.S.L.P.'s of Class 3 firms, there is no second stage sampling design. Individual shipments are not selected from the files of the D.S.L.P.'s. Instead, aggregated data are collected at the D.S.L.P. level.

11. FIELD OPERATIONS

The field operations are different for class 1 and class 2 firms than for class 3 firms. For class 1 and class 2 firms, the operations consist of selecting shipments from the files of their D.S.L.P.'s and of transcribing the characteristics of the selected shipments on coding sheets. For class 3 firms, they consist of obtaining aggregated data over the telephone about their trucking operations.

This section discusses the activities that involve the Statistics Canada Regional Operations staff; namely the training of the Regional Operations project managers, the planning of the collection, the collection at the D.S.L.P.'s of class 1 and class 2 firms, the collection from the class 3 firms and finally the profiling of class 1 and 2 D.S.L.P.'s.

11.1 Training of the regional operation project managers

Every year, the Statistics Canada regional operations project managers are trained on all aspects of the survey. The training session is four days long and is conducted during the month of March. It is broken down into two components: an in-class-training and an on-the-job training. The in-class training consists of a series of talks and exercises given by the survey project manager and the methodologist(s). The on-the-job training consists of having groups of three to four people visiting a D.S.L.P. and applying and discussing the knowledge acquired during the in-class training.

11.2 Planning of the collection

Having been trained, the regional operations project managers recruit the interviewers and administer a thorough training program. Then the interviewers with the advice of their regional operations project manager schedule their work and plan their itineraries for their visits to D.S.L.P.'s of class 1 and class 2 firms. The itineraries are drawn to avoid unnecessary travel and to achieve maximum productivity. The interviewers mail to the D.S.L.P. officials introductory letters which provide a brief explanation of the survey. Subsequently, the interviewers telephone D.S.L.P. officials for appointments. The collection of the data takes place between May and September for the survey covering the previous calendar year.

11.3 Overall description of the collection in the D.S.L.P.'s of Class 1 and Class 2 firms

At the time of the appointment, the interviewer conducts an interview with the D.S.L.P. officials. During the interview, he/she explains the survey, describes the uses of the data, estimates the time required to do the work and asks information about the firm. This information concerns mainly revisions to the names and addresses, changes of ownership, type(s) of document and filing system used and aggregated data about the operations of the D.S.L.P. during the reference year.

The most common types of shipping documents are the probills, bills of lading, load manifests, trip reports, and invoices. A firm may use any combination of these.

The types of filing system include: in complete numeric sequence; in broken numeric sequence; in chronological order; in alphabetical order (e.g. by customer name); by terminals; by commodity type or in no order at all. The documents may even be cross-filed; for example, by serial number and by customer's name. Within a filing system, documents may be kept in a set of file drawers, in sets of binders or shannon files, on shelves, in drawers, or even in books.

The aggregated data about the operations of the D.S.L.P. cover several variables among which are the total transportation revenue earned; the total tonnage carried; the total number of shipments transported; the percentages of each of these three items represented by intercity shipments and the percentages represented by international shipments; the types of commodities carried and the percentage each type represents in the total transportation revenue.

Often the interviewer has a choice of filing systems which provide information on the items needed in the survey. The interviewer assesses the completeness of the various filing systems with regard to the information on the five principal characteristics and on the reference year, and then chooses the system having the smallest under-coverage. However, if two or more systems have the same under-coverage (if any), the interviewer selects the one that includes the smallest number of out-of-scope records or the one that allows out-of-scope records to be removed from the file or not to be counted.

Next, the interviewer selects the sample of shipments as follows. Using the number of shipments reported by the official of the D.S.L.P., he/she gets from a table the corresponding sampling interval and random start. In some instances, the interval and the random start may have already been pre-determined by Statistics Canada Head Office. Next, he/she adds the random start and/or the interval to the document numbers to get the selected shipments in numeric filing system. Otherwise, he/she has to count a number of documents equal to the random start and/or to the interval to get the selected shipments.

Once a shipment is selected, the interviewer transcribes its characteristics. The transcribing operation is often difficult because it can be hard to understand the various documents and the coding used on some documents. This is especially true for the commodity names. The interviewer must avoid the use of brand names, proper names and names which have more than one meaning. The interviewer often has to interpret the information on the documents and to enter on the coding sheets the data in a format that would be accepted by the computer system.

11.4 Overall description of the collection in the Class 3 firms

The interviewer mails an introductory letter two to three weeks prior to any attempt to contact the firm by phone. Subsequently, he/she contacts the official in the firm that is best suited to provide the required information. This may, however, take several phone calls. The interviewer then conducts an interview over the phone.

During the interview, he/she will ask questions similar to the questions asked for class 1 and 2 D.S.L.P.'s. There is a major difference however; no questions are asked about the types of documents utilized and the filing systems used by the firm. Once this first part of the interview is completed, the interviewer proceeds to have the respondent describe his types of shipments. For each type of shipment, the description is to be made in terms of province of origin; province of destination and name of commodity carried. Then the official is asked to report an estimate of the number of shipments, the average weight and the average revenue of each type of shipment.

It is a general subject matter belief that the operations of any given class 3 firms are fairly homogeneous. Therefore each has only a few types of shipments to report. The coverage obtained throught this approach is also believed to be acceptable from a user point of view. No testing was done of this hypothesis.

11.5 "Profiling" of Class 1 and 2 D.S.L.P.'s

It sometimes happens that a class 1 or 2 D.S.L.P. cannot provide any documents, does not keep documents suitable for sampling or cannot provide a portion of the shipping documents and that this portion cannot be represented by the available documents. The latter may happen for example when the missing documents represent specific contracts that have been removed for audit purposes. In these cases, the interviewer has to "profile" the missing documents. The profiling consists in having a D.S.L.P. official describe the types of shipments on the missing documents. The profile is similar to the description

of the types of shipments of the class 3 firms with the exception that the precise origin and the precise destination of the shipments (i.e. the village, town, city, etc.) is wanted in this case.

The profiling activity can be long in some D.S.L.P.'s because their operations can be quite extensive. It requires good cooperation from the D.S.L.P.'s official.

12. DATA PROCESSING

12.1 Manual processing

The completed documents are sent to Statistics Canada Head Office in Ottawa. Upon receipt, the documents are logged in and the identification numbers verified. Two short tasks are also undertaken at this point.

First, a brief scan is conducted to identify and code closings of D.S.L.P.'s, death of firms, out-of-scope firms and abortions. Out-of-scope firms are active firms for which the in-scope revenue is nil for the reference year. Abortions are D.S.L.P.'s for which no information was collected although it was known that the D.S.L.P. had in-scope revenue for the reference year. As examples, a firm found in the field to have earned its revenue 100% from local shipments would be an out-of-scope firm while a single D.S.L.P. firm that refuses to cooperate or is on strike would be an abortion.

Secondly, the profile data of the class 1 and 2 D.S.L.P.'s are examined to determine the number of shipments that should have been transcribed for each reported type of shipments if the documents had been available. These numbers are determined by performing calculations using the total number of shipments covered by the profile, the random start and the sampling interval that should have been used if the documents had been available. These numbers are then coded so that the computer could generate the required number of transcription records for each type of shipments as if transcriptions were obtained.

12.2 Data capture

The forms are next sent to data capture. The capture is done on a mini-computer which allows edits and other processing to be performed on-line.

There are many edits performed on the mini. Some edits generate error messages and require corrective action; others generate warning messages that require verification of the entered data and corrective action only if necessary Some edits consider the validity of each response individually while others consider the relationships between valid characteristics of the same shipment. The operators of the mini-computer are expected to possess subject matter knowledge to perform corrections on-line. Manual imputations are performed when necessary because there is no automated imputation performed for class 3 D.S.L.P.'s.

As part of the other processing, the weight is converted to metric and the rate to $/100 kilograms. Also, the origin and destination names (i.e. village, town, etc.) if present, are matched against a municipality library to obtain a Standard Geographical Classification (S.G.C.) code, a latitude and a longitude. Whenever there is a nonmatch, the operator is instructed to enter a synonym. Similarly, the commodity name is matched against a commodity library to get a 3-digit Standard Commodity Classification (S.C.C.) code. Whenever there is a nonmatch, the operator uses a synonym or enters "unknown". There is therefore always an S.C.C. code for each shipment. Also, the mini generates the required number of transcription records for each type of shipments from the profile data of the class 1 and 2 D.S.L.P.'s.

Finally, the data are unloaded from the minis and two data sets are created; a data set of shipments of class 1 and 2 firms and a data set of type of ship-data of class 3 firms. The principal difference between the two data sets are that the first one is at the shipment level while the second one is at an aggregated level. Note also that the first one has more variables (e.g. rate, place of origin rather than province of origin, etc.) than the second one.

12.3 Main frame edits and imputations

A road distance between the origin and the destination of each in-scope class 1 or 2 shipment has to be obtained in order to be able to provide tonne-kilometres estimates for class 1 and 2 firms. Therefore, the origin S.G.C. - destination S.G.C. pair of each in-scope class 1 or 2 record is matched against a distance library to get a road distance in kilometres between the two locations. Whenever there is a nonmatch, an aerial distance (X) is calculated using the latitudes and longitutes of the origin and of the destination. using the latitudes and longitudes of the origin and of the destination. Then X is converted to a road distance Y using the simple linear regression model

$$Y = a X + b$$

where a and b vary according to 12 regions of origin and 12 regions of destination. The road distance is assigned to the record.

Missing data of partially transcribed shipments of class 1 and 2 firms are also imputed. The imputation technique used depends on the missing variable or the pair of missing variables. Major imputations are performed using fixed relationships between reported figures, unit weight conversion factors and pro-rate tables. An example of a fixed relationship between reported figures is

$$\text{weight} = \frac{\text{revenue} \times 100}{\text{rate}}$$

This relationship can be used to impute weight when revenue and rate are present or revenue when weight and rate are present. Unit weight conversion factors are coefficients determined by unit type (e.g. case, bag, litre, etc.) by S.C.C. code. Knowing the unit and the S.C.C. code of the commodity, the proper conversion factor can be applied to the quantity of units to derive the weight. Finally, pro-rate tables show rates by commodity section, by distance block and by revenue or weight group. These tables are based on the previous years' data modified by incoming valid current-year data. The pro-rate tables are used to calculate the weight when the revenue is present or the revenue when the weight is present.

In cases where too many characteristics of a shipment have to be imputed, the shipment is flagged as not usable.

Expansion edits are subsequently performed. For class 1 and 2 firms, these edits consist of weighting up crudely the number of shipments transcribed, the transcribed revenue and the transcribed tonnage and comparing the results to the total number of shipments, revenue and tonnage provided during the interview by the D.S.L.P. official. Similar edits are performed for class 3 firms. Discrepancies in both cases are followed up.

13. ESTIMATION PROCEDURES

For the estimation procedures, it was decided to consider the second stage systematic sampling in the class 1 and 2 firms as simple random sampling without replacement (S.R.S.W.O.R.). This decision was made because first the documents were considered to be in random order and secondly the use of S.R.S.W.O.R. allows the computation of an estimate of the sampling variance.

As the first step of the estimation procedures, weights are calculated. There are first stage and second stage weights for class 1 and 2 records but only first stage weights for class 3 records. In general, first stage weights correspond to the inverse of the probability of selecting of a D.S.L.P. in its stratum and second stage weights correspond to the inverse of the probability of selecting of a shipment in its D.S.L.P. supposing S.R.S.W.O.R. was used. First stage weights are adjusted by computer to reflect the contribution of abortions. No adjustments are made for the closing of D.S.L.P.'s, deaths of firms and out-of-scope firms because they are considered as having generated no shipments. Final weights are attached to each record on the data set of class 1 and 2 firms and on the data set of class 3 firms.

Detailed diagnostic reports are produced. These reports are tables which present the data under various aggregates. They are useful tools to analyse the data and to perform final quality checks.

The data set of class 1 and 2 firms is cleared by discarding out-of-scope shipments. Some types of out-of-scope shipments are shipments to or from the U.S.A.; shipments transported 15 miles or less from origin to destination; shipments which were off-highway; shipments which would be double counted as a result of interlining between road carriers; shipments which would be double counted because they were recorded by household goods movers who are van line agents and by the van lines themselves; shipments which did not generate any intercity transportation revenue; and records which relate to non-transportation services such as storage, packing, equipment rental, labour loading and unloading.

Estimates of revenue, tonnage and tonne-kilometres for the publication are finally generated by summing the weighted data over the appropriate domains. Measures of error such as the coefficients of variation are also provided with the estimates. The coefficients of variation are obtained from the formula derived from the sample design but supposing the systematic sample of shipments is a simple random sample of shipments.

14. USE OF THE DATA AND METHODS OF DATA DISSEMINATION

14.1 Use of data

Requests for estimates yielded from the old survey came from a wide variety of sources. The nature of these requests has also varied a great deal. It is expected that the nature of the demand for data from the new survey will be similar to that in the past.

The estimates have been used extensively to satisfy five main requirements, namely, to measure the volume of domestic trade transported by intercity for-hire carriers provincially and interprovincially; to measure the rate of industrial growth reflected by intercity commodity movements; to provide information on regional development; to assist in transportation studies; and to support the presentation of briefs, submissions and other inquiries to

regulatory authorities and commissions.

One specific use of the data was to define the characteristics of trucking markets using variables such as commodities carried, average lengths of haul and shipment weight. Another specific study examined and analyzed selected aspects of performance of carriers operating in regulated and unregulated environment. The cost behaviour of these carriers was examined by using traffic characteristics such as shipment size and average length of haul.

In the past, special requests for estimates from this survey have come from sources such as government departments concerned with trade, transport regulatory officials at both federal and provincial levels; carriers; university consultants; industry associations; and many other organizations and individuals who share a common interest in transportation.

14.2 Methods of dissemination

The redesigned survey will provide information in three modes similar to the old survey.

First the publication will present the estimates that are generated by the regular system of the survey. Measures of error such as the coefficients of variation will be given with the estimates. Secondly, special requests will be processed subject to cost and reliability constraints. Finally, the data base of shipments generated by the survey may be made available on magnetic tape to selected users subject to constraints of confidentiality.

15. FUTURE WORK

As mentioned earlier in this paper, the survey accepts three types of input, one of which is computer tape from selected respondents. This type of input has been found difficult to handle and, although work has commenced on this

subject, progress so far is disappointing. Extensive negotiations are required with the firms to obtain the requested data on tape and then a further analysis is needed to evaluate the data. For reference year 1981, only one tape will be used for a firm which handled about 5 million shipments during reference year 1981. More firms will provide data on tape in subsequent years. Agreements are presently being reached with 5 additional companies for reference year 1982.

Nevertheless, when a firm's computer tape is finally obtained and found to meet our requirements, extensive systems manipulation will still be required to handle the tape. Also, manual interventions will be necessary to handle non matches to the various libraries. Therefore, records will most likely have to be sampled on each tape using the same second stage sampling design as for the sampling of documents of class 1 and 2 firms.

Another area where future work is needed is on having firms themselves sample their documents. As an example, a company could photocopy the pro-bills ending in a given number when the pro-bills are issued, and send the photocopies to Statistics Canada on a monthly basis.

Finally, major efforts will be made to evaluate thoroughly the various phases of the survey and to formulate recommendations for improvements. These recommendations will hopefully be implemented for the 1982 reference year survey.

REFERENCES

[1] Statistics Canada, For-hire Trucking Survey, Catalogue 53-224, Annual.

[2] "Report on the Findings and Recommendations of the Working Group on the For-hire Trucking Survey Phase I Review"; prepared by the Transportation and Communications Division of Statistics Canada, July 7, 1978.

[3] "Report of the Interdepartmental Working Group on For-hire Trucking - Phase II Review"; prepared by the Transportation and Communications Division of Statistics Canada, April 1979.

[4] Statistics Canada Motor Carriers - Freight and Household Goods Movers, Catalogue No. 53-227, Annual.

[5] Lussier, R. (1981), "For-hire Trucking Survey: Survey Design", Survey Methodology, Statistics Canada, Vol. 7, No. 1, pp. 74-92.

THE METHODOLOGY OF THE CANADIAN AIR SCHEDULED INTERNATIONAL PASSENGER ORIGIN AND DESTINATION ESTIMATION SYSTEM[1]

Greg Hunter and Lisa DiPiétro[2]

The Air Scheduled International Passenger Origin and Destination (ASIPOD) estimation system uses the data from two air traffic surveys to produce origin-destination estimates of international passengers. The "assignment technique" is the solution to the problem caused by the non-coverage of non-interlining traffic. The assumptions of the technique are sufficiently questionable to warrant an evaluation of the bias of the estimates. However, major improvements will be made in the new system which will decrease the bias in the estimates. Also, estimates of reliability will be produced. And as a result, knowledge of the strength of the inferences made with respect to air traffic markets from these estimates will be improved in international bilateral air negotiations.

1. INTRODUCTION

In 1979, Statistics Canada embarked on a revision of the federal aviation statistics program by inviting Transport Canada and the Canadian Transport Commission (i.e. the two "user departments") to form an interdepartmental revision team. The ASIPOD estimation system is one of several projects in the revision program.

The ASIPOD estimation system uses the data from two air traffic surveys to produce estimates of the number of passengers on scheduled international flights between Canadian and foreign markets for various origin-destination combinations. The first of these two surveys, the revenue passenger origin and destination survey, provides a sample of origin-destination data on international journeys with Canadian carriers on at least one leg of the itinerary.

---

[1] Presented at the Joint Statistical Meetings of the American Statistical Association in Cincinnati, August 1982.

[2] Greg Hunter, Business Survey Methods Division, Statistics Canada.
Lisa DiPiétro, Transportation and Communications Division, Statistics Canada.

A coverage problem exists, since no data are available on those international journeys with foreign carriers on all legs of the itinerary. The second survey, the airport activity survey, counts all passengers entering or leaving Canada on all Canadian or foreign scheduled carriers without consideration of the passenger's origin or destination.

This paper first outlines the requirements users have for international passenger origin and destination estimates. Then, the relevant aspects of the two air traffic surveys and the non-coverage problem are presented. And, finally, the paper describes how the ASIPOD estimation system will produce estimates of the number of passengers and the associated coefficients of variation for various international origin-destination pairs for the portions of the international market both covered and not covered by the first survey.

2. USER REQUIREMENTS

The users require estimates of international scheduled commercial air service passengers by origin and destination for bilateral air negotiations.

An international scheduled commercial air service is defined to be an operation which is between points in Canada and points in any other country, and which provides public transportation of persons, goods or mail by aircraft in accordance with a schedule and at a toll or charge per unit of traffic. Such a service is referred to as a "unit toll" service.

Before an international scheduled commercial air service can be operated into and out of Canada, some form of formal agreement must exist between the Government of Canada and the government of the second country. The formal agreement between countries may take the form of an interim diplomatic exchange of notes or of a complete negotiated Air Transport Agreement.

International bilateral air negotiations involve officials of the Canadian government from External Affairs, the Canadian Transport Commission, Transport Canada and the Ministry of Industry, Trade and Commerce. Negotiations may last several months or several years.

The routes for scheduled air services are normally the major item for negotiation, but there are many others. Some of the items, or articles written into air transport agreements, may be:

- rights to fly across, or to make stops for non-traffic purposes in, a given territory
- designation of the airline to operate each route
- compliance with laws and regulations of each country, dealing with such issues as entry, clearance, immigration, passports, customs and quarantine
- airworthiness, certificates of competency, and licences
- exchange of statistical information
- tariffs
- transfer of funds
- exemption from taxation of income

In order to negotiate these items, and particularly for an exchange of routes, the negotiating officials must know where air traffic markets are and whether they are growing. Analyses of the costs and benefits to Canada and to foreign countries of various international routes for Canadian and foreign air carriers must be available to the negotiating officials. To ensure that Canada can negotiate a fair market share, the provision of international passenger origin and destination estimates is a necessity. A crude and indirect indication of the value of such data to the Canadian economy is that the revenue[3] generated from all international air routes to and from Canada in 1980 was about 2.3 billion dollars (Canadian).

[3] Taken from tabulations internal to the Canadian Transport Commission.

3. NATURE OF THE NON-COVERAGE PROBLEM

An underlying assumption of the redesign is that the same basic methodology, as implemented in the existing system, is to be used. As a result, most of the feasibility work involved identifying desirable improvements, ranking their desirability and determining how much could be done within cost and time constraints. This "same basic methodology" provided direction with respect to the calculation of estimates of the number of international passengers, but not to the calculation of estimates of reliability of the passenger estimates.

The target population of the international estimation system is the set of all tickets with an international (i.e. between a foreign country and either Canada or the United States) journey. An exchange program on passenger O & D data is maintained between Canada and the United States, whereby the United States gives Canada those records detailing the complete itineraries of the tickets collected in their survey on which:

(i) a U.S. and Canadian point is shown in the routing, or
(ii) a U.S. carrier is recorded as having flown to or from a Canadian point, or
(iii) a Canadian carrier is recorded as having flown to or from a U.S. point.

As a result of this exchange agreement, the expression "foreign" and "foreign (non-U.S.)" are both used in this paper to denote "neither Canadian nor American".

The revenue passenger origin and destination survey collects tickets issued for international journeys, but only major Canadian carriers participate in the survey. Each participating carrier selects a flight coupon on a ticket

with a serial number ending in '0', if that carrier is the first participating carrier to fly on a leg of that ticket. Hence, the survey reports a 10 percent sample of unique flight coupons on which there is at least one participating Canadian or American carrier.

Some information concerning the markets of foreign (non-U.S.) carriers is obtained from the revenue passenger origin and destination survey. For example, if a passenger travels from Ottawa to Montréal with Air Canada, then connects with Air France for Paris, the revenue passenger origin and destination survey will capture the trip because a Canadian carrier participated somewhere in the journey. The Canadian carrier would report the complete carrier and routing detail, including the Air France segment.

The revenue passenger origin and destination survey, however, does not cover coupons with foreign carriers on all legs of an itinerary. An example of such an itinerary would be that of a passenger flying on Air France from Paris to Montréal and then back to Paris on Air France. If this itinerary were the passenger's total journey, this journey would not be reported to the revenue passenger origin and destination survey. However, the itineraries of such passengers are in the target population of international journeys.

This incomplete coverage of the target population is the non-coverage problem for the ASIPOD estimation system. The coverage problem seems to be "non-coverage" as opposed to "undercoverage", since it is not even possible to include a large portion of the universe in the frame.

The existing system takes the revenue passenger origin and destination survey data and the airport activity survey data and applies a method called "the assignment technique" in order to produce total market estimates.

The airport activity survey counts passengers, on a census basis by flight, entering and leaving each Canadian airport. The survey covers all Canadian, American and foreign scheduled carriers, but it does not consider the passengers' initial origin or final destination.

Hence, the airport activity survey provides a count of the total volume of passengers for all carriers in the target population. The assignment technique is a method of estimating the non-coverage volume of passengers and assigning it to origin-destination pairs. However, in order to explain the assignment technique, a somewhat more thorough description of the two air traffic surveys is required.

4. RELEVANT ASPECTS OF THE TWO SURVEYS

The authorizing agency for the two survey programs is the Air Transport Committee of the Canadian Transport Commission, in co-operation with Transport Canada. The data are collected from the air carriers on behalf of the Air Transport Committee by the Aviation Statistics Centre (ASC) of Statistics Canada. Under the authority of the Air Carrier Regulations of the Aeronautics Act, reporting by the carriers on the ASC statements (ie. questionnaires) is compulsory.

The Revenue Passenger Origin and Destination (O & D) statistics are reported to the ASC via Statement 35. The reported data items, among others, include:

- ticket origin and ticket destination

- points of intra- and interlining (i.e. routing)

- carrier on each flight coupon stage

The revenue passenger origin and destination data are submitted monthly by major Canadian unit toll air carriers conducting scheduled passenger services. Since January 1, 1982 the seven Canadian carriers contributing information to this survey have been Air Canada, CP Air, Eastern Provincial Airways, Nordair, Pacific Western Airlines, Air Ontario and Quebecair. The American data are collected by the Civil Aeronautics Board from all certificated United States' air carriers, except helicopter operators and intra-Alaska carriers. The data for the three months of each quarter are combined, and duplicates are

eliminated; so that a file of complete itineraries, Ticket Origin and Destination (TOD) records, is obtained for the quarter.

However, passenger origin and destination statistics are compiled using the Directional Origin and Destination (DOD) concept. The DOD concept can be defined as "points of initial departure and ultimate destination named in the sequence which indicates the direction of travel". DOD's are pieces of itineraries which are broken up such that each component piece defines a reasonably consistent direction. To create DOD's, "open-jaw" and return itineraries, such as "symmetrical" and "circle" itineraries, must be broken into pieces which are essentially one-way trips. To obtain the DOD's, the TOD's are passed through the breakpoint routines. This breakpoint process is automated within the Passenger Origin and Destination System, and involves the calculation of various point-to-point distances within the itinerary and the comparison of these distances to the total itinerary length. As a general rule, itineraries are broken at the farthest point from the origin. Each DOD formed is recycled through breakpoint routines until no further breakpoints can be assigned.

The airport activity data are filed on Statement 32. The relevant items included for each flight are:

- the reporting carrier
- the reporting airport
- the point of origin and final scheduled destination of the flight
- the last station arrived from, for arrivals; or next station departed to, for departures
- the number of deplaned or enplaned revenue passengers

The airport activity data are submitted monthly by Canada's transcontinental

(Air Canada and CP Air) and regional air carriers (Eastern Provincial, Quebecair, Nordair and Pacific Western Airlines), by Norcanair and by all foreign carriers (including American carriers) operating scheduled international flights into and out of Canada. Since January 1, 1982, there have been 10 American and 21 other foreign carriers filing reports for each Canadian airport they served. Each new foreign carrier, granted a licence to serve Canada on a scheduled basis, is automatically included as a participant in the airport activity data collection system.

From the airport activity data, the census traffic flow data are obtained. "Traffic flow" can be defined as a count, over a certain period of time, of the number of persons who are flying on a specific carrier between a Canadian reporting airport and an adjacent point. The adjacent point is called the next stop or the last stop. For the purposes of the assignment technique only the traffic data for foreign (non-U.S.) carriers are input into the system. The data elements extracted from this survey, and used to determine the O & D international markets, are the number of revenue passengers enplaned and deplaned in Canada, the Canadian gateway carrier, and the Canadian gateway. In this survey the concept of "Canadian gateway" is defined to be that reporting airport at which a foreign (non-U.S.) carrier enters or leaves Canada.

However, in the revenue passenger O & D survey, the Canadian gateway for Canadian and U.S. carriers is the first/last Canadian point in the itinerary for a flight entering/leaving Canada. For foreign carriers the Canadian gateway refers to the point inside Canada where the passenger enters or leaves the foreign carrier. Consider the following fictitious example:

Assume that Air France flies Toronto - Montréal - Paris. Some passengers enplaned in Toronto, and some enplaned in Montréal.

Assume also that the matching single crossing DOD's are:

Winnipeg - Air Canada - Toronto - Air France - Montréal - Air France - Paris.

Toronto - Air France - Montréal - Air France - Paris - British Airways - London

The Canadian gateway in the above example would be Toronto because Toronto is the point at which the passengers enter the foreign carrier.

5. THE OBJECTIVES OF THE REVISION

The four main objectives, the fourth of which will be discussed in detail in this paper, are as follows:

(i) to eliminate problems which have been identified in the existing system.

The existing system does not impute for illegible carrier codes on flight coupons, so that "unknown carriers" becomes the third largest carrier in tabulations. Also, there is no check on the coding which indicates whether a carrier is flying to and from airports at which it actually has landing rights. As a result of even existing tabulation requirements, additional edits and imputations, to handle illegible or incorrect data on international flight coupons, are required over and above those required for domestic flight coupons alone.

Other nonsampling errors have been identified, but can not be easily corrected by an estimation system. Examples of such errors are misinterpretation by participating carriers of instructions for selecting flight coupons; systematic errors in serial numbers, used for sample selection, on ticket stock; errors in the carriers' processing systems, etc. The control of these nonsampling errors for which it is not easy to correct is not an objective of this revision.

(ii) to develop a simple computing system which is easy to use, and produces summary diagnostic information.

Some 625,000 origin and destination records, and some 820,000 airport activity records must be processed annually with minimal manual intervention. Since large volumes of passengers are dispersed across a large number of origin-destination pairs, the diagnostics at each stage of the system must summarize the processing, and still be able to point out potential problems.

(iii) to tabulate the international passenger estimates, regularly and on an ad hoc basis, in ways which will simplify the analyses undertaken by users.

(iv) to produce quantifiably reliable estimates of the number of air scheduled international passengers by origin and destination.

Although the reliability of these statistics has been thought to be variable in the past, it has been, in fact, unknown to date. Estimating the reliability of these data will improve the knowledge of the strength of the inferences that can be made from analyses of these data. Inferences made without knowledge of the reliability of the data could actually be quite misleading.

6. SOLUTION OF THE NON-COVERAGE PROBLEM

6.1 Magnitude of the Problem

As in other surveys, non-coverage is a problem for the ASIPOD methodology, since no sample data are available on the origin-destination patterns of the non-coverage portion of the target population.

The following table gives an indication of the volume of passengers travelling between Canada and nine world areas. (These data are 1979 annual estimates. The world areas are not identified because these data are confidential.)

Table 1 - Estimated Number of International Passengers - 1979

Between Canada and ...	Revenue Passenger Origin and Destination	Non-coverage (percentage of Total in brackets)		Total
World Area #1	116,050	495	(0.4)	116,545
World Area #2	325,200	2,020	(0.6)	327,220
World Area #3	99,010	12,628	(11.3)	111,638
World Area #4	67,200	14,558	(17.8)	81,758
World Area #5	575,010	126,076	(18.2)	703,086
World Area #6	205,180	101,464	(33.1)	306,644
World Area #7	1,221,040	908,876	(42.7)	2,129,916
World Area #8	54,410	56,807	(51.1)	111,217
World Area #9	45,330	57,267	(55.8)	102,597
Total World	2,708,430	1,282,191	(32.1)	3,990,621

From the percentage non-coverage figures, it is evident that the non-coverage problem is a major concern.

The same table as above, but between Eastern Canada and the same nine world areas, would tend to have a higher percentage non-coverage for each world area. Hence, a lower level of geographic aggregation in origin-destination pairs generally implies a higher percentage non-coverage. For example, the non-coverage for Eastern Canada to World Area #7 is 55 percent, compared to the 42.7 percent tabulated for all of Canada to World Area #7 (as above). To clarify the idea that a higher level of geographic aggregation in origin-destination pairs generally implies a lower percentage non-coverage, consider the fact that non-coverage exists only for foreign traffic terminating at Canadian gateways. There is complete coverage of all traffic for which the Canadian end of the origin-destination pair is not a Canadian gateway. Hence, as the level of geographic aggregation becomes higher, more and more interlining traffic is included, and the percentage non-coverage becomes lower.

6.2 The Assignment Technique

The assignment technique estimates the non-coverage volume of passengers and then allocates this volume to origin-destination pairs.

From the airport activity survey, b benchmark counts of passengers entering and leaving Canada at Canadian gateway airports are tabulated by carrier. The value of b, the number of assignment groups, can be determined as follows:

$$b = 2 \times \sum_{i=1}^{g} n_i$$

where the '2' accounts for the fact that there is one count each for passengers entering and leaving Canada,

g is the number of Canadian gateway airports, and

n_i is the number of foreign (non-U.S.) carriers with landing rights at the ith Canadian gateway airport.

From the revenue passenger origin and destination survey, a corresponding number of inbound and outbound passengers on international DOD's can be tabulated by crossing carrier and Canadian gateway airport. Hence, there are also b such counts from the sample survey data.

In the first stage of the assignment technique the non-coverage volume, A_i, of passengers in assignment grouping i can be estimated as follows:

$$A_i = C_i - (1/f) \times D_i \quad (i = 1, \ldots b).$$

where C_i is the airport activity census count in assignment grouping i.

f is the revenue passenger O & D survey sampling fraction (i.e. 1/10).

D_i is the sample number of international passengers in assignment grouping i.

A_i is, then, an estimate of the number of passengers, carried on the foreign carrier in the ith of b assignment groupings, for which there is no origin-destination information.

The next stage allocates the non-coverage volume to origin-destination pairs. Such pairs in the non-coverage portion are called non-interlining DOD's, since they are DOD's, flown by foreign carriers, which do not interline with a participating Canadian carrier. The assignment technique imputes non-interlining DOD's in the ith assignment grouping as follows:

(i) All of the DOD's contributing passengers to D_i are identified. (These would be DOD's which match on Canadian gateway, foreign carrier and direction.)

(ii) The domestic portion of such DOD's (i.e. the portion from the Canadian point to the Canadian gateway city) is eliminated. (The domestic portion of such DOD's would be on a Canadian carrier, and, therefore, would be picked up in the revenue passenger O & D survey. The resultant "truncated DOD's" are, then, non-interlining.)

(iii) The non-coverage volume, i.e. A_i, is assigned to the resultant "sample DOD's" in proportion to their original contribution to D_i.

New DOD records consisting of "assigned passengers" are produced.

The assignment technique assumes that the truncated DOD's are representative of the non-interlining traffic. As a result, some original sample DOD's are receiving more weight than they would in the revenue passenger O & D survey alone. Hence, the estimator, $\hat{d}_j^T$, for the total market number of

passengers for the jth origin-destination pair can be derived by adjusting the sampling fraction as follows:

$$\hat{d}_j^{\,T} = (1/f_j) \times d_j$$

$$\text{where } f_j = \frac{d_j}{(d_j/f) + a_j} \quad (1)$$

and where f_j is the adjusted sampling fraction associated with the jth origin-destination pair, from the revenue passenger O & D sample survey.

d_j is the sample number of international passengers in the jth origin-destination pair, from the revenue passenger O & D sample survey.

a_j is the number of passengers assigned to the jth origin-destination pair.

Note that a_j, according to point (iii) above,

is $$a_j = \frac{d_j}{D_i} \times A_i \qquad j \varepsilon i$$

where $$\sum_{j \varepsilon i} d_j = D_i$$

and $$\sum_{j \varepsilon i} a_j = A_i$$

6.3 An Example of the Assignment Technique

A simple example will illustrate how the assignment technique works. Assume that the airport activity data from British Airways indicated that 120 passengers enplaned at Montréal and went to London. Therefore, C_i = 120. Assume that the only two DOD's for this assignment grouping from the revenue passenger origin and destination survey are:

Sample Number of Passengers	Estimate of the Number of Passengers	DOD's
1	10	YWG-AC-YMX-BA-LON-LO-WAW
2	20	YYZ-CP-YMX-BA-LON-LH-HAM

where the codes are to be interpreted as:

Code	Denotes
YWG	Winnipeg
YMX	Montréal (Mirabel)
LON	London
WAW	Warsaw
YYZ	Toronto
HAM	Hamburg
AC	Air Canada
BA	British Airways
LO	LOT
CP	CP Air
LH	Lufthansa

Therefore, $D_i = 3$, and $A_i = 120 - (1/.1) \times 3 = 90$.

The truncated DOD's, the proportion of their contribution to D_i, the resultant number of assigned passengers and total market estimates are then:

Truncated DOD's	Proportions	Assigned Passengers(a_j)
YMX - BA - LON - LO - WAW	1/3	30
YMX - BA - LON - LH - HAM	2/3	60

6.4 Shortcomings of the Assignment Technique

The assignment technique has some recognized shortcomings.

The basic assumption of the assignment technique is that the truncated DOD's are representative of the non-interlining traffic. Consider the hypothetical example above in order to determine whether this assumption is intuitively reasonable. The assignment technique presumes that passengers flying on a particular foreign air carrier and originating in Montréal would have the same ultimate destination as passengers originating in Winnipeg or Toronto who fly through Montréal. Hence, the travel patterns of ethnic communities, for example the Polish and German communities in Toronto and Winnipeg respectively, might be used to impute for travel patterns of the more predominantly French communities in Montréal. And, in fact, the assumption that interlining and non-interlining travel patterns are the same was proven empirically to be suspect in a pilot test (Rosen and Conroy (1977)) conducted by the Canadian Transport Commission in 1977. Therefore, there is not only some intuitive but also some empirical evidence against the basic assumption of the assignment technique.

The accuracy of the estimates of the number of passengers by origin-destination pair is jeopardized by any violations of the assumption that truncated DOD's are representative of non-interlining traffic. Large volumes of passengers are allocated to origin-destination pairs, as was seen in Table 1 above, based on a small "effective" sample size. For example, the "effective" sampling fraction of the Canada - World Area #7 market is, not 10% as in the domestic survey, but 5.7% (ie. (100% - 42.7%) x 10%), because of the non-coverage of non-interlining traffic. Hence, a smaller than 10% sample of DOD's is used to allocate a large volume of passengers. Violations of the aforementioned assumption, then, would cause a potentially large bias in the estimates.

Since airport activity census counts are used as benchmark figures for traffic volumes, their accuracy is very important. Although no evaluations have been

undertaken to investigate the magnitude of bias from nonsampling errors in the airport activity census counts, aviation statistics economists feel that this is a survey in which such errors would be small. There are currently, however, ongoing discussions with the major Canadian air carriers on how the reporting requirements of government agencies can be minimized. As a result of these discussions, the airport activity survey could become a sample survey. If a sample is to be designed, the accuracy of the activity counts of gateway airport passengers would be an important design consideration.

The assignment technique also assumes that all of the non-coverage is accounted for by the non-interlining traffic. An interesting way of validating this assumption would be to compare, for the same reference period, the airport activity census count for a Canadian air carrier to the analogous estimate from the origin-destination sample survey. This analogous estimate would be the sum of all passengers on DOD's with the same Canadian gateway airport and crossing carrier. It would be necessary to be able to determine whether differences in the estimates were ascribable to differences in the concepts of the two surveys, and if so, and whether these differences have been taken into account in the ASIPOD estimation system. If such differences have not been accounted for, then it could be that there is a problem with the assumption that all of the non-coverage is accounted for by non-interlining traffic.

Many of these shortcomings should be investigated. However, there are major problems in the existing system, and no alternative solutions, which are superior to the same basic methodology of the assignment technique and which can be implemented within time and cost constraints, have been found. Furthermore, the improvements in the estimates of the number of passengers by air traffic market in the new ASIPOD system will be substantial.

The use of the assignment technique to estimate for international markets is an innovative solution to a large problem. It does not completely solve the

non-coverage problem, but it is a major step in the right direction, as will be the production of estimates of the variance of the estimates. These variance estimates should take account of the assignment technique and its assumptions, and, at the same time, give a meaningful measure of the reliability of the DOD estimates.

7. ESTIMATION OF VARIANCE

The estimator of the variance of the international origin-destination estimates is a simple extension of the variance estimator for the revenue passenger origin and destination survey.

7.1 Variance of the Estimate of Interlining Traffic

The development of the estimator of the variance of the revenue passenger origin-destination estimates is dependent upon the way in which tickets contribute passengers to origin-destination pairs (i.e. the domains of interest). Recall that each ticket is selected with probability 0.1, and that each ticket may be broken up into several segments or DOD's. Each ticket may contribute 0,1,2, etc. passengers to a given domain of interest. For example, the itinerary

YWG - AC - LON - BA - YYZ

would be broken, via the breakpoint routines, into the two DOD's

YWG - AC - LON
and
LON - BA - YYZ.

Consider the inbound plus outbound estimates which are total passenger figures, independent of direction. For such estimates this ticket would add passengers to, among others, the following domains of interest:

Domains of Interest	Passenger Count
Winnipeg - London	1
Toronto - London	1
Canada - London	2
Canada - Europe	2
Eastern Canada - Europe	1
Western Canada - Europe	1

Note that the number of passengers per ticket depends on the geographic level of aggregation, and, therefore, on the particular origin-destination estimate (i.e. domain of interest).

The estimate, $\hat{d}$, of the number of passengers from the revenue passenger origin and destination survey can be developed for a particular domain of interest as follows:

$$\hat{d} = \sum_{i=1}^{n} \frac{x_i}{f}$$

where x_i is the number of DOD's belonging to the domain of interest on the ith ticket.

n is the number of sample tickets selected, and

N is the population number of tickets in the revenue passenger origin and destination survey.

f is the sampling fraction, i.e. $n/N = .1$

The revenue passenger origin and destination survey sample is effectively a 10% simple random sample because

(i) the selection of coupons with serial numbers ending in the digit '0' produces a systematic sample, and

(ii) there is no cycle associated with the distribution of tickets which would cause a relationship between the survey estimates and the last digit of the serial number.

The estimate of the variance can be written, then, as

$$\hat{var}(\hat{d}) = N^2 \frac{1}{n} (1 - f) v_s$$

$$\text{where } v_s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{and } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i ,$$

and the coefficient of variation, as

$$\hat{cv}(\hat{d}) = \sqrt{\hat{var}(\hat{d})}/\hat{d} .$$

Note that $\hat{var}(\hat{d})$ can also be written as

$$\hat{var}(\hat{d}) = \frac{(1-f)}{f^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (2)$$

where n is assumed to be sufficiently large for n/(n-1) to be approximately equal to 1.

7.2 Variance of Total Market Estimates

The method for calculating total market coefficients of variation has to recognize that the assigned data are a function of the sample data. In other words different samples will produce different assigned data and, thereby, different values of the total market estimate. The re-use of certain portions

of the sample has an effect on the sampling distribution of the estimates.

The method which will be used adjusts the sampling fraction from 10% to be the percentage for which sampled records for a particular domain of interest are actually accounting. Since the use of the assignment technique is a given, it has to be assumed that truncated DOD's from the revenue passenger origin and destination survey are representative of the non-interlining traffic. The measure of reliability will be a measure of the precision of the DOD estimates, only to the extent to which this assumption is valid. As a result, it is reasonable to adjust the weights of the sample DOD's to take into account the non-interlining DOD's. The sampling fraction, then, for the estimation for the jth domain of interest would be f_j as developed in equation (1) above. f_j would replace f in the formula for $\hat{var}(\hat{d})$ in equation (2) above in order to yield the formula for the variance of the total market estimates:

$$\hat{var}(\hat{d}_j^T) = \frac{(1-f_j)}{f_j^2} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

And, the coefficient of variation for the total market estimate would be

$$\hat{cv}(\hat{d}_j^T) = \hat{var}(\hat{d}_j^T) \,/\, \hat{d}_j^T.$$

Note that $f_j < .10$ for $a_j > 0$. This means that less than a 10% sample is achieved when it is necessary to include assigned data in a total market estimate. Hence, by using f_j instead of .10 in the expression for the variance of a DOD estimate, the coefficient of variation is adjusted to relate to the total market estimate.

This method gives credit to the use of sample data in the assignment technique; but it is dependent, as is the determination of the estimates themselves, upon the assumption that truncated DOD's are representative of non-interlining traffic.

8. FUTURE CONSIDERATIONS

Earlier, the need for data on air traffic markets in international bilateral air negotiations was explained. The exchange of statistical information is one of the negotiable articles in air transport agreements. Currently, agreements for the exchange of statistical information exist with several countries. The concepts and quality of the data from some of these countries indicate that these data could be used in the ASIPOD estimation system. Such data would provide sample information on the non-coverage portion of the international universe of tickets. Feasibility work is currently underway to determine whether the number of countries for which these data can be used would improve the accuracy of a sufficient number of estimates to justify the expansion of the ASIPOD system to use "exchange data".

9. ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the Transportation and Communications Division and the Business Survey Methods Division of Statistics Canada, and in particular Robin Dunn, for their many helpful comments during the development of this paper.

REFERENCES

[1] Burchell, J.M.: "International Air Passenger Origin/Destination Project"; presentation to the Transportation and Communications Division of Statistics Canada, March 1981.

[2] "International Air Passenger O & D Statistics System - User Requirements Specifications"; prepared in the Aviation Statistics Centre (ASC) of the Transportation and Communications Division of Statistics Canada, revised March 1981.

[3] "International Air Passenger O & D Statistics System - Feasibility Study Report"; prepared in the ASC, January 1981.

[4] "International Air Passenger O & D Statistics System - Requirements for Functional Specifications (Concepts)"; prepared in the ASC, revised September 1981.

[5] Rosen, F.G. and Conroy, P.F.: "A Test of the Assignment Technique Based on a Survey of International Air Travellers at Montreal and Toronto"; a confidential report prepared for the Passenger and Aviation Economics Directorate of the Research Branch of the Canadian Transport Commission, February 1977.

[6] Carpenter, R.: "Specifications for Estimates of the Reliability for the Air Scheduled International Passenger Origin and Destination Estimates"; prepared in the Business Survey Methods Division of Statistics Canada, revised January 1982.

SOME ASPECTS OF QUALITY OF CANCER MORTALITY AND INCIDENCE STATISTICS

D. Binder and A. Malhotra[1]

Statistics Canada, Canada's central statistical agency, has been compiling national mortality statistics, including those on cancer mortality since 1921. Also, cancer incidence data are available from 1969.

The data quality of these files may be assessed in a variety of ways. Ratios of cancer mortality to incidence give some information on coverage errors. Micro-data matches between incidence and mortality files give an indication of misclassifications. As well, multiple registrations for cancer incidence may be duplicates. Completeness and availability of data items are also important for special studies.

In this paper, the feasibility of using these measures of data quality and the implications of these measures are discussed.

1. INTRODUCTION

Population based cancer statistics are the basis of epidemiological research into the distribution and determinants of cancer and underlie health programmes for the prevention, diagnosis and treatment of cancer. Statistics Canada, Canada's central statistical agency, compiles two such types of data on cancer.

1. National mortality data which are based on reports from provincial vital statistics registration systems. These data date back to 1921.

2. National cancer incidence data which are based on notifications from provincial cancer registries. This data series was established in 1969.

[1] D. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada and A. Malhotra, Health Division, Statistics Canada.

Good statistics which provide reliable information on risk differentials depend on completeness and accuracy of cancer registration and comparability of the data between different registration areas and time periods. Cancer incidence and mortality data as reflections of the true cancer risk each have their own merits and limitations.

Cancer incidence data are particularly suitable for the epidemiological study of cancer because they provide information on all cancers, not only those that are fatal, because they can provide an early warning of emerging problems and because the diagnostic information is usually detailed and of a high quality. For example, the publication Cancer Incidence in Five Continents [1] emphasizes the role that international comparisons of cancer incidence play in yielding clues about the causes of cancer in spite of certain well known limitations of the data. These limitations include difficulty in achieving complete registration of new cancer cases and differences between registries in the extent to which this is achieved. Major factors influencing coverage are the number and types of data sources used, how active case finding is, the length of time a registry has been in operation and whether or not the reporting of cancer is a legal requirement in the registration area. Canadian cancer registries are quite heterogeneous in their data collection methodology but all attempt to follow international [2] as well as national [3] guidelines for the standardized recording of cancer incidence data. Differences in sources and techniques of registration not only influence coverage but also other aspects of data quality such as detail of socio-demographic and geographic information that is provided. Also, cancer incidence data are sensitive to such factors as mass screening programmes which result in the inclusion of previously undiagnosed prevalent cases.

The editors of Cancer Incidence in Five Continents use and discuss a number of indices which may be useful in assessing completeness of registration and reliability of the data [4]. These include cancer mortality-incidence ratios as indicators of completeness of registration (see Sections 2.1 and 3.1 of this paper).

The reporting of deaths is a legal requirement in most developed countries so that coverage error is assumed to be small. Known and suspected limitations

of mortality data for purposes of epidemiologic cancer research include a lack of information on non-fatal cancers, less precise diagnostic information and more frequent misclassification due to assignment and coding of underlying cause of death resulting in less accuracy compared with the diagnostic information in cancer incidence statistics.

A quality assessment of vital statistics [5] which was undertaken as a pilot study gives some indication of the quality of Canadian mortality data (all causes), particularly on error rates in the coding of underlying cause of death. This error rate was 7.2% in the data year 1976. About two-thirds of the errors involved the first or second digit of the 4-digit cause of death code. Variation in the error rate by specific causes of death was not investigated.

This paper is concerned with assessing the feasibility of measuring certain aspects of quality of the two cancer data files at Statistics Canada which are used in epidemiological studies, namely the cancer incidence file and subset of records from the mortality file with an underlying cause of death of cancer.

The aspects of data quality selected for investigation were:

1. Completeness of registration of new cancer cases through a comparison of cancer mortality with cancer incidence in the same period. This comparison is a crude but commonly used indicator of completeness of registration.

2. Consistency of assignment of diagnosis and cause of death codes through matching individual records on the two files.

3. Availability and completeness of data items through an analysis of how often valid values are present on the files.

4. Registration of multiple primary cancers on the incidence file through a matching of records within the file.

The period covered by the study is 1969-1978, the period for which cancer incidence data are available. Ontario was excluded from all investigations because the National Cancer Incidence Reporting system includes data for this province for 1969-1971 only[2].

A discussion of the approach taken in the investigations is contained in Section 2 of this report. In Section 3 the findings are discussed.

2. DESCRIPTION OF MEASURES

In this section we describe some methods for studying the data quality of the cancer mortality and incidence files.

2.1 Mortality-Incidence Ratios

In order to study relative rates of undercoverage among cancer incidence registrations, we consider the ratios of deaths to incidents of cancer. Since the Mortality System registers all deaths in Canada by cause of death and the concept of the National Cancer Incidence Reporting System (NCIR) is to register all new incidents of primary malignant neoplasms[3], if the two registration systems were of the same quality for all reporting registries, one would expect that the ratio of mortality rate to the incidence rate for a particular site would be fairly consistent within a population of given age and sex over sufficiently long periods of time. (We compute these ratios for deaths and incidents over 5-year and 10-year periods to reduce the effect of the time lag between the reporting of a cancer incident and death). Inconsistency of this ratio would arise if any of the following rates differ across reporting registries:

(a) rates of survival or sudden increase rate of incidence

(b) mortality rates from other competing risks,

(c) rates of error in coding of underlying cause of death,

---

[2] Ontario developed a passive registration system which makes use of reports on cancer patients made for other purposes. Data for recent years are currently being prepared by the province.

[3] By exception, metastatic cancers are registered when the primary site is unknown.

(d) rates of error in classifying cancer site for new cancers,

(e) rates of under-reporting or over-reporting of cancer incidents.

If the mortality-incidence ratios are grossly different for most sites, then differing rates in (a) and (b) can be discounted. With respect to the error rates in (c), studies on the coding error for underlying cause of death have yielded error rates of less than 10% [5]. In an unpublished report it was found that these errors rates could vary from 3% to 18% across reporting registries. However, the observed differences in the mortality-incidence ratios (see Tables 1 and 2) cannot be completely explained by these error rates. Also, since about 90% of the registrations of new cancers are confirmed histologically, the error rate in (d) would be small. Therefore, these mortality-incidence ratios do give an indication of the coverage error in the NCIR System.

Concentrating on those sites with leading diagnosis count (excluding skin cancer), based on the NCIR file, for each sex, we show in Tables 1 and 2 the national mortality-incidence ratios as well as provinces with largest and smallest ratios, for two five-year periods, broken down by age-groups. We omit Prince Edward Island from consideration because the number of observed events is too small for valid comparison.

2.2 Matching Mortality and Incidence Records

In order to assess the feasibility of evaluating errors in cause of death classification, or errors in cancer site classification, a sample or records can be selected from either the Mortality File or Cancer Incidence File and then the other file can be searched for matching records. The manual search does not guarantee that all true matches will be found, and, in fact, the rate of successfully matching may be different across reporting registries, because the level of detail of matching variables can vary from one registry to another. (See Section 2.3 for a study on availability of data).

Records with malignant neoplasms of the lung or bronchus (ICDA-8[4] is 162.1) for the years 1969-1978 were selected as starting points from both files. This choice was based on considerations of high incidence, high mortality and short survival times so that conditions were favourable for finding matching records on both files. In spite of this, because of the time difference between diagnosis and death, it is true that cancer deaths in the earlier years and newly diagnosed cancers in the later years are less likely to have a corresponding record on the matched file. The analysis of the results would be improved in future studies if the sample design controlled for year of death and year of diagnosis.

Two independent samples were selected: one from Mortality File and the other from the NCIR file.

2.2.1 Mortality to Incidence

All deaths from cancer between 1969 and 1978 should have, at least conceptually, a corresponding record on the NCIR system. Noteworthy exceptions to this rule are that the cancer was first diagnosed in Ontario or outside Canada or that the cancer was first diagnosed before 1969. Besides the exceptions, a lack of a corresponding record on the NCIR system is an indicationn of under-coverage. This, of course, assumes that the underlying cause of death in the Mortality File is error-free.

Therefore, if a sample of deaths from cancer are selected and matched to the NCIR system, we have a number of possible outcomes:
(a) no matching record is found,
(b) a match was found with a record having a different cancer site,
(c) a match was found with a record having the same cancer site.

If no match is found, this is an indication of under-coverage, or that the cancer was first diagnosed in Ontario or outside Canada or prior to 1969, or that the death was not really a death due to cancer. Alternatively, as

[4] International Classification of Diseases, Adapted for Use in the United States, Eighth Revision.

previously mentioned, the matching process itself is not perfect. If a match is found, but the records have different cancer sites, this may be an indication of an error on either the Mortality File or the NCIR System. As mentioned previously, it is generally believed that the NCIR system yields the more accurate disease classification, because of the high rate of histological confirmation.

A small scale study was undertaken to measure this phenomenon. A random sample of 56 records with underlying cause of death reported as a malignant neoplasm of the lung or bronchus (ICDA-8 is 162.1) was selected from each of the provinces except Ontario yielding a total sample of 504 records. Only deaths between 1969 and 1978 were selected.

The national rate of successful matches was 82.3%. The rates varied between 73.2% and 96.4% across the nine provinces. Among those with successful matches, 92.5% had the same 4-digit ICDA-8 classification. These rates varied from 74.4% to 100.0%. For those provinces with the lowest and highest rates of matches with the same ICDA-8 classification, we give in Table 3 the breakdown of the observed disease classifications.

2.2.2 Incidence to Mortality

A sample of records from the NCIR System was also selected and matched to the Mortality File. Fifty-six records with malignant neoplasms of the lung and bronchus (ICDA-8 is 162.1) from each of the nine reporting registries were randomly selected yielding a total sample of 504 records. The matched records were then checked for underlying cause of death on the complete Mortality File for 1969 to 1978. We did not check the cause of death on the original death certificate for this study, although this would be feasible for future studies.

The outcomes from this manual match may be classified as follows:

(a) no matching record is found,

(b) a match was found with a cause of death other than cancer,

(c) a match was found with a cause of death being cancer but not cancer of the lung or bronchus,

(d) a match was found with the same cause of death.

If no match is found, this is an indication of the inadequacy of the matching process, unless death occurred after 1978 or outside Canada or the person is still alive.

A match found with a different underlying cause of death is an indication of one of the following:

(a) a competing risk took precedence,

(b) cancer was a contributing cause of death but not the underlying cause of death or

(c) the underlying cause of death on the Mortality Data Base was incorrect, or the cancer site was incorrect on the NCIR system (the latter being assumed less likely).

The average rate of successful matches was 69.4%. The rates varied between 55.4% and 80.4% across the nine provinces. Among those with successful matches, 92.0% had the same 4-digit ICDA-8 classification. These rates varied from 85.3% to 100.0%. For those provinces with the lowest and highest rates of matches with the same ICDA-8 classification, we give in Table 4 the breakdown of the observed cause of death classifications.

2.3 Availability and Completeness of the Data

One simple measure of the quality of the data files is the relative frequency of valid data for specific items. For the National Cancer Incidence Reporting System and the cancer deaths on the Mortality File, we concentrate on the following items:

- date of birth (day, month, year)
- age
- place of birth
- county and subdivision of residence

We chose these items to exemplify how easy or difficult it would be to match records from other files (e.g. Section 2.2), or to create special tabulations, such as small area statistics. For each item we classify the data as being

valid or invalid. Besides blank values, invalid data would arise when alphabetic characters are found in a numeric variable or the numeric value is out of range. We have aggregated the relative frequencies into two five-year groupings (1969-1973 and 1974-1978) so that we can see whether the quality has changed significantly in the later years.

In Tables 5 and 6 we report the national averages for the two data bases as well as show the values for the provinces with largest deviation from the national average. For the Mortality File, we give the results only for cancer deaths in the nine provinves outside of Ontario so that the comparison with the NCIR system is more meaningful.

2.4 Multiple Registrations on the Cancer Incidence System

The concept of the National Cancer Incidence Reporting System is to register all new incidents of malignant neoplasms. An individual should be registered more than once when multiple malignant neoplasms develop. To avoid duplicate registration of the same incident or duplicate reporting of patients registered in more than one province, all provincial cancer registries follow routine procedures. In spite of this, duplicate reporting of the same cancer incident may occur. In order to evaluate the extent of the duplication, we searched for records which are likely duplicates. The search was nowhere near exhaustive, so that the number of potential duplicates found is an underestimate. Of the 457,158 records, we removed the records with invalid surnames or years of diagnosis. For those with missing birthyear, we calculated the birthyear from the age when available. We also removed skin cancer records (ICDA-8 is 173) since this is known to have multiple occurrences.
Of the remaining records, we found those cases where all the following occurred:

- birthyear or calculated birthyear matched exactly,
- surname matched exactly,
- first four letters of first given name agreed,
- the three digit ICDA-8 code agreed.

Of these records, we identified multiple registration as follows:

- year, month and day of birth was present and agreed, or
- day of birth was not present on at least one record but month of birth agreed.

We also manually verified all groups with at least 3 individuals where the month and day information did not agree and all groups of 2 individuals where the month or day information was missing on at least one record.

In all, this resulted in identifying 6113 records which where potential duplicates. These records correspond to 5947 individuals. (Note that some individuals were duplicated more than once.) We did not make the judgment as to whether these were legitimate multiple registrations or actual duplicates.

For each 3-digit ICDA-8 value, we show in Table 7 the breakdown of these potential duplicates according to whether the records came from the same reporting registry or different registries, as well as how many potential duplicates have the same fourth digit of the ICDA-8 classification.

3. DISCUSSION

3.1 Mortality - Incidence Ratios (Tables 1 and 2)

Ratios of cancer mortality to incidence can provide an indication of completeness of registration. The ratios will vary with cancer site (the highest ratios occur for sites with the lowest survival), age and sex for all registries. However, if a comparison of the ratios for different registries shows major differences within a given site, sex and age group, differences in completeness of registration of new cases of cancer must be suspected. A higher ratio, which means a higher proportion of deaths compared with newcases in the same period may indicate less complete registration of new cases.

In both time periods there were two registries which consistently had the highest ratios for all sites combined and for most of the major sites shown. There is little doubt that these high ratios do reflect underregistration of new cases - the registries are the only ones which do not use death

notifications as one of their sources of registration. In addition, one of these registries uses only a single data source, hospital reports, to register cancer cases. This registry had previously reported the results of a special study which showed that it was receiving notifications for only an estimated 70% of new cancer patients admitted to hospitals up to the end of 1976. Following this, major changes were made to improve the notification system. Since 1977 the registry has been reporting a higher number of cancer cases which is reflected in a marked reduction in mortality-incidence ratios.

All other Canadian cancer registries use multiple sources of registration which is considered essential to achieve good coverage and which could also have a positive impact on the completeness and quality of individual data items. The completeness of reporting of data items was examined in this study (see Subsection 2.3). However, it turns out that the one registry that uses only one data source actually ranks quite highly in terms of completeness of information for many data items.

A possible drawback associated with using multiple sources of registration is that duplicate registration may result. However an analysis of multiple registrations for the same individual and the same cancer site does not bear this out. In general, registries using a larger number of different sources of registration do not have more multiple registrations for the same site than registries using fewer sources of registration.

Cancer mortality-incidence ratios for the other six registries were more similar to each other. For these registries there was no consistent pattern of one registry always having higher or lower ratios for all sites and both time periods.

There are many factors that can influence variations in the observed ratios by cancer site. Factors which tend to result in less complete registration of new cases and therefore higher mortality-incidence ratios includes difficulty in diagnosing the cancer (e.g. in deep-seated organs) and lack of access to specific data sources (e.g. haematology reports confirming a diagnosis of leukaemia).

Factors which may lead to overregistration of new cases and lower ratios are mass screening programmes (which may lead to the inclusion of prevalent cases, especially, for slow-growing tumours), duplicate registration, inclusion of in-situ cases and inclusion of latent cancers discovered only at autopsy (this particularly affects cancer of the prostrate). In addition, differences in the accuracy of assignment of diagnosis or cause of death may lead to artefactual differences. For example, death certificates may state "cancer of the uterus, unspecified" or "leukaemia, unspecified" as the cause of death whereas a cancer registry will often have more precise information and will assign more precise codes [6]. An analysis of mortality-incidence ratios at the level of the more detailed diagnosis would therefore show gross discrepancies.

Of the leading cancer sites that were examined for males, cancer of the lung and stomach were associated with the highest mortality-incidence ratios for all registries but interprovincial variation in the ratios was greatest for cancer of the colon (excluding rectum) prostate and bladder. Use of cancer incidence data in studies designed to identify differences in cancer risk by geographic area (province) would therefore be more reliable for the former two cancer sites.

For females, of the leading sites examined, cancer of the colon and ovary had the highest ratios for all registries. Interprovincial variation in the ratios was greatest for cancer of the uterus and cervix uteri as well as for cancer of the colon. For purposes of interprovincial comparisons, incidence data for breast cancer and cancer of the ovary would therefore be more reliable.

In the case of the sites of cancer of the uterus (other than cervix) and cancer of the cervix, there are large interprovincial variations in the ratios if the sites are considered separately. This variation is greatly reduced if the two sites are combined, suggesting that there are differences in the accuracy of diagnosis and cause of death assignment for these sites.

The site-specific mortality-incidence ratios were examined for major age groups. The highest ratios consistently occur at older ages (65 and over) for

all registries and all sites shown. This is as expected, since the risk of death increases with age so that proportionately more death than new cases occur at older ages. It is also recognized that diagnosis and registration of cancers in older persons is generally more difficult. However, the relative increase in the ratios at older ages is much greater for the registries which have the highest average ratios to start with. This indicates that while all registries may have some difficulty in registering older patients, under-coverage of the older population is greater for registries which in general have less complete registration systems.

The Canadian data therefore lend support to recommendations made by the International Union against Cancer (1970) and the International Agency for Research on Cancer (1976) and the reiteration of this recommendation in a recent paper by Doll and Peto [7] that "reasonably reliable comparisons of cancer incidence are obtained only if comparisons are limited to men and women in middle life".

3.2 Matching Mortality and Incidence Records(Tables 3 and 4)

Since Statistics Canada is responsible for managing both the cancer incidence and the mortality data files, it is possible to compare reports for individuals who are listed in the two separate data files to verify the reported information.

In this part of the investigation of data quality, the accuracy of assignment of diagnosis and cause of death was of particular interest. Within the scope of the study it is only possible to describe the results - the reasons underlying the discrepancies found remain unknown. However, it is felt that the findings are revealing and do indicate, in the case of the particular cancer selected for analysis, lung cancer, that agreement on diagnosis between the two files is generally high, over 90%.

The study also indicates that a larger scale match would be feasible to assess the accuracy of diagnosis and cause of death codes. Of course, if a larger scale study were based on a sample, it would be preferable to stratify the sample by year of diagnosis or year of death. In theory, if computerized

matching techniques were used, this type of analysis is possible for all cancer sites. If such an undertaking were to be supplemented with, for example, studies on accuracy of coding of cause of death and diagnosis in the field, such as described in two U.S. reports [8] [9], interpretation of epidemiological research findings would be facilitated.

3.2.1 Mortality to Incidence File Search

Of the sample of 504 death records with an underlying cause of death of lung cancer from 1969-1978, 415 (82%) corresponding records on the cancer incidence file for the years of diagnosis 1969-1978 were found[5]. The rate of unsuccessful matches varied from 3.6% to 26.8% across the nine provinces. This rate is influenced by four main factors: (a) that the cancer was first diagnosed prior to 1969, (b) that it reflects underregistration of new cases, (c) that identifying information was not adequate to permit matching of records, or (d) that the cause of death code is incorrectly given as cancer. There were insufficient data to allow assessment of the relative contribution of each of these factors.

Of the 415 death records for which a corresponding record was found on the incidence file, there was agreement on the diagnosis, primary cancer of the lung, in 92.5% of cases. There was 95.2% agreement that a cancer of the respiratory system was present. The small number of remaining records, had diagnoses for sites other than respiratory cancer on the incidence file. It is generally accepted that the diagnostic information on cancer registry files is more accurate than the cause of death information on death certificates. However, given the scope of this study it is not possible to determine if misclassification on either of the files (or perhaps the fact that a lung cancer was first diagnosed prior to 1969 followed by a subsequent registration for another primary cancer) account for the disagreement. Across the provinces, the rate of agreement on diagnosis varied from 74.4% to 100%.

---

5 In six cases more than one corresponding record for the same individual existed on the incidence file. Only one of these records was counted as a successful match.

Interestingly, for the province with total agreement on diagnosis there was also the highest success rate (96.4%) of locating corresponding records on the two files. This could possibly indicate close liaison between the provincial vital statistics office and the cancer registry. In the reverse match of a sample of cancer incidence records to mortality records (described in Section 3.2.2) it was the same province that had the highest rate of successful matches as well as complete agreement on diagnosis.

3.2.2 Incidence to Mortality File Search

The reverse search using the incidence file for the years 1969-1978 as a starting point and attempting to locate a corresponding record on the complete mortality file for the same period was successful for 69.4% of the selected sample of 504 records with a diagnosis of primary lung cancer. The rate of unsuccessful matches was higher than in the match from mortality to incidence records for all provinces and varied from 19.6% to 44.6%. Possible reasons for not finding a match include (a) that the patient was still alive at year end 1978, or (b) that identifying information was not adequate to permit a match. It is in general less likely that one will find a corresponding record in the search from incidence to mortality file since some persons diagnosed to have lung cancer do survive this whereas all persons who die from lung cancer should be registered as new cases either prior to death or at time of death.

For the records that were successfully matched there was agreement that the diagnosis was a primary lung cancer in 91.4% of cases, a rate very similar to that found in the reverse comparison. The samples for the two comparisons were chosen independently so the consistency of the findings concerning agreement on diagnosis is reassuring. Of the remaining cases, 4.3% had the underlying cause of death classified to cancer sites other than the respiratory system, and 3.7% had an underlying cause of death which was not cancer. For this latter group it is possible that cancer was mentioned on the death certificate as a contributing cause of death. This analysis is possible but was n-not carrier out.

3.2.3 Availability and Completeness of the Data (Tables 5 and 6)

One measure of quality and usefulness of the data files is the frequency of valid information for specific data items. This measure is crude because "valid" as defined here means valid according to computerized edit checks and does not preclude that imputation of missing information or errors in definition or classification of the data item render the information invalid.

Subject to the above caveats the measure may be useful in showing if and where there are improvements in reporting of data items over the years, and whether or not particular analyses of the data are feasible. For example, information on date of birth is important for purely statistical (age-specific) analysis of the data as well as for medical follow-up analyses which depend on good identifying information.

On the cancer incidence file, a complete birthdate (i.e., day, month and year) is on average present on only 68% of the records in the period 1969-1978. If the two time periods, 1969-1973 and 1974-1978 are considered separately some improvement in the more recent period becomes evident.

On cancer mortality records for the same time period a complete birthdate is present in over 95% of cases. However, at least part of this high rate is due to the fact that the mortality system imputes a date of birth from age and date of death when the birthdate is not reported. In 1976 the imputation rate was 11.5% [5]. No such imputations are carried out in the cancer incidence system[6].

Small area analyses of cancer occurrence require complete and detailed residence information. Cancer mortality data are much more useful for these purposes because census division (county) of residence codes are present on 99.8% of records and census subdivision (city, town, village) codes are present on 96.2% of records. In contrast, on the cancer incidence file, census division codes are present on 89.6% of records and census subdivision

---

6 Imputations may be useful for statistical purposes but are actually mis-leading in medical follow-up studies unless it is made clear that the information is based on an imputation.

codes on only 25.2% of records. On the incidence file, there is improvement in the reporting of census subdivision information in the second time period.

3.2.4 Multiple Registrations on the Cancer Incidence System (Table 7)

Comparability of cancer incidence data is affected if there are differences in the reporting of multiple primary cancers in the same individual and in inadvertent duplicate registration.

The rules for reporting of multiple primary cancers are difficult to interpret, so some provincial differences in their application are expected.

Inadvertent duplicate registration of the same cancer incident may occur if a provincial registry cannot determine if the same case has been registered previously (perhaps because identifying information is inadequate) or if the same incident is reported by two different registries[7]. The search for multiple primary cancers was restricted to multiple entries for the same cancer site (at the 3-digit level of the ICDA code).

No attempt was made to separate duplicate registrations from multiple primaries, although it can be speculated that the majority of cases reported by two seperate registries may be duplicates whereas the cases reported by the same registry are more likely to be valid multiple primaries, particularly those that differ in the 4th digit of the diagnosis code.

Using very strict matching criteria and excluding skin cancers (other than melanoma of the skin), 1.7% (6113) of records on the 1969-1978 cancer incidence file were identified as multiple entries.

By province, this rate varied from 0.5% to 1.9%. Only 0.4% of multiple records were reported by two different registries. There was agreement down to the 4th digit level of the diagnosis code in 88.4% of cases.

---

7 The national Cancer incidence System does not carry out routine checks on such duplication.

By cancer site, if only sites with more than fifty records are considered, the rate of multiple primaries varied from 1.0% for cancer of the stomach and pancreas to 3.6% for breast cancer. The high rate for breast cancer is not surprising since the current rules for reporting of multiple primary cancers require separate reports for cancers in both sides of (most) bilateral organs.

On the whole it is felt that while there is some inconsistency arising from multiple primary and duplicate reporting, this is very small compared with that arising from undercoverage.

4. SUMMARY

The techniques described in this paper have been successful at identifying differing levels of quality of cancer incidence and mortality data. It has been found that the mortality-incidence ratios, in particular, can be used to assess coverage errors, which are one of the major concerns of a high quality cancer incidence system. The data quality for those who are registered on the incidence system is sufficiently high that it is possible to assess the quality of the cause of death classification on the mortality system through a micro-data match. In fact a computerized micro-data match could be used to evaluate the undercoverage because the number of cancer deaths without previous registration on the NCIR system could be ascertained.

ACKNOWLEDGEMENTS

The authors are grateful to J. Gorman, D. Lawrence, K. McClean, S. Moore and P. Walsh for their assistance in preparing the data for this paper. We are also grateful to J. Silins for his comments. The referee's comments are also very much appreciated.

REFERENCES

[1] Waterhouse, J., Muir, C., et al., Editors, CANCER INCIDENCE IN FIVE CONTINENTS, VOLUME III, WHO, International Agency for Research on Cancer, Lyon, 1976, p.3.

[2] WHO HANDBOOK FOR STANDARDIZED CANCER REGISTRIES, World Health Organization, Geneva, 1976.

[3] MANUAL FOR CANCER RECORDS OFFICERS, National Cancer Institute of Canada.

[4] Op. cit., in reference 1 , pp. 45-51

[5] Nagnur, D.N., Currie, S.G., Heath, B., QUALITY ASSESSMENT OF VITAL STATISTICS (A pilot study), Statistics Canada, Health Division, Ottawa, 1981.

[6] King, H.S., Wigle, D.T., Hill, G.B., Silins, J., MORTALITY TRENDS FOR CANCERS OF THE CORPUS UTERI AND CERVIX UTERI, Alberta 1969-1978, CMAJ.

[7] Doll, R., Peto, R., THE CAUSES OF CANCER: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today, JNCI, Vol. 66, No. 6, June 1981.

[8] Percy, C., COMPARISON OF THE CODING OF DEATH CERTIFICATES RELATED TO CANCER IN SEVEN COUNTRIES, Public Health Reports, Vol. 93, No. 4, July-August, 1978.

[9] Feigl, P., Breslow, N.E., Laszlo, J., Priore, R.L., Taylor, W.F., U.S. CENTRALIZED CANCER PATIENT DATA SYSTEM FOR UNIFORM COMMUNICATION AMONG CANCER CENTRES, JNCI, Vol. 67, No. 5, November, 1981, p. 1019.

Table 1

Cancer Mortality - Incidence Ratios

Deaths in Period (Mortality) as Percentage of New Cases Registered (Incidence)
Canada (excluding Ontario) and Provinces with the Highest and Lowest Ratios.

Males	Leading Sites by Age Group and Sex						
Cancer Site	Age	1969 - 1973			1974 - 1978		
		Canada	Highest Ratio	Lowest Ratio	Canada	Highest Ratio	Lowest Ratio
Lung (162)	**Total**	**96**	**110**	**83**	**95**	**101**	**80**
	0-24	27	37*	-	70	-	100*
	25-44	89	96	83	85	75	82
	45-64	94	106	81	88	86	78
	65+	100	118	83	101	118	82
Prostate (185)	**Total**	**44**	**55**	**35**	**39**	**54**	**33**
	0-24	71*	133	-	71*	-	100*
	25-44	39	54	-	22	-	-
	45-64	27	32	33	24	30	20
	65+	48	62	35	43	60	36
Colon (153)	**Total**	**76**	**101**	**54**	**67**	**86**	**56**
	0-24	56	60	-	40	-	-
	25-44	61	76	40	53	87	45
	45-64	66	86	41	58	61	56
	65+	84	115	61	73	100	58
Bladder (188)	**Total**	**35**	**42**	**26**	**28**	**35**	**23**
	0-24	20	25*	33*	3	-	-
	25-44	7	9	10	6	-	-
	45-64	23	30	16	18	22	13
	65+	44	53	32	36	45	29
Stomach (151)	**Total**	**104**	**124**	**82**	**90**	**112**	**80**
	0-24	29*	-	-	86*	-	-
	25-44	82	102	54	74	40	77
	45-64	93	102	79	80	89	80
	65+	112	144	86	96	130	80
All Cancers (140-209) excluding skin (173)	**Total**	**69**	**83**	**56**	**64**	**77**	**56**
	0-24	58	64	45	48	49	56
	25-44	51	62	39	45	55	37
	45-64	65	78	53	60	65	50
	65+	74	94	59	69	90	60

- Either mortality, incidence or both are zero.
* Ratios based on fewer than 10 cases for both mortality and incidence.

Table 2

Cancer Mortality - Incidence Ratios

Deaths in Period (Mortality) as Percentage of New Cases Registered (Incidence)
Canada (excluding Ontario) and Provinces with the Highest and Lowest Ratios.

Females	Leading Sites by Age Group and Sex						
Cancer Site	Age	1969 - 1973			1974 - 1978		
		Canada	Highest Ratio	Lowest Ratio	Canada	Highest Ratio	Lowest Ratio
Breast (174)	**Total**	**40**	**47**	**31**	**37**	**46**	**31**
	0-24	14	13	-	6	50*	-
	25-44	28	33	26	23	30	17
	45-64	37	44	31	35	43	29
	65+	50	62	35	46	55	38
Colon (153)	**Total**	**73**	**94**	**50**	**67**	**80**	**50**
	0-24	15	20	-	27	33*	-
	25-44	48	49	41	41	64	34
	45-64	63	77	43	55	57	42
	65+	83	114	55	74	95	55
Uterus (182)	**Total**	**27**	**38**	**17**	**20**	**26**	**15**
	0-24	33	40*	20*	-	-	-
	25-44	13	19	7	11	15	9
	45-64	16	25	9	12	15	10
	65+	49	64	33	35	43	26
Cervix Uteri (180)	**Total**	**39**	**45**	**29**	**32**	**48**	**27**
	0-24	12	20*	17*	5	-	4
	25-44	21	17	12	15	25	10
	45-64	39	54	32	34	49	33
	65+	66	64	51	53	65	50
Ovary (183)	**Total**	**69**	**79**	**54**	**69**	**78**	**63**
	0-24	20	33*	-	28	100*	-
	25-44	43	41	26	30	38	20
	45-64	68	74	52	64	72	55
	65+	87	105	77	95	98	106
All Cancers (140-209) excluding skin (173	**Total**	**57**	**67**	**46**	**53**	**64**	**45**
	0-24	46	49	43	37	44	28
	25-44	33	39	27	27	35	22
	45-64	48	56	38	44	51	39
	65+	75	94	58	69	85	58

- Either mortality, incidence or both are zero.
* Ratios based on fewer than 10 cases for both mortality and incidence.

Table 3

Match of Mortality with Incidence Records

Disease Classification on the Cancer Incidence File for a Sample of Lung Cancer Deaths (Percentage Distribution)

Canada (excluding Ontario) and
Provinces with the Highest and Lowest Rates of Lung Cancer Incidents among the Matches.

ICDA-8	CLASSIFICATION ON THE INCIDENCE FILE	CANADA	HIGHEST	LOWEST
	1. CANCER OF THE RESPIRATORY SYSTEM	95.2	100.0	83.7
162.1	(a) LUNG; Primary	92.5	100.0	74.4
160-163	(b) OTHER RESPIRATORY SYSTEM; Primary	1.0		4.7
197.0-197.3	(c) RESPIRATORY SYSTEM; Secondary	1.7		4.7
	2. OTHER CANCERS	4.8		16.3
174	(a) BREAST	0.7		2.3
200-209	(b) LYMPHATIC PHEMATOPOIETIC SYSTEM Primary	0.7		0.0
196	Secondary	1.0		4.6
	(c) OTHER SPECIFIED PRIMARY SITE	1.9		9.3
195,199	(d) ILL DEFINED OR UNDEFINED SITE	0.5		0.0
	TOTAL	**100.0**	**100.0**	**100.0**

SAMPLE SIZE FOR MATCHES	415	54	43
MATCH SUCCESS RATE (%)	82.3	96.4	76.8

Table 4

Match of Incidence with Mortality Records

Cause of Death Classification for a Sample of Lung Cancer Cases from the Cancer Incidence File (Percentage Distribution)

Canada (excluding Ontario) and
Provinces with the Highest and Lowest Rates of Lung Cancer Deaths among the Matches.

ICDA-8	CLASSIFICATION ON THE MORTALITY FILE	CANADA	HIGHEST	LOWEST
	1. CANCER OF THE RESPIRATORY SYSTEM	92.0	100.0	85.3
162.1	(a) LUNG; Primary	91.4	100.0	82.9
160-163	(b) OTHER RESPIRATORY SYSTEM; Primary	0.3		0.0
197.0-197.3	(c) RESPIRATORY SYSTEM; Secondary	0.3		2.4
	2. OTHER CANCERS	4.3		9.8
174	(a) BREAST	0.3		0.0
200-209	(b) LYMPHATIC PHEMATOPOIETIC SYSTEM Primary	0.3		0.0
196	Secondary	-		0.0
	(c) OTHER SPECIFIED PRIMARY SITE	3.1		9.8
195,199	(d) ILL DEFINED OR UNDEFINED SITE	0.6		0.0
	3. NOT CANCER	3.7		4.9
	TOTAL	**100.0**	100.0	**100.0**

SAMPLE SIZE FOR MATCHES	350	45	41
MATCH SUCCESS RATE (%)	69.4	80.4	73.2

Table 5

Cancer Incidence
Availability and Completeness of Data Items

Canada (excluding Ontario) and
Provinces with the Highest and Lowest Percentages of Data Completeness

DATA ITEM	YEAR OF DIAGNOSIS	CANADA	HIGHEST PERCENT	LOWEST PERCENT
DATE OF BIRTH				
Day	1969 - 1973	63.7	99.8	0.0
	1974 - 1978	71.6	99.8	4.0
	1969 - 1978	68.0	99.3	2.2
Month	1969 - 1973	65.6	98.8	0.0
	1974 - 1978	73.3	99.8	4.1
	1969 - 1978	69.8	99.4	2.2
Year	1969 - 1973	92.9	100.0	15.9
	1974 - 1978	96.3	100.0	21.9
	1969 - 1978	94.7	100.0	19.1
Complete Birthdate	1969 - 1973	63.6	98.7	0.0
	1974 - 1978	71.6	99.8	4.0
	1969 - 1978	68.0	99.3	2.2
AGE	1969 - 1973	99.4	100.0	98.6
	1974 - 1978	100.0	100.0	100.0
	1969 - 1978	99.7	100.0	99.4
BIRTHPLACE (Country or Province)	1969 - 1973	15.2	19.8	0.0
	1974 - 1978	24.4	71.1	0.0
	1969 - 1978	20.2	46.0	0.0
RESIDENCE Census Division	1969 - 1973	89.6	100.0	4.1
	1974 - 1978	89.6	100.0	82.8
	1969 - 1978	89.6	100.0	49.0
Census Subdivision	1969 - 1973	16.6	43.2	0.0
	1974 - 1978	32.3	76.4	0.0
	1969 - 1978	25.2	61.6	0.0

Table 6

Cancer Mortality
Availability and Completeness of Data Items

Canada (excluding Ontario) and
Provinces with the Highest and Lowest Percentages of Data Completeness

DATA ITEM	YEAR OF DEATH	CANADA	HIGHEST PERCENT	LOWEST PERCENT
DATE OF BIRTH				
Day	1969 - 1973	96.6	99.4	0.0
	1974 - 1978	97.8	99.7	43.9
	1969 - 1978	97.2	99.6	23.0
Month	1969 - 1973	97.0	99.7	0.0
	1974 - 1978	98.1	99.9	44.2
	1969 - 1978	97.6	99.8	23.2
Year	1969 - 1973	100.0	100.0	99.2
	1974 - 1978	99.9	100.0	99.6
	1969 - 1978	99.9	100.0	99.4
Complete Birthdate	1969 - 1973	96.6	99.4	0.0
	1974 - 1978	97.8	99.7	43.9
	1969 - 1978	97.2	99.6	23.0
AGE	1969 - 1973	100.0	100.0	100.0
	1974 - 1978	99.9	100.0	99.6
	1969 - 1978	100.0	100.0	99.8
BIRTH PLACE (Country or Province)	1969 - 1973	98.3	100.0	98.9
	1974 - 1978	52.2	99.9	0.0
	1969 - 1978	73.9	99.9	46.2
RESIDENCE Census Division	1969 - 1973	99.8	100.0	99.6
	1974 - 1978	99.7	100.0	99.8
	1969 - 1978	99.8	100.0	99.7
Census Subdivision	1969 - 1973	92.2	99.9	51.9
	1974 - 1978	99.7	99.5	99.4
	1969 - 1978	96.2	99.7	77.1

Table 7

Cancer Incidence
1969 - 1978
Multiple Primaries Within Each Site
(Canada excluding Ontario)

ICDA Cancer Site	Multiple Primaries			
	Number	Percent of Incidence	Percent Same Registry	Percent With Same 4am Digit ICDA Code
Total All Sites (except skin, 173)	**6,113**	**1.7**	**76.8**	**88.4**
140 Lip	87	1.4	90.8	83.9
141 Tongue	30	1.6	63.3	63.3
142 Salivary Gland	8	0.6	12.5	75.0
143 Gum	8	1.5	100.0	62.5
144 Floor of Mouth	18	1.8	94.4	N.A.
145 Mouth, Other and Unspecified	16	1.4	75.0	62.5
146 Oropharynx	14	1.0	92.9	78.6
147 Nasopharynx	8	1.1	62.5	N.A.
148 Hypopharynx	3	0.5	100.0	33.3
149 Pharynx, Unspecified	3	1.3	66.7	N.A.
150 Oesophagus	37	1.1	75.7	N.A.
151 Stomach	178	1.0	71.9	74.2
152 Small Intestine	12	1.2	91.7	91.7
153 Lge. Intestine Excl. Rectum	623	1.9	82.3	44.0
154 Rectum	256	1.4	77.0	74.6
155 Liver	11	0.6	63.6	90.9
156 Gall Bladder	22	0.7	77.3	77.3
157 Pancreas	96	1.0	70.8	54.2
158 Peritoneum	6	0.7	100.0	100.0
159 Unspec. Digestive Organs	1	0.3	100.0	N.A.
160 Nose, Etc.	5	0.7	80.0	80.0
161 Larynx	94	2.0	77.7	57.4
162 Trachea, Bronchus, Lung	795	1.9	75.0	99.5
163 Resp. Organs, Other & NOS	7	0.7	85.7	85.7
170 Bone	28	2.0	78.6	82.1
171 Connective Tissue	32	1.3	81.3	75.0
172 Melanama of Skin	61	1.4	78.7	54.1
174 Breast	1,791	3.6	84.4	N.A.

Table 7 (concl'd)

Cancer Incidence
1969 - 1978
Multiple Primaries Within Each Site
(Canada excluding Ontario)

ICDA Cancer Site	Multiple Primaries			
	Number	Percent of Incidence	Percent Same Registry	Percent With Same 4απ Digit ICDA Code
180 Cervix Uteri	133	1.4	60.9	N.A.
181 Chorionepithelioma	1	1.0	0.0	N.A.
182 Other, of Uterus	149	1.1	73.8	91.3
183 Ovary, Etc.	126	1.5	80.2	97.6
184 F. Genital Organs, Other	31	1.5	71.0	83.9
185 Prostate	464	1.6	69.2	N.A.
186 Testis	27	1.5	44.4	N.A.
187 M. Genital Organs, Other	9	1.4	88.9	100.0
188 Bladder	274	1.6	77.7	N.A.
189 Urinary Org., Other & NOS	119	1.5	72.3	79.8
190 Eye	13	1.1	46.2	N.A.
191 Brain	93	1.6	46.2	N.A.
192 Other Nervous System	6	0.4	33.3	66.7
193 Thyroid Gland	49	1.5	55.1	N.A.
194 Other Endocrine Glands	5	0.6	80.0	100.0
195 Ill - Defined Sites	4	0.6	100.0	100.0
196 Sec. & Unspec. Lymph Nodes	5	0.3	100.0	100.0
197 Sec., Resp. & Digestive	9	0.3	88.9	77.8
198 Other Secondary	1	0.1	100.0	100.0
199 Without Spec. of Site	12	0.3	75.0	83.3
200 Lymphosarcoma, Etc.	68	1.1	60.3	97.1
201 Hodgkin's Disease	85	2.2	55.3	N.A.
202 Other of Lymphoid Tissue	22	0.8	90.9	72.7
203 Multiple Myeloma	45	1.3	64.4	N.A.
204 Lymphatic Leukaemia	64	1.5	62.5	84.4
205 Myeloid Leukaemia	32	1.0	53.1	90.6
206 Monocytic Leukaemia	4	1.0	75.0	100.0
207 Other & Unspec. Leukaemia	11	0.7	81.8	90.9
208 Polycythemia Vera	1	0.2	100.0	N.A.
209 Myelofibrosis	1	0.4	100.0	N.A.

ESTIMATING MONTHLY GROSS FLOWS IN LABOUR FORCE PARTICIPATION[1]

Stephen E. Fienberg and Elizabeth A. Stasny[2]

The Canadian Labour Force Survey is a household survey conducted each month for the purpose of producing point-in-time estimates of the number of persons employed, unemployed and not in the labor force. The survey has a rotating panel design in which all individuals in a sampled household location are interviewed each month, for six consecutive months. In the past, little use has been made of this longitudinal structure, although considerable interest has been expressed in the month-to-month gross flows (transitions) amongst the labour force status categories. In this paper we discuss methods being considered by Statistics Canada for the production of gross flow estimates, but from a model-based perspective.

1. INTRODUCTION

The Canadian Labour Force Survey is a monthly household survey used to produce cross-sectional or point-in-time estimates of labour force participation. This survey, however, like the Current Population Survey in the United States and many other large scale sample surveys, is designed using a panel structure so that the subjects are interviewed a number of times before being dropped from the sample. Although the survey is used mainly to obtain cross-sectional estimates, it has long been recognized that information from the repeated interviewing of subjects provides an additional longitudinal data base that could be exploited to give estimates of change over time for a very small additional cost (see, for example Kalachek, 1979, and Fienberg and Tanur, 1983).

[1] This research was supported in part by a contract with Statistics Canada. The authors wish to thank Murray Lawes, Larry Swain and Richard Veevers for their help in understanding the Labour Force Survey Methodology and the resulting data, as well as the Editor and a referee for helpful comments and suggestions.

[2] Stephen E. Fienberg and Elizabeth A. Stasny, Carnegie-Mellon University, Pittsburgh, Pennsylvania, PA. 15213.

Some attempts have been made to use the longitudinal data obtained from panel surveys. For example, the longitudinal data from the Current Population Survey has been used since 1948 to produce tables showing gross movements of individuals between labor force categories from one month to the next. Although these tables are produced each month, they have not been published since 1952 because of statistical problems. Smith and Vanski (1979) discuss the production of gross change data using the longitudinal structure of the Current Population Survey.

Recently, Statistics Canada has initiated an investigation of possible uses of the longitudinal data available as a by-product from the Canadian Labour Force Survey. They, too, would like to find a method for producing reliable estimates of gross movements between labour force categories. In this paper, we discuss the methods being considered by Statistics Canada for the production of such gross change data.

In Section 2, we give a brief description of the coverage and design of the Labour Force Survey, and we describe the structure of the resulting data. Then in Section 3 we outline the proposed method for gross flow estimation developed by Statistics Canada, which depends on the used of sample-based weights, adjustment for inflows to and outflows from the population of interest, consistency adjustments, and bias correction for misclassification error. By developing some simple models for the gross flow process, we explore in Section 4 the implications of Statistics Canada's proposed method. Finally, in Section 5 we describe some work on handling non-response in gross flow estimation.

2. DESCRIPTION OF THE LABOUR FORCE SURVEY

2.1 Survey Coverage

Approximately 56,000 households, chosen from the ten provinces of Canada, are included in the Labour Force Survey sample each month. Questionnaires are

completed for all civilian, non-institutionalized members of sampled households who are 15 years of age and older. The survey questions primarily relate to the subjects' work related activities during the reference week which is the week prior to the survey week and usually contains the fifteenth day of the month. Responses to the survey questions are used to classify subjects as employed, unemployed, or not in the labour force. For a discussion on classification of labour force status, see Guide to Labour Force Survey Data, Statistics Canada, (1979).

2.2 Survey Design

The labour Force Survey was designed to enable estimation of levels and rates of employment and unemployment for each of the ten provinces separately. Thus, except for the constraint on the total sample size, each province is sampled independently.

Economic regions (ER's), areas of similar economic structure, form the primary strata within provinces. ER's are divided into self-representing units (SRU's) and non-self-representing units (NSRU's). SRU's are large urban centers and NSRU's are generally composed of a small urban center and a rural area. Sampling is carried out separately in SRU's and NSRU's.

SRU's are sampled using a stratified, two-stage sampling design. NSRU's are sampled using a stratified multi-staged sampling scheme. In addition to the SR and NSR areas, some sample units are selected from an apartment frame and a special area frame. The final sampling units for the Labour Force Survey are households. A detailed description of the sampling plan for the survey can be found in Methodology of the Canadian Labour Force Survey 1976, Statistics Canada (1977).

Households selected for the Labour Force Survey are included in the survey for six consecutive months and are then dropped from the sample. For example, households rotated into the survey in January are interviewed for six consecutive months, and then dropped from the sample after the June interview. Each

group of households that rotated into and out of the sample together makes up a panel. In any one month, individuals from six different panels are included in the Labour Force Survey sample.

2.3 Sampled-Based Weights

The cross-sectional data, information for a given month from all subjects in the six panels interviewed in that month, is used to produce monthly estimates of labour force participation. The monthly estimates are weighted averages of values for each person in the sample. A weighted average is used because each sampled person is thought of as "representing" a number of people in the population of interest. The weight assigned to an individual's record corresponds to the number of persons in the population that the person in the sample represents.

Let $W_{t,i}$ be the weight assigned to surveyed individual i in the month t. If individual i is classified as outside the population of interest in month t, then $W_{t,i} = 0$. Otherwise, the assigned weights are determined by the probabity of selecting the cluster, the probability of selecting the household within the cluster, nonresponse within the month, rural/urban factors, sub-sampling adjustment for fast-growing areas, and ratio adjustments for province/age/sex factors.

An individual's assigned weight can change from month to month because of replacement of 1/6 of the sample each month, non-response, and, to a lesser extent, because of changes in the size of the population of interest. Thus, for any one individual, i, it may be the case that $W_{t-1,i} = W_{t,i}$.

2.4 Longitudinal Structure and Gross Flow Estimation

Although the main prupose for the Labour Force Survey is to produce point-in-time estimates of labour force participation, the panel structure of the survey design results in a longitudinal data base, with approximately 5/6 of the

household locations sampled in any one month being in the sample for the following month. Naturally, fewer than 5/6 of the surveyed individuals or families are the same in two consecutive months due to non-response and moving. However, Statistics Canada is interested in the possibility of using information from those individuals who do respond in two consecutive months to produce estimates of gross flows among labour force categories.

Estimates of gross flows are useful for answering questions such as a) How much of the increase in unemployment is due to persons losing jobs and how much is due to persons formerly not in the labour force starting to look for jobs? or b) How many unemployed persons become discouraged and leave the labour force?

We discuss the problem of estimating gross flows among labour force categories in the next two sections.

3. STATISTICS CANADA'S PROPOSED METHOD FOR GROSS FLOW ESTIMATION

In this section we describe the multi-stage estimation procedure for gross flows developed by Statistics Canada (e.g. see Macredie and Veevers, 1977; Wong, 1983). Our description of the procedure includes various interpretations of the impact of individual stages.

3.1 Data Needed to Estimate Gross Flows

Statistics Canada has proposed estimating gross flows using a 4×4 matrix as shown below:

Labour Force Status in Month t

		E	U	N	O
Labour Force Status in Month t-1	E	X_{EE}	X_{EU}	X_{EN}	X_{EO}
	U	X_{UE}	X_{UU}	X_{UN}	X_{UO}
	N	X_{NE}	X_{NU}	X_{NN}	X_{NO}
	O	X_{OE}	X_{OU}	X_{ON}	X_{OO}

where E = employed
U = unemployed
N = not in the labour force
O = outside the population of interest, and
x_{ij} = estimated number with labour force status i in month t-1 and status j in month t.

The final monthly Labour Force Survey files from two consecutive months can be used to obtain the data for estimating the 4×4 matrix of gross flows. In order to use these data to produce gross flow estimates, Statistics Canada must match individual records from the two consecutive monthly files using the unique identification numbers assigned to sampled individuals for the duration of their time in the study.

An individual appearing on the data file for one month may be missing from the file for the other month due to rotation into or out of the sample or because the person moved, was not at home, or refused to respond. The sample weights described in Section 2.4 include an adjustment for non-response within each month. When dealing with gross flows, we also need to consider the month-to-month non-response. Statistics Canada proposes to reweight records for individuals who responded in both months t-1 and t to compensate for this additional non-response.

After the reweighting is completed, Statistics Canada will have a single data file that includes information for all persons who were respondents in two consecutive months. That file will contain geographic and demographic information for each individual as well as the individual's labour force status and assigned weights for both month t-1 and month t.

3.2 Differences in Weights

As we noted in Section 2.3, an individual's sample-based weight can change from month to month because of the rotation replacement structure, non-response, and changes in the size of the population of interest. Even when the adjustment factor for non-response is computed on the basis of month-to-month data, for any individual i, it may still be the case that $W_{t-1,i} \neq W_{t,i}$. Some method is needed for handling this difference in weights if data are to be used for estimating gross flows.

Statistics Canada proposes to resolve this dilemma by assuming that differences in the two weights occur only as a result of inflows to and outflows from the population of interest. Thus, differences in weights are added to the appropriate cell in either the last row or last column of the gross flow matrix. This procedure depends heavily on the interpretation of the weights suggested in Section 2.3, namely that sample individual i represents $W_{t,i}$ persons in the population in month t.

As an illustration of the procedure, suppose an individual is classified as employed in both months t-1 and t but $W_{t-1,i} = 300$ and $W_{t,i} = 305$. The minimum weight, 300, is added to the EE cell of the gross flow table. The difference, $W_{t,i} - W_{t-1,i}$, of 5 is added to the OE cell since those 5 would be thought of as having been outside the population of interest during the month t-1 and then having moved in the population as employed persons for month t.

If, on the other hand, the individual is employed in both months but $W_{t-1,i}$ = 305 and $W_{t,i}$ = 300, then 300 is again added to the EE cell but the excess weight of 5 is added to the EO cell. Here, the difference between weights represents 5 persons who were employed in month t-1 and then moved outside the population of interest in month t.

An individual who is classified as outside the population of interest in month t-1 and is then, say, employed in month t will have $W_{t-1,i}$ = 0. If $W_{t,i}$ =300, then 300 is added to the OE cell. Individuals classified as outside the population of interest in month t are treated similarly with $W_{t-1,i}$, being added to the appropriate cell in the last column of the gross flow matrix.

Because persons outside the population of interest are assigned a weight of zero, a person who is classified as such in both months t-1 and t would have $W_{t-1,i}$ = 0 and $W_{t,i}$ = 0. Therefore, X_{OO}, the entry in the OO cell of the gross flow matrix, must always be zero.

3.3 Adjustment of the Inflow and Outflow Cells

Adding differences in weights to the inflow and outflow cells of the gross flow matrix provides a method for handling the changes in sample-based weights from one month to the next and gives estimates of inflows to and outflows from the population of interest. Independent estimates of inflows and outflows, available from Census data, suggest that this method overestimates the actual amount of movement into and out of the population of interest. Thus, Statistics Canada plans to adjust the X_{OE}, X_{OU}, X_{ON}, X_{EO}, X_{UO}, and X_{NO} entries in the gross flow matrix. These cells will be proportionally adjusted so that total inflows and outflows shown in the gross flow matrix equal the Census estimates of inflows and outflows respectively.

Let I be the independent census estimate of inflows to the population of interest and F be the census estimate of outflows from the population. Call the

sum of estimated inflows $X_{0+} = X_{0E} + X_{0U} + X_{0N}$ and the sum of estimated outflows $X_{+0} = X_{E0} + X_{U0} + X_{N0}$. The proportionally adjusted inflows are:

$$Y_{0j} = X_{0j}I/X_{0+} \qquad \text{for } j = E,U,N. \qquad (1)$$

The proportionally adjusted outflows are

$$Y_{j0} = X_{j0}F/X_{+0} \qquad \text{for } j = E,U,N. \qquad (2)$$

3.4 Consistency of Gross Flow Estimates With Monthly Totals

Statistics Canada would like their gross flow estimates to be consistent with the published monthly estimates of labour force participation totals. Thus, the row totals for the gross flow matrix must be the month t-1 estimates of labour force participation and the column totals must be the month t cross-sectional estimates. The marginal totals of the gross flow matrix constructed as described above are not consistent with the monthly labour force totals.

Statistics Canada plans to use the method of iterative proportional scaling, originally proposed by Deming and Stephan (1940), and described in detail by Bishop, Fienberg, and Holland (1975), to adjust the gross flow matrix to agree with the monthly labour force totals. When used to adjust the gross flow matrix, iterative proportional scaling alternatively 1) constrains the rows of the matrix to sum to the month t-1 estimates and then 2) constrains the columns to sum to the month t estimates. Steps 1) and 2) are repeated until the entries in the matrix do not change from one step to the next.

Testing at Statistics Canada has shown that cell changes resulting from the application of iterative proportional scaling were both absolutely and relatively small and fell roughly within the bounds of sampling variability associated with the cells. This suggests that the consistency adjustment does not seriously distort the gross flow estimates.

3.5 Bias Correction for Misclassification Error

Statistics Canada proposed method for estimating gross flows also includes a step correcting for misclassification bias. This is the bias that results from the incorrect assignment of an individual's labour force status. A technique developed by Fred Wong (1983) at Statistics Canada uses reinterview data to correct for the misclassification bias.

4. IMPLICATIONS OF STATISTICS CANADA'S PROPOSED METHOD

4.1 Modeling Gross Flows

Each step of Statistics Canada's proposed method described in the previous section is a logical attempt to correct problems that arise concerning the production of good estimates of gross flows. It is not clear, however, what effect the various adjustments have on the final estimated gross flow matrix. In order to better understand Statistics Canada's proposal to treat differences in weights as being due to inflows to and outflows from the population of interest, in this section we develop a model for the gross flow process. Our discussion centers on the quantities in the inflow and outflow cells of the estimated gross flow matrix since the problems with Statistics Canada's proposed method seem to occur in those cells. Because the design of the Canadian Labour Force Survey is quite complex, we begin with a set of simplifying assumptions. In the following we assume that

1. a single stage stratified sample is chosen,
2. there is no response error, and
3. non-response occurs only because random individuals move between strata or because of rotation into or out of the sample.

4.2 Allocation of Net Population Changes to Inflow and Outflow Cells

Suppose that the population of interest is divided into S strata indexed by $s = 1,2, \ldots S$. Let

N_k^s = population size in strata s in month k.

Each month, a simple random sample is chosen from each stratum for the survey and sampled individuals are interviewed for six consecutive months before being dropped from the survey. Our goal is to estimate gross flows from month t-1 to month t.

For the purpose of estimating gross flows, only individuals who are interviewed in both month t-1 and t will be used. This excludes individuals who rotate into or out of the sample and persons who move between strata. Let

$r_{t-1,t}^s$ = number of sampled individuals from stratum s who were interviewed in both months t-1 and t.

Each of the $r_{t-1,t}^s$ respondents in stratum s is assigned the following weights in months t-1 and t respectively for the purpose of gross flow estimation:

$$W_{t-1}^s = N_{t-1}^s \,/\, r_{t-1,t}^s \text{ and } W_t^s = N_t^s \,/\, r_{t-1,t}^s. \qquad (3)$$

As long as movements between strata and selection of panels are "random processes," these weights represent the inverse of the probability that an individual within a stratum is interviewed in both months t-1 and t. Since all individuals within a stratum have the same weight in any given month, aggregates for each stratum may be used. Therefore, we let

n_{ij}^s = number of sampled individuals from stratum s classified as having labour force status i in month t-1 and status j in month t for i,j = E,U,N,O.

The methodology proposed by Statistics Canada requires that the minimum of the months t-1 and t weights for each individual be added to the appropriate cell in the gross flow matrix. The difference is added to the appropriate inflow

cell if the month t weight is greater than the weight in month t-1 and to the appropriate outflow cell otherwise. Thus, for example, the stratum s entry in the EE (employed to employed) cell of the gross flow matrix is:

$$\min(W^s_{t-1},W^s_t)n^s_{EE} = \min\left[(N^s_{t-1} / r^s_{t-1,t}),\ (N^s_t / r^s_{t-1,t})\right] n^s_{EE}$$

$$= \min\ (N^s_{t-1},\ N^s_t)\ n^s_{EE} / r^s_{t-1,t}$$

$$= \min\ (N^s_{t-1},\ N^s_t)\ f^s_{EE} \qquad (4)$$

where f^s_{EE} = fraction of all individuals from stratum s, interviewed in both months t-1 and t, who where employed in both months.

The contribution from stratum s to the OE cell for the matrix from individuals employed in month t is:

$$\max(0,W^s_t - W^s_{t-1})n^s_{EE} = \max\left[0,(N^s_t - N^s_{t-1} / r^s_{t-1,t})\right] n^s_{EE}$$

$$= \max(0,N^s_t - N^s_{t-1})\ n^s_{EE} / r^s_{t-1,t}. \qquad (5)$$

Differences from individuals falling in the UE and NE cells will also contribute to the OE cell. Thus, the total contribution to the OE cell from stratum s is:

$$\max(0,N^s_t-N^s_{t-1})\ \{(n^s_{EE} / r^s_{t-1,t}) + (n^s_{UE} / r^s_{t-1,t}) + (n^s_{NE} / r^s_{t-1,t})\}$$

$$= \max(0,N^s_t - N^s_{t-1})\ n^s_{+E} / r^s_{t-1,t}$$

$$= \max(0,N^s_t - N^s_{t-1})\ f^s_{+E} \qquad (6)$$

where f^s_{+E} = fraction of all individuals from stratum s, interviewed in both months t-1 and t, who were employed in month t.

We obtain total for all cells in the gross flow matrix in a similar manner. The resulting gross flow matrix is as follows:

Gross Flow Matrix - Month t-1 to Month t

Month t-1 \ Month t	E	U	N	O	
E	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{EE}$	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{EU}$	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{EN}$	$\sum_{s=1}^{S} \max(0, N^s_{t-1} - N^s_t) f^s_{E+}$	$\sum_{s=1}^{S} N^s_{t-1} f^s_{E+}$
U	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{UE}$	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{UU}$	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{UN}$	$\sum_{s=1}^{S} \max(0, N^s_{t-1} - N^s_t) f^s_{U+}$	$\sum_{s=1}^{S} N^s_{t-1} f^s_{U+}$
N	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{NE}$	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{NU}$	$\sum_{s=1}^{S} \min(N^s_{t-1}, N^s_t) f^s_{NO}$	$\sum_{s=1}^{S} \max(0, N^s_{t-1} - N^s_t) f^s_{N+}$	$\sum_{s=1}^{S} N^s_{t-1} f^s_{N+}$
O	$\sum_{s=1}^{S} \max(0, N^s_t - N^s_{t-1}) f^s_{+E}$	$\sum_{s=1}^{S} \max(0, N^s_t - N^s_{t-1}) f^s_{+U}$	$\sum_{s=1}^{S} \max(0, N^s_t - N^s_{t-1}) f^s_{+N}$	0	
	$\sum_{s=1}^{S} N^s_t f^s_{+E}$	$\sum_{s=1}^{S} N^s_t f^s_{+U}$	$\sum_{s=1}^{S} N^s_t f^s_{+N}$		

Notice that each term in the summations for the nine in-population cells (the cells showing gross flows between employed, unemployed, and not in the labour force) is the product of the net size of the strata and the observed fraction of subjects who had the various labour force classifications in months t-1 and t. The out-of-population to employed cell of the gross flow matrix contains a sum of terms from each strata that grew from month t-1 to month t. Each term is the product of the net increase in size for the strata and the fraction of the subjects in the strata who reported being employed in month t.The out-of-population to unemployed and not in the population of interest cells

contain the sums of similar terms except that the net increase in size for each strata is multiplied by the fraction of subjects in the strata who were unemployed or not in the labour force in month t respectively. In other words, the net increase in size for each strata is proportionally allocated to the three inflow cells of the gross flow matrix based on the observed fractions of employed, unemployed, and not in the labour force in month t. Similarly, the net decrease in size for each strata that shrank between months t-1 and t is proportionally allocated to the outflow cells of the matrix based on the observed fractions of employed, unemployed, and not in the labour force in month t-1.

In this model we have assumed that the only way for counts to appear in the inflow and outflow cells is as a result of differences in weights. In practice, a small number of individuals who move in and out of the population-of-interest show up in the sample and their assigned weights are added to the appropriate inflow and outflow cells. The effect of such individuals on the estimates is very small.

The fractions f^s_{+E}, f^s_{+U}, f^s_{+N}, f^s_{E+}, f^s_{U+}, and f^s_{N+} are estimated using individuals who appear in the sample in both months. Almost all individuals classified, for example, as OE could not be respondents in both months because they are not sampled by design or because they are movers. Thus, these people who could not have been respondents in both months are represented by people who did respond in both months. To the extent that these groups differ, the proportional allocation of net increases and decreases in strata size may result in biased estimates in the inflow and outflow of the gross flow matrix.

4.3 Effects of Movements Between Strata

The weights used for the purpose of gross flow estimation, as shown in expression (3), are determined by the number of respondents in both months t-1 and t, a quantity that remains constant for the two months, and by the stratum population. The population of a stratum changes if a) individuals enter from

outside the population of interest, such as when persons reach their 15th birthday or leave the full-time military, b) individuals move outside the population of interest, as when persons enter the military or an institution, or c) individuals in the population of interest move between strata. This subsection describes the effects of such changes in population size on the quantities in the gross flow matrix.

As in the preceding subsection, we suppose that the population of interest is divided into S strata. Again, individuals are sampled at random from each stratum every month, interviewed for six consecutive months, and then dropped from the sample. Let $r^s_{t-1,t}$ be, as before the number of individuals from stratum s who are interviewed in both months t-1 and t.

Next we suppose that there are N^s_{t-1} individuals in stratum s in month t-1. Let movements into and out of strata between months t-1 and t be denoted by

$m_{u,v}$ = number of individuals who move from u to v, $u \neq v$, between interviews for months t-1 and t where u and v may take on the values.

s = stratum s for s = 1,2, ... S and

0 = outside the population of interest.

Using this notation, the population in stratum s in month t is

$$N^s_t = N^s_{t-1} + \Sigma_{u \neq s} (m_{u,s} - m_{s,u}). \qquad (7)$$

The weights assigned to individuals in stratum s in months t-1 and t respectively are

$$W^s_{t-1} = N^s_{t-1} / r^s_{t-1,t} \text{ and } W^s_t = N^s_{t-1} + \Sigma_{u \neq s} (m_{u,s} - m_{s,u}) / r^s_{t-1,t}. \qquad (8)$$

Since our focus in this section is on movement into and out of the population of interest, it is not necessary for us to divide those in the population of interest into employed, unemployed, and not in the labour force. Thus, the gross flow matrix used here is a 2×2 matrix formed by collapsing the first three rows and columns of the 4×4 gross flow matrix used in the preceding subsection.

The entry for stratum s in the in-population to in-population cell is

$$\begin{aligned}\min(W^s_{t-1},W^s_t)r^s_{t-1,t} &= \min\{N^s_{t-1}/\ r^s_{t-1,t},\ [N^s_{t-1}+\ \Sigma_{u\neq s}(m_{u,s}-m_{s,u})]/r^s_{t-1,t}\}r^s_{t-1,t}\\ &= \min[N^s_{t-1},\ N^s_{t-1}+\ \Sigma_{u\neq s}\ (m_{u,s}-m_{s,u})]\\ &= N^s_{t-1}+\min[0,\ \Sigma_{u\neq s}\ (m_{u,s}-m_{s,u})]. \qquad (9)\end{aligned}$$

The entry for stratum s in the out-of-population to in-population, or inflow, cell is

$$\begin{aligned}\max(0,W^s_t-W^s_{t-1})r^s_{t-1,t} &= \max[0,\Sigma_{u\neq s}\ (m_{u,s}-m_{s,u})/r^s_{t-1,t}]\ r^s_{t-1,t}\\ &= \max[0,\Sigma_{u\neq s}\ (m_{u,s}-m_{s,u})]. \qquad (10)\end{aligned}$$

The entry for the in-population to out-of-population, or outflow, cell is found similarly. Thus, the 2×2 gross flow matrix is as follows:

		Month t: In-Population	Month t: Out-of-Population	
Month t-1	In population	$\sum_{s=1}^{S} \{N_{t-1}^{s} + \min[0, \Sigma_{u\neq s}(m_{u,s} - m_{s,u})]\}$	$\sum_{s=1}^{S} \max[0, \Sigma_{u\neq s}(m_{u,s} - m_{s,u})]$	$\sum_{s=1}^{S} N_{t-1}^{s}$
	Out-of-population	$\sum_{s=1}^{S} \max[0, \Sigma_{u\neq s}(m_{u,s} - m_{s,u})]$	0	

$$\sum_{s=1}^{S} \{N_{t-1}^{s} + \Sigma_{u\neq s}(m_{u,s} - m_{s,u})\}$$

Let us consider the quantity in the inflow cell of this gross flow matrix. This cell should contain the net increase in population from outside the population of interest, $m_{0,s} - m_{s,0}$, for each stratum that gained members from outside the population. What the cell does contain is $\Sigma_{u\neq s}(m_{u,s} - m_{s,u})$ for each stratum s that grew as a result of movements between strata and from outside the population of interest. The summation, $\Sigma_{u\neq s}(m_{u,s} - m_{s,u})$, does include the quantity $m_{0,s} - m_{s,0}$ but it may also contain other terms.

For example, suppose the population is made up of three strata called A, B, and C. If strata A and B grew from month t-1 to month t and stratum C lost members, then the inflow cell contains

$$\Sigma_{u\neq A}(m_{u,A} - m_{A,u}) + \Sigma_{u\neq B}(m_{u,B} - m_{B,u}) = m_{0,A} - m_{A,0} + m_{B,A} - m_{A,B} + m_{C,A} - m_{A,C}$$

$$+ m_{0,B} - m_{B,0} + m_{A,B} - m_{B,A} + m_{C,B} - m_{B,C}$$

$$= m_{0,A} - m_{A,0} + m_{0,B} - m_{B,0} + m_{C,A} - m_{A,C} + m_{C,B} - m_{B,C} \qquad (11)$$

Note that the movements between strata A and B cancel out but the terms showing the movement between strata A and C and strata B and C remain in the summation.

In general, the inflow cell contain extra terms of the form $m_{v,u}-m_{u,v}$ for each stratum v that loses population while stratum u gains population. Similarly, the outflow cell contains extra terms of the form $m_{x,y}-m_{y,x}$ for each stratum y that grows while stratum x loses population.

In the inflow cell, the quantity $\Sigma_{u\neq s}(m_{u,s}-m_{s,u})$ for each strata s that gains population from month t-1 to t will be positive, although each individual term in the summation need not be positive. If

$$\Sigma_{u\neq s}\ (m_{u,s} - m_{s,u}) > m_{0,s} - m_{s,0} \tag{12}$$

then the contribution to stratum s is more than the inflow to stratum s from outside the population of interest. This excess comes from terms of the form $m_{u,v} - m_{v,u}$ as described above. That is, the overestimate is due to movements between strata within the population. A similar result holds for the in-population to out-of-population cell of the matrix.

Statistics Canada staff report that the method they proposed for handling differences in weights from month to month does appear to give overestimates in the inflow and outflow cells of the gross flow matrix. Although they are based on simplifying assumptions, the results here give a possible explanation for the overestimation, i.e. the overestimation may be due to movements within the population of interest.

Finally we note that, in the 2×2 gross flow matrix shown above, the in-population to in-population cell must contain an underestimate equal to the overestimate in the outflow cell. Whatever the amount of underestimation, it is spread over the nine in-population to in-population cells in the 4×4 gross flow matrix. Moreover, the size of the overestimation is small in comparison

to the total size of the nine in-population cells.

4.4 Comments on the Proposed Gross Flow Estimation Method

The results described in the preceding two subsections illustrate problems with the proposed method of handling month-to-month differences in weights for the purpose of gross flow estimation. These results do not come as a surprise to Statistics Canada. Because of their experience with Labour Force Survey methods and data, they realized that the movements in individuals within the population might explain some of the overestimation in the inflow and outlfow cells of the gross flow matrix. The results obtained by modelling the process reinforce their beliefs and make it clear just how the movements of individuals effects the estimates. In addition, the modelling brought to light a problem about which Statistics Canada had not been aware: the compensating underestimation spread over the nine in-population to in-population cells of the gross flow matrix.

In section 4.2, we saw that the net increases in strata are allocated to the inflow cells while the net decreases are allocated to the outflow cells according to the fractions of observed individuals classified as employed, unemployed, and not in the labour force in month t and month t-1, respectively. The estimation of inflows and outflows in this manner is valid only if individuals who move in and out of the population of interest are a random sample of individuals and, hence, "the same" as individuals who remain within the population of interest. Sampled individuals who are classified as outside the population of interest appear in the sample by accident rather than by design; the Labour Force Survey is not designed to estimate numbers of persons outside the population of interest. If we want to obtain reasonable estimates for the inflow and outflow cells of the matrix, it may be necessary to include individuals outside the population of interest in the Labour Force Survey sample or to use a special, supplementary sample.

In section 4.3, we saw that the overestimates in the inflow and outflow cells

could be a result of movements of individuals between a strata whose population grew and a strata whose population shrank. The fact that it was movements between strata that caused the problem is a result of the simplifying assumptions which we made. We assumed that the final sample was randomly chosen from within each strata. Hence, the weights assigned to individuals sampled from a single strata were equal. If, instead, we assumed that the strata had been divided into clusters and random samples of individuals had been chosen from within the clusters, then all individuals sampled from a single cluster would have been assigned the same weight and the overestimate would come as a result of movements between clusters.

To correct for the overestimate, and corresponding underestimate, directly in the case where final samples are chosen at random from within strata, we would need estimates of the number of movers between each pair of strata where one strata grew and the other strata lost population. If the final samples are chosen at random from within clusters, similar estimates would be required for each pair of clusters. This is a considerable amount of information. A further complication is that, in practice, the ratio adjustments applied to the weights make it possible for individuals within one household to have unequal weights.

As was suggested earlier, if individuals outside the population of interest were included in the sample, we could obtain estimates of movement into and out of the population of interest directly. One other possibility that should be considered is to discard the monthly weights for the purpose of gross flow estimation and derive a longitudinal weight for each individual in the Labour Force Survey sample in either of the two months.

As statisticians, we are quite comfortable with estimates of gross flows that do not have the published monthly labour force participation totals as marginal totals; however, we realize the problems that might arise if gross flow estimates, that were not consistent with the monthly totals, were published. Nevertheless, it should not be assumed automatically that the monthly estimates are correct and that the problem lies solely in the gross flow estimates.

As we noted in section 3.5, the gross flow matrix is adjusted to correct for misclassification errors. The monthly estimates, however, are not corrected for misclassification bias. Thus, when iterative proportional fitting is used to adjust the gross flow matrix to agree with the monthly totals, the matrix is being altered to be consistent with biased values. We feel that it would be more appropriate to address the problem of misclassification of labour force status in the monthly data where it occurs rather than just in the estimates of gross flows.

5. NON RESPONSE AND GROW FLOW ESTIMATION

Statistics Canada's proposed method for gross flow estimation compensates for non-response by adjusting the sample-based weights of respondents. This method of handling non-response is appropriate if the missing data are missing at random (e.g., see Rubin, 1976). In order to explore the assumption of random non-response, we used a longitudinal file for a single panel to produce the data in Table 1. This table shows the unweighted percentages of individuals reporting being employed or unemployed in zero to six months according to the number of months in which they responded to the survey.

Consider the probabilities underlying the observed percentages shown in part (a) of Table 1. Let

π_i = probability that an individual is employed in i out of 6 months for i = 0,1,... 6.

Under the assumption that non-response occurs at random, the probabilities corresponding to the first column of that table can be written as

P(observing 0 months employed out of 6-k months responding)

$$= \sum_{j=1}^{k} \pi_j \Big/ \binom{6}{j}, \quad \text{for } k = 0,1,\ldots 5. \tag{13}$$

Notice that these probabilities increase from the first row of the column to the last row.

In a similar manner, it can be shown that, if data are missing at random, then the underlying probabilities must increase from the top to bottom of each column in both tables. The first column of each table deviates from this pattern quite noticeably. In both cases, the observed percentages decrease through the first four rows of the table and then increase in the last two rows. It does not seem likely that sampling variability alone could be responsible for such a pattern in both tables. Thus, it appears as if there is some evidence that non-response does not occur at random.

Of course, the above analysis is based on just a single panel of Labour Force Survey data. However, in a larger study using data from 1980 and 1981, Paul and Lawes (1982) also found evidence of a relationship between employment status and non-response. Therefore, there is a need to consider methods for gross flow estimation that do not require the assumption that non-response occurs at random.

Statistics Canada's proposed method for gross flow estimation only makes use of the information from individuals who responded in both of the months. There is also information available from those individuals who responded in just one of the two months. Stasny (1983) presents a method for month-to-month gross flow estimation that makes use of the information available from individuals who are respondents in only one of the two months and that can be used when non-response is related to time or employment status. For this method, we take the observed gross flow data to be the end result of a two-stage process. In the first stage of the process, which we do not get to observe, individuals are allocated to the sixteen cells of the gross flow matrix according to a single multinomial sampling scheme. Then, in the second stage, each individual may lose either the month t-1 or month t labour force classification with some probability. The probability of losing a month's classification can be modeled to depend on the month, or labour force status, or both. Maximum likelihood estimates for the parameters of the multinomial

distribution of the first stage and the probabilities of losing a month's classification are obtained using iterative methods.

When these models were fit to Labour Force Survey data from a single panel, Stasny (1983) found that the model where the probability of losing a month's classification depends on labour force status provides a reasonable fit to the data for all gross flow matrices with the exception of the months 1-2 matrix. For the data from month 1 to month 2, the probability of losing a month's classification appears to depend on the month. This may be due to the fact that there is higher non-response in the first month a panel is in the survey. We believe that it would be worthwhile to fit this type of model to additional data from the Labour Force Survey to see if similar results are obtained over other panels.

Clearly, the problem of obtaining good estimates of gross flows from the Labour Force Survey is not a simple one. The survey is designed to give data for the production of monthly estimates of labour force participation, not estimates of gross flows. A survey designed specifically for the purpose of estimating gross flows among labour force categories would certainly be different from the current Labour Force Survey. Thus, the longitudinal data from the survey is not ideal for gross flow estimation. The data, however, are available and, if they can be used to give reasonable estimates of gross flows, then additional, useful information is produced for a relatively small cost.

REFERENCES

[1] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. Discrete Multivariate Analysis: Theory and Practice. The MIT Press, 1975.

[2] Deming, W.E. and Stephan, F.F. On A Least Squares Adjustment Of a Sampled Frequency Table When the Expected Marginal Totals Are Known. Annals of Mathematical Statistics, 1940, 11, 427-444.

[3] Fienberg, S.E. amd Tanur, J.M. The Design and Analysis of Longitudinal Surveys: Controversies and Issues of Cost and Continuity. Technical Report 289, Department of Statistics, Carnegie-Mellon University, May 1983.

[4] Kalachek, E. Longitudinal Surveys and Labor Market Analysis. Data Collection, Processing and Presentations: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 160-189.

[5] Macredie, I.D. and Veevers, R. Discussion on Smith, R.E. and Vanski, J.E. Gross Change Data: The Neglected Data Base. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 153-158.

[6] Paul, E.C. and Lawes, M. Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey. Survey Methodology, 1982, 8, 48-85.

[7] Rubin, D.B. Inference and Missing Data. Biometrika, 1976, 63, 581-592.

[8] Smith, R.E. and Vanski, J.E. Gross Change Data: The Neglected Data Base. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 132-150.

[9] Stasny, E.A. Estimating Gross Flows in Labor Force Participation Using Information From Individuals With Incomplete Classifications. Technical Report 272, Department of Statistics, Carnegie-Mellon University, Jan. 1983.

[10] Statistics Canada, Guide to Labour Force Survey Data. Catalogue 71-528 Occasional, July 1979.

[11] Statistics Canada, Methodology of the Canadian Labour Force Survey 1976. Catalogue 71-526 Occasional, Oct. 1977.

[12] Wong, F.A. Technique To Correct The Response Bias In the 4×4 Labour Force Gross Flow Matrix, Technical Reprot, Statistics Canada, 1983.

Table 1

a)

Months of Data Present	Months Employed: 0	1	2	3	4	5	6
6	52.23	3.22	1.93	2.21	2.45	3.55	34.41
5	51.48	3.04	2.17	3.36	4.22	35.73	
4	49.26	4.15	3.16	4.83	38.60		
3	46.31	6.18	5.62	41.89			
2	51.40	8.32	40.28				
1	52.87	47.13					

b)

Months of Data Present	Months Unemployed: 0	1	2	3	4	5	6
6	92.43	3.63	1.65	0.93	0.57	0.43	0.35
5	91.23	4.46	2.08	1.13	0.72	0.38	
4	89.28	5.76	2.17	1.43	1.36		
3	89.00	6.42	2.81	1.77			
2	91.33	6.01	2.66				
1	91.43	8.57					

REDESIGN OF THE NIAGARA TENDER FRUIT OBJECTIVE YIELD SURVEY

J. Kovar[1]

The peach, sour cherry and the grape objective yield surveys have been carried out annually in the Niagara Peninsula since 1964 in order to forecast the magnitude of change in marketable fruit production from the previous year. Timeliness of the estimates is essential in order to enable the Ontario Tender Fruit Growers Marketing Board (OTFGMB) and the Ontario Grape Growers Marketing Board (OGGMB) to establish the marketing strategies well ahead of the harvest. This paper summarizes the major changes due to the second redesign initiated in 1982. In particular, the sample design, data collection operation and modifications of the estimation procedures are elaborated upon.

1. INTRODUCTION

The decision to switch from a list frame to an area frame survey was made in the first redesign in 1974 primarily due to the lack of an adequate list of commercial growers in the Niagara Peninsula. However, in 1981 the Ontario Ministry of Agriculture and Food (OMAF) has conducted a Tree Fruit Census and a Grape Vine Census. The availability of the census data makes it possible to redesign the survey for the second time in order to reflect the changes that the industry has undergone in the last eight years. Based on discussions with OMAF, it was decided that the census lists of growers are complete and accurate and that they contain sufficient information to form the sampling frame for the Tender Fruit Surveys. As a result, the peach, sour cherry and grape surveys will be conducted employing three independent samples selected from the 1981 OMAF census lists.

The object of all three surveys is to forecast the total amount of fruit actually sold (as fresh fruit or to processors). These forecast are made by

[1] J. Kovar, Business Survey Methods Division, Statistics Canada. This work was done while the author was in the Institutional and Agriculture Survey Methods Division, Statistics Canada.

estimating a ratio of the number of pieces of marketable fruit in the current year to the corresponding total for the previous year and applying this ratio to the previous year's figure of actual amount sold reported as a tonnage by the Ontario Fruit and Vegetable Statistics Committee. Thus an assumption of high correlation of fruit weight and fruit count must be made. Secondly due to the time lag between the surveys and the harvest, it has to be assumed that any loss of fruit between these two times is consistent from year to year.

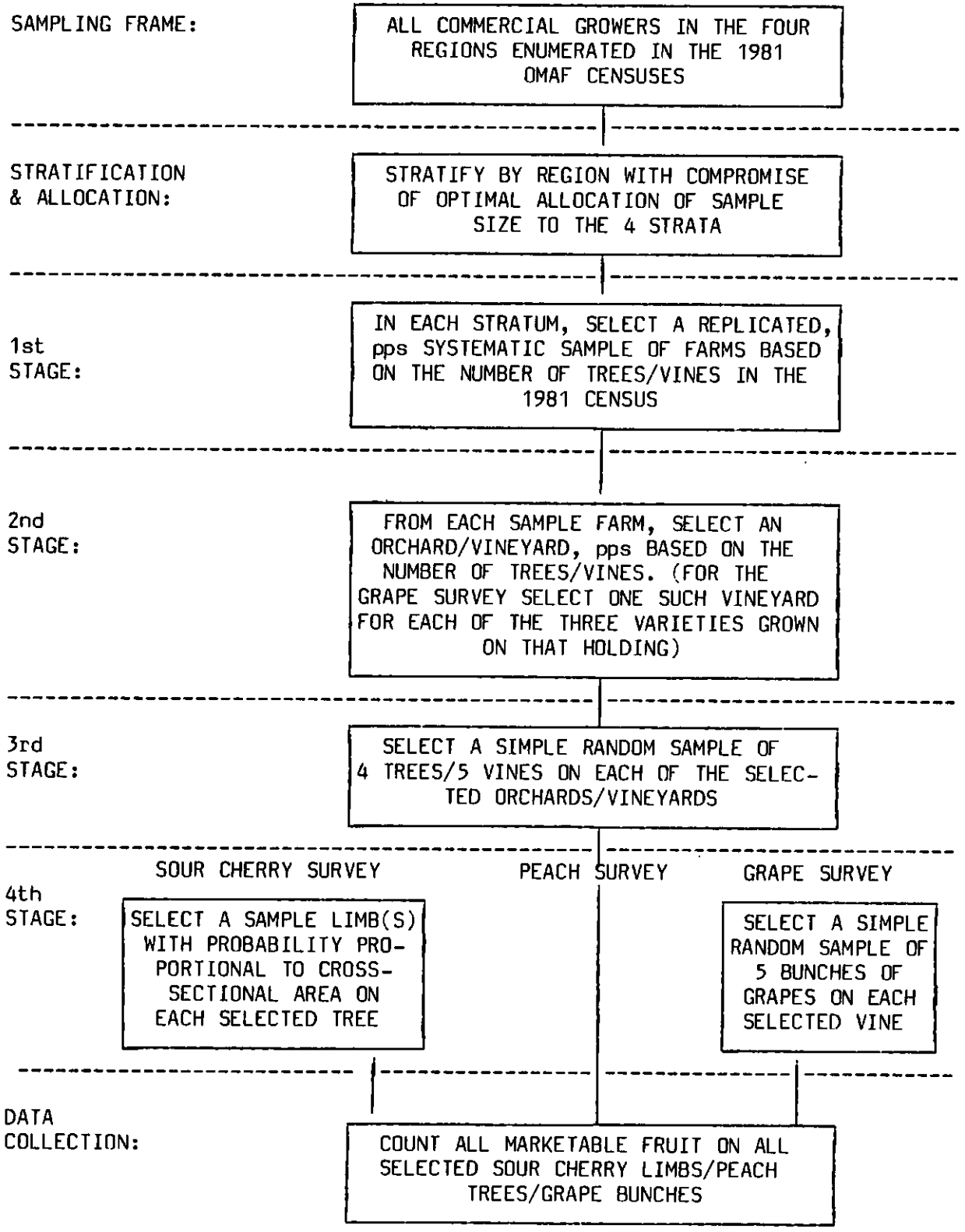
2. OVERVIEW OF THE SAMPLE DESIGN

The samples of the three objective yield surveys were selected independently according to a multistage, stratified (by geographical region), replicated, pps (farms and orchards/vineyards were selected with probability proportional to size), nearly self-weighting (all trees/vines have an approximately equal probability of selection) sample design. Figure 1 provides a visual summary of the sampling strategy. Note that due to the fact that the weight variables are collected at various points in time, the design is not exactly self-weighting.

2.1 Target Population, Sampling Frames and Total Sample Size

The target population for the three objective yield surveys comprises all commercial growers of the respective fruit in the Niagara Peninsula. Commercial growers for the three surveys were defined by OMAF as operators of those holdings which reported more than 200 peach trees, 200 sour cherry trees or 5000 grape vines respectively in the 1981 Tree Fruit or Grape Vine Censuses. Using the above definition, a separate frame was created for each of the three surveys. The lists for the peach, the sour cherry and the grape surveys contain 423, 145 and 552 commercial growers respectively. The total sample size (number of orchards/vineyards to be enumerated) for each survey was determined by OMAF's budget constraints to be in the neighbourhood of 60 for the peach survey, 55 for the sour cherry survey and 155 for the grape survey. Since for the grape survey all available varieties of interest are to be sampled on a selected farm, the final sample size for the grape survey is unknown. However, based on the 1981 Grape Vine Census, it is estimates that 62 farms will generate a sample of approximately 155 vineyards.

FIGURE 1: Sample Design for the Tender Fruit Objective Yield Surveys



2.2 Stratification and Sample Size Allocation to Regions

The Niagara Peninsula was divided into four regions for which separate estimates are required. These were defined as follows (based on the 1976 Census boundaries):

Region 1: Town of Grimsby in the Niagara Regional Municipality and township of Saltfleet in the Regional Municipality of Hamilton-Wentworth. (Township 8 of county 29 and township 4 of county 17).

Region 2: City of St. Catherines and the Town of Lincoln in the Niagara Regional Municipality. (Townships 5 and 9 of county 29).

Region 3: Town of Pelham and the Town of Thorold in the Niagara Regional Municipality. (Townships 11 and 12 of county 29).

Region 4: City of Niagara Falls and the Town of Niagara-on-the-Lake in the Niagara Regional Municipality. (Townships 3 and 10 of county 29).

Due to the increasing demand of crop production estimates by geographic area, an independent sample of farms was drawn in each of the four regions. An attempt was made to allocate the resources (i.e. number of farms sampled) optimally between regions. However, due to the unusually small population size in some regions (see Table 1) a compromise between proportional allocation, optimal allocation and a rule of "minimum of 2 farms per region per replicate" was made. The latter rule was deemed appropriate in order to diminish the possibility of complete nonresponse in a given replicate (as could be the case if only one farm per replicate was selected). The number of trees/vines in each farm was used as a measure-of-size variable for the purposes of allocation as well as for pps selection in the first and second stages. Previous results [6] indicate that other proxy variables (such as area under cultivation) are likely to be no more efficient than the tree-count variable.

TABLE 1: Population and Sample Sizes for the Tender Fruit Surveys of Commercial Growers by Region

	REGION 1		REGION 2		REGION 3		REGION 4		TOTAL	
	POPL'N	SAMPLE	POPL'N	SAMPLE	POPL'N	SAMPLE	POPL'N	SAMPLE	POPL'N	SAMPLE
PEACH	15	4	198	22	15	4	195	30	423	60
SOUR CHERRY	20	4	55	20	30	20	40	12	145	56
GRAPE	67	4	275	32	46	4	164	22	552	62

2.3 First Stage Design

Within each region, for each survey, two independent replicates of farms were selected systematically (in order to obtainm a representative sample) with probability proportional to the total number of trees/vines on the holding as of the 1981 Censuses. The total sample sizes for the two replicates are displayed in Table 1 by region. Since the two replicates are selected independently and since large farms are more likely to be selected in the sample, it is to be expected that a certain amount of overlap between replicates will exist. In fact, some farms are so large, that not only are they guaranteed to be in the sample, but they can appear more than once in the same replicate [4]). Each such appearance is treated as a separate event and one orchard/vineyard is selected without replacement every time the farm is selected. The actual number of distinct farms in the sample is therefore decreased as indicated in Table 2.

TABLE 2: Total Number of Distinct Farms in the Sample by Region

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PEACH	3	22	3	27	55
SOUR CHERRY	4	17	15	10	46
GRAPE	4	30	4	20	58

2.4 Second Stage Design

From the second stage on, the sampling strategies involve some field operation. Once an initial contact with the farmer is made (in the spring of 1983) it is imperative that every effort be made to obtain the respondent's cooperation. It is as this time that the farmer will be requested to aid the enumerator in listing all orchards and establishing the current size (i.e. number of trees) of each for the peach and sour cherry surveys. For the grape survey a similar listing must be prepared for each of the three varieties of interest: Concord, DeChaunac and "Other". (Note that since some varieties are grown together, one vineyard can appear on several of the lists. However, its size for a particular variety listing would be measured by the number of vines of that variety only).

On each holding, for the peach and sour cherry surveys, one orchard will be selected with probability proportional to size. For the grape survey one vineyard will be selected independently for each of the three varieties actually grown on that holding, again with probability proportional to size. It cannot be overemphasized that these procedures must be followed faithfully in order not to jeopardize the validity of the estimates. Selection procedures should be monitored carefully to ensure there is no bias in the selection towards small orchards or single variety vineyards, which admittedly would be easier to enumerate.

To avoid an overlap of orchards/vineyards on farms that are selected in both replicates or more than once in the same replicate, all orchards/vineyards for a given holding are to be selected at the same time, using a pps systematic sampling method. The assignment to replicates is to be performed at random after this selection. (Note that for the grape survey, on farms which appear in both replicates, two vineyards of each variety grown are to be selected).

2.5 Third Stage Design

Once an orchard/vineyard is selected, its current count of producing trees/vines is determined and a simple random sample of four producing trees/five producing vines is selected without replacement. The trees/vines

are marked for future identification since the same units are enumerated from year to year. (In subsequent years, if a sampled tree/vine has been destroyed, pulled up or has died, a replacement tree/vine is selected and enumerated. However, it does not contribute to the estimate until its second year in the sample). Also, each year the producing tree/vine count of selected orchards/vineyards is reestablished in order that the industry's growth (decline) can be monitored. (Note that in the grape survey this implies that only vines of the particular variety sampled are to be counted in each vineyard).

2.6 Fourth Stage Design

This stage exists only for the sour cherry and grape surveys.

2.6.1 The Sour Cherry Survey

It is operationally impossible to count all sour cherries on a selected tree. To estimate the total marketable fruit count, a sample limb (or limbs) is selected with probability proportional to the cross-sectional area of the limb. A method of selecting a limb in this way is described by Jessen [3]. It consists of selecting a limb at the initial (or primary) branching point of the trunk with probability proportional to the cross-sectional area and following the selected limb to the next branching point. This is repeated until the cross-sectional area of a subsequently selected limb is within five to fifteen percent of the primary limbs cumulative cross-sectional area total. As it is not always possible to select one such limb, in some instances two limbs will have to be enumerated. The selected limb(s) on each sample tree are then marked for future identification since the same limbs are enumerated from year to year.

2.6.2 The Grape Survey

As for sour cherries, it is equally impossible to count all marketable berries on a sample vine. Thus to estimate this total, the number of bunches of grapes (i.e. those clusters containing more than five berries) is counted and 5 bunches are selected at random without replacement in order to be enumerated.

As with the other surveys, the vines are marked and are to be visited the following year.

3. DATA COLLECTION

The actual enumeration will be performed roughly four weeks before harvest each year. It is of great importance that the selected sample vines, trees and limbs as well as the orchards and vineyards be well identified in order to enable the enumerators to complete their job in the short time available. The enumerators will be required to count all marketable fruit (i.e. excluding culls which are immature or damaged fruit that will not be harvested) on the sample peach trees and the selected cherry limbs. The fruit on the entire peach tree is counted primarily due to the fact that the fruit tends to be distributed much more unevenly on a peach tree than on a sour cherry tree [2], precluding the possibility of merely enumerating sample limbs.

For the grape survey, all berries on the five selected bunches are to be counted, excluding culls. Since most bunches are very tightly packed, this will, in most cases, involve picking the fruit. For this reason and due to time constraints, it is impossible to enumerate the entire sample vine.

4. REPLACEMENTS

Even though every attempt will be made to return to the same trees, limbs or vines in the following years, there arise cases when this is impossible. (For example, branches are sawn off, trees or vines are pulled up or are otherwise destroyed). If a sour cherry tree limb was sawn off, an attempt will be made to select another limb on the same tree using the same procedures as before. In the event that this is not possible, then just as in the case of peach trees and grape vines, a new tree/vine will be selected at random in the same orchard/vineyard. In the case that the whole orchard/vineyard has been destroyed a new orchard/vineyard will be selected on the same holding using the same procedures as described in Section 2.4. In all these cases, the newly

selected sample limbs, trees, vines, orchards or vineyards will be enumerated, however they will will not contribute to the estimate until the following year's data is collected, as only matched observations are considered.

For those hopefully rare, cases where the farmer has ceased to grow the fruit of interest entirely or where the initial contact resulted in a refusal, a third "replicate" of much smaller size was selected without replacement for each of the surveys. The procedures for selecting the orchard/vineyard and the sample of trees, limbs and vines for each replacement farm are the same as those described above. The limbs, trees and vines will be enumerated every year but will contribute to the estimate only when it is necessary to rotate one of them into the sample. The sizes of the replacement sample are indicated in Table 3 by region.

Table 3: Sample Sizes of the Replacement Sample for the Tender Fruit Survey by Region

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PEACH	1	2	1	3	7
SOUR CHERRY	1	2	2	2	7
GRAPE	1	3	1	2	7

5. ESTIMATION FORMULAE

5.1 Estimates of Fruit Count per Tree/Vine

Denote by y_τ the total number of marketable fruit on a tree (vine) τ.

Then for the peach survey, y_τ is estimated by $\hat{y}_\tau$, the total number of marketable peaches counted on a sample tree τ. For the sour cherry survey, y_τ is estimated by

$$\hat{y}_\tau = \hat{y}_{\tau\ell}/p_\ell \tag{5.1.1}$$

where $\hat{y}_{\tau\ell}$ is the total number of marketable sour cherries counted on the sample limb(s) ℓ of the selected sample tree τ;

and p_ℓ is the probability of selecting the sample limb(s) ℓ. Finally for the grape survey, y_τ is estimated by

$$\hat{y}_\tau = \frac{N_\tau}{n_\tau} \sum_{\ell=1}^{n_\tau} \hat{y}_{\tau\ell} \tag{5.1.2}$$

where N_τ is the total number of bunches of grapes on the sample vine τ;

n_τ is the number of bunches of grapes that were enumerated on the sample vine τ (typically $n_\tau = 5$);

and $\hat{y}_{\tau\ell}$ is the number of berries on a bunch ℓ of the sample vine τ.

5.2 Regional Estimates of Fruit Count by Replicate

Denote by $\hat{Y}_{ar}$ the estimated total number of marketable fruit in replicate r of region (area) a in the current year. For the grape survey,

$$\hat{Y}_{ar} = \sum_{v=1}^{3} \hat{Y}_{arv} \tag{5.2.1}$$

where $\hat{y}_{arv}$ is the estimated total grape count of variety v in replicate r of region a.

For the purpose of uniformity of the following formulae, for the sour cherry and peach surveys $\hat{Y}_{ar}$ and $\hat{Y}_{arv}$ can be used interchangeably, since there is only one variety of sour cherries and peaches to be estimated. (In other words, the subscript v can be ignored for the sour cherry and peach surveys). Then $\hat{Y}_{arv}$ can be estimated by

$$\hat{Y}_{arv} = \frac{1}{n^*_{arv}} \sum_{f=1}^{n_{arv}} \sum_{b=1}^{n_{arvf}} \frac{N_a^{81}}{N_{arf}^{81}} \times \frac{N_{arvf}^{83}}{N_{arvfb}^{83}} \times \frac{N_{arvfb}^{C}}{n_{arvfb}} \sum_{\tau=1}^{n_{arvfb}} \hat{Y}_{arvfb\tau} \qquad (5.2.2)$$

with $\hat{Y}_{arvfb\tau}$ = the current year's estimated total number of marketable fruit on the tree/vine τ (variety v) in orchard/vineyard b, on farm f, in replicate r, of region a (i.e. $\hat{y}_\tau$);

n_{arvfb} = the number of trees/vines (of variety v) sampled in the current year in orchard/vineyard b, of farm f, in replicate r, of region a (typically n_{arvfb} = 4 for the sour cherry and peach surveys and 5 for the grape survey);

n_{arvf} = the number of orchards/vineyards (sampled for variety v in the current year) on farm f, in replicate r of region a (typically n_{arvf} = 1 except for duplicates, i.e. large farms selected more than once in the <u>same</u> replicate);

n_{arv} = the total number of distinct farms on which variety v was sampled in the current year) in replicate r, of region a;

n^*_{arv} = the total number of orchards/vineyards (sampled for variety v in the current year) in replicate r of region a

(i.e. $n^*_{arv} = \sum_{f=1}^{n_{arv}} n_{arvf}$);

N^C_{arvfb} = the total current count of producing trees/vines (of variety v) in orchard/vineyard b, on farm f, in replicate r, of region a;

N^{83}_{arvfb} = the total count of trees/vines (of variety v), in orchard /vineyard b, on farm f, in replicate r of region a as of the 1982/1983 mapping operation;

N^{83}_{arvf} = the total count of trees/vines (of variety v), on farm f, in replicate r, of region a as of the 1983 listing operation;

N^{81}_{arf} = the 1981 Census count of total trees/vines (all varieties) on farm f replicate r, of region a (supplied with the sample listing)

and N^{81}_{a} = the 1981 Census count of all trees/vines in region a (as per Table 4)

Table 4: 1981 Census Counts of Trees/Vines (N^{81}_{a}) for the Tender Fruit Surveys by Region

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PEACH	13,094	389,157	10,271	411,697	824,219
SOUR CHERRY	9,496	50,888	55,449	32,536	148,369
GRAPE	1,142,067	5,660,008	946,509	3,975,202	11,729,786

5.3 Regional Estimates of Ratio of Change and their Precision

The ratio of change in production in region a from the previous year, denoted by R_a, is estimated by

$$\hat{R}_a = \hat{Y}_a/\hat{X}_a \qquad (5.3.1)$$

where $\hat{Y}_a$ = the estimated total marketable fruit count (of peaches, sour cherries, grapes or grapes of variety v) in region a in the current year is given by

$$\hat{Y}_a = \frac{1}{2}\sum_{r=1}^{2} \hat{Y}_{ar} \qquad (5.3.2)$$

in the case of peaches, sour cherries, and <u>total</u> grapes of all varieties, and by

$$\hat{Y}_a = \hat{Y}_{av} = \frac{1}{2}\sum_{r=1}^{2} \hat{Y}_{arv} \qquad (5.3.3)$$

in the case of grapes by variety, and where $\hat{X}_a$, $\hat{X}_{ar}$, $\hat{X}_{arv}$ are the corresponding previous year's estimates.

(Note that the subscript v can now be dropped as all estimates are treated in the same manner, be it peach, sour cherry, total grape or grapes by variety estimates.)

Define the variances of $\hat{Y}_a$ and $\hat{X}_a$ and their covariance by

$$V(\hat{Y}_a) = (\hat{Y}_{a1} - \hat{Y}_{a2})^2/4 = D_{ya}^2/4 \qquad (5.3.4)$$

$$V(\hat{X}_a) = (\hat{X}_{a1} - \hat{X}_{a2})^2/4 = D_{xa}^2/4 \qquad (5.3.5)$$

$$Cov(\hat{Y}_a,\hat{X}_a) = (\hat{Y}_{a1}-\hat{Y}_{a2})(\hat{X}_{a1}-\hat{X}_{a2})/4 = D_{ya}D_{xa}/4$$

where the numeric subscripts refer to the replicate number. Then the variance, $V(\hat{R}_a)$, of the ratio of change estimate, $\hat{R}_a$, can be estimated by [1].

$$\hat{V}(\hat{R}_a) = \frac{1}{\hat{X}_a^2} \{V(\hat{Y}_a) - 2\,\hat{R}_a\,\mathrm{Cov}(\hat{Y}_a,\hat{X}_a) + \hat{R}_a^2\,V(\hat{X}_a)\}$$

$$= \{\frac{D_{ya}}{S_{xa}} - \frac{S_{ya}\,D_{xa}}{S_{xa}^2}\}^2 \qquad (5.3.7)$$

with $S_{ya} = \hat{Y}_{a1} + \hat{Y}_{a2}$

$D_{ya} = \hat{Y}_{a1} - \hat{Y}_{a2}$, etc (5.3.8)

The coefficient of variation of $\hat{R}_a$ is then estimated by

$$C\hat{V}(\hat{R}_a) = \frac{\{\hat{V}(\hat{R}_a)\}^{\frac{1}{2}}}{\hat{R}_a} \times 100\% \qquad (5.3.9)$$

5.4 Regional Estimates of Total Fruit Production and their Precision

Denoting by X_a^T the previous year's actual yield (tonnage) in region a and by $\hat{Y}_a^T$ the corresponding current year estimate, then $\hat{Y}_a^T$ is given by

$$\hat{Y}_a^T = X_a^T\hat{R}_a \qquad (5.4.1)$$

with its coefficient of variation estimated by

$$C\hat{V}(\hat{Y}_a^T) = \frac{X_a^T\,\{\hat{V}(\hat{R}_a)\}^{\frac{1}{2}}}{X_a^T\,\hat{R}_a} \times 100\% = C\hat{V}(\hat{R}_a) \qquad (5.4.2)$$

5.5 Estimates of Total Fruit Production and their Precision

Denote by $\hat{Y}^T$ the estimated total fruit production over all four regions in the current year. Then $\hat{Y}^T$ is given by,

$$\hat{Y}^T = \sum_{a=1}^{4} \hat{Y}_a^T \qquad (5.5.1)$$

with a coefficient of variation estimated by

$$C\hat{V}(\hat{Y}\ ) = \frac{\{\sum_{a=1}^{4} (X_a^T)^2\ V(\hat{R}_a)\}^{\frac{1}{2}}}{\hat{Y}^T} \times 100\% \qquad (5.5.2)$$

6. SUMMARY

The first enumeration will take place in 1983, however, it will not be until the summer of 1984 that the first estimates from the redesigned survey will be produced. For this reason, it will be necessary to conduct both the old and the new surveys in 1983 so that estimates will be available for that year.

Even though the survey was designed to be self-weighting, it is only approximately so due to the time differences of the 1981 Censuses, the 1983 initial listing operation and the subsequent enumerations. The estimation formulae presented in the previous section take these time differences into account. However, due to the appealing simplicity of the self-weighting estimate, an investigation of its performance has been proposed once the data becomes available.

REFERENCES

[1] Cochran, W.G. (1963). Sampling Techniques, John Wiley and Sons, New York.

[2] Davidson, G. (1977). Redesign of the sour cherry, peach and grape objective yield surveys in the Niagara Peninsula, Survey Methodology 3, 38-61.

[3] Jessen, R.J. (1955). Determining the fruit count on a tree by randomized branch sampling, Biometrics 11, 99-109.

[4] Kish, L. (1965). Survey Sampling, John Wiley and Sons, New York.

[5] Murthy, M.N. (1967). Sampling Theory and Methods, Statistical Publishing Society, Calcutta.

[6] Singh, U. and Sukhatme, B.V. (1980). Sampling for estimating production of fruit crops, Sankhya C42, 17-30.

A TIMELY AND ACCURATE POTATO ACREAGE ESTIMATE FROM LANDSAT: RESULTS OF A DEMONSTRATION[1]

R.A. Ryerson, J.-L. Tambay, R.J. Brown
and
L.A. Murphy, B. McLaughlin[2]

This paper describes the procedures used and results of a joint Canada Centre for Remote Sensing (CCRS) and Statistics Canada project to provide a timely potato acreage estimate for New Brunswick, a major potato producing province in Canada. The project has demonstrated that satellite imagery combined with more traditional potato area estimation procedures can lower respondent burden, produce timely crop distribution maps and produce reliable estimates for subregions.

1. INTRODUCTION

Earlier satellite remote sensing work in St. John Valley, New Brunswick by the Canada Centre for Remote Sensing (CCRS) and the New Brunswick Department of Agriculture has proved that both an accurate and low cost estimate of potato crop area could be made using satellite data (Mosher et al. 1978; Ryerson et al. 1979; Ryerson et al. 1980). Interest in this and other CCRS work on rapeseed-canola (Brown et al. 1980) resulted in Statistics Canada initiating a real-time demonstration using data from Landsat satellite in the 1980 crop year. Statistics Canada, the Federal agency responsible for crop data collection, wanted to compare traditional and satellite-derived estimates of crop area in the same region. Potatoes were selected as the subject crop, and the St. John Valley was the region.

---

[1] Originally presented at the fifteenth International Symposium on Remote Sensing of Environment, Ann Harbor, MI, May 1981.

[2] R.A. Ryerson and J. Brown, Canada Centre for Remote Sensing (CCRS), E.M.R., J.-L. Tambay, Business Survey Methods Division (this work was done while the author was in the Institutional and Agriculture Survey Methods Division), Statistics Canada, L.A. Murphy, Agriculture Statistics Division, Statistics Canada, and B. McLaughlin, Agriculture Statistics Division, Regional office at Truro, Nova Scotia, Statistics Canada.

The particular benefits of satellite remote semsing which are of interest to Statistics Canada are improving accuracy of estimates obtained from their regular surveys, possibly at more local levels, lowering of respondent burden by reducing the number and/or size of questionnaires, and the possibility of providing maps of small areas containing speciality crops to better plan sampling methods.

Following a summary of the main results in the next section, the balance of this paper outlines the remote sensing methodology used in this project, and describes the existing Statistics Canada data collection system, the project region, the ground data sample and data collection, the analysis of remotely sensed data and the verification and analysis of results.

2. MAIN RESULTS

Data collected by satellite were used to produce estimates of the potato area in the St. John Valley region of New Brunswick. These estimates, expanded to the provincial level, were within two percentage points of the Statistics Canada published estimate of 52,000 acres. The published estimate was based on the results of three independent surveys in the province.

The analysis of satellite data was done in real-time (almost instantaneously) at CCRS, as much of the work could be initiated prior to data acquisition. The Agriculture Enumerative Survey (A.E.S.) area sample provided the ground data needed to calibrate the system, and was used to obtain ratio and regression estimates which corrected for biases in the satellite classification of potato fields. Although the demonstration was not carried out in a production environment, the final estimates could have been produced less than two weeks after the satellite pass over the test region.

Problems in the satellite classification were caused by the presence of clouds (satellite nonresponse), and by the confusion of potatoes with "similar appearing" features on the analysis system. The first problem caused loss of

data, and required some imputation. The second problem was partly resolved by adjustments to the classification, and by the use of ratio and regression estimators.

A comparison of interviewer - collected ground data with data collected through aerial photographs of sample fields showed that some fields were missed by the A.E.S. interviewers, as these were not required for A.E.S. purposes. This resulted in the aerial photography data being used instead of A.E.S. data for the 1980 satellite estimates. The 1981 A.E.S. enumeration procedure were changed to accommodate both A.E.S. and remote sensing requirements.

As a result of the success of this demonstration, the experiment was repeated in 1981. In addition, a similar experiment was undertaken that year to estimate the canola acreage in the Peace River District of Alberta and British Columbia.

3. REMOTE SENSING USING THE LANDSAT SATELLITE

Remote sensing is the sensing or measuring of the characteristics of an object from a distance, usually from an aircraft or satellite. When satellite data are used, complete coverage of large areas can be provided quickly at a relatively low cost. Possible areas of application include agriculture, forestry, land utilisation, ice formations and general map making.

The United States National Aeronautics and Space Administration Landsat satellites provided the satellite coverage for this, and an earlier experiment in New Brunswick. Each Landsat satellite orbits the earth 14 times a day in a sun-synchronous orbit (permitting coverage of the earth to be done at the same local sun time). Light reflected from the ground is recorded on four narrow bands of the spectrum using a Multispectral Scanner (MSS). The data are transmitted in Canada to one of two receiving stations in Prince Albert,

Saskatchewan and in Shoe Cove, Newfoundland. A point on earth is covered once every 18 days by a Landsat satellite (every nine days if two satellites are used).

The CCRS Image Analysis (CIAS) analyses the data, received on standard products such as Computer Compatible Tapes (CCT's) covering areas of 25,600 square kilometers. The smallest units for which image data are defined are called picture elements, or pixels. Each pixel carries its own spectral signature, a measurement of its reflectance on the four spectral bands. The spectral signature will depend on the features present in the pixel (roads, crops, etc.), each of which carries its own signature. Crop signature is a function of plant structure, type of soil background visible, crop maturity, height, and leaf density, among other factors.

To obtain estimates of crop areas, it is necessary to identify each pixel belonging to a crop of interest. Large known fields of the crop of interest are located to train the system to identify the crop's signature. All pixels are then classified as belonging to the crop or not, based on their spectral signatures.

Areas for specific regions are obtained by counting the number of pixels inside the regions that are classified as belonging to the crop. Additional training may be done to cover pixels "missed" in the initial classification, or to further separate confusion crops, that is, crops whose spectral signature closely resembles that of the crop of interest.

Accurate ground data are needed, first, to locate large training fields for the crop of interest and second, to correct for any biases in the satellite classification. These data can be obtained by trained ground enumerators, or by using airborne imagery, which is interpreted by image analysts.

4. CURRENT STATISTICAL DATA COLLECTION SYSTEM

Historically, Statistics Canada has used data obtained from annual mail non-probability surveys as the primary input into its crop estimation system. While these surveys are relatively inexpensive and can be completed quickly, they are limited by varying response rates and possible non-representativeness of respondents. Probability enumerative surveys were introduced in the mid-1970's to overcome some of these problems. These involve enumeration of a random sample of farmers by personal interviews. In 1980, Statistics Canada's estimates of potato area in New Brunswick were based on the results of three surveys: the Mail Survey, the Objective Potato Yield Survey (O.P.Y.S.), and the Agriculture Enumerative Survey (A.E.S.).

The Mail Survey questionnaires are mailed out in early June to all farmers listed on a Farm Register maintained by the Agriculture Statistics Division. Replies are compiled on a county basis and county estimates are derived by linking annual changes in reported potato acreages to census potato acreages for the county. The county estimates are aggregated to give provincial estimates by late June.

The O.P.Y.S. is a specialized mail and enumerative survey designed to estimate potato area and yields in the potato growing region of New Brunswick. The Survey is conducted in mid-July on a random sample of potato farmers selected from the Farm Register and potato area estimates are generated by mid-August.

The A.E.S. is a multi-purpose enumerative survey designed to estimate crop, livestock and farm expemse data at provincial levels. The A.E.S. is a multiple frame survey consisting of a random list sample of farmers selected from the Farm Register and a random area sample of segments. Enumerators visit the sampled farms in late June and early July. Acreage estimates from the survey are available in early August. Each year about 20% of the segments are changed.

During the growing season, two potato crop area estimates are published. The first, in late June, is based on the mail survey results. The second estimate, in early September, is based on a review of the estimates from all three surveys and discussion with provincial authorities. The date of the second estimate was the target date for generation of a satellite-derived estimate.

5. PROJECT REGION

The area for which an estimate was required is located in the upper St. John Valley in New Brunswick. It starts in the south at Woodstock in Carleton County and follows the St. John River for 200 kilometers northwest through Victoria County to Claire in Madawaska County.

The region is heavily wooded, of varied, rolling terrain. There are some problems related to stoniness and drainage. Within the area are 70,000 hectares of improved cropland, of which about 20,000 are usually potatoes. Other major crops are grains, hay and processing vegetables such as peas, broccoli and brussels sprouts. Parcel sizes range from 0.1 hectare seed plots to 40 hectare fields.

6. 1980 GROUND SAMPLE AND DATA COLLECTION

The area sample for the A.E.S. in New Brunswick was considered a suitable vehicle for obtaining ground data for interpreting remote sensing data. This sample is selected in two stages. At the first stage, census Enumeration Areas (EA's), which had farm headquarters in them in the 1976 census (called Census Agricultural EA's), were stratified based on their potato acreages, cattle, and pig numbers (1976 census data). Within each stratum, two replicated simple random samples of EA's were selected. Each sampled EA was segmented into identifiable area units of about 5 to 8 square kilometers using maps, and a simple random sample of one in 10 segments was selected per EA. A.E.S. enumerators working in the study region were supplied with old enlarged

aerial photos (scale 1" - 832') for each sampled segment. The photos were obtained from provincial sources. Most of them had been taken in 1976. While contacting the farmers operating land inside the segment, they were required to show the photograph to the farmer and identify on it all potato and corn[3] fields and note their areas as reported by the farmer. Written instructions on procedures to be followed were included in the interviewers' manual and interviewers and supervisors were trained on procedures to be followed.

7. ANALYSIS OF REMOTELY SENSED DATA

7.1 Previous Work

Work in the same region using 1975 data has been reported elsewhere (Mosher et al. 1978), and a detailed description of the approach is available (Ryerson et al. 1980). In the 1975 work, a test area which contained about twenty percent of the province's potato crop was selected for detailed analysis from the potato growing region. This was supported by ground data collection for the entire test area.

The 125 square kilometer test area and two sub-areas were located on the colour video display screen of the CCRS Image Analysis System (CIAS) (Goodenough, 1979). A very simple supervised training scheme was used to gather the statistics of pixels in three contiguous potato fields in the form of four one-dimensional histograms. A four-dimensional parallelepiped in feature space was generated as defined by the limits of each of the histograms to serve as a decision boundary. All points within the parallelepiped were classified as potatoes, and those outside the region as "other".

One of the major problems was the proper classification of boundary pixels. These present special problems, as they fall on the border of two different

[3] Corn fields were also required to be identified since earlier remote sensing work indicated that corn was a confusion crop for potatoes (Ryerson et al. 1980).

fields. Their reflectance is a function of the amount of each field within the pixel and the reflectance of each cover material in the two fields. To attempt to compute the percentage of each cover type present in such pixels is generally very complicated. However, it was found that by modifying the original decision boundary through adding a second parallelepiped formed by training on a number of boundary pixels which appeared to be in potato fields, reasonable area estimates could be achieved. Selection of the appropriate boundary pixels to be classed as potatoes was on the basis of subjective visual interpretation of the display (data from three of the four spectral bands were merged to form a colour display, with colours simulating those of a colour infrared film).

Less then four hours of CIAS time were required to perform the area estimate for the entire potato belt. Location, display and analysis of the primary and sub-test areas required just over one hour. Selection, display and analysis of the subsequent five subscenes required two and one-half hours, while location of the New Brunswick border and elimination of data from outside the province required another hour.

Compared to total area of potato fields interpreted and measured from low altitude aerial photographs taken at the same time, the 1975 satellite estimates were 95% accurate (i.e., 95% of the estimated true value), in the sub-area containing the training site, 80% accurate in the second sub-area, and 88% accurate over the whole primary test area. On repeated tests using different training fields, the accuracy over the primary test area ranged from 85% to 97%. The province wide accuracy was 84.5%. Some of the error in the provincial estimate resulted from the fact that some potatoes are grown outside the potato belt. Other factors contributing to the error are discussed below.

7.2 PROCEDURES TO IMPROVE ESTIMATES

Although the previous work was successful, potential sources of error were identified for applications requiring accuracies greater than say 85%. The

major problems arise in the subjective selection of the potato field boundary pixels, in the handling of small fields, and in the resolution of difficulties with the crops confused with potatoes.

With regard to confusion crops, ideally one should know the spectral reflectances of potatoes and all of its confusion crops throughout the growing season. With such information, it is possible to specify the phenological window during which potatoes can be reliably separated from other crops. Unfortunately, such a data set does not exist, although knowledge of the region's crops and the cultural practices does provide some general guidelines. In this case, based on the field experience of the authors, it was hypothesized that the optimum date for separation of potatoes from other crops in this region would be from mid-July to mid-August. To test this hypothesis and provide an indication of the degree of separability of potatoes from other crops, an analysis was performed of a Landsat MSS Computer Compatible Tape (CCT) acquired over the St. John Valley on August 8, 1975. Figure 1 shows the Landsat band radiance values for potatoes, corn, peas, hay, broccoli, pasture, buckwheat, bare soil and grains. It can be seen from this that potatoes are easily separable from the other crops except for the peas, which are usually harvested by mid to late August. It would therefore appear that the analysis of data collected late in the growing season is likely to lead to the separation and identification of potatoes.

The problem of small fields and boundary pixels can be handled by an approach which uses available ground information over a limited area to produce a more accurate crop area estimate over an extended area. Given the size of the potato growing region here, the area of potatoes within each of ten to fifteen segments is required. Each segment is of the order of five to eight square kilometers in size. The whole area is still classified as well as possible, but the subjective boundary pixel class is not produced. The classification result for each segment is then used, along with the available ground information (the aerial photograph data in 1980, A.E.S. data for future years) to obtain a regression relationship which is applied to the entire area estimate to

produce a revised estimate (Hanuschak et al. 1979). A ratio estimate, based on the total area estimates obtained from satellite and aerial photograph data, is also produced.

7.3 GENERATION OF THE 1980 ESTIMATE USING LANDSAT

The generation of a satellite-based potato area estimate can be described as a three part process: ordering of old and new data, pre-location of A.E.S. segment boundaries, and image analysis.

The Landsat data ordered where Digital Image Correction (DICS) Computer Compatible Tapes (CCT's) using the Sin × / × interpolation for geometric correction (Guertin et al. 1979) and Cal 3 radiometric correction (Ahern and Murphy, 1978). Each CCT covers four 1:50,000 National Topographic System (NTS) map sheets with a resampled square pixel of 50M. Four CCT's are required to cover the region. Existing data were ordered for delivery in May of 1980, while new data were ordered for the appropriate satellite passes from mid-July to mid-August. The ordering process proceeded smoothly for existing data, but was complicated for the real-time data by the failure of Landsat-III. A Landsat-II pass on August 17 was used to create DICS CCT's which were delivered to the analysis facility on August 22, well ahead of schecule.

The pre-location of A.E.S. segments was done in the spring of 1980 using the polygon cursor option on the CCRS Image Analysis System (Goodenough, 1977). A.E.S. segment boundaries were provided by Statistics Canada on 1:50,000 map sheets and on photocopies of the airphoto enlargements given to the A.E.S. enumerators. Although some segments had boundaries which were easy to locate (streams, forested edges, lakes, etc.), others based on political or census boundaries proved very complex. Segments whose boundaries were a combination of major roads and/or major rivers could be located, bounded and stored after less than five minutes work on the enlarged colour display of 128 x 128 50M pixels on a 512 x 512 monitor. The most complex took up to an hour - with an

average of 20-25 minutes. Once located, the segment was stored by specific DICS pixel coordinates so that it could be overlaid on new data as it arrived to locate both training data and inputs for the estimator. Because of the location of some segments on the boundary between two DICS CCT's and other similar problems, a number of segment boundaries were not located in the preparatory phase. Software is now being written to shorten the time required for the entire project, especially the pre-location phase. Use of original photographs in place of photocopies planned for the 1981 project's new segments (15 rotated in for 1981), should also shorten the time required.

Once the 1980 ground data were delivered to the analysis center, potential training sites (based on field size) were selected. Several fields were selected from one segment in the north of the region (near St. André) while several others were selected from a segment in the south (near Hartland).

Upon receipt of the 1980 satellite data, it became a relatively simple task to recall the segment boundaries, overlay them, locate training fields and begin classification using methods described in 7.1. In addition to the selected training areas, another group of large fields was selected as were areas of known potato fields which appeared brighter red on the monitor than those in the training set. As classification results were available, crop areas were recorded for each 512 x 512 pixel subscene and for each of seventeen A.E.S. segments.

There were four problems encountered during the classification; one involving imputation of crop under scattered cloud and cloud shadow and the other three related to confusions. The method of imputation of potatoes under cloud was quite straightforward. It was assumed that the percent area of crop under the clouds was similar to the area of crop in an adjacent "like-appearing" region. A simple formula was used to determine potato area under cloud, PC:

$$PC = \frac{P_M}{T_A - C_A} \times C_A$$

where: P_M = potato area measured in the cloud free region; C_A = area in cloud and shadow in the region and T_A = total area in the region. These areas were incorporated into the total satellite estimate, no A.E.S. segments under cloud were included in the ratio or regression analyses. The problems with confusions were, for the most part, solved through careful modification of the classification parameters. In one case, an unknown form of widely scattered natural regrowth in forest cut-overs was confused. In another case, hay fields with regularly spaced piles of stone were similar to potatoes. Only a few fields of clover in one segment and some of the fairways of a golf course in another remained as confusions after modifying the classification.

The areas calculated for the region are presented and discussed in more detail below.

8. RESULTS AND ANALYSIS OF THE 1980 NEW BRUNSWICK POTATO PROJECT

8.1 INTRODUCTION

The analysis of data from the 1980 New Brunswick potato project was done at the regional, segment, and field levels. Two types of estimates were obtained for the total potato area of the test region in the St. John Valley using satellite data and ground verified measures from high altitude aerial photograph data (the aerial photography data were obtained and analysed by CCRS). The estimates and their variances were then compared to other estimates from Statistics Canada surveys in New Brunswick. Segment potato acreages reported by the A.E.S. and by satellite were then compared to the aerial photograph acreages (which are considered here to be closest to the actual values) to determine the strength of their relationship at the segment level. This analysis was complemented by examining the variation in A.E.S. reporting of field acreages for each interviewer's assigned area.

8.2 ESTIMATION OF POTATO AREAS USING SATELLITE DATA

A ratio and a regression estimate of the total potato area in the test region in New Brunswick were obtained using satellite data and the aerial photograph segment data. Estimates, along with their variances, were calculated using the A.E.S. sample design. The A.E.S. in New Brunswick is a multiple-frame stratified replicated two-stage sample of segments, designed to give accurate estimates of various items at the provincial level (see sections 4 and 6). Since the A.E.S. strata did not coincide with the test region boundaries, the technique of post-stratification was used for estimation, treating the EA's as a with-replacement-sample. Segments with missing satellite data were not included in the sample, nor was one outlier (see Figure 3). These estimates were based on 40 segments. Finally, since A.E.S. enumerators did not always collect data on all farms inside the segment (see 8.3), aerial photograph segment acreages were used for estimation.

Let the label x represent the reported satellite potato data and y represent aerial photograph data. The ratio and regression estimates of Y, the total potato area in the test region were calculated as:

$$\hat{Y}_{Ratio} = \hat{R}\,X = \frac{\hat{Y}}{\hat{X}}\,X, \text{ and}$$

$$\hat{Y}_{Reg.} = \hat{Y} + \hat{B}\,(X-\hat{X}),$$

where $\hat{Y} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$ is the design estimate of Y;

$\hat{X}$ is the design estimate of X, the total uncorrected satellite potato area in the test region. $\hat{X}$ is obtained by substituting x_{hij} for y_{hij} in the formula for $\hat{Y}$;

y_{hij} and x_{hij} are the observed values for the jth selected segment in the ith selected first-stage unit (EA) from stratum h;

N_h and n_h are the total and sampled number of EA's in stratum h, h=1, ..., L;

M_{hi} and m_{hi} are the total and sampled number of second-stage units (segments) in the ith selected EA of stratum h;

$\hat{R}$ is an estimate of Y/X; and,
$\hat{B}$ = cov $(\hat{Y},\hat{X})$/ var $(\hat{X})$ is the linear regression coefficient.

The variance estimates of $\hat{Y}_{Ratio}$ and $\hat{Y}_{Reg.}$ are given by:

$$v(\hat{Y}_{Ratio}) = v(\hat{Y}) - 2\hat{R}\ \mathrm{cov}\ (\hat{Y},\hat{X}) + \hat{R}^2\ v(\hat{X}), \text{ and}$$

$$v(\hat{Y}_{Reg.}) = v(\hat{Y}) - 2\hat{B}\ \mathrm{cov}\ (\hat{Y},\hat{X}) + \hat{B}^2\ v(\hat{X}),$$

where $$v(\hat{Y}) = \sum_{h=1}^{L} \frac{N_h^{\ 2}}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left[\frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} - \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \right]^2$$

and $v(\hat{X})$ is calculated by substituting x_{hij} for y_{hij} in $v(\hat{Y})$.

It can be seen that $\hat{B}$ = cov $(\hat{Y},\hat{X})/v(\hat{X})$ is the value of B which minimizes $v(\hat{Y}_{Reg.})$.

Table 1 shows the ratio and regression estimates, along with other Statistics Canada estimates of the potato area in New Brunswick. The ratio and regression estimates, pro-rated to the provincial level, are both very close to the Statistics Canada published figure of 52,000 acres. There is very little difference between the two estimates, and they both have the same coefficient of variation. In order to give an idea of the gain in efficiency obtained by the ratio estimate, the variance of the ratio estimate was as low as one-fifth of that of $\hat{Y}$, design estimate of Y based on the post-stratified design (area sample only). It may be noted that the C.V.'s of the ratio and regression esti-

mates are of the same order as that of the A.E.S. multiple-frame estimate, although the latter is based on a larger sample size.

8.3 COMPARISON OF DATA AT THE SEGMENT AND FIELD LEVEL

Figures 2 and 3 show plots of segment potato acreages, as reported by the A.E.S. enumerators and by satellite, against aerial photograph acreages. Not all sample segments were used in the analysis. Satellite data were missing for 16 segments due to cloud cover and image location. Eight A.E.S. segments, containing survey non-respondents and very large farms whose field data were not to be collected by the enumerators, were not used. The two outliers in the plots are not used in the calculations here and in Section 8.2.

The plots show a strong linear relationship between A.E.S. and aerial photograph data, as well as between satellite and aerial photograph data, at the segment level, with correlations[4] of .991 and .968 respectively[5]. There is a tendency for both the A.E.S. enumerators and the satellite classification to underestimate the acreages. This is less marked for the A.E.S. Causes of discrepancies of satellite acreages were explained in Section 7. Some segments with little or no potatoes were over-estimated because of confusion crops. (The satellite outlier had confused a large hay field with rocks for a potato field - this error could have been removed by modifying the classification). One major cause of A.E.S. underestimation was that the interviewers did not enumerate some farms in the segment because of the multiple-frame procedures (these include farms which appeared on the list frame as well as farms that had land in more than one sample segment). Specific instructions have been written up in the 1981 A.E.S. field procedures to ensure that <u>all</u> farms in the segment are enumerated for their potato acreages next year when the test is to be repeated. This is expected to bring the A.E.S. reported acre-

[4] The correlations were estimated using the sample design weights.

[5] A plot of satellite data against the A.E.S. data looked very similar to Figure 3, with a correlation of .957.

ages closer to the aerial photograph acreages. The strong relationship between the A.E.S. and the aerial photograph data for the sampled segments is encouraging and supports making adjustments to the satellite estimates using the data collected by the A.E.S. enumerators.

Another cause of A.E.S. discrepancy with aerial photograph segment data was the mis-reporting of field boundaries and field sizes. A.E.S. field acreages were obtained by interviewing farm operators, and thus, were frequently reported in multiple of five acres. Plots of A.E.S. reported field acreages against aerial photograph acreages by interviewer assignment areas indicate that there may be a difference in field reporting between the assignment areas. This could be caused by the interviewers themselves, but also by other factors such as geographic location and structures of fields within sample segments, or by variation in reporting errors. More variation was observed in the region east of the St. John River, where average field sizes were generally larger. The relationship between the A.E.S. data and aerial photogrpah data was stronger at the segment level than at the field level.

9. SUMMARY AND CONCLUSION

Satellite data were used in 1980 to generate a highly accurate potato area estimate for the three major potato producing counties of New Brunswick. Through the project described here, refinements have been identified for field procedures and analysis methods which should provide even more accurate estimates of crop area with satellite data, reduce respondent burden, and provide detailed spatial information previously available only in Census years.

10. ACKNOWLEDGEMENTS

The authors would like to acknowledge the technical advice and editorial comments of Mrs. N. Chinnappa of Statistics Canada and of Drs. J. Cihlar and F.Ahern of the Canada Centre for Remote Sensing.

11. REFERENCES

[1] Ahern, F.J. and J. Murphy, 1978, "Radiometric Calibration and Correction of Landsat 1, 2, and 3 MSS Data", CCRS Research Report 78-4, Energy, Mines and Resources, Ottawa, Canada.

[2] Brown, R.J., Ahern, F.J., Ryerson, R.A., Thomson, K.P.B., Goodenough, D.G., McCormick, J.A. and Teillet, P.M., 1980, "Rapeseed: Guidelines for Operational Monitoring", 6th Canadian Symposium on Remote Sensing, Halifax, Nova Scotia, Canada, 321-330.

[3] Cochran, W.G., 1977, Sampling Techniques (3rd Edition), John Wiley and Sons Inc., New York.

[4] Goodenough, D.G., 1979, "The Image Analysis System (CIAS) at the Canada Centre for Remote Sensing". Canadian Journal of Remote Sensing 5(1) May 1979, 3-17.

[5] Guertin, F.G., T.J. Butlin and R.G. Jones, 1979, "Correction geometrique des images Landsat au Centre canadien de télédétection", Canadian Journal of Remote Sensing, 5(2), December 1979.

[6] Hanuschak, G., R. Sigman, M. Craig, M. Ozga, R. Luebbe, P. Cook, D. Kleweno, and C. Miller, 1979, "Crop - Area Estimates with Landsat: Transition from Research and Development to Timely Results", Proceedings of the Fifth Machine Processing of Remotely Sensed Data Symposium, Lafayette, Indiana.

[7] Mosher, P.N., R.A. Ryerson and W.M. Strome, 1978, "New Brunswick Potato Area Estimation from Landsat", 12th International Symposium on Remote Sensing of Environment, Manila, Philippines, April 1978, 1415-1419.

[8] Ryerson, R.A., P. Mosher, V. Wallen and N. Stewart, "Three Tests of Agricultural Remote Sensing in Eastern Canada: Results, Problems and Prospects", Canadian Journal of Remote Sensing, 5(1) 1979, 53-66.

[9] Ryerson, R.A., P.N. Mosher and J. Harvie, 1980, "Potato Area Estimation Using Remote Sensing Methods", CCRS Users Manual 80-2, Energy, Mines and Resources, Ottawa, February 1980.

TABLE 1

SURVEY ESTIMATES OF THE TOTAL POTATO AREA FOR THE TEST REGION AND FOR THE PROVINCE OF NEW BRUNSWICK[1]

Survey and/or Estimate	Test Region Estimate	New Brunswick Estimate	Coefficient of Variation (%)
	(Acres)	(Acres)	
Satellite			
Uncorrected	42,354	N/A	N/A
Ratio	49,504	51,524	5.5
Regression	49,115	51,119	5.5
Mail Survey	N/A	50,800	N/A
A.E.S. Multiple Frame	N/A	53,854	5.2
O.P.Y.S.	47,203	49,129	4.59
Statistics Canada Published (Sept. 5)	N/A	52,000	N/A

[1] The test region, composed of the counties of Carleton, Madawaska, and Victoria, accounts for about 96.08% of the total potato area of New Brunswick.

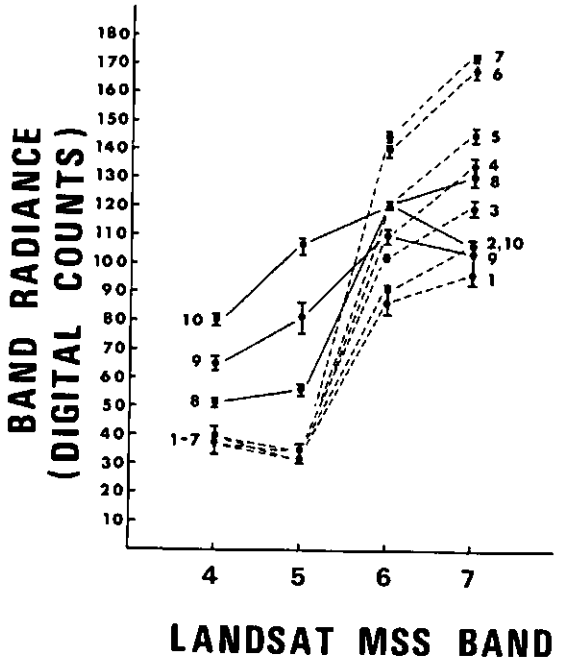


FIGURE 1 Comparison of Satellite Band Radiances for Various Crops.

Landsat MSS band radiance (August 8, 1975) for grains (1), buckwheat (2), pasture (3), hay (4), corn (5), peas (6), potatoes (7), broccoli (8), bare soil (9), weeds (10). The bands are numbered 4 to 7. The relative distances between the digital counts in each column indicate how separable the crops are for the given band.

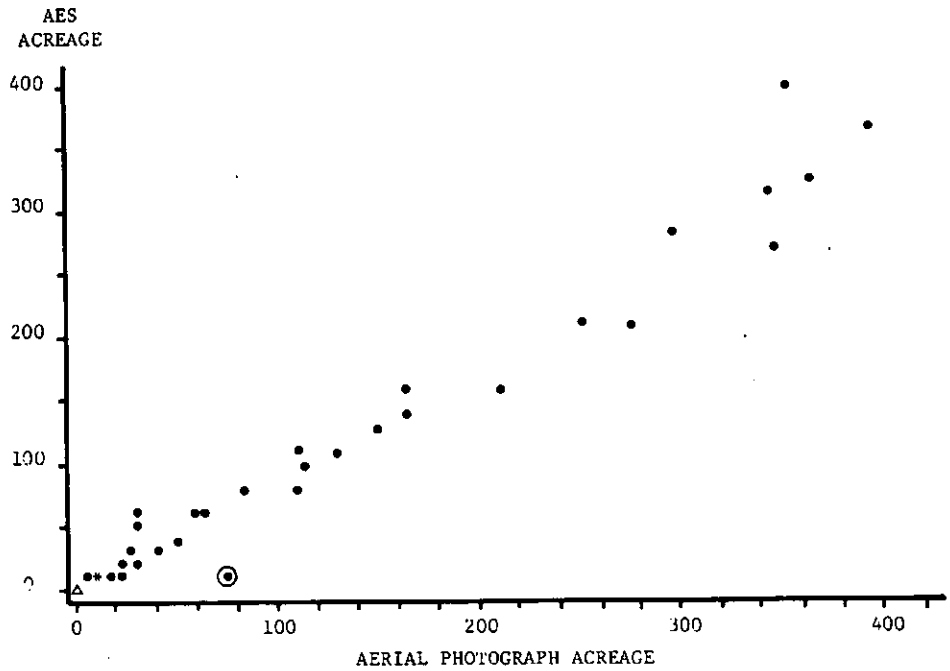


FIGURE 2

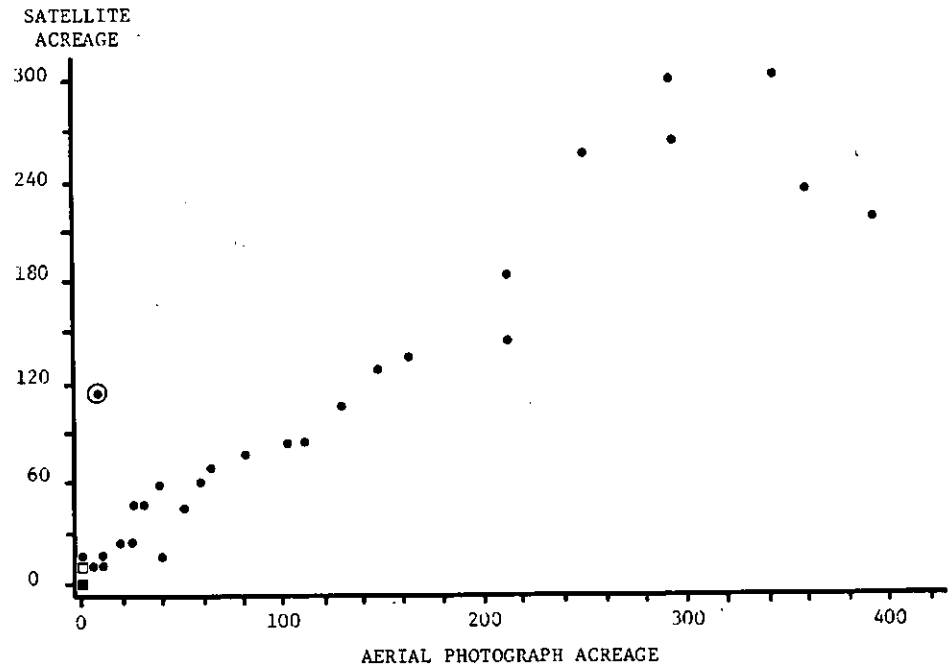


FIGURE 3

FIGURES 2 and 3 - Plots of reported A.E.S. and satellite potato acreages vs. low altitude aerial photograph acreages for sampled segments in New Brunswick. The circled observations represent outliers.
Legend: .=1 obs., □ =2 obs., *=3 obs., ■ =11 obs., and Δ =15 obs.

SAMPLING ON TWO OCCASIONS WITH PPSWOR

G.H. Choudhry and Jack E. Graham[1]

A theory of sampling on two occasions with unequal probabilities and without replacement is presented. Fellegi's (1963) method, which yields the same selection probabilities for a given unit on each occasion, is used to select the units for the rotation sample. The variances of composite estimators of the population total on the second occasion are developed. Numerical results are presented for small sample sizes and efficiency comparisons are made with a competing strategy.

1. INTRODUCTION

In surveys of a repetitive nature there are advantages to using a partial replacement sampling scheme both from the point of view of efficiency of estimation as well as reduction of respondent's burden. Essentially, after each sampling occasion a fraction of the units is rotated out of the sample and is replaced by a fresh subsample from the population. The literature abounds with discussions of sampling procedures and estimators when sampling on two or more occasions with equal probability. But of particular practical importance is the situation where units are selected on a given occasion with unequal probabilities. Thus, consider a finite population of N units $\{1,2,\ldots,N\}$ and two sampling occasions 1 (the previous occasion) and 2 (the current occasion). Let y_{1i} and y_{2i} denote the values of a characteristic y borne by the i-th unit on occasions 1 and 2 and let Y_1 and Y_2 denote the respective population totals. A size measure x_i is known for each of the units in the population.

Raj (1965) considered the following pps (probabilities proportional to size) sampling scheme: on the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values and with replacement

[1] G.H. Choudhry, Census & Household Survey Methods Division, Statistics Canada and Jack E. Graham, Carleton University.

(wr). On the second occasion a simple random sample s_1 of m units is selected from s without replacement (wor) and an independent pps sample s_2 of $u = n-m$ is selected wr from the entire population. Y_1 and Y_2 are then respectively estimated by

$$\hat{Y}_1 = \sum_s y_{1i}/(np_i) \text{ and } \hat{Y}^*_{2R} = Q^*\hat{Y}_{2u} + (1-Q^*)\hat{Y}'_2,$$

where $$\hat{Y}_{2u} = \sum_{s_2} y_{2i}/(up_i),\ \hat{Y}'_2 = \hat{Y}_1 + \sum_{s_1} (y_{2i} - y_{1i})/(mp_i),$$

and Q^* is a weight, $0 \leq Q^* \leq 1$.

The minimum variance of $\hat{Y}^*_{2R}$ was developed under the assumption that

$$v_{pps}(y_t) = \sum_{i=1}^{N} p_i(y_{ti}/p_i - Y_t)^2$$

is the same for t = 1 and 2.

The problem of sampling with ppswor on one occasion has attracted considerable attention in the literature. A major difficulty lies in the specification of feasible procedures which lead to specified probabilities at each and every draw. Fellegi (1963) has proposed a method such that the probability that unit i is selected on each of the n draws is p_i by determining n-1 sets of "working probabilities". This is an extremely desirable feature for rotating samples where it is essential that the usual pps estimator be unbiased for Y_2; this will not be true for any partial replacement design that does not feature a constant p_i for each of the n draws. The calculations inherent in Fellegi's scheme have, until recently, been prohibitive for $n > 2$. Choudhry (1981) has developed an iterative procedure for implementing Fellegi's scheme and prepared a computer program to evaluate the working probabilities when $n \leq 5$. Although the convergence is fast in terms of the number of iterations, the amount of computation increases at the rate N^n.

The program also computes the joint probabilities for the inclusion of both units i and j in the sample for variance calculation purposes.

Rao, Hartley and Cochran (1962) devised the "random group method" for selecting a sample with ppswor. The population of N units is split into n groups of sizes N_1, N_2, ..., N_n where $\Sigma N_h = N$ and a sample of one unit is drawn independently from each group with probabilities proportional to the p_i's. Ghangurde and Rao (1969) extended the random group method to sampling on two occasions. For simplicity, the N units were split into n groups each of size N/n (assumed to be an integer). On occasion 1, one unit is drawn from each random group as above, giving a sample s of n units. On the second occasion a simple random sample s_1 of m matched units is selected from the n units wor and an independent sample s_2 of u=n-m units is drawn from the whole population by the method used in obtaining s. They form a composite estimator $\hat{Y}'_{2G}$ of Y_2 and obtain its minimum variance under an optimum choice of the weight Q. The optimum value of $\lambda = m/n$ is then determined. The authors remarked that it would likely be more efficient to select s_2 from the N-n units in the population that are not included in s.

Chotai (1974) modified the Ghangurde-Rao (G-R) design on the second occasion; the n units in s are split at random into m groups of size n/m (assumed to be an integer). One unit is selected from each of the m groups with probability proportional to p_i, yielding a sample s_1. A sample s_2 is obtained as in the G-R method. The optimum variance of his composite estimator $\hat{Y}'_{2c}$ is derived, the optimum λ determined and relative efficiency comparisons of $\hat{Y}'_{2c}$ with respect to G-R's and Raj's optimal estimators are made. $\hat{Y}'_{2c}$ was found to be always more efficient than $\hat{Y}^*_{2R}$ and, in many cases, $\hat{Y}'_{2G}$ as well. A brief discussion of the case when n/m is not an integer is provided. It is worth noting that because λ is not a continuous function, the optimum λ should really be determined using integer programming methods. In what follows a sampling strategy is developed which often results in greater gains in efficiency over ppswor sampling than previously proposed schemes.

2. SAMPLING STRATEGY

2.1 SAMPLING PROCEDURE

From the population of N units, (1,2,...,N), select a sample of n+u units, $u < n$, draw by draw and without replacement using Fellegi's Method such that the probability of selecting the i-th unit at each draw is p_i, i=1,2,...N, $\Sigma\, p_i = 1$. On the first of the two occasions, the first n units are observed from the n+u selected; on the second occasion the first u units are dropped from the sample and the unused set of u units is rotated into the sample. Thus $m = n-u$ units are observed on both occasions. The n units observed on the first occasion are referred to as s, those units observed on both occasions as s_1 (where $s_1 \subset s$) and the set of unmatched units observed only on the second occasion as s_2. Note that Fellegi's scheme guarantees that the selection probabilities for a given unit i are the same on each draw and hence the same on both occasions. By restricting his attention to a sub-class of non-homogeneous linear model-design unbiased estimators, Chaudhuri (1980) has shown that the foregoing sampling scheme yields an optimal strategy. This is a further motivation for using Fellegi's method.

2.2 ESTIMATION THEORY

In what follows, composite estimators of Y_2, the current occasion total, are proposed and their variances determined using an indicator variable approach.

Let ${}_ra_i = 1$ if unit i, i=1,2,...,N, is selected at draw r, r=1,2,...,n+u and ${}_ra_i = 0$ otherwise. Since the expectation of ${}_ra_i$ is p_i, an unbiased estimator of the first occasion population total Y_1 is

$$\hat{Y}_1 = \frac{1}{n} \sum_{r=1}^{n} \sum_{i=1}^{N} {}_ra_i\, y_{1i}/p_i .$$

Then $$\hat{Y}'_2 = \hat{Y}_1 + \frac{1}{m} \sum_{r=u+1}^{n} \sum_{i=1}^{N} {}_r a_i \, (y_{2i} - y_{1i})/p_i$$

is an unbiased estimator of the second occasion total Y_2. An unbiased estimator of Y_2 based on the current observations is

$$\hat{Y}_2 = \frac{1}{n} \sum_{r=u+1}^{u+n} \sum_{i=1}^{N} {}_r a_i \, y_{2i}/p_i$$

A composite estimator of Y_2 is the weighted sum

$$\hat{Y}_{2c} = Q\hat{Y}'_2 + (1-Q)\hat{Y}_2,$$

where $0 \leq Q \leq 1$.

The variance of $\hat{Y}_{2c}$, $Var(\hat{Y}_{2c}) = Q^2 Var(\hat{Y}'_2) + (1-Q)^2 Var(\hat{Y}_2) + 2Q(1-Q)Cov(\hat{Y}'_2, \hat{Y}_2)$, is derived by using the following properties of the indicator variable ${}_r a_i$:

$$Var({}_r a_i) = p_i(1-p_i), \ (i=1,2,\ldots,N, \ r=1,2,\ldots,n+u),$$

$$Cov({}_r a_i, {}_t a_i) = -p_i^2, \ (r \neq t),$$

$$Cov({}_r a_i, {}_r a_j) = -p_i p_j, \ (i \neq j),$$

$$Cov({}_r a_i, {}_t a_j) = E({}_r a_i \cdot {}_t a_j) - p_i p_j, \text{ otherwise},$$

where E(.) denotes the expected value with respect to the probability design. Now $E({}_r a_i \cdot {}_t a_j) = P({}_r a_i \cdot {}_t a_j = 1) = P({}_r a_i = 1, {}_t a_j = 1)$ where P(.) denotes probability.

Let $\Sigma_{(k-2;\ i,j)}$ denote summation over all possible ordered (k-2)-tuples of different units $\{i_1, i_2, \ldots, i_{r-1}, i_{r+1}, \ldots, i_{k-2}, i_{k-1}\}$ included in the sample from the first k draws selected from the N-2 units in the set $\{1,2,\ldots,i-1, i+1,\ldots,j-1, j+1,\ldots,N\}$ such that the i-th unit is selected at draw r and the j-th unit at draw k. There are (N-2)(N-3)...(N-k+1) terms involved in the summation.

As in Fellegi (1963), let $\{p_i(\ell);\ i=1,2,\ldots,N\}$ be the set of "working probabilities" for selecting a unit at draw ℓ, $\ell=1,2,\ldots,n+u$. For draws k and r with $k > r$,

$$E({}_r a_i \cdot {}_k a_j) = \Sigma_{(k-2;i,j)}\ p_{i_1}(1)\ \frac{p_{i_2}(2)}{1-p_{i_1}(2)} \cdots \frac{p_{i_{r-1}}(r-1)}{1-\sum_{\ell=1}^{r-2} p_{i_\ell}(r-1)}$$

$$\times \frac{p_i(r)}{1-\sum_{\ell=1}^{r-1} p_{i_\ell}(r)} \times \frac{p_{i_{r+1}}(r+1)}{1-\sum_{\ell=1}^{r-1} p_{i_\ell}(r+1)-p_i(r+1)} \times$$

$$\cdots \times \frac{p_j(k)}{1-\sum_{\ell=1}^{r-1} p_{i_\ell}(k) - p_i(k) - \sum_{\ell=r+1}^{k-1} p_{i_\ell}(k)} .$$

Now

$$\mathrm{Var}\ (\hat{Y}_2') = \frac{1}{n^2}\ \mathrm{Var}\left[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i\right] + \frac{1}{m^2}\ \mathrm{Var}\left[\sum_{r=u+1}^{n}\sum_{i=1}^{N} {}_r a_i (y_{2i}-y_{1i})/p_i\right]$$

$$+ \frac{2}{mn}\ \mathrm{Cov}\left[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i,\ \sum_{r=u+1}^{n}\sum_{i=1}^{N} {}_r a_i (y_{2i} - y_{1i})/p_i\right]$$

Using the previously cited properties of the indicator variables ${}_r a_i$ it may be verified that

$$\frac{1}{n^2}\ \mathrm{Var}\left[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i\right] = \frac{1}{n}\ \sum_{i=1}^{N} p_i z_{1i}^2 + \frac{1}{n^2}\ \sum_{i\neq j}\sum P(i,j\varepsilon s) z_{1i} z_{1j} - Y_1^2 ,$$

$$\frac{1}{m^2}\ \mathrm{Var}\left[\sum_{r=u+1}^{n}\sum_{i=1}^{N} {}_r a_i (y_{2i} - y_{1i})/p_i\right] = \frac{1}{m}\ \sum_{i=1}^{N} p_i (z_{2i} - z_{1i})^2$$

$$+\frac{1}{m^2}\sum_{i\neq j}\sum P(i,j\ \varepsilon s_1)(z_{2i}-z_{1i})(z_{2j}-z_{1j})-(Y_2-Y_1)^2\ ,$$

where $z_{tj}=y_{tj}/p_j$, t=1,2 and n-u=m.

Also,

$$\frac{1}{mn}\,\mathrm{Cov}\left[\sum_{r=1}^{n}\sum_{i=1}^{N} {}_r a_i y_{1i}/p_i,\ \sum_{u+1}^{n}\sum_{1}^{N} {}_r a_i(y_{2i}-y_{1i})/p_i\right]$$

$$=\frac{1}{n}\sum_{1}^{N} p_i z_{1i}(z_{2i}-z_{1i})+\frac{1}{mn}\sum_{i\neq j}\sum P(i\varepsilon s,\ j\varepsilon s_1)z_{1i}(z_{2j}-z_{1j})-Y_1(Y_2-Y_1).$$

Combining the foregoing 3 terms gives

$$\mathrm{Var}(\hat{Y}_2')=\sum_{1}^{N} p_i\left[\frac{z_{2i}^2}{n}+(z_{2i}-z_{1i})^2\left(\frac{1}{m}-\frac{1}{n}\right)\right]+\sum_{i\neq j}\sum\left[\frac{P(i,j\varepsilon s)}{n^2}z_{1i}\,z_{1j}\right.$$

$$+\frac{P(i,j\ \varepsilon s_1)}{m^2}(z_{2i}-z_{1i})(z_{2j}-z_{1j})$$

$$\left.+\frac{2P(i\varepsilon s,j\varepsilon s_1)}{nm}z_{1i}\,(z_{2j}-z_{1j})\right]-Y_2^2\ . \qquad (1)$$

Also,

$$\mathrm{Var}(\hat{Y}_2)=\frac{1}{n}\sum_{i}p_i z_{2i}^2+\frac{1}{n^2}\sum_{i\neq j}\sum P(i,j\varepsilon s^*)z_{2i}z_{2j}-Y_2^2, \qquad (2)$$

where s* is the set of n units observed on the second occasion, and

$$\mathrm{Cov}(\hat{Y}_2',\hat{Y}_2)=\sum_{i\neq j}\sum\left[\frac{P(i\varepsilon s,\ j\varepsilon s^*)}{n^2}z_{1i}z_{2j}+\frac{P(i\varepsilon s_1,\ j\varepsilon s^*)}{nm}(z_{2i}-z_{1i})z_{2j}\right]$$

$$+ \frac{1}{n} \sum_i p_i z_{2i} \left(z_{2i} - \frac{u}{n} z_{1i}\right) - Y_2^2 \qquad (3)$$

Expressions (1), (2) and (3), when combined, yield $Var(\hat{Y}_{2c})$.

The optimum value of the weight Q which minimizes $Var(\hat{Y}_{2c})$ is

$$Q_{opt} = [Var(\hat{Y}_2)-Cov(\hat{Y}_2',\hat{Y}_2)]/[(Var(\hat{Y}_2')+Var(\hat{Y}_2)-2\ Cov(\hat{Y}_2',\ \hat{Y}_2)].$$

The corresponding minimum variance is

$$Var(\hat{Y}_{2c})=[Var(\hat{Y}_2').Var(\hat{Y}_2)-(Cov(\hat{Y}_2',\hat{Y}_2))^2]/[Var(\hat{Y}_2')+Var(\hat{Y}_2)-2\ Cov(\hat{Y}_2',\hat{Y}_2)].$$

An alternative composite estimator $\hat{Y}_{2c}^*$ of Y_2 is

$$\hat{Y}_{2c}^* = Q^*\hat{Y}_2' + (1-Q^*)\hat{Y}_{2u},$$

where

$$\hat{Y}_{2u} = \sum_{r=n+1}^{n+u} \sum_{i=1}^{N} {}_r a_i y_{2i}/(up_i).$$

The variance of $\hat{Y}_{2c}^*$ is found by combining (1) with

$$Var(\hat{Y}_{2u})= \frac{1}{u} \sum_i p_i z_{2i}^2 + \frac{1}{u^2} \sum_{i\neq j}\sum P(i,j\varepsilon s_2)z_{2i}z_{2j} - Y_2^2 ,$$

$$Cov(\hat{Y}_2',\ \hat{Y}_{2u}) = \frac{1}{nu} \sum_{i\neq j}\sum P(i\varepsilon\ s,j\varepsilon\ s_2)\ z_{1i}z_{2j}$$

$$+ \frac{1}{mu} \sum_{i\neq j}\sum P(i\varepsilon s_1,j\varepsilon s_2)(z_{2i} - z_{1i})\ z_{2j} - Y_2^2 .$$

2.3 SPECIAL CASE

As a check on the calculations, consider the case of simple random sampling without replacement.

Then $\hat{Y}_2' = N(\bar{y}_1 + (\bar{y}_{2m} - \bar{y}_{1m}))$ where $\bar{y}_1, \bar{y}_{1m}$ are, respectively, the sample means based on all the sampled units and all matched units on the first occasion, and $\bar{y}_{2m}$ is the sample mean based on all matched units on the second occasion. A direct evaluation gives

$$\mathrm{Var}(\hat{Y}_2') = N^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right) (S_1^2 - 2S_{12}) + \left(\frac{1}{m} - \frac{1}{N}\right) S_2^2 \right]$$

where, e.g.,

$$S_{12} = \sum_{i=1}^{N} (y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2)/(N-1).$$

This agrees with the result given by (1) with $p_i = 1/N$ and $P(i,j \in s) = n(n-1)/N(N-1)$ $(i \neq j)$.

Also, under simple random sampling, $\hat{Y}_2 = N\bar{y}_2$ (where $\bar{y}_2$ is the sample mean based on all n sampled units on the second occasion) with variance

$$\mathrm{Var}(\hat{Y}_2) = N(N-n)S_2^2 .$$

An evaluation of $\mathrm{Var}(\hat{Y}_2)$ from (2) gives the same result. Finally,

$$\mathrm{Cov}(\hat{Y}_2', \hat{Y}_2) = -NS_2^2$$

from either a direct evaluation or from (3). Similarly, $\mathrm{Var}(\hat{Y}_{2c}^*)$ may also be

checked.

3. NUMERICAL EXAMPLES

The composite estimators $\hat{Y}_{2c}$ and $\hat{Y}^*_{2c}$ with their optimum Q and Q* values which minimize their respective variances are compared in efficiency with the pps estimator $\hat{Y}_2$ which is based on the current occasion information only. Because closed forms for $Var(\hat{Y}_{2c})$ and $Var(\hat{Y}^*_{2c})$ are not available to permit analytic comparisons to be made, small populations of variate values were employed to affect these contrasts. (The populations studied were necessarily small, like those one might encounter in stratified sampling, since differential effect of sampling with and without replacement is evident only when the sampling fractions are not negligible). Four rotation sampling plans were applied to each population: (n,m) = (2,1), (3,2), (3,1) and (4,3) where n is the number of units in the sample on each occasion and m is the number of units in the sample common between the two occasions. Two of the populations are given in Murthy (1967) where his single population of 34 villages was subdivided into two populations of sizes 16 and 17 (one outlier unit being discarded). The size measure characteristic is x = cultivated acreage in 1961 with y_1 and y_2 being the acreage under wheat in 1963 and 1964 respectively. A third population is a set of 14 farms in the province of Saskatchewan with x = 1980 farm acreage and y_1 and y_2 the 1980 and 1981 cropland acreages respectively. Two additional real data sets relating to populations of sizes 15 and 16 respectively are also analyzed.

Table 1 reports the relative efficiencies of $\hat{Y}_{2c}$ and $\hat{Y}^*_{2c}$ with respect to $\hat{Y}_2$ for each of these 5 populations and 4 sampling plans. A crucial parameter in each comparison is the correlation ρ_z between $z_{1i} = y_{1i}/p_i$ and $z_{2i} = y_{2i}/p_i$:

$$\rho_z = \frac{\sum_{i=1}^{N} p_i z_{1i} z_{2i} - Y_1 Y_2}{\sqrt{\sum_{i=1}^{N} p_i z_{1i}^2 - Y_1^2}\ \sqrt{\sum_{i=1}^{N} p_i z_{2i}^2 - Y_2^2}} .$$

The populations studied yielded ρ_z values ranging from 0.940 to 0.213. The optimum Q and Q* values are also cited. An investigation of the efficiencies of $\hat{Y}_{2c}$ and $\hat{Y}^*_{2c}$ under non-optimum choice of Q and Q* is planned.

We note the following from these empirical studies: (1) The optimum Q values tend to be larger when ρ_z is large and as ρ_z decreases, the optimum Q tends to decrease in both $\hat{Y}_{2c}$ and $\hat{Y}^*_{2c}$. (2) The optimum Q value for $\hat{Y}^*_{2c}$ always exceeds that for $\hat{Y}_{2c}$. (3) As ρ_z decreases, the efficiency of $\hat{Y}_{2c}$ with respect to $\hat{Y}_2$ decreases (as expected), approaching unity as a lower bound under an optimum choice of Q. On the other hand, no such distinct behaviour for $\hat{Y}^*_{2c}$ is evident since $\mathrm{Var}(\hat{Y}^*_{2c})$ is not a monotone function of ρ_z. For small ρ_z values, small efficiency gains and losses relative to $\hat{Y}_2$ are both recorded. (4) $\hat{Y}^*_{2c}$ is more efficient than $\hat{Y}_{2c}$ for larger ρ_z values whereas $\hat{Y}_{2c}$ is more efficient than $\hat{Y}^*_{2c}$ for smaller ρ_z values. (5) If $\lambda = m/n$ is small, e.g., the (3,1) plan, then large efficiency gains using $\hat{Y}^*_{2c}$ in preference to $\hat{Y}_2$ result for large ρ_z's. For smaller ρ_z values, the (4,3) plan yields the largest gains using $\hat{Y}_{2c}$; the other three schemes give about the same gains.

It is worth remarking that even if one has a good correlation between the y_{1i} and y_{2i} values, composite estimation using $\hat{Y}^*_{2c}$ can still lead to efficiency losses compared to the use of the pps estimator $\hat{Y}_2$ based on current occasion data only. The critical factor is the correlation ρ_z between the z_{1i} and the z_{2i} values. One cannot lose using $\hat{Y}_{2c}$ under an optimum choice of Q for $\hat{Y}_{2c} = \hat{Y}_2$ with Q = 0.

Table 2 provides the relative efficiencies of $\hat{Y}_{2c}$ and $\hat{Y}^*_{2c}$ in the ppswor design with the estimator $\hat{Y}^*_{2R}$ used by Raj (1965) in his ppswr design described earlier. For more valid comparisons, it was not assumed that $V_{pps}(y_t)$ was the same for occasions t = 1 and 2; the optimum Q* values for the given (n,m) combinations were utilized. In all cases, as expected, the estimators $\hat{Y}_{2c}$ and $\hat{Y}^*_{2c}$ in the wor design are more efficient than $\hat{Y}^*_{2R}$ in Raj's design. As n increases for a given ρ_z, the efficiency gain using the wor strategy increases. Finally, we note that Raj realized efficiency gains compared with no matching only when $\rho_z > 0.5$ whereas efficiency gains always resulted using $\hat{Y}_{2c}$ for any ρ_z in the wor situation.

ACKNOWLEDGEMENT

We wish to thank Professor J.N.K. Rao for suggestions and comments made during the course of this study. The referee's helpful comments are also much appreciated.

REFERENCES

[1] CHAUDHURI, A. (1980), "On Optimal and Related Strategies for Sampling on Two Occasions with Varying Probabilities", Preprint, Indian Statistical Institute.

[2] CHOTAI, J. (1974), "A Note on the Rao-Hartley-Cochran Method for PPS Sampling Over Two Occasions", Sankhya, 36, C, 173-180.

[3] CHOUDHRY, G.H. (1981), "Construction of Working Probabilities and Joint Selection Probabilities for Fellegi's PPS Sampling Scheme", Survey Methodology, 7, No.1, 93-108.

[4] FELLEGI, I.P. (1963), "Sampling with and Without Replacement: Rotating and Non-rotating Samples", J. Amer. Stat. Ass., 58, 183-201.

[5] GHANGURDE, P.D. and Rao, J.N.K. (1969), "Some Results on Sampling Over Two Occasions", Sankhya, 31, A, 463-472,

[6] MURTHY, M.N. (1967), "Sampling Theory and Methods", Calcutta: Statistical Publishing Society.

[7] RAJ, D. (1965), "On Sampling Over Two Occasions with Propability Proportional to Size", Ann. Math. Statist., 36, 327-330.

[8] RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962), "On a Simple Procedure of Unequal Probability Sampling without Replacement", J.R. Statist. Soc. B., 24, No. 2, 482-491.

TABLE 1

Efficiencies of composite wor estimators relative to ppswor estimator

Population	N	(n,m)	ρ_z	Q_{opt}	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}_{2c})}$	Q^*_{opt}	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}^*_{2c})}$
Murthy	17	(2,1)	0.940	0.443	1.337	0.640	1.476
set 1		(3,2)		0.456	1.230	0.738	1.367
		(3,1)		0.419	1.524	0.545	1.656
		(4,3)		0.462	1.188	0.793	1.309
Murthy	16	(2,1)	0.867	0.377	1.205	0.615	1.364
set 2		(3,2)		0.402	1.151	0.727	1.296
		(3,1)		0.336	1.282	0.499	1.459
		(4,3)		0.431	1.191	0.786	1.255
Acreages	14	(2,1)	0.546	0.181	1.083	0.466	0.927
		(3,2)		0.215	1.070	0.646	0.927
		(3,1)		0.137	1.092	0.295	0.933
		(4,3)		0.279	1.117	0.736	0.931
Data set 1	15	(2,1)	0.392	0.113	1.019	0.506	1.013
		(3,2)		0.140	1.017	0.670	1.013
		(3,1)		0.082	1.020	0.340	1.013
		(4,3)		0.330	1.170	0.752	1.012
Data set 2	16	(2,1)	0.213	0.061	1.007	0.451	0.898
		(3,2)		0.078	1.007	0.636	0.896
		(3,1)		0.042	1.007	0.280	0.908
		(4,3)		0.285	1.142	0.730	0.901

TABLE 2

Efficiencies of composite wor estimators relative to Raj's composite estimator

Population	N	(m,m)	ρ_z	$Var(\hat{Y}^*_{2R})/Var(\hat{Y}_{2c})$	$Var(\hat{Y}^*_{2R})/Var(\hat{Y}^*_{2c})$
Murthy set 1	17	(2,1)	0.940	1.038	1.146
		(3,2)		1.127	1.252
		(3,1)		1.215	1.320
		(4,3)		1.248	1.375
Murthy set 2	16	(2,1)	0.867	1.001	1.133
		(3,2)		1.106	1.246
		(3,1)		1.130	1.287
		(4,3)		1.309	1.380
Acreages	14	(2,1)	0.546	1.244	1.065
		(3,2)		1.330	1.151
		(3,1)		1.351	1.154
		(4,3)		1.507	1.257
Data set 1	15	(2,1)	0.392	1.095	1.089
		(3,2)		1.197	1.192
		(3,1)		1.200	1.192
		(4,3)		1.522	1.317
Data set 2	16	(2,1)	0.213	1.197	1.068
		(3,2)		1.293	1.152
		(3,1)		1.280	1.154
		(4,3)		1.589	1.254

SURVEY METHODOLOGY

1982 Vol. 8 Nos. 1 & 2

A Journal produced by Methodology Staff, Statistics Canada

CONTENTS