

CS12-001/9-1



Statistics Canada Statistique Canada

SURVEY METHODOLOGY

1983
Volume 9
Number 1

UNIVERSITY OF P.E.I.

LIBRARY USE ONLY

A Journal produced by
Statistics Canada

Canada

SURVEY METHODOLOGY

1983

Vol. 9

No. 1

A Journal produced by Statistics Canada.

CONTENTS

The Redesign of a Survey to Measure Commodity Origin and Destination Movements by the For-Hire Trucking Industry in Canada ROBERT LUSSIER and STEVEN MOZES.....	1
The Methodology of the Canadian Air Scheduled International Passenger Origin and Destination Estimation System GREG HUNTER and LISA DIPIÉTRO.....	27
Some Aspects of Quality of Cancer Mortality and Incidence Statistics D. BINDER and A. MALHOTRA.....	50
Estimating Monthly Gross Flows in Labour Force Participation STEPHEN E. FIENBERG and ELIZABETH A. STASNY.....	77
Redesign of the Niagara Tender Fruit Objective Yield Survey J. KOVAR.....	103
A Timely and Accurate Potato Acreage Estimate from Landsat; Results of a Demonstration R.A. RYERSON, J.-L. TAMBAY, R.J.BROWN, L.A. MURRAY and B. MCLAUGHLIN.....	119
Sampling on Two Occasions with PPSWOR G.H. CHOUDHRY and JACK E. GRAHAM.....	139

8-3200-501

Reference No.

Z - 079

ISSN: 0714-0045

SURVEY METHODOLOGY

1983

Vol. 9

No. 1

A Journal produced by Statistics Canada.

Editorial Board:	G.J.C. Hole	
	C. Patrick	
	R. Platek	- Chairman
	M.P. Singh	- Editor
	P.F. Timmons	
	H. Lee	- Assistant Editor

Editorial Policy:

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

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 CANADA¹

Robert Lussier and Steven Mozes²

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 selected 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 serious 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 implementation 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 to the 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 Survey³.

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 Survey⁴ 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 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:

$$n_{1h} = \frac{N_h^2 S_h^2}{N_h S_h^2 + Y_h^2 (C.V._h)^2}$$

where n_{1h} : 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 possible⁵.

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 through 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 shipment 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 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 SYSTEM¹

Greg Hunter and Lisa DiPiétro²

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 revenue³ 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 i^{th} 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:

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

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 i^{th} 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 i^{th} 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 j^{th} 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 j^{th} origin-destination pair, from the revenue passenger O & D sample survey.

d_j is the sample number of international passengers in the j^{th} origin-destination pair, from the revenue passenger O & D sample survey.

a_j is the number of passengers assigned to the j^{th} origin-destination pair.

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

$$\text{is } a_j = \frac{d_j}{D_i} \times A_i \quad j \in i$$

$$\text{where } \sum_{j \in i} d_j = D_i$$

$$\text{and } \sum_{j \in i} a_j = A_i$$

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:

<u>Sample Number of Passengers</u>	<u>Estimate of the Number of Passengers</u>	<u>DOD's</u>
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:

<u>Code</u>	<u>Denotes</u>
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:

<u>Truncated DOD's</u>	<u>Proportions</u>	<u>Assigned Passengers(^a_j)</u>
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. $(100\% - 42.7\%) \times 10\%$), 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:

<u>Domains of Interest</u>	<u>Passenger Count</u>
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 i th 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

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

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

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

and the coefficient of variation, as

$$c\hat{v}(\hat{d}) = \sqrt{\hat{\text{var}}(\hat{d})/\hat{d}} .$$

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

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

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 j^{th} domain of interest would be f_j as developed in equation (1) above. f_j would replace f in the formula for $\hat{\text{var}}(\hat{d})$ in equation (2) above in order to yield the formula for the variance of the total market estimates:

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

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

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

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. Malhotra¹

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 only².

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 neoplasms³, 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 error 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 of 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-8⁴ 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 indication 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 provinces 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 were 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 prostate). 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 found⁵. 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 not carried 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 system⁶.

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 registries⁷. 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 separate 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	1969 - 1973	89.6	100.0	4.1
Census Division	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	1969 - 1973	99.8	100.0	99.6
Census Division	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 4 th 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 Melanoma 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 4 th 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 PARTICIPATION¹

Stephen E. Fienberg and Elizabeth A. Stasny²

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 use 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 probability 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 purpose 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 4x4 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 4x4 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_{00} , the entry in the 00 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:

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

The proportionally adjusted outflows are

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

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 $s = 1, 2, \dots, S$. 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:

$$w_{t-1}^s = N_{t-1}^s / r_{t-1,t}^s \text{ and } w_t^s = N_t^s / r_{t-1,t}^s. \quad (3)$$

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:

$$\begin{aligned} \min(W_{t-1}^s, W_t^s) n_{EE}^s &= \min [(N_{t-1}^s / r_{t-1,t}^s), (N_t^s / r_{t-1,t}^s)] n_{EE}^s \\ &= \min (N_{t-1}^s, N_t^s) n_{EE}^s / r_{t-1,t}^s \\ &= \min (N_{t-1}^s, N_t^s) f_{EE}^s \end{aligned} \quad (4)$$

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

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

$$\begin{aligned} \max(0, W_t^s - W_{t-1}^s) n_{EE}^s &= \max[0, (N_t^s - N_{t-1}^s / r_{t-1,t}^s)] n_{EE}^s \\ &= \max(0, N_t^s - N_{t-1}^s) n_{EE}^s / r_{t-1,t}^s. \end{aligned} \quad (5)$$

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:

$$\begin{aligned} \max(0, N_t^s - N_{t-1}^s) \{ (n_{EE}^s / r_{t-1,t}^s) + (n_{UE}^s / r_{t-1,t}^s) + (n_{NE}^s / r_{t-1,t}^s) \} \\ &= \max(0, N_t^s - N_{t-1}^s) n_{+E}^s / r_{t-1,t}^s \\ &= \max(0, N_t^s - N_{t-1}^s) f_{+E}^s \end{aligned} \quad (6)$$

where f_{+E}^s = 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			
		E	U	N	O
Month $t - 1$	E	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{EE}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{EU}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{EN}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{E+}^s$
	U	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{UE}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{UU}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{UN}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{U+}^s$
	N	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{NE}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{NU}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{NN}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{N+}^s$
	O	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+E}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+U}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+N}^s$	0
		$\sum_{s=1}^S N_t^s f_{+E}^s$	$\sum_{s=1}^S N_t^s f_{+U}^s$	$\sum_{s=1}^S N_t^s f_{+N}^s$	

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_{+E}^S , f_{+U}^S , f_{+N}^S , f_{E+}^S , f_{U+}^S , and f_{N+}^S 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 $r_{t-1,t}^s$ 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_{t-1}^s 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 $s = 1, 2, \dots, S$ and

0 = outside the population of interest.

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

$$N_t^s = N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u}). \quad (7)$$

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

$$w_{t-1}^s = N_{t-1}^s / r_{t-1,t}^s \text{ and } w_t^s = N_t^s / (N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})) / r_{t-1,t}^s. \quad (8)$$

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 2x2 matrix formed by collapsing the first three rows and columns of the 4x4 gross flow matrix used in the preceding subsection.

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

$$\begin{aligned} \min(W_{t-1}^s, W_t^s) r_{t-1,t}^s &= \min \{ N_{t-1}^s / r_{t-1,t}^s, [N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})] / r_{t-1,t}^s \} r_{t-1,t}^s \\ &= \min [N_{t-1}^s, N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})] \\ &= N_{t-1}^s + \min [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]. \end{aligned} \quad (9)$$

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

$$\begin{aligned} \max(0, W_t^s - W_{t-1}^s) r_{t-1,t}^s &= \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u}) / r_{t-1,t}^s] r_{t-1,t}^s \\ &= \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]. \end{aligned} \quad (10)$$

The entry for the in-population to out-of-population, or outflow, cell is found similarly. Thus, the 2x2 gross flow matrix is as follows:

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

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

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 $\sum_{u \neq s} (m_{u,s} - m_{s,u})$ for each stratum s that grew as a result of movements between strata and from outside the population of interest. The summation, $\sum_{u \neq s} (m_{u,s} - m_{s,u})$, 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

$$\begin{aligned} \sum_{u \neq A} (m_{u,A} - m_{A,u}) + \sum_{u \neq B} (m_{u,B} - m_{B,u}) &= m_{0,A} - m_{A,0} + m_{B,A} - m_{A,B} + m_{C,A} - m_{A,C} \\ &\quad + m_{0,B} - m_{B,0} + m_{A,B} - m_{B,A} + m_{C,B} - m_{B,C} \\ &= m_{0,A} - m_{A,0} + m_{0,B} - m_{B,0} + m_{C,A} - m_{A,C} + m_{C,B} - m_{B,C} \end{aligned} \quad (11)$$

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 $\sum_{u \neq s} (m_{u,s} - m_{s,u})$ 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

$$\sum_{u \neq s} (m_{u,s} - m_{s,u}) > m_{0,s} - m_{s,0} \quad (12)$$

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 2x2 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 4x4 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 outflow 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 $i = 0, 1, \dots, 6$.

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

$P(\text{observing } 0 \text{ months employed out of } 6-k \text{ months responding})$

$$= \frac{\sum_{j=1}^k \pi_j}{\binom{6}{j}}, \quad \text{for } k = 0, 1, \dots, 5. \quad (13)$$

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. and 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 4x4 Labour Force Gross Flow Matrix, Technical Reprot, Statistics Canada, 1983.

Table 1

		Months Employed						
a)		0	1	2	3	4	5	6
Months of Data Present	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					
		Months Unemployed						
b)		0	1	2	3	4	5	6
Months of Data Present	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. Kovar¹

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.

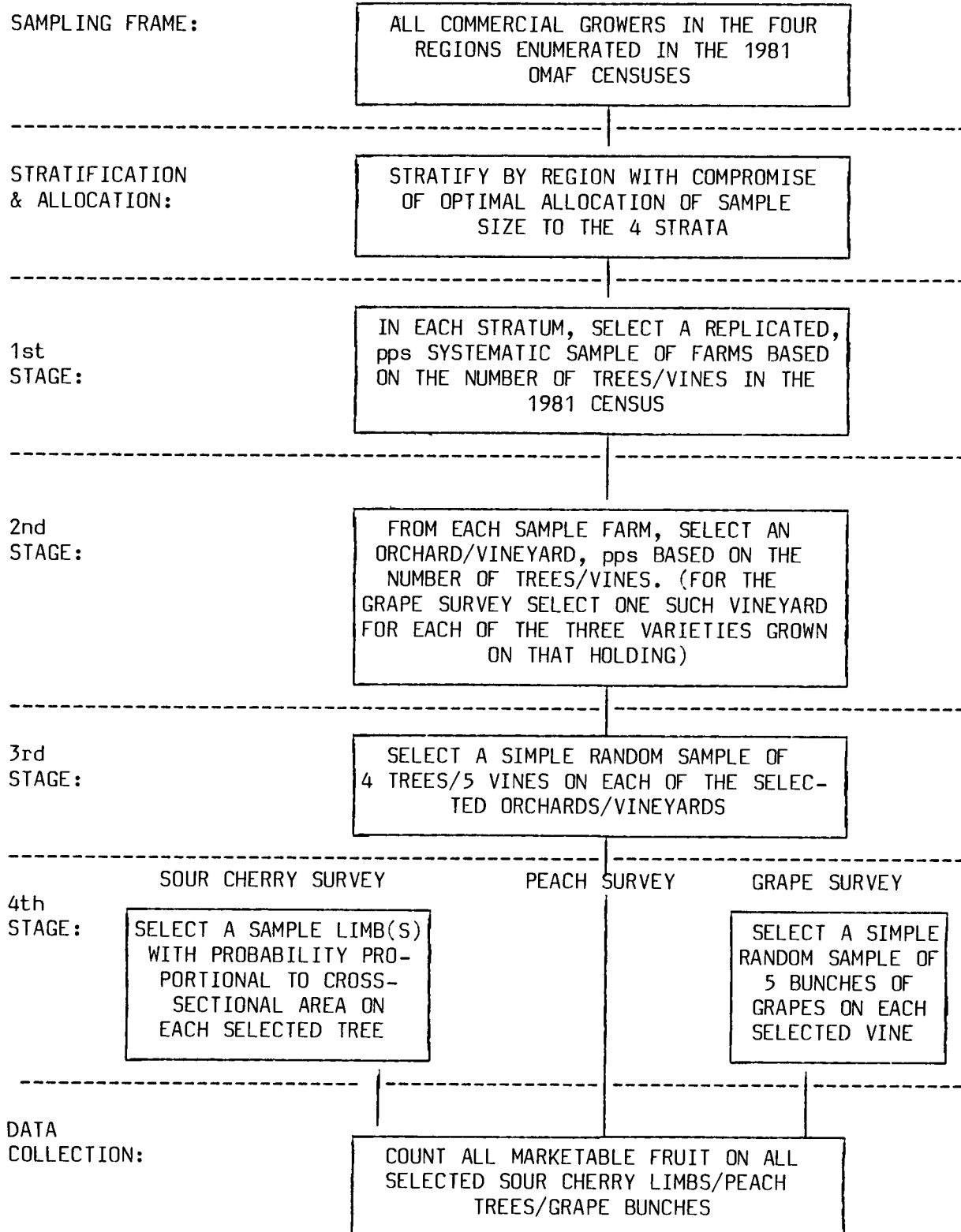
2. OVERVIEW OF THE SAMPLE DESIGN

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 estimated 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. Catharines 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 obtain 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 at 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

$$\hat{y}_\tau = \hat{y}_{\tau\ell} / p_\ell \quad (5.1.1)$$

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

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

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,

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

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

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

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 same 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

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

N_{arvfb}^C = 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_{arvfb}^{83} = 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_{arvf}^{83} = 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_{arf}^{81} = 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_a^{81} = the 1981 Census count of all trees/vines in region a (as per Table 4)

Table 4: 1981 Census Counts of Trees/Vines (N_a^{81})
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

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

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

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

in the case of peaches, sour cherries, and total grapes of all varieties, and by

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

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

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

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

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

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].

$$\begin{aligned}\hat{V}(\hat{R}_a) &= \frac{1}{\hat{\chi}_a^2} \{V(\hat{Y}_a) - 2 \hat{R}_a \text{Cov}(\hat{Y}_a, \hat{X}_a) + \hat{R}_a^2 V(\hat{X}_a)\} \\ &= \left\{ \frac{D_{ya}}{S_{xa}} - \frac{S_{ya} D_{xa}}{S_{xa}^2} \right\}^2\end{aligned}\quad (5.3.7)$$

with

$$\begin{aligned}S_{ya} &= \hat{Y}_{a1} + \hat{Y}_{a2} \\ D_{ya} &= \hat{Y}_{a1} - \hat{Y}_{a2}, \text{ etc}\end{aligned}\quad (5.3.8)$$

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

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

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

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

with its coefficient of variation estimated by

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

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,

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

with a coefficient of variation estimated by

$$CV(\hat{Y}) = \frac{\left\{ \sum_{a=1}^4 (\hat{Y}_a^T)^2 V(\hat{R}_a) \right\}^{\frac{1}{2}}}{\hat{Y}^T} \times 100\% \quad (5.5.2)$$

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 DEMONSTRATION¹

R.A. Ryerson, J.-L. Tambay, R.J. Brown

and

L.A. Murphy, B. McLaughlin²

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 sensing 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 expense 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 corn³ 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 than 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 $\sin x / x$ 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 schedule.

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:

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

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

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 j th selected segment in the i th 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 , $h=1, \dots, L$;

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

$\hat{R}$ is an estimate of Y/X ; and,

$\hat{B} = \text{cov}(\hat{Y}, \hat{X}) / \text{var}(\hat{X})$ is the linear regression coefficient.

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

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

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

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

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

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

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 correlations⁴ of .991 and .968 respectively⁵. 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 all 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 photograph 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 BRUNSWICK¹

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.

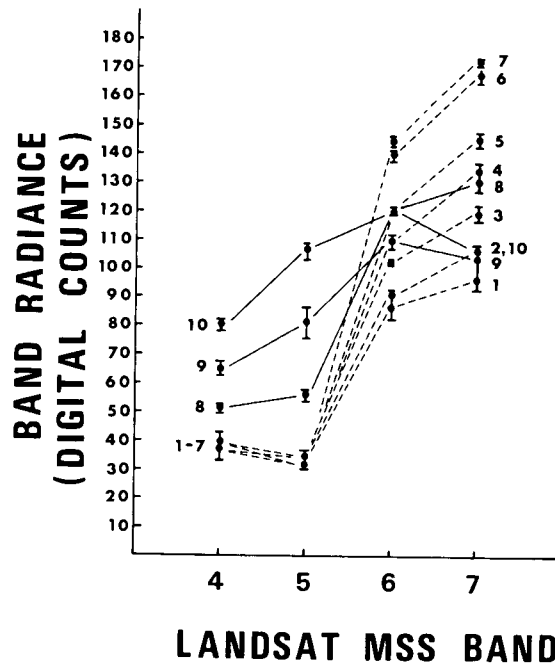
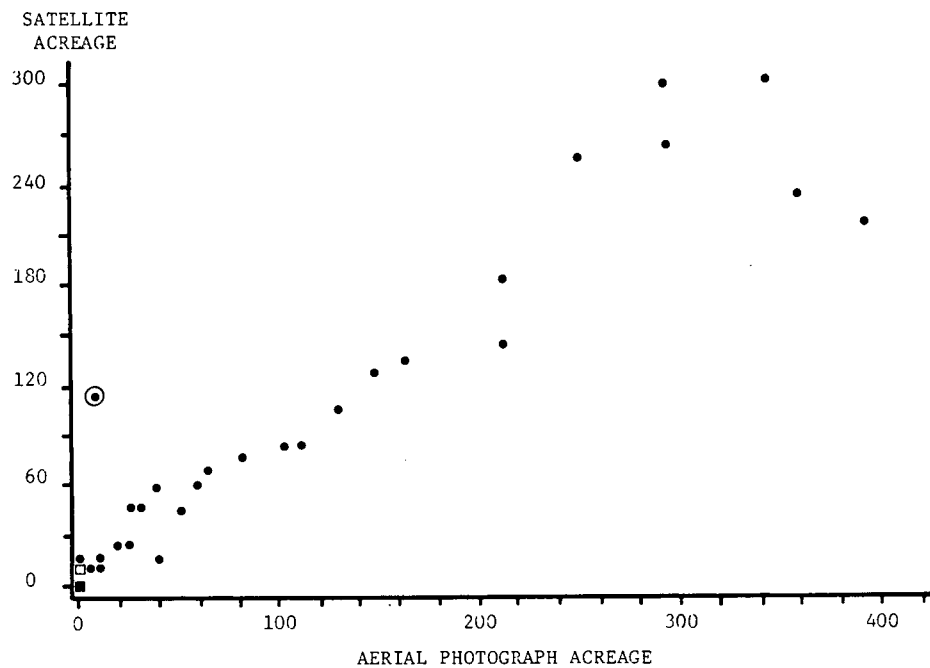
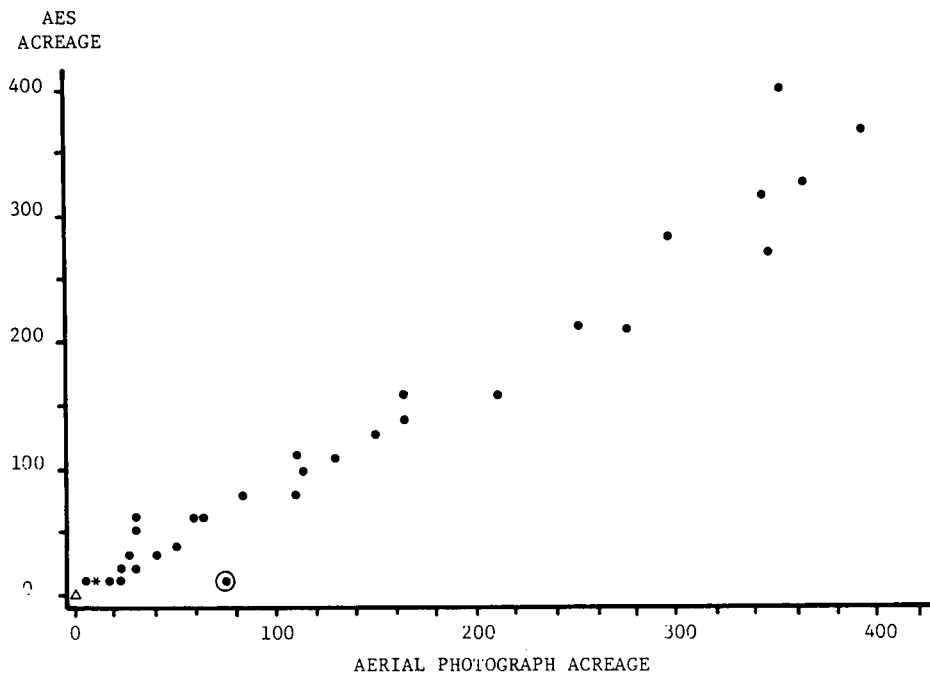


FIGURE 1 Comparison of Satellite Band Radiances for Various Crops.

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.



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. Graham¹

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 $\{1, 2, \dots, N\}$ 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 $N_1, N_2, \dots, N_n$ where $\sum N_h = N$ 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, $(1, 2, \dots, N)$, 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 , $i=1, 2, \dots, N$, $\sum p_i = 1$. 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 ${}_r a_i = 1$ if unit i , $i=1, 2, \dots, N$, is selected at draw r , $r=1, 2, \dots, n+u$ and ${}_r a_i = 0$ otherwise. Since the expectation of ${}_r a_i$ is p_i , an unbiased estimator of the first occasion population total Y_1 is

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

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

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

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

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

$$\text{Var}(a_{ri}) = p_i(1-p_i), (i=1,2,\dots,N, r=1,2,\dots,n+u),$$

$$\text{Cov}(a_{ri}, a_{ti}) = -p_i^2, (r \neq t),$$

$$\text{Cov}(a_{ri}, a_{rj}) = -p_i p_j, (i \neq j),$$

$$\text{Cov}(a_{ri}, a_{tj}) = E(a_{ri} \cdot a_{tj}) - p_i p_j, \text{ otherwise,}$$

where $E(\cdot)$ denotes the expected value with respect to the probability design.

$$\text{Now } E(a_{ri} \cdot a_{tj}) = P(a_{ri} \cdot a_{tj} = 1) = P(a_{ri} = 1, a_{tj} = 1)$$

where $P(\cdot)$ denotes probability.

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

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

$$E(a_{r i k} \cdot a_{j k}) = \sum_{(k-2; i, j)} p_{i_1}^{(1)} \frac{p_{i_2}^{(2)}}{1 - p_{i_1}^{(2)}} \cdots \frac{p_{i_{r-1}}^{(r-1)}}{1 - \sum_{\ell=1}^{r-2} p_{i_\ell}^{(r-1)}} \\ \times \frac{p_i^{(r)}}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}^{(r)}} \times \frac{p_{i_{r+1}}^{(r+1)}}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}^{(r+1)} - p_i^{(r+1)}} \times \\ \cdots \times \frac{p_j^{(k)}}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}^{(k)} - p_i^{(k)} - \sum_{\ell=r+1}^{k-1} p_{i_\ell}^{(k)}} .$$

Now

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

Using the previously cited properties of the indicator variables $r a_i$ it may be verified that

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

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

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

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

Also,

$$\begin{aligned} & \frac{1}{mn} \text{Cov} \left[\sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i, \sum_{u=1}^n \sum_{j=1}^N r a_j (y_{2j} - y_{1j}) / p_j \right] \\ &= \frac{1}{n} \sum_{i=1}^N p_i z_{1i} (z_{2i} - z_{1i}) + \frac{1}{mn} \sum_{i \neq j} \sum P(i \in s, j \in s_1) z_{1i} (z_{2j} - z_{1j}) - Y_1 (Y_2 - Y_1). \end{aligned}$$

Combining the foregoing 3 terms gives

$$\begin{aligned} \text{Var}(\hat{Y}_2') &= \sum_{i=1}^N p_i \left[\frac{z_{2i}^2}{n} + (z_{2i} - z_{1i})^2 \left(\frac{1}{m} - \frac{1}{n} \right) \right] + \sum_{i \neq j} \left[\frac{P(i, j \in s)}{n^2} z_{1i} z_{1j} \right. \\ &\quad + \frac{P(i, j \in s_1)}{m^2} (z_{2i} - z_{1i})(z_{2j} - z_{1j}) \\ &\quad \left. + \frac{2P(i \in s, j \in s_1)}{nm} z_{1i} (z_{2j} - z_{1j}) \right] - Y_2^2. \end{aligned} \quad (1)$$

Also,

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

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

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

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

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

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

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

The corresponding minimum variance is

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

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

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

where

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

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

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

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

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

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

$$\text{Var}(\hat{Y}'_2) = N^2 \left[\left(\frac{1}{m} - \frac{1}{N} \right) (S_1^2 - 2S_{12}) + \left(\frac{1}{m} - \frac{1}{N} \right) S_2^2 \right]$$

where, e.g.,

$$S_{12} = \frac{N}{\sum_{i=1}^N} (y_{1i} - \bar{y}_1) (y_{2i} - \bar{y}_2) / (N-1).$$

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

$$\text{Var}(\hat{Y}_2) = N(N-n)S_2^2.$$

An evaluation of $\text{Var}(\hat{Y}_2)$ from (2) gives the same result. Finally,

$$\text{Cov}(\hat{Y}'_2, \hat{Y}_2) = -NS_2^2$$

from either a direct evaluation or from (3). Similarly, $\text{Var}(\hat{Y}_{2c}^*)$ 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 $\text{Var}(\hat{Y}_{2c})$ and $\text{Var}(\hat{Y}_{2c}^*)$ 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 $\text{Var}(\hat{Y}_{2c}^*)$ 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 ppswr 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 set 1	17	(2,1)	0.940	0.443	1.337	0.640	1.476
		(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 set 2	16	(2,1)	0.867	0.377	1.205	0.615	1.364
		(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	$\text{Var}(\hat{Y}_{2R}^*)/\text{Var}(\hat{Y}_{2c})$	$\text{Var}(\hat{Y}_{2R}^*)/\text{Var}(\hat{Y}_{2c}^*)$
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

The Role of the Questionnaire in Survey Design R. PLATEK and D. ROYCE	1
Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey J.D. DREW, M.P. SINGH and G.H. CHOUDHRY	17
Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey ELIZABETH CLAYTON PAUL and MURRAY LAWES	48
Rotation Group Bias in the LFS Estimates P.D. GHANGURDE	86
Computerization of Complex Survey Estimates M.A. HIDIROGLOU	112

5. FORMULES D'ESTIMATION

5.1 Estimations du nombre de fruits par arbre ou par vigne

On représente par y_τ le nombre total de fruits commercialisables sur un arbre (vigne) τ . Dans l'enquête sur les pêches, on estime y_τ par $\hat{y}_\tau$ nombre total de pêches commercialisables que l'on compte sur un arbre-échantillon τ . Dans l'enquête sur les griottes, on estime y_τ de la façon suivante :

$$\hat{y}_\tau = \hat{y}_{\tau\ell} / p_\ell \quad (5.1.1)$$

où $\hat{y}_{\tau\ell}$ est le nombre total de griottes commercialisables que l'on compte sur une ou des branches-échantillon ℓ de l'arbre-échantillon τ choisi;

et p_ℓ représente les probabilités de sortie de la ou des branches-échantillon ℓ . Enfin, dans l'enquête sur les raisins, on estime y_τ comme suit:

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

où N_τ est le nombre total de grappes de raisins sur la vigne-échantillon τ ;

n_τ est le nombre de grappes de raisins qui ont été comptées sur la vigne-échantillon τ (généralement, $n_\tau = 5$);

et $\hat{y}_{\tau\ell}$ est le nombre de raisins sur une grappe ℓ de la vigne-échantillon τ .

5.2 Estimations, par région, du nombre de fruits par échantillon répété

On représente par $\hat{Y}_{ar}$ le nombre total estimé de fruits commercialisables dans un échantillon répété r d'une région a , pour l'année courante. Dans l'enquête sur les raisins.

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

où $\hat{Y}_{arv}$ est le nombre total estimé de raisins de la variété v dans l'échantillon répété r de la région a .

Aux fins d'uniformité, on peut utiliser $\hat{Y}_{ar}$ et $\hat{Y}_{arv}$ l'un pour l'autre pour les enquêtes sur les griottes et les pêches, les estimations ne portant que sur une seule variété de griottes et de pêches. (En d'autres mots, on peut omettre l'indice inférieur v pour les enquêtes sur les griottes et les pêches.) Puis, on estime $\hat{Y}_{arv}$ de la façon suivante :

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

où $\hat{Y}_{arvfb\tau}$ = le nombre total estimé de fruits commercialisables sur l'arbre ou la vigne τ (variété v), dans le verger ou le vignoble b , sur la ferme f , dans l'échantillon répété r , dans la région a , pour l'année en cours (c'est-à-dire, $\hat{y}_{\tau}$);

n_{arvfb} = le nombre d'arbres ou de vignes (variété v) qui ont fait l'objet d'un échantillonnage pendant l'année en cours, dans le verger ou le vignoble b, sur la ferme f, dans l'échantillon répété r, dans la région a (généralement, $n_{arvfb} = 4$ dans les enquêtes sur les griottes et les pêches et 5 dans l'enquête sur les raisins);

n_{arvf} = le nombre de vergers ou de vignobles (où la variété v a fait l'objet d'un échantillonnage dans l'année en cours) sur la ferme f, dans l'échantillon répété r, dans la région a (généralement, $n_{arvf} = 1$, sauf dans le cas des unités choisies plus d'une fois dans le même échantillon répété, par exemple, les grosses fermes);

n_{arv} = le nombre total de fermes distinctes (où la variété v a fait l'objet d'un échantillonnage dans l'année en cours), dans l'échantillon répété r, dans la région a;

n_{arv}^* = le nombre total de vergers ou de vignobles (où la variété v a fait l'objet d'un échantillonnage dans l'année en cours), dans l'échantillon répété r, dans la région a

$$(c'est-à-dire, n_{arv}^* = \sum_{f=1}^{n_{arv}} n_{arvf});$$

N_{arvfb}^C = le nombre total, à ce jour, d'arbres producteurs ou de vignes productrices (variété v) dans le verger ou le vignoble b, sur la ferme f, dans l'échantillon répété r, dans la région a;

N_{arvfb}^{83} = le nombre total d'arbres ou de vignes (variété v), dans le verger ou le vignoble b, sur la ferme f, dans l'échantillon répété r, dans la région a, selon les cartes établies en 1982-1983;

N_{arvf}^{83} = le nombre total d'arbres ou de vignes (variété v), sur la ferme f, dans l'échantillon répété r, dans la région a, selon les listes établies en 1983;

N_{arf}^{81} = le nombre total d'arbres ou de vignes (toutes les variétés), sur la ferme f, dans l'échantillon répété r, dans la région a, selon le recensement de 1981 (accompagné de la liste-échantillon);

et N_a^{81} = le nombre de tous les arbres ou de toutes les vignes dans la région a, selon le recensement de 1981 (voir le tableau-4).

TABEAU 4: Nombre d'arbres ou de vignes (N_a^{81}) par région, établi lors du recensement de 1981 et utilisé pour les enquêtes sur les fruits tendres

	RÉGION 1	RÉGION 2	RÉGION 3	RÉGION 4	TOTAL
PÊCHES	13,094	389,157	10,271	411,697	824,219
GRIOTTES	9,496	50,888	55,449	32,536	148,369
RAISINS	1,142,067	5,666,008	946,509	3,975,202	11,729,786

5.3 Estimations, par région, du taux de changement et précision de ces estimations

On estime le taux de changement applicable à la production de la région a par rapport à l'année précédente, représenté par R_a comme suit:

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

où $\hat{Y}_a$ = le nombre total estimé de fruits commercialisables (pêches, griottes, tous les raisins ou raisins de la variété v) dans la région a, pour l'année en cours (ce nombre correspond à

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

dans le cas des pêches, des griottes et de tous les raisins de toutes les variétés, et à

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

dans le cas des raisins d'une variété donnée); et où $\hat{X}_a$, $\hat{X}_{ar}$, $\hat{X}_{arv}$ représentent les estimations correspondantes de l'année précédente.

(Il est à noter qu'on peut omettre l'indice inférieur v, toutes les estimations étant calculées de la même façon, qu'il s'agisse des pêches, des griottes, de tous les raisins de toutes les variétés ou des raisins d'une variété donnée.)

On définit les variances de $\hat{Y}_a$ et de $\hat{X}_a$ ainsi que leur covariance comme suit:

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

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

$$\text{Cov}(\hat{Y}_a, \hat{X}_a) = (\hat{Y}_{a1} - \hat{Y}_{a2}) (\hat{X}_{a1} - \hat{X}_{a2})/4 = D_{ya} D_{xa}/4 \quad (5.3.6)$$

où les indices inférieurs numériques correspondent au numéro de l'échantillon répété. Puis, on estime la variance, $V(\hat{R}_a)$, du taux de changement estimé, $\hat{R}_a$, de la façon suivante [1] :

$$\begin{aligned} \hat{V}(\hat{R}_a) &= \frac{1}{\hat{X}_a^2} \{V(\hat{Y}_a) - 2 \hat{R}_a \text{Cov}(\hat{Y}_a, \hat{X}_a) + \hat{R}_a^2 V(\hat{X}_a)\} \\ &= \left\{ \frac{D_{ya}}{S_{xa}} - \frac{S_{ya} D_{xa}}{S_{xa}^2} \right\}^2 \end{aligned} \quad (5.3.7)$$

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

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

On peut alors établir le coefficient de variation de $\hat{R}_a$ comme suit :

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

5.4 Estimations, par région, de la production fruitière totale et précision de ces estimations

On représente par X_a^T le rendement réel (en tonnes de l'année précédente, dans la région a, et par $\hat{Y}_a^T$ les estimations correspondantes de l'année courante. Puis, on estime $\hat{Y}_a^T$ comme suit :

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

Le coefficient de variation est calculé de la façon suivante :

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

5.5 Estimations de la production fruitière totale et précision de ces estimations

On représente par $\hat{Y}^T$ la production fruitière totale dans les quatre régions, pour l'année courante, que l'on estime comme suit :

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

Le coefficient de variation est calculé de la façon suivante :

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

6. RÉSUMÉ

Le premier recensement aura lieu en 1983, bien que ce n'est pas avant l'été 1984 que paraîtront les premières estimations de l'enquête remaniée. Pour cette raison, il sera nécessaire de mener à la fois l'ancienne enquête et la nouvelle enquête en 1983 si l'on veut disposer d'estimations pour cette année-là.

Bien qu'on ait conçu l'enquête de façon à ce qu'elle soit autopondérée, elle ne l'est pas complètement en raison des écarts temporels entre les recensements de 1981, la confection initiale des listes en 1983 et les recensements subséquents. Les formules d'estimation présentées à la section précédente tiennent compte de ces écarts temporels. Cependant, étant donné la grande simplicité de l'estimation autopondérée, nous avons proposé que l'on étudie son efficacité, une fois les données disponibles.

BIBLIOGRAPHIE

- [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. et Sukhatme, B.V. (1980), Sampling for estimating production of fruit crops, Sankhya C42, 17-30.

**UNE ESTIMATION PRÉCISE ET RAPIDE DE LA SUPERFICIE
PLANTÉE EN POMMES DE TERRE GRÂCE À LANDSAT:
RÉSULTATS D'UNE DÉMONSTRATION¹**

R.A. Ryerson, J.-L. Tambay, R.J. Brown

et

L.A. Murphy, B. McLaughlin²

Cette communication décrit les procédures et les résultats d'un projet conjoint de Statistique Canada et du Centre canadien de télédétection (CCT) visant à fournir une estimation rapide de la superficie plantée en pommes de terre au Nouveau-Brunswick, province qui est un important producteur de pommes de terre au Canada. Le projet a démontré que l'imagerie obtenue par satellite et combinée aux méthodes plus usuelles d'estimation des superficies plantées en pommes de terre peut alléger le fardeau des répondants, fournir rapidement des cartes de la distribution des cultures et permettre l'établissement d'estimations fiables pour les sous-régions.

1. INTRODUCTION

Les premières applications de la télédétection (par satellite) dans la vallée de la rivière Saint-Jean (Nouveau-Brunswick) par le Centre canadien de télédétection et le ministère de l'Agriculture du Nouveau-Brunswick ont démontré que les données obtenues par satellite pouvaient fournir des estimations précises et peu coûteuses de la superficie plantée en pommes de

¹ Communication initialement présentée au Quinzième symposium international sur la télédétection de l'environnement, Ann Arbor (Michigan), mai 1981.

² R.A. Ryerson et R.J. Brown, Centre canadien de télédétection (CCT), EMR; J.-L. Tambay, Division des méthodes d'enquête-entreprises (ce travail a été fait lorsque l'auteur était à la Division des méthodes d'enquête - institutions et agriculture), Statistique Canada; L.A. Murphy, Division de la statistique agricole, Statistique Canada; et B. McLaughlin, Division de la statistique agricole, Bureau régional de Truro (Nouvelle-Écosse), Statistique Canada.

terre (Mosher et al, 1978; Ryerson et al, 1979; Ryerson et al, 1980). L'intérêt manifesté dans cette étude et dans d'autres travaux du CCT sur le colza (Brown et al, 1980) a amené Statistique Canada à mettre sur pied un projet de démonstration en temps réel utilisant les données recueillies par le satellite Landsat au cours de l'année-récolte 1980. Statistique Canada, qui est l'organisme fédéral chargé d'obtenir des données sur les récoltes, désirait comparer les estimations des superficies plantées obtenues par des méthodes traditionnelles avec les estimations obtenues par satellite pour une même région. On a retenu la pomme de terre comme récolte-cible, et la vallée de la rivière Saint-Jean comme région d'étude.

La télédétection par satellite présente plusieurs attraits pour Statistique Canada: précision accrue des estimations établies à partir d'enquêtes régulières (probablement à des niveaux plus locaux), allègement du fardeau imposé aux répondants (ce qui pourrait se traduire par une réduction du nombre de questionnaires et des questions) et possibilité de dresser des cartes pour de petites régions où poussent des récoltes spéciales, ceci afin de mieux planifier les méthodes d'échantillonnage.

La section suivante résume les principaux résultats de cette démonstration, et le reste de la communication indique en gros les méthodes de télédétection qui ont été utilisées dans ce projet. On décrit aussi le système actuel de saisie des données de Statistique Canada, la région visée par l'étude, l'échantillon au sol et la saisie des données sur le terrain, l'analyse des données obtenues par télédétection et, en dernier lieu, la vérification et l'analyse des résultats.

2. PRINCIPAUX RÉSULTATS

Les données recueillies par satellite ont servi à estimer la superficie des cultures de pommes de terre dans la vallée de la rivière Saint-Jean, au Nouveau-Brunswick. Ces estimations, extrapolées au niveau provincial, correspondaient à 2 % près aux estimations de 52,000 acres publiées par Statistique Canada. Ces estimations étaient basées sur les résultats de trois enquêtes indépendantes menées dans la province.

L'analyse des données obtenues par satellite a été faite en temps réel (c'est-à-dire presque instantanément) au CCI, car la plupart des travaux de préparation avaient été effectués avant l'acquisition des données. La zone d'échantillonnage retenue pour l'Enquête descriptive sur l'agriculture (EDA) a fourni les données au sol nécessaires pour l'étalonnage du système; ces données ont permis d'établir des estimations par le quotient et par régression pour la correction des erreurs systématiques dans la classification par télédétection des champs de pommes de terre. Bien que cette démonstration n'ait pas eu lieu dans un cadre opérationnel, on aurait pu produire les estimations définitives moins de deux semaines après le passage du satellite au-dessus de la région d'étude.

La classification par satellite a été entravée par la présence de nuages (non-réponse au niveau du satellite) et par la similitude entre la signature spectrale des pommes de terre et celle d'autres objets présents dans la région étudiée. Le premier de ces problèmes s'est traduit par une perte de données, et sa correction a nécessité une certaine dose d'imputation. Quant au second problème, on l'a résolu en partie en corrigeant la classification et en utilisant des estimateurs par le quotient et par régression.

La comparaison des données recueillies au sol (par les enquêteurs), à partir des photographies aériennes des champs échantillonnés, a révélé que certains champs ont été omis par les enquêteurs de l'EDA, car les données sur ces champs n'étaient pas requises pour les besoins de l'EDA. Aussi a-t-on dû se servir des données tirées de photographies aériennes au lieu des données de l'EDA pour les estimations par satellite en 1980. En 1981, les procédures d'énumération de l'EDA ont été modifiées pour satisfaire tant aux exigences de l'EDA qu'à celles de la télédétection.

À la suite du succès rencontré par cette démonstration, l'expérience a été reprise en 1981. De plus, une expérience similaire a aussi été entreprise en 1981 afin d'estimer la superficie des récoltes de colza dans le district de Peace River, à la frontière de l'Alberta et de la Colombie-Britannique.

3. TÉLÉDÉTECTION À L'AIDE DES SATELLITES LANDSAT

La télédétection consiste à mesurer les caractéristiques d'un objet à distance, habituellement à partir d'un aéronef ou d'un satellite. Avec les données captées par satellite, on peut obtenir rapidement la couverture complète de grandes superficies, et ce, à un coût relativement faible. Parmi les différents domaines d'application de ces données, mentionnons l'agriculture, l'exploitation forestière, l'utilisation des terres, la formation des glaces et la cartographie en général.

Les satellites américains de la série Landsat, lancés par la NASA, ont fourni les données pour cette expérience, et celles utilisées auparavant pour une expérience précédente au Nouveau-Brunswick. Chaque satellite Landsat fait le tour de la terre 14 fois par jour, en orbite héliosynchrone (ce qui permet ainsi de couvrir les différents points de la terre à la même heure solaire locale). La lumière réfléchie par le sol est enregistrée dans quatre bandes spectrales étroites, à l'aide du balayeur multispectral (MSS). Les données transmises au Canada sont captées à l'une des deux stations de réception situées, respectivement, à Prince Albert (Saskatchewan) et à Shoe Cove (Terre-Neuve). Chaque point de la terre est survolé tous les 18 jours par un satellite Landsat (si deux satellites sont utilisés, cette fréquence de couverture est de neuf jours).

Les données sont analysées par le Système d'analyse d'images du CCT (CIAS); ces données étant codées sur des produits standard comme les bandes d'ordinateur, chaque scène imagée couvrant une superficie de 25,600 kilomètres carrés. La plus petite unité de surface imagée est appelée élément d'image, ou pixel (de l'anglais "picture element"). Chaque pixel est identifié par une signature spectrale qui lui est propre, cette signature étant une mesure de la réflectance de la zone imagée dans les quatre bandes spectrales. La signature spectrale dépend des objets présents dans le pixel (routes, cultures, etc.), chaque objet ayant une signature particulière. La signature des terres cultivées dépend de la structure des plantes, du type de sol apparent, de la maturité des récoltes, de la hauteur des plantes, de la densité des feuilles et de nombreux autres facteurs.

Pour estimer la superficie des récoltes, on doit identifier chaque pixel qui appartient à un type de récolte donné. Des grands champs, que l'on sait appartenir au type de récolte qui nous intéresse sont localisés pour "l'apprentissage" du système. Les données de ces champs permettent d'identifier la signature particulière à cette récolte. Tous les pixels sont ensuite classés comme appartenant ou n'appartenant pas à cette récolte, selon leurs signatures spectrales.

Pour déterminer la superficie d'une récolte dans une région donnée, on compte le nombre de pixels se trouvant à l'intérieur de cette région et ayant été identifiés comme appartenant à la récolte. Un "apprentissage" additionnel peut être exécuté afin d'identifier les pixels "manqués" dans la classification initiale, ou encore pour réduire la confusion entre les récoltes, c'est-à-dire pour mieux séparer les récoltes dont la signature spectrale ressemble fortement à celle de la récolte d'intérêt.

Des données au sol précises sont requises pour deux raisons: en premier lieu, pour localiser les champs d'apprentissage étendus et, en second lieu, pour corriger toute erreur systématique dans la classification des données obtenues par satellite. Ces données au sol peuvent être obtenues par des recenseurs spécialisés travaillant sur le terrain, ou encore en utilisant une imagerie aérienne qui est interprétée par des analystes d'image.

4. SAISIE DES DONNÉES STATISTIQUES

Au fil des ans, Statistique Canada a utilisé les données obtenues par les enquêtes postales annuelles comme intrants principaux pour son système d'estimation des récoltes. Bien que ces enquêtes soient relativement peu coûteuses et qu'elles peuvent être réalisées rapidement, elles sont limitées par les taux de réponses variables et par la non-représentativité toujours possible des répondants. Les enquêtes descriptives probabilistes ont été introduites au cours des années 1970, afin de pallier à certains de ces problèmes. Ces procédures font appel à des enquêteurs sur le terrain pour le dénombrement d'un échantillon aléatoire d'agriculteurs. En 1980, Statistique Canada a fondé ses estimations des superficies plantées en pommes de terre au Nouveau-Brunswick sur les résultats de trois enquêtes:

l'enquête postale, l'enquête objective sur le rendement des pommes de terre (EORPT) et l'enquête descriptive sur l'agriculture (EDA).

Les questionnaires de l'enquête postale sont expédiés au début de juin à tous les agriculteurs répertoriés dans le registre des fermes dressé par la Division de la statistique agricole. Les réponses sont compilées par comté, et l'on obtient des estimations par comté en reliant la variation annuelle des superficies indiquées à la superficie totale de pommes de terre déterminée d'après le recensement pour chaque comté. Les estimations par comté sont ensuite additionnées afin de donner les estimations provinciales pour la fin juin.

L'enquête objective sur le rendement des pommes de terre est une enquête spécialisée, à la fois postale et descriptive, conçue pour estimer la superficie des récoltes de pommes de terre, ainsi que le rendement de la pomme de terre dans son habitat du Nouveau-Brunswick. L'enquête est réalisée à la mi-juillet sur un échantillon aléatoire de producteurs de pommes de terre obtenu du registre des fermes; les estimations des superficies de pommes de terre sont produites vers la mi-août.

L'enquête descriptive sur l'agriculture (EDA) est, comme son nom l'indique, une enquête descriptive et polyvalente conçue pour estimer les récoltes, le bétail et les dépenses agricoles au niveau provincial. L'EDA est une enquête à bases multiples réalisée à partir d'un échantillonnage aléatoire d'agriculteurs choisis dans le registre des fermes et à partir d'un échantillon aréolaire et aléatoire de segments (parcelles de terre). Les agents recenseurs visitent les fermes échantillonnées à la fin juin et au début de juillet. Les estimations des superficies sont disponibles au début d'août. Chaque année, environ 20 % des segments sont renouvelés.

Au cours de la saison de croissance, on publie deux estimations de la superficie plantée en pommes de terre. La première, publiée au début de juin, est basée sur les résultats de l'enquête postale. La seconde estimation, publiée au début de septembre, est basée sur l'examen des estimations provenant des trois enquêtes et de discussions avec les autorités provinciales. Comme date cible pour la publication des estimations obtenues par

satellite, on a retenu la date de publication de la seconde estimation.

5. RÉGION D'ÉTUDE

La région visée par l'étude est située dans la vallée supérieure de la rivière Saint-Jean, au Nouveau-Brunswick. Cette région prend naissance au sud de Woodstock, dans le comté Carleton, et suit la rivière Saint-Jean en direction nord-ouest sur environ 200 kilomètres, à travers le comté Victoria, jusqu'à Claire, dans le comté de Madawaska.

La région est densément boisée, de topographie variée et accidentée. Le drainage et la nature pierreuse du sol présentent certains problèmes dans cette région. On y compte quelque 70,000 hectares de terres agricoles améliorées, dont environ 20,000 servent habituellement à la culture de la pomme de terre. Les autres cultures importantes de ces régions sont les céréales, le foin et les légumes de traitement comme les pois, le brocoli et le chou de Bruxelles. La taille des terrains varie des parcelles de semences de 0.1 hectare aux champs de 40 hectares.

6. SAISIE DES DONNÉES ET ÉCHANTILLONNAGE AU SOL EN 1980

Dans le cadre de l'EDA au Nouveau-Brunswick, l'échantillon aréolaire a été jugé approprié pour obtenir les données au sol requises pour l'interprétation des données de télédétection. Cet échantillon a été composé en deux phases. Dans la première phase, les secteurs de dénombrement (SD) du recensement (lesquels comprenaient les sièges des fermes du recensement de 1976 - appelés SD agricoles du recensement) ont été stratifiés selon la superficie des champs de pommes de terre, l'importance du bétail et le nombre de porcs (données du recensement de 1976). À l'intérieur de chaque strate, on a choisi deux échantillons aléatoires simples par réplikat de secteurs de dénombrement. Chaque secteur échantillonné a été ensuite segmenté à l'aide de cartes, en superficie identifiable d'environ 5 à 8 kilomètres carrés, et on a choisi, comme échantillon aléatoire simple, un segment sur 10 par secteur de dénombrement. On a fourni aux agents recenseurs de l'EDA travaillant dans la région d'étude des agrandissements d'anciennes photo-

graphies aériennes (échelle de 1" pour 832') de chaque segment échantillonné. Les photographies provenaient de sources provinciales. La plupart d'entre elles avaient été prises en 1976. Ces agents recenseurs devaient visiter les agriculteurs exploitant des terres à l'intérieur de chaque segment afin de leur présenter les photographies, pour qu'ils puissent identifier tous leurs champs de pommes de terre et de maïs³; les agents recenseurs devaient noter les superficies indiquées par les agriculteurs. Le manuel des enquêteurs contenait une section sur les procédures à suivre, les procédures faisant partie de leur entraînement pour l'EDA.

7. ANALYSE DES DONNÉES OBTENUES PAR TÉLÉ DÉTECTION

7.1 TRAVAUX PRÉCÉDENTS

Les travaux réalisés dans cette même région à l'aide de données de 1975 ont été rapportés ailleurs (Mosher et al) et la description détaillée de la méthodologie utilisée a été publiée (Ryerson et al, 1980). Dans les travaux de 1975, on a retenu une zone d'étude contenant environ 20 % de toutes les récoltes de pommes de terre de la province. À l'appui de cette étude, on a obtenu des données de terrain pour toute la région d'étude.

La région d'étude de 125 kilomètres carrés et deux sous-régions ont été localisées sur l'écran couleurs du Système d'analyse d'images du CCT (CIAS - Goodenough, 1979). On a eu recours à une méthode très simple d'apprentissage dirigé pour dresser les données des pixels dans trois champs de pommes de terre sous forme de quatre histogrammes unidimensionnels. Un parallélépipède quadridimensionnel a été défini par les limites de chaque histogramme et a servi à borner les zones de décisions. Tous les points qui se trouvent à l'intérieur du parallélépipède ont été classés comme des pommes de terre, et ceux qui se trouvent à l'extérieur de cette région ont été classés sous la rubrique "autres".

Un des principaux problèmes rencontrés dans l'analyse des données était la classification adéquate des pixels situés sur la bordure des champs. Ces

³ On devait aussi identifier les champs de maïs, car les travaux antérieurs en télédétection ont indiqué que le maïs pouvait être confondu avec les pommes de terre (Ryerson et al, 1980).

pixels présentent un problème spécial, car, comme nous venons de le mentionner, ils chevauchent deux champs différents. La réflectance de ces pixels dépend de la proportion de la superficie du pixel couverte par chaque champ et de la réflectance du matériau superficiel dans chaque champ. Il est normalement très difficile de calculer le pourcentage de chaque type de matériau présent dans ces pixels. Toutefois, on a pu obtenir des estimations acceptables des superficies en modifiant la limite originale des zones de décision: on y a joint un second parallélogramme formé par apprentissage sur un certain nombre de ces pixels de chevauchement qui semblaient appartenir aux champs de pommes de terre. Pour choisir les pixels de chevauchement appropriés et les inclure dans la classe des pommes de terre, on s'est fondé sur une interprétation visuelle et subjective de la scène affichée à l'écran (les données de trois des quatre bandes spectrales étaient fusionnées pour former une image dont les couleurs ressemblaient à celle d'une émulsion de couleurs pour l'infrarouge).

L'estimation de la superficie pour toute la ceinture de la pomme de terre a nécessité moins de quatre heures d'utilisation du système CIAS. La localisation, l'affichage et l'analyse du secteur primaire et des sous-secteurs ont pris un peu plus d'une heure. La sélection, l'affichage et l'analyse des cinq sous-scènes choisies ultérieurement ont pris deux heures et demi, et une autre heure a été nécessaire pour localiser la frontière du Nouveau-Brunswick et éliminer les données appartenant à l'extérieur de la province.

Par rapport à la superficie totale plantée de pommes de terre qui avait été déterminée et mesurée à partir des photographies aériennes à basse altitude prises en même temps, les estimations obtenues par satellite en 1975 étaient précises à 95 % (c'est-à-dire qu'elles correspondaient à 95 % de la valeur véritable estimée) dans la sous-région contenant le site d'apprentissage, à 80 % dans la deuxième sous-région et à 88 % pour l'ensemble de la région d'étude. Lors d'essais répétés utilisant des champs d'apprentissage différents, la précision est passée de 85 à 97 % pour la région d'étude primaire, alors que la précision pour l'ensemble de la province s'établissait à 84.5 %. Une partie de la marge d'erreur, dans l'estimation provinciale, était due au fait qu'un certain pourcentage des champs de pommes de terre se trouvaient à l'extérieur de la ceinture dite de la pomme de

terre. D'autres facteurs ayant contribué à cette marge d'erreur sont discutés ci-dessous.

7.2 AMÉLIORATION DES ESTIMATIONS

Bien que les travaux précédents aient connu un certain succès, on a identifié les sources potentielles d'erreur pour des applications nécessitant une précision supérieure à, disons, 85 %. Les principaux problèmes sont associés au choix subjectif des pixels de chevauchement des champs de pommes de terre, à la manipulation des champs de petite taille et à l'élimination des récoltes présentant des caractéristiques spectrales similaires à celles des pommes de terre.

En ce qui concerne ce dernier problème, l'idéal serait de connaître précisément la réflectance spectrale des pommes de terre et celle des autres récoltes tout au cours de la saison de croissance. À l'aide de ces renseignements, il serait alors possible d'identifier la fenêtre phénologique pendant laquelle les pommes de terre peuvent être distinguées avec certitude des autres récoltes. Malheureusement, un tel ensemble de données n'existe pas, bien que les connaissances actuelles sur les récoltes et les pratiques agricoles de la région fournissent des indications générales. Dans ce cas particulier, l'expérience des auteurs sur le terrain leur a permis de supposer que la date optimale pour distinguer les pommes de terre des autres récoltes dans cette région serait entre la mi-juillet et la mi-août. Pour vérifier cette hypothèse et fournir une indication quant au degré de séparabilité des pommes de terre et des autres récoltes, on a analysé des données MSS (Landsat) sur bande d'ordinateur obtenues au dessus de la vallée de la rivière Saint-Jean le 8 août 1975. La Figure 1 indique les valeurs relatives de radiance dans chaque bande Landsat pour les pommes de terre, le maïs, les pois, le foin, le brocoli, le pâturage, le sarrasin, le sol nu et les céréales. On voit sur cette figure que les pommes de terre sont aisément séparables des autres récoltes, sauf les pois qui sont habituellement récoltés entre la mi-août et la fin août. Il semble donc que l'analyse des données recueillies vers la fin de la saison de croissance permettent la séparation et l'identification des pommes de terre.

Quant aux problèmes des champs de petite taille et des pixels de chevauchement, on pourrait les résoudre en utilisant davantage de données au sol dans quelques petites surfaces, afin d'obtenir des estimations plus précises sur la superficie des récoltes dans une région. Vu l'étendue de la région où la pomme de terre est cultivée, on doit connaître la superficie des champs de pommes de terre à l'intérieur de dix à quinze segments, chaque segment ayant une superficie qui varie de cinq à huit kilomètres carrés. Toute la région serait alors classifiée le mieux possible, mais sans se servir de la classe subjective des pixels de chevauchement. La classification de chaque segment serait ensuite utilisée, avec les données au sol disponibles (les photographies aériennes prises en 1980, les données de l'EDA pour les années à venir), pour obtenir une relation de régression appliquée par la suite à l'estimation de toute la région afin de produire une estimation révisée (Hanuschak et al, 1979). On établirait aussi une estimation par le quotient, basée sur les estimations de toute la région obtenues par photographies aériennes et par données de satellite.

7.3 PRODUCTION DES ESTIMATIONS DE 1980 À PARTIR DES DONNÉES LANDSAT

On peut décrire l'estimation de la superficie plantée en pommes de terre, à partir des données satellite, comme un processus en trois étapes: l'obtention des données anciennes et nouvelles, la prélocalisation des frontières des segments de l'EDA, et, en dernier lieu, l'analyse des images.

Les données Landsat utilisées pour l'estimation étaient codées sur des bandes pour ordinateur et traitées sur le système de correction des images numériques par l'interpolation ($\sin x/x$) pour la correction géométrique (Guertin et al, 1979) et par la correction radiométrique CAL 3 (Ahern et Murphy, 1978). Chaque bande pour ordinateur couvre quatre cartes du Système topographique national au 1:50,000^e et comporte des pixels carrés de 50 m. Quatre bandes pour ordinateur sont nécessaires pour couvrir la région. Des données existantes avaient été commandées et elles devaient être livrées en mai 1980, alors que de nouvelles données avaient été commandées pour les passages du satellite au-dessus de la région entre la mi-juillet et la mi-août. La commande n'a posé aucun problème pour les données existantes, mais elle s'est avérée plus compliquée pour les données

en temps réel en raison de la panne de Landsat-III. Un passage de Landsat-II le 17 août a été utilisé pour produire une bande d'ordinateur traitée sur le système DICS et livrée au Centre d'analyse le 22 août, bien avant la date prévue.

La prélocalisation des segments de l'EDA a été réalisée au printemps de 1980 à l'aide de la fonction "curseur de polygonation" sur le Système d'analyse des images du CCT (Goodenough, 1977). Les frontières des segments de l'EDA ont été fournies par Statistique Canada sur des cartes au 1:50,000^e et sur des photocopies d'agrandissement de photographies aériennes, remises aux agents recenseurs de l'EDA pour leur énumération. Bien que les limites de certains segments furent faciles à localiser (ruisseaux, lisières de boisés, lacs, etc.), d'autres frontières géographiques ou de recensement, étaient beaucoup plus complexes. Pour les segments dont les limites consistaient en une combinaison de routes principales et (ou) des rivières importantes, on pouvait les localiser, les délimiter et les mettre en mémoire en moins de cinq minutes de travail sur l'image couleurs de 128×128 pixels de 50m agrandie sur un écran de 512×512 lignes. Les segments les plus complexes ont nécessité jusqu'à une heure de travail - la moyenne s'établissant entre 20 et 25 minutes. Une fois localisé, la frontière du segment était mise en mémoire selon des coordonnées de pixel propres au système DICS, de sorte qu'elle pouvait être superposée sur de nouvelles données lorsque venait le temps de localiser les données d'apprentissage et les intrants des estimateurs. Lors de la phase préparatoire, on n'a pu localiser qu'un certain nombre de frontières de segment, car les autres segments chevauchaient deux bandes d'ordinateur DICS ou présentaient d'autres problèmes similaires. On travaille actuellement sur un logiciel qui permettra de réduire le temps requis pour tout le projet, particulièrement pour la phase de prélocalisation. L'utilisation des photographies originales au lieu des photocopies, prévue pour le volet 1981 du projet (avec 15 nouveaux segments en 1981), devrait aussi permettre de réduire le temps requis.

Une fois les données au sol de 1980 acheminées au centre d'analyse, on procéda au choix des sites potentiels d'apprentissage (selon la taille des champs). On a retenu plusieurs champs dans un segment situé au nord de la

région (près de Saint-André), ainsi que plusieurs autres champs dans un segment situé au sud de la région (près de Hartland).

Après la réception des données satellite, il fut relativement aisé de rappeler les frontières de segment, de les superposer, de localiser les champs d'apprentissage et de commencer la classification selon les méthodes décrites à la section 7.1. En plus du choix des régions d'apprentissage, on a retenu un autre groupe de champs de grande superficie, ainsi que des secteurs de champs connus de pommes de terre qui apparaissaient plus rouges sur l'écran que les champs faisant partie de l'ensemble d'apprentissage. À mesure que les résultats de la classification devenaient disponibles, les superficies cultivées étaient enregistrées pour chaque sous-scène de 512 x 512 pixels et pour 17 segments de l'EDA.

Les analystes ont rencontré quatre problèmes au cours de la classification; un de ces problèmes avait trait à l'interpolation des récoltes se trouvant sous des nuages épars et sous l'ombre des nuages, alors que les trois autres problèmes avaient trait au brouillage par d'autres récoltes. La méthode d'interpolation pour les champs de pommes de terre se trouvant sous les nuages était relativement simple. On a supposé que la proportion de récoltes sous les nuages était similaire à celle de la région adjacente qui semblait contenir des pommes de terre. Pour déterminer la superficie des champs de pommes de terre sous les nuages (la valeur PC), on a utilisé la formule suivante:

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

où: P_M = superficie des champs de pommes de terre mesurée dans la région sans nuage; C_A = superficie recouverte par les nuages et l'ombre des nuages dans la région et T_A = superficie totale de la région. Les estimations totales par satellite ont incorporé ces régions, alors qu'on n'a inclus aucun segment EDA sous les nuages dans les analyses pour les

estimations par le quotient et par régression. Quant aux problèmes du brouillage par d'autres récoltes, on les a, du moins pour la plupart, résolus par une légère modification des paramètres de classification. Dans un cas, le brouillage était dû à une forme non-identifiée de repousse naturelle fortement diffusée dans des régions de coupes forestières. Dans d'autres cas, des champs de foin dans lesquels se trouvaient des amoncellements de pierres espacés régulièrement ont présenté une signature spectrale similaire à celle des pommes de terre. Après modification de la classification, les seuls éléments de brouillage qui demeuraient étaient dus à quelques parcelles de trèfles dans un segment et à certaines allées d'un terrain de golf.

Le calcul des superficies est présenté plus en détail dans les paragraphes suivants.

8. RÉSULTATS ET ANALYSE DU PROJET D'ÉVALUATION DE LA SUPERFICIE PLANTÉE EN POMMES DE TERRE AU NOUVEAU-BRUNSWICK (1980)

8.1 INTRODUCTION

Dans le cadre du projet d'évaluation de la superficie plantée en pommes de terre au Nouveau-Brunswick, réalisé en 1980, on a analysé les données au niveau de la région, des segments et des champs. On a obtenu deux types d'estimation pour la superficie totale des récoltes de pommes de terre dans la région cible de la vallée de la rivière Saint-Jean. Les estimations proviennent des données recueillies par satellite et de données obtenues par photographies aériennes en haute altitude, ces dernières ayant fait l'objet de vérification au sol (les photographies aériennes ont été obtenues et analysées par le CCT). Les estimations et leurs variances ont ensuite été comparées aux autres estimations obtenues par Statistique Canada lors d'enquêtes réalisées au Nouveau-Brunswick. On a ensuite comparé les superficies plantées en pommes de terre calculées par l'EDA et par satellite à celles qui furent calculées à partir des photographies aériennes (et qui représentaient, dans ce cas-ci, la meilleure approximation des valeurs réelles) afin de déterminer le degré de concordance de ces deux es-

timations au niveau du segment. Comme complément d'analyse, on a examiné la variation des superficies indiquées lors de l'EDA, pour chaque secteur assigné à un enquêteur.

8.2 ESTIMATION DE LA SUPERFICIE PLANTÉE EN POMMES DE TERRE À L'AIDE DES DONNÉES OBTENUES PAR SATELLITE.

Pour les estimations par le quotient et par régression, on a estimé la superficie totale des champs de pommes de terre dans la région d'étude du Nouveau-Brunswick, recourant pour ce faire à des données obtenues par satellite et à des données sur les segments obtenues par photographie aérienne. Les estimations, ainsi que leurs variances, ont été calculées selon le plan d'échantillonnage de l'EDA. Pour le Nouveau-Brunswick, l'EDA est une enquête à base multiple utilisant un échantillon stratifié à deux degrés par réplikat de segments, cet échantillon étant conçu pour donner des estimations précises sur divers caractères au niveau provincial (voir les sections 4 et 6). Comme les strates de l'EDA ne coïncidaient pas avec les limites de la région d'étude, on a eu recours à la technique de la post-stratification afin de calculer les estimations, traitant les secteurs de dénombrement comme des échantillons avec remplacement. On n'a pas incorporé dans l'échantillon les segments pour lesquels les données par satellite étaient manquantes, ni un segment représentant une valeur aberrante (voir la Figure 3). Ces estimations ont été basées sur 40 segments. En dernier lieu, comme les agents recenseurs de l'EDA ne recueillaient pas toujours les données sur toutes les fermes se trouvant à l'intérieur du segment (voir la section 8.3), on a utilisé la superficie du segment calculée d'après les photographies aériennes.

Soit "x" la donnée sur les pommes de terre obtenue par satellite et "y" une donnée obtenue par photographie aérienne. Les estimations par le quotient et par régression de Y, qui représente la superficie totale des champs de pommes de terre dans la région d'essai, ont été calculées à l'aide des formules suivantes:

$$\hat{Y}_{\text{quot.}} = \hat{R} X = \frac{\hat{Y}}{\hat{X}} X, \text{ et}$$

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

où
$$\hat{Y} = \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$$
 est l'estimation théorique de Y;

$\hat{X}$ est l'estimation théorique de X, qui représente la superficie totale des champs de pommes de terre obtenue par satellite et non corrigée pour la région d'essai. $\hat{X}$ est obtenu en substituant x_{hij} pour y_{hij} dans la formule pour $\hat{Y}$;

y_{hij} et x_{hij} sont les valeurs observées pour le j-ème segment choisis dans la i-ème unité (SD) de premier degré choisie pour la strate h;

N_h et n_h sont respectivement le nombre total et le nombre échantillonné de secteur de dénombrement dans la strate h, h prenant les valeurs de 1 à L;

M_{hi} et m_{hi} sont, respectivement, le nombre total et le nombre échantillonné d'unités (segments) de second degré dans le i-ème SD choisi dans la strate h;

$\hat{R}$ est une estimation de Y/X ; et, finalement,

$\hat{B} = \text{cov}(\hat{Y}, \hat{X}) / \text{var}(\hat{X})$ est le coefficient de régression linéaire.

Les estimations de variance de $\hat{Y}_{\text{quot.}}$ et $\hat{Y}_{\text{rég.}}$ sont données par les relations suivantes:

$$v(\hat{Y}_{\text{quot.}}) = v(\hat{Y}) - 2\hat{R} \text{cov}(\hat{Y}, \hat{X}) + \hat{R}^2 v(\hat{X}), \text{ et}$$

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

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

et $v(\hat{X})$ est calculé en substituant x_{hij} pour y_{hij} dans l'expression pour $v(\hat{Y})$.

On peut voir que $\hat{B} = \text{cov}(\hat{Y}, \hat{X})/v(\hat{X})$ est la valeur de B qui minimise $v(\hat{Y}_{\text{rég.}})$.

Le Tableau I illustre les estimations par le quotient et par régression, ainsi que les estimations de Statistique Canada pour la superficie plantée en pommes de terre au Nouveau-Brunswick. Les estimations par le quotient et par régression, calculées au prorata du niveau provincial, sont très près de la superficie calculée par Statistique Canada, laquelle se chiffre à 52,000 acres. En fait, il y a très peu de différence entre les deux estimations, et elles possèdent toutes deux le même coefficient de variation. Afin de donner une idée du meilleur rendement obtenu grâce à l'estimation par le quotient, la variance de cette estimation ne représentait que le 1/5 de celle de $\hat{Y}$, qui est l'estimation théorique de Y basé sur le plan d'échantillon avec post-stratification (échantillon aréolaire seulement). On peut noter que le coefficient de variation des estimations par le quotient et par régression sont du même ordre que celui de l'estimation à bases multiples obtenu par l'EDA, bien que cette dernière soit basée sur un échantillon plus important.

8.3 COMPARAISON DES DONNÉES AU NIVEAU DES SEGMENTS ET DES CHAMPS

Les Figures 2 et 3 contiennent les graphes des superficies plantées en pom-

mes de terre par segment et calculés selon les données recueillies par les agents recenseurs de l'EDA et par satellite, respectivement en fonction des superficies obtenues d'après les photographies aériennes. On n'a pas utilisé tous les segments d'échantillons dans l'analyse. Il n'y avait pas de données par satellite pour 16 segments, en raison de la couverture nuageuse et de la localisation des images. De plus, on n'a pas utilisé 8 segments de l'EDA, car ces segments contenaient des non-répondants à l'enquête et des fermes d'une certaine importance pour lesquelles les agents recenseurs n'avaient pas à recueillir de données sur les champs. Pour les calculs de cette section et ceux de la section 8.2, nous n'avons pas utilisé les deux valeurs aberrantes de ces graphes.

Ces graphes indiquent une très forte relation linéaire entre les données de l'EDA et celles obtenues par les photographies aériennes, mais il en est de même pour les données par satellite comparées aux données obtenues par photographie aérienne au niveau de segment: les coefficients de corrélation⁴ sont respectivement⁵ de 0.991 et de 0.968. On note une tendance à la sous-estimation des superficies dans le cas des agents recenseurs de l'EDA et dans celui des données obtenues par satellite. Cette tendance est cependant moins marquée pour l'EDA. On a déjà expliqué à la section 7 les causes des écarts dans le cas des superficies calculées par les données satellite. La superficie de certains segments contenant peu ou pas de pommes de terre a été surestimée, à cause du brouillage par d'autres récoltes (certaines données satellite ont été rejetées, car elles confondaient un grand champ de foin parsemé de pierre avec un champ de pommes de terre - on aurait pu éliminer cette erreur en modifiant la classification). Quant à la sous-estimation dans le cas de l'EDA, une des principales causes était que les enquêteurs n'avaient pas énumérés certaines fermes se trouvant dans les segments à cause des procédures à bases multiples (parmi les fermes

⁴ Les coefficients de corrélation ont été calculés en utilisant la pondération du plan d'échantillonnage.

⁵ La représentation graphique des données obtenues par satellite par rapport aux données de l'EDA a été très similaire à celle de la Figure 3, le coefficient de corrélation s'établissant dans ce cas à 0.957.

omises, on comptait les fermes qui étaient exclues de la base aréolaire, ainsi que des fermes dont les terrains couvraient plus d'un segment d'échantillon). Pour corriger cette situation, le guide des procédures sur les terrains pour l'EDA de 1981 contenait des instructions spécifiques pour s'assurer que toutes les fermes soient répertoriées, en vue des estimations de superficie qui seront réalisées l'an prochain, lors de la reprise du projet de démonstration. On croit que ces modifications permettront à l'EDA de fournir des superficies correspondant davantage aux superficies estimées d'après les photographies aériennes. La forte relation entre les données de l'EDA et celle des photographies aériennes, pour les segments échantillonnés, est encourageante et justifie la correction des estimations par satellite grâce aux données recueillies par les agents recenseurs de l'EDA.

Une description erronée des limites et de la superficie des champs fut une autre cause de l'écart entre les données EDA et celles des photographies aériennes. Les données de l'EDA sur la superficie des champs ont été obtenues en interrogeant les exploitants agricoles, et ainsi, elles ont été souvent indiquées en multiples de 5 acres. Les graphes des données EDA en fonction des superficies estimées d'après les photographies aériennes, par zone assignée aux enquêteurs, indiquent une différence possible dans les déclarations des répondants, selon le secteur d'affectation. Ces différences pourraient être attribuables aux enquêteurs, mais aussi à d'autres facteurs comme l'emplacement géographique, la structure des champs à l'intérieur des segments d'échantillon, ou bien à la variabilité des erreurs dans les déclarations. On a observé davantage de variations dans la région sise à l'est de la rivière Saint-Jean, où la superficie moyenne des champs était généralement plus grande. La relation entre les données EDA et les données recueillies par photographie aérienne était plus forte au niveau des segments qu'au niveau des champs.

9. RÉSUMÉ ET CONCLUSIONS

En 1980, on a utilisé des données recueillies par satellite pour estimer

assez précisément la superficie plantée en pommes de terre dans trois importants comtés producteurs de cette culture, au Nouveau-Brunswick. Tout au long du projet décrit dans cette communication, les auteurs ont raffiné les méthodes d'analyse et les procédures sur le terrain, afin d'accroître la précision des estimations, de réduire le fardeau imposé aux répondants et de fournir des données spatiales détaillées qui n'étaient disponibles auparavant que pour les années de recensement.

10. REMERCIEMENTS

Les auteurs désirent remercier pour leurs conseils techniques et leurs commentaires rédactionnels Mme N. Chinnappa, de Statistique Canada, et J. Cihlar et F. Ahern du Centre canadien de télédétection.

11. BIBLIOGRAPHIE

- [1] Ahern, F.J. et J. Murphy, 1978, "Radiometric Calibration and Correction of Landsat 1, 2, 3, MSS Data", CCRS Research Report 78-4, Énergie, Mines et Ressources, Ottawa, Canada.
- [2] Brown, R.J., Ahern, F.J., Ryerson, R.A., Thomson, K.P.B., Goodenough, D.G., McCormick, J.A. et Teillet, P.M., 1980, "Rapeseed: Operational Monitoring", 6ieme Symposium canadien sur la télédétection, Halifax, Nouvelle-Écosse, Canada pp. 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), mai 1979, 3-17.
- [5] Guertin, F.G., 1979, Butlin, T.J. and Jones, R.G., 1979, "Correction géométrique des images Landsat au Centre canadien de télédétection", Canadian Journal of Remote Sensing, 5(2), décembre 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", Proceeding of the Fifth Machine Processing of Remote Sensed Data Symposium, Lafayette, Indiana.

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

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

- [9] Ryerson, R.A., P.N. Mosher et Harvie J., 1980, "Potato Area Estimation Using Remote Sensing Methods", CCRS Users Manual 80-2, Énergie, Mines et Ressources, Ottawa, février 1980.

TABLEAU I

ESTIMATIONS DE LA SUPERFICIE TOTALE PLANTÉE EN POMMES DE TERRE
DANS LA RÉGION D'ÉTUDE ET POUR LA PROVINCE DU NOUVEAU-BRUNSWICK¹

Enquête et/ou moyen d'estimation	Estimation pour la région d'étude (en acres)	Estimation pour le Nouveau-Brunswick (en acres)	Coefficient de variation (%)
Satellite			
Données non corrigées	42,354	n.d.	n.d.
Estimation par le quotient	49,504	51,524	5.5
Estimation par régression	49,115	51,119	5.5
Enquête postale	n.d.	50,800	n.d.
EDA à bases multiples	n.d.	53,854	5.2
EORPT	47,203	49,129	4.59
Estimation de Statistique Canada publiée le 5 septembre	n.d.	52,000	n.d.

¹ La région d'essai, regroupant les comtés de Carleton, de Madawaska et le comté Victoria, représente environ 96.08 % des terrains producteurs de pommes de terre au Nouveau-Brunswick.

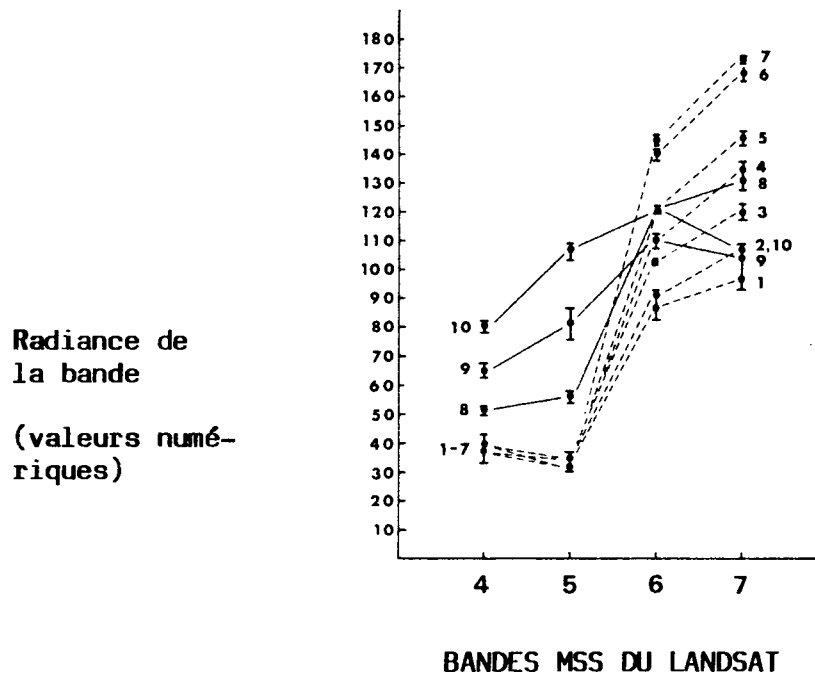
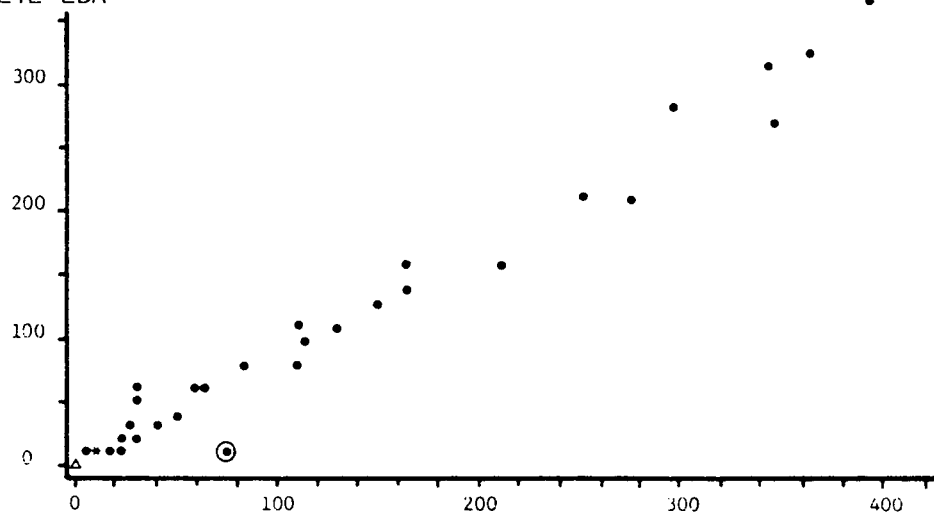


FIGURE I COMPARAISON DES DIVERSES BANDES LANDSAT POUR DIFFÉRENTES RÉCOLTES

Radiance des bandes MSS de Landsat (8 août 1975) pour les céréales (1), le sarrasin (2), les pâturages (3), le foin (4), le maïs (5), les pois (6), les pommes de terre (7), le brocoli (8), le sol nu (9), les herbes (10). Les bandes sont numérotées de 4 à 7. L'espacement relatif entre les valeurs numériques dans chaque colonne indique le degré de séparabilité des récoltes pour une bande donnée.

SUPERFICIE D'APRÈS
L'ENQUÊTE EDA

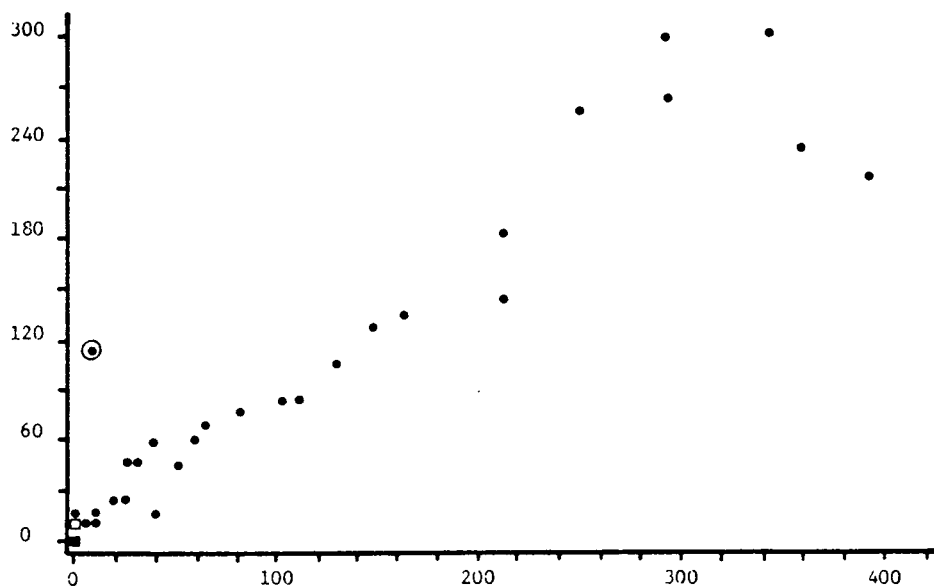
FIGURE 2



SUPERFICIE ESTIMÉE D'APRÈS LES PHOTOGRAPHIES AÉRIENNES

SUPERFICIE ESTIMÉE
PAR SATELLITE

FIGURE 3



SUPERFICIE ESTIMÉE D'APRÈS LES PHOTOGRAPHIES AÉRIENNES

FIGURES 2 et 3 - Graphes des superficies estimées, (en acres), respectivement, par l'EDA et par satellite en fonction des superficies estimées d'après les photographies aériennes à basse altitude, pour des segments échantillonnés du Nouveau-Brunswick. Les points encerclés représentent les valeurs aberrantes.

Légende: • = 1 observation, □ = 2 observations,
* = 3 observations, ■ = 11 observations et Δ =
15 observations

ÉCHANTILLONNAGE À DEUX REPRISES AVEC PPTSR

G.H. Choudhry et Jack E. Graham¹

On décrit une théorie d'échantillonnage à deux reprises avec probabilités inégales et sans remise. La méthode de Fellegi (1963), où à chaque reprise la même probabilité de sélection est attribuée à une unité donnée, est utilisée pour choisir les unités de l'échantillon de renouvellement. On examine ensuite le développement mathématique de la variance des estimateurs composites de la valeur globale de la population à la deuxième reprise. Des résultats quantitatifs pour des échantillons de petite taille sont présentés et on compare l'efficacité de cette méthode à celle d'une technique différente.

1. INTRODUCTION

Dans les enquêtes à caractère répétitif, l'utilisation d'un plan d'échantillonnage avec remise partielle présente certains avantages tant du point de vue de l'efficacité que de la réduction du fardeau du répondant. Essentiellement, après chaque tirage d'un échantillon, une fraction des unités sont supprimées et remplacées par un nouvel sous-échantillon de la population. Un très grand nombre d'ouvrages statistiques examinent les méthodes et les estimateurs que l'on utilise pour l'échantillonnage à deux reprises ou plus avec probabilités égales. Mais les cas où les probabilités de sélection sont inégales présentent encore plus d'importance du point de vue de la pratique. Ainsi, prenons une population finie de N unités $\{1, 2, \dots, N\}$ qui est échantillonnée à deux reprises, soit en période 1 (l'échantillon précédent) et en période 2 (l'échantillon courant). Soient y_{1i} et y_{2i} les valeurs d'une caractéristique y de la i ème unité dans les échantillons 1 et 2 et Y_1 et Y_2 les valeurs globales de la population calculées à partir des échantillons correspondants. Une mesure de taille, x_i , est connue pour toutes les unités de la population.

¹ G.H. Choudhry, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, et Jack E. Graham, Université Carleton.

Raj (1965) a étudié le plan d'échantillonnage suivant avec PPT (probabilité proportionnelle à la taille) : un premier échantillon s de taille n est constitué avec probabilités p_i proportionnelles aux valeurs de x_i et avec remise (AR). Dans une deuxième période, un échantillon aléatoire simple s_1 de m unités est tiré de s sans remise (SR) et un échantillon indépendant s_2 de taille $u = n - m$ est sélectionné avec PPT à partir de l'ensemble de la population. Les valeurs de Y_1 et Y_2 sont alors estimées à l'aide des équations :

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

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

et Q^* est un poids, $0 \leq Q^* \leq 1$.

La variance de $\hat{Y}_{2R}^*$ a été réduite au minimum à partir de l'hypothèse selon laquelle l'expression

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

est identique pour les temps $t = 1$ et $t = 2$.

Les ouvrages statistiques ont prêté une attention considérable au problème de l'échantillonnage à une reprise avec PPTS. Une des difficultés principales est la spécification de méthodes pratiques pour obtenir certaines probabilités données à chaque tirage. Fellegi (1963) a proposé une technique où la probabilité de sélectionner l'unité i à chacun des n tirages est égale à p_i si l'on détermine $n - 1$ ensembles de "probabilités de travail". Ce procédé est extrêmement important pour le renouvellement d'échantillons, où il est essentiel que l'estimateur habituel de Y_2 pour les échantillons constitués avec PPT ne soit pas biaisé; cet objectif ne sera pas atteint si p_i n'est pas constant à chacun des n tirages d'un plan d'échantillonnage avec renouvellement partiel. Jusqu'à récemment, toutefois, les calculs nécessaires pour la technique de

Fellegi étaient beaucoup trop compliqués pour $n > 2$. Choudhry (1981) a conçu une méthode itérative pour mettre en oeuvre la technique de Fellegi et il a élaboré un programme informatique qui évalue les probabilités de travail pour $n \leq 5$. Même si la convergence est rapide du point de vue du nombre d'itérations, le nombre de calculs augmente par une proportion égale à N^n . Le programme indique aussi pour le calcul de la variance la probabilité composée que les unités i et j soient incluses en même temps dans l'échantillon.

Rao, Hartley et Cochran (1962) ont mis au point la "méthode des groupes aléatoires" pour la sélection d'un échantillon avec PPSR. La population de N unités est divisée en n groupes de taille $N_1, N_2, \dots, N_n$, où $\sum N_h = N$, et un échantillon d'une unité est prélevé indépendamment de chaque groupe avec probabilité proportionnelle aux valeurs de p_i . Ghangurde et Rao (1969) ont adapté la méthode des groupes aléatoires à l'échantillonnage à deux reprises. Pour simplifier la description, les N unités sont réparties en n groupes de taille N/n (nombre qu'on suppose entier). Dans un premier temps, une unité est tirée de chaque groupe aléatoire, tel qu'il a été mentionné plus haut, pour former un échantillon s de n unités. À un deuxième moment, un échantillon aléatoire simple s_1 de m unités appariées est sélectionné SR parmi les n unités et un échantillon indépendant s_2 de $u = n - m$ unités est recueilli à partir de l'ensemble de la population par la même méthode appliquée pour produire l'échantillon s . Ghangurde et Rao utilisent un estimateur composite $\hat{Y}'_{2G}$ de Y_2 pour réduire au minimum la variance de cette caractéristique à l'aide d'une valeur optimale du poids Q . Ensuite, la valeur optimale de $\lambda = m/n$ est déterminée. Ces auteurs font remarquer qu'il serait probablement plus efficace de tirer s_2 parmi les $N - n$ unités de la population non incluses dans s .

Chotai (1974) a modifié la méthode de Ghangurde et Rao (G-R) pour l'échantillonnage à la deuxième reprise; les n unités de s sont choisies au hasard pour former m groupes de taille n/m (nombre qu'on suppose entier). Une unité est tirée de chacun des m groupes avec probabilité proportionnelle aux valeurs de p_i , pour constituer un échantillon s_1 . Ensuite, un échantillon s_2 est prélevé à l'aide de la méthode G-R. Une fois que la variance optimale de

l'estimateur composite $\hat{Y}'_{2c}$ est calculée, Chotai détermine la valeur optimale de λ et compare l'efficacité relative de $\hat{Y}'_{2c}$ par rapport aux estimateurs optimaux de Ghangurde et Rao et de Raj. Il s'est avéré que $\hat{Y}'_{2c}$ est toujours plus efficace que $\hat{Y}^*_{2R}$ et, dans bien des cas, plus efficace que $\hat{Y}'_{2G}$. Chotai examine brièvement le cas où n/m n'est pas un nombre entier. Il convient de noter que, étant donné que λ n'est pas une fonction continue, la valeur optimale de λ devrait être calculée par des méthodes de programmation en nombres entiers. Dans la section suivante, on présente une méthode d'échantillonnage qui conduit souvent à des résultats plus efficaces que ceux obtenus au moyen des techniques d'échantillonnage avec PPTSR proposées jusqu'à présent.

2. MÉTHODE D'ÉCHANTILLONNAGE

2.1 Sélection de l'échantillon

Dans la population de N unités $(1, 2, \dots, N)$, choisissons un échantillon qui contient $n + u$ unités ($u < n$) tirées une à la fois sans remise selon la méthode de Fellegi, où la probabilité de sélectionner la $i^{\text{ème}}$ unité à chaque tirage est p_i , $i = 1, 2, \dots, N$, $\sum p_i = 1$. Lors du premier échantillonnage, les n premières unités des $n + u$ unités sont examinées, tandis qu'au deuxième échantillonnage, les u premières unités sont supprimées de l'échantillon et les u unités non étudiées la première fois sont incluses. Ainsi, $m = n - u$ unités sont observées dans les deux cas. Les n unités examinées la première fois composent l'échantillon s , les unités étudiées à deux reprises constituent l'échantillon s_1 (où $s_1 \subset s$) et l'ensemble d'unités introduites seulement à la deuxième période représentent l'échantillon s_2 . Soulignons que la méthode de Fellegi garantit que la probabilité de sélection d'une unité i est la même à chaque tirage et donc au cours des deux périodes. Chaudhuri (1980) a étudié une sous-catégorie d'estimateurs non biaisés de modèles linéaires non homogènes et il a démontré que le plan d'échantillonnage exposé plus haut conduit à

une méthode optimale. Ce résultat offre une raison de plus d'utiliser la méthode de Fellegi.

2.2 Méthode d'estimation

Dans cette section, on propose des estimateurs composites de Y_2 , la valeur globale courante, et on évalue leur variance à l'aide d'une variable nominale.

Posons que $r a_i = 1$ si l'unité i , $i = 1, 2, \dots, N$ est sélectionnée au tirage r , $r = 1, 2, \dots, n+u$ et que $r a_i = 0$ autrement. Comme l'espérance mathématique de $r a_i$ est égale à p_i , un estimateur non biaisé de la valeur globale Y_1 de la population calculée la première fois s'écrit :

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

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

est un estimateur non biaisé de la valeur globale Y_2 obtenue la deuxième fois. Un estimateur non biaisé de Y_2 qui est une fonction des observations recueillies à la deuxième reprise est

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

Un estimateur composite de Y_2 correspond à la moyenne pondérée

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

où $0 \leq Q \leq 1$.

La variance de $\hat{Y}_{2c}$, $\text{Var}(\hat{Y}_{2c}) = Q^2 \text{Var}(\hat{Y}'_2) + (1-Q)^2 \text{Var}(\hat{Y}_2) + 2Q(1-Q) \text{Cov}(\hat{Y}'_2, \hat{Y}_2)$,

est le résultat des propriétés suivantes de la variable nominale ${}_ra_i$:

$$\text{Var}({}_ra_i) = p_i (1 - p_i), (i = 1, 2, \dots, N, r = 1, 2, \dots, n+u),$$

$$\text{Cov}({}_ra_i, {}_ta_i) = -p_i^2, (r \neq t),$$

$$\text{Cov}({}_ra_i, {}_ra_j) = -p_i p_j, (i \neq j),$$

$$\text{Cov}({}_ra_i, {}_ta_j) = E({}_ra_i \cdot {}_ta_j) - p_i p_j, \text{ autrement,}$$

où $E(.)$ correspond à l'espérance mathématique de l'expression entre parenthèses dans le plan d'échantillonnage probabiliste. Or $E({}_ra_i, {}_ta_j) = P({}_ra_i \cdot {}_ta_j = 1) = P({}_ra_i = 1, {}_ta_j = 1)$, où $P(.)$ représente une probabilité.

Soit $\Sigma_{(k-2; i, j)}$ la sommation pour toutes les valeurs possibles des $(k-2)$ -uples formées de différentes unités $\{i_1, i_2, \dots, i_{r-1}, i_{r+1}, \dots, i_{k-2}, i_{k-1}\}$ incluses dans l'échantillon lors des k premiers tirages à partir des $N-2$ unités de l'ensemble $\{1, 2, \dots, i-1, i+1, \dots, j-1, j+1, \dots, N\}$, où la i ème unité est sélectionnée au tirage r et la j ème unité au tirage k . Cette sommation comporte $(N-2)(N-3)\dots(N-k+1)$ termes.

Comme dans l'analyse de Fellegi (1963), définissons $\{p_i(\ell); i=1, 2, \dots, N\}$ comme étant l'ensemble des "probabilités de travail" pour la sélection d'une unité au tirage ℓ , $\ell = 1, 2, \dots, n+u$. Aux tirages k et r , où $k > r$:

$$E({}_ra_i \cdot {}_ka_j) = \Sigma_{(k-2; i, j)} p_{i_1}(1) \frac{p_{i_2}(2)}{1 - p_{i_1}(2)} \dots \frac{p_{i_{r-1}}(r-1)}{1 - \sum_{\ell=1}^{r-2} p_{i_\ell}(r-1)} \\ \times \frac{p_i(r)}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}(r)} \times \frac{p_{i_{r+1}}(r+1)}{1 - \sum_{\ell=1}^{r-1} p_{i_\ell}(r+1) - p_i(r+1)} \times$$

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

Or

$$\begin{aligned} \text{Var}(\hat{Y}'_2) &= \frac{1}{n^2} \text{Var} \left[\sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i \right] + \frac{1}{m^2} \text{Var} \left[\sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right] \\ &+ \frac{2}{mn} \text{Cov} \left[\sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i, \sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right] \end{aligned}$$

À l'aide des propriétés des variables nominales $r a_i$ présentées plus haut, on peut démontrer que :

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

$$\begin{aligned} \frac{1}{m^2} \text{Var} \left[\sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right] &= \frac{1}{m} \sum_{i=1}^N p_i (z_{2i} - z_{1i})^2 \\ &+ \frac{1}{m^2} \sum_{i \neq j} \sum P(i, j \in s_1) (z_{2i} - z_{1i})(z_{2j} - z_{1j}) - (Y_2 - Y_1)^2, \end{aligned}$$

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

En outre,

$$\begin{aligned} \frac{1}{mn} \text{Cov} \left[\sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i, \sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right] \\ = \frac{1}{n} \sum_{i=1}^N p_i z_{1i} (z_{2i} - z_{1i}) + \frac{1}{mn} \sum_{i \neq j} \sum P(i \in s, j \in s_1) z_{1i} (z_{2j} - z_{1j}) - Y_1(Y_2 - Y_1). \end{aligned}$$

Si l'on substitue les trois expressions présentées ci-dessus dans l'équation de $\text{Var}(\hat{Y}'_2)$, on obtient la formule :

$$\begin{aligned} \text{Var}(\hat{Y}'_2) = & \sum_1^N p_i \left[\frac{z_{2i}^2}{n} + (z_{2i} - z_{1i})^2 \left(\frac{1}{m} - \frac{1}{n} \right) \right] + \sum_{i \neq j} \left[\frac{P(i, j \in s)}{n^2} z_{1i} z_{1j} \right. \\ & + \frac{P(i, j \in s_1)}{m^2} (z_{2i} - z_{1i})(z_{2j} - z_{1j}) \\ & \left. + \frac{2P(i \in s, j \in s_1)}{nm} z_{1i} (z_{2j} - z_{1j}) \right] - Y_2^2. \end{aligned} \quad (1)$$

En outre,

$$\text{Var}(\hat{Y}_2) = \frac{1}{n} \sum_i p_i z_{2i}^2 + \frac{1}{n^2} \sum_{i \neq j} P(i, j \in s^*) z_{2i} z_{2j} - Y_2^2, \quad (2)$$

où s^* est l'ensemble des n unités observées à la deuxième reprise et

$$\begin{aligned} \text{Cov}(\hat{Y}'_2, \hat{Y}_2) = & \sum_{i \neq j} \left[\frac{P(i \in s, j \in s^*)}{n^2} z_{1i} z_{2j} + \frac{P(i \in s_1, j \in s^*)}{nm} (z_{2i} - z_{1i}) z_{2j} \right] \\ & + \frac{1}{n} \sum_i p_i z_{2i} (z_{2i} - \frac{u}{n} z_{1i}) - Y_2^2 \end{aligned} \quad (3)$$

Si l'on combine les expressions (1), (2) et (3), on obtient $\text{Var}(\hat{Y}_{2c})$.

La valeur optimale du poids Q qui réduit au minimum $\text{Var}(\hat{Y}_{2c})$ est

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

La variance minimum ainsi obtenue est

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

Un autre estimateur composite $\hat{Y}_{2c}^*$ de Y_2 est

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

où

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

Pour calculer la variance de $\hat{Y}_{2c}^*$, on utilise la formule (1) et les expressions suivantes :

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

$$\begin{aligned} \text{Cov}(\hat{Y}_2', \hat{Y}_{2u}) &= \frac{1}{nu} \sum_i \sum_{i \neq j} P(i \in s, j \in s_2) z_{1i} z_{2j} \\ &+ \frac{1}{mu} \sum_i \sum_{i \neq j} P(i \in s_1, j \in s_2) (z_{2i} - z_{1i}) z_{2j} - Y_2^2 . \end{aligned}$$

2.3 Cas spécial

Pour vérifier les résultats des calculs précédents, prenons le cas de l'échantillonnage aléatoire simple sans remise.

Ainsi, $\hat{Y}_2' = N(\bar{y}_1 + (\bar{y}_{2m} - \bar{y}_{1m}))$, où $\bar{y}_1$ et $\bar{y}_{1m}$ sont les moyennes de l'échantillon calculées respectivement pour toutes les unités échantillonnées et pour toutes les unités appariées à l'échantillon précédent et $\bar{y}_{2m}$ est la moyenne de l'échantillon calculée pour toutes les unités appariées de l'échantillon courant. Une évaluation directe conduit au résultat :

$$\text{Var}(\hat{Y}'_2) = N^2 \left[\left(\frac{1}{m} - \frac{1}{n} \right) (S_1^2 - 2S_{12}) + \left(\frac{1}{m} - \frac{1}{N} \right) S_2^2 \right]$$

où, par exemple :

$$S_{12} = \sum_{i=1}^N (y_{1i} - \bar{y}_1) (y_{2i} - \bar{y}_2) / (N-1).$$

Cette expression concorde avec l'équation (1) dans le cas où $p_i = 1/N$ et $P(i, j) = n(n-1)/N(N-1)$ ($i \neq j$).

De plus, dans l'échantillonnage aléatoire simple, $\hat{Y}_2 = N\bar{y}_2$ (où $\bar{y}_2$ est la moyenne de l'échantillon calculée à partir de toutes les n unités de l'échantillon constitué à la deuxième reprise) et la variance de $\hat{Y}_2$ est

$$\text{Var}(\hat{Y}_2) = N(N-n)S_2^2.$$

Une évaluation de $\text{Var}(\hat{Y}_2)$ au moyen de l'expression (2) produit le même résultat. Enfin, on peut obtenir l'expression suivante soit par une évaluation directe ou à partir de l'équation (3) :

$$\text{Cov}(\hat{Y}'_2, \hat{Y}_2) = -NS_2^2$$

De façon semblable, on peut calculer $\text{Var}(\hat{Y}_{2c}^*)$.

3. EXEMPLES QUANTITATIFS

On compare ici l'efficacité des estimateurs composites $\hat{Y}_{2c}$ et $\hat{Y}_{2c}^*$, auxquels correspond une valeur optimale, Q et Q^* respectivement, qui réduit au minimum la variance de l'estimation, et celle de l'estimateur $\hat{Y}_2$ calculé à partir des renseignements courants recueillis auprès d'un échantillon sélectionné avec

PPT. Comme on ne dispose pas d'une forme fermée de $\text{Var}(\hat{Y}_{2c})$ et de $\text{Var}(\hat{Y}_{2c}^*)$ pour permettre des comparaisons analytiques, on a utilisé de petites populations pour obtenir des valeurs de variables et mesurer les contrastes entre les estimateurs. (Les populations étudiées devaient être de petite taille, comme celles qu'on trouve souvent dans les échantillons stratifiés, puisque l'effet de l'échantillonnage avec ou sans remise se voit seulement quand les fractions de sondage sont non négligeables.) Quatre plans d'échantillonnage avec renouvellement ont été appliqués à chaque population : $(n,m) = (2,1)$, $(3,2)$, $(3,1)$ et $(4,3)$, où n est le nombre d'unités de l'échantillon à chaque reprise et m est le nombre d'unités incluses dans l'échantillon à la première et à la deuxième reprise. Deux populations proviennent de l'ouvrage de Murthy (1967), qui a divisé une seule population de 34 villages en deux populations de 16 et de 17 villages (une unité extrême a été exclue). La caractéristique x qui sert à mesurer la taille est le nombre d'acres cultivés en 1961; y_1 et y_2 sont le nombre d'acres de blé en 1963 et en 1964 respectivement. Une troisième population est un ensemble de 14 fermes dans la province de la Saskatchewan où x est le nombre d'acres des fermes en 1980 et y_1 et y_2 sont le nombre d'acres ensemencés en 1980 et en 1981 respectivement. On analyse aussi deux autres ensembles de données recueillies parmi une population de 15 unités et une autre de 16 unités.

Le tableau 1 montre l'efficacité relative de $\hat{Y}_{2c}$ et $\hat{Y}_{2c}^*$ par rapport à $\hat{Y}_2$ pour chacune des cinq populations et les quatre plans d'échantillonnage. Un paramètre très important dans chaque comparaison est la corrélation ρ_z entre $z_{1i} = y_{1i}/p_i$ et $z_{2i} = y_{2i}/p_i$:

$$\rho_z = \frac{\sum_{i=1}^N p_i z_{1i} z_{2i} - Y_1 Y_2}{\sqrt{\sum_{i=1}^N p_i z_{1i}^2 - Y_1^2} \sqrt{\sum_{i=1}^N p_i z_{2i}^2 - Y_2^2}} .$$

Les valeurs de ρ_z calculées pour les populations étudiées varient entre

0.940 et 0.213. Le tableau 1 indique également les valeurs optimales de Q et Q^* . On prévoit effectuer une étude ultérieure qui portera sur l'efficacité de $\hat{Y}_{2c}$ et $\hat{Y}_{2c}^*$ lorsque les valeurs choisies pour Q et Q^* ne sont pas optimales.

Les conclusions suivantes se dégagent de ces données empiriques: 1) La valeur optimale de Q est généralement plus grande quand ρ_z est grand et que la valeur optimale de Q diminue, que l'estimateur soit $\hat{Y}_{2c}$ ou $\hat{Y}_{2c}^*$. 2) La valeur optimale de Q pour $\hat{Y}_{2c}^*$ dépasse toujours celle de Q pour $\hat{Y}_{2c}$. 3) Lorsque ρ_z diminue, l'efficacité de $\hat{Y}_{2c}$ par rapport à $\hat{Y}_2$ diminue aussi (tel que prévu) et se rapproche de l'unité comme limite inférieure quand la valeur de Q est optimale. Par contre, ce genre de phénomène ne s'observe pas clairement dans le cas de $\hat{Y}_{2c}^*$ puisque $\text{Var}(\hat{Y}_{2c}^*)$ n'est pas une fonction monotone de ρ_z . Pour de petites valeurs de ρ_z , on enregistre de faibles gains et de faibles pertes d'efficacité par rapport à $\hat{Y}_2$. 4) $\hat{Y}_{2c}^*$ est plus efficace que $\hat{Y}_{2c}$ quand la valeur de ρ_z est élevée, alors que $\hat{Y}_{2c}$ est plus efficace que $\hat{Y}_{2c}^*$ quand la valeur de ρ_z est basse. 5) Si $\lambda = m/n$ est faible, comme dans le plan (3,1), on peut réaliser d'importants gains d'efficacité si l'on utilise $\hat{Y}_{2c}^*$ au lieu de $\hat{Y}_2$ quand la valeur de ρ_z est élevée. Quand ρ_z est faible, le plan (4,3) conduit aux meilleurs gains d'efficacité si l'on se sert de $\hat{Y}_{2c}$; les résultats des trois autres plans d'échantillonnage sont à peu près les mêmes.

Il convient de noter que même s'il existe une bonne corrélation entre les valeurs de y_{1i} et y_{2i} , l'estimateur composite $\hat{Y}_{2c}^*$ peut toujours causer des pertes d'efficacité en comparaison de l'estimateur $\hat{Y}_2$ calculée à partir de données recueillies à une seule reprise au moyen de l'échantillonnage avec PPT. Le facteur critique est la corrélation ρ_z entre les valeurs de z_{1i} et z_{2i} . La meilleure solution est d'utiliser $\hat{Y}_{2c}$ avec la valeur optimale de Q puisque $\hat{Y}_{2c} = \hat{Y}_2$ quand $Q = 0$.

Le tableau 2 indique l'efficacité relative de $\hat{Y}_{2c}$ et $\hat{Y}_{2c}^*$ dans un plan d'échantillonnage avec PPTSR par rapport à l'estimateur $\hat{Y}_{2R}^*$ utilisé par Raj (1965) dans son plan d'échantillonnage avec PPTAR décrit plus haut. Pour que les comparaisons soient plus valables, on n'a pas supposé que $V_{ppt}(y_t)$ était égal aux deux reprises $t = 1$ et $t = 2$; on s'est servi des valeurs optimales de Q^* pour chaque couple (n,m) . Dans tous les cas, tel que prévu, les estimateurs $\hat{Y}_{2c}$ et $\hat{Y}_{2c}^*$ dans le plan d'échantillonnage SR s'avèrent plus efficaces que $\hat{Y}_{2R}^*$ dans le plan de Raj. Pour une valeur donnée de ρ_z , plus n est grand, plus le gain d'efficacité entraîné par l'échantillonnage SR est élevé. Enfin, on constate que Raj a réalisé des gains d'efficacité par rapport aux échantillons sans appariement seulement quand $\rho_z > 0.5$, alors que qu'on réalise toujours des gains d'efficacité si l'on utilise $\hat{Y}_{2c}$ dans l'échantillonnage SR, quelle que soit la valeur de ρ_z .

REMERCIEMENTS

Nous remercions le Professeur J.N.K. Rao pour ses suggestions et ses observations au cours de cette étude, ainsi que l'arbitre pour ses remarques pertinentes.

BIBLIOGRAPHIE

- [1] CHAUDHURI, A. (1980). "On Optimal and Related Strategies for Sampling on Two Occasions with Varying Probabilités", pré tirage, 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", Techniques d'enquête 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. et 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 Probability Proportional to Size", Ann. Math. Statist. 36, 327-330.
- [8] RAO, J.N.K., HARTLEY, H.O. et COCHRAN, W.G. (1962). "On a Simple Procedure of Unequal Probability Sampling Without Replacement", J.R. Statist. Soc. B. 24, no. 2, 482-491.

TABLEAU 1

Efficacité des estimateurs composites des plans d'échantillonnage SR par rapport à celle de l'estimateur du plan d'échantillonnage avec PPTSR

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 Groupe 1	17	(2,1)	0.940	0.443	1.337	0.640	1.476
		(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 Groupe 2	16	(2,1)	0.867	0.377	1.205	0.615	1.364
		(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
Nombre d'acres	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
Ensemble de données 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
Ensemble de données 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

TABLEAU 2

Efficacité des estimateurs composites des plans d'échantillonnage SR par rapport à celle de l'estimateur composite de Raj

Population	N	(n,m)	ρ_z	$\text{Var}(\hat{Y}_{2R}^*)/\text{Var}(\hat{Y}_{2C})$	$\text{Var}(\hat{Y}_{2R}^*)/\text{Var}(\hat{Y}_{2C}^*)$
Murthy Groupe 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 Groupe 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
Nombre d'acres	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
Ensemble de données 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
Ensemble de données 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

TECHNIQUES D'ENQUÊTE

1982

Vol. 8

No. 1 & 2

Préparé par les méthodologistes de Statistique Canada

TABLE DES MATIÈRES

Importance du questionnaire dans le plan de sondage R. PLATEK, D. ROYCE	1
Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada J.D. DREW, M.P. SINGH, G.H. CHOUDHRY	19
Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active du Canada ELIZABETH CLAYTON PAUL, MURRAY LAWES	33
Le biais de renouvellement de l'échantillon dans les estimations de l'EPA P.D. GHANGURDE	94
Informatisation du calcul d'estimation pour les enquêtes complexes M.A. HIDIROGLOU	112

4. REMPLACEMENTS

Bien que l'agent recenseur doive s'efforcer de recenser les mêmes arbres, branches ou vignes les années subséquentes, cela peut s'avérer impossible dans certains cas (par exemple, lorsque des branches ont été coupées, des vignes ou des arbres arrachés ou détruits d'une autre façon). Si une branche de griottier a été coupée, on doit choisir une autre branche, sur le même arbre, en suivant la méthode décrite précédemment. Si cela ne peut se faire, on doit, tout comme pour les pêcheurs et les vignes, choisir de façon aléatoire un nouvel arbre dans le même verger. Si tout le verger ou tout le vignoble a été détruit, on doit choisir un nouveau verger ou un nouveau vignoble, sur la même ferme, en suivant la méthode décrite à la section 2.4. Dans tous les cas, on fera le recensement des branches, arbres, vignes, vergers ou vignobles nouvellement sélectionnés; toutefois, les résultats obtenus ne serviront pas à l'établissement des estimations de l'année courante, mais bien seulement de l'année suivante, car seules les données appariées sont considérées.

Dans les cas (peu fréquents, espérons-le) où le producteur a cessé complètement la culture du fruit visé ou refuse de collaborer dès le contact initial, on doit choisir, pour chacune des enquêtes, un troisième échantillon répété de taille beaucoup moins importante, sans remplacement. Les méthodes à suivre pour choisir le verger ou le vignoble et les arbres, branches ou vignes-échantillon pour chaque remplacement sont les mêmes que celles qui sont décrites dans les paragraphes précédents. Chaque année, on recense ces branches, ces arbres et ces vignes, mais les données obtenues ne servent à établir les estimations que lorsqu'il est nécessaire d'inclure une de ces unités dans l'échantillon. L'effectif de l'échantillon de remplacement, par région, figure au tableau 3.

Tableau 3 : Effectif de l'échantillon de remplacement, par région, pour les enquêtes sur les fruits tendres

	RÉGION 1	RÉGION 2	RÉGION 3	RÉGION 4	TOTAL
PÊCHES	1	2	1	3	7
GRIOTTES	1	2	2	2	7
RAISINS	1	3	1	2	7

2.6.2 Enquête sur les raisins

Comme pour les griottes, il est également impossible de compter tous les raisins commercialisables sur une vigne-échantillon. Par conséquent, pour estimer le total des raisins, on compte les grappes de raisins (celles qui ont plus de 5 raisins) et on choisit au hasard, sans remplacement, 5 grappes dont on compte les fruits un à un. Comme pour les autres enquêtes, les vignes sont marquées pour identification future, les mêmes unités étant recensées d'une année à l'autre.

3. COLLECTE DES DONNÉES

Le recensement lui-même a lieu environ quatre semaines avant la récolte, chaque année. Il est très important que les vignes, les branches et les arbres qui ont été sélectionnés, de même que les vergers et les vignobles, soient bien identifiés afin de permettre aux agents recenseurs de faire leur travail dans le court laps de temps dont ils disposent. Les agents recenseurs doivent compter tous les fruits commercialisables (c'est-à-dire, tous les fruits à l'exception des fruits de qualité inférieure non parvenus à maturité ou des fruits endommagés qui ne seront pas récoltés) sur les pêcheurs-échantillon et sur les branches de griottiers qui ont été sélectionnées. On compte tous les fruits d'un pêcher principalement parce que la distribution des fruits sur cet arbre tend à être beaucoup plus inégale que celle des fruits sur un griottier [2], ce qui exclut la possibilité de simplement recenser des branches-échantillon.

Pour l'enquête sur les raisins, on compte tous les raisins composant les cinq grappes choisies, à l'exception des fruits de qualité inférieure. Comme les raisins sont souvent groupés de façon très serrée sur la grappe, il faut, la plupart du temps, récolter les fruits. Cette raison, ajoutée à des contraintes temporelles, fait qu'il est impossible de recenser toute la vigne-échantillon.

identification future, car les mêmes unités sont recensés d'une année à l'autre. (Les années subséquentes, si une vigne ou un arbre échantillonné a été détruit ou arraché ou qu'il est mort, on choisit un arbre ou une vigne de remplacement qu'on inclut dans le recensement. Cependant, les données sur cet arbre ou cette vigne de remplacement ne servent à établir les estimations qu'à partir de la deuxième année.) En outre, chaque année, on recompte les vignes ou les arbres producteurs faisant partie des vignobles ou des vergers sélectionnés afin de rendre compte de l'évolution de l'industrie. (Il est à noter que, dans le cas de l'enquête sur les raisins, on compte seulement les vignes de la variété échantillonnée, dans chaque vignoble.)

2.6 Plan d'échantillonnage - 4^e degré

Cette étape ne s'applique qu'aux enquêtes sur les griottes et les raisins.

2.6.1 Enquête sur les griottes

Il est techniquement impossible de compter toutes les griottes sur un arbre sélectionné. Afin d'estimer le nombre total de fruits destinés à la vente, on choisit une ou plusieurs branches-échantillon, avec probabilité proportionnelle à la surface de la coupe transversale d'une branche. Jessen [3] décrit cette façon de choisir une branche. Cela consiste à sélectionner une branche au point initial (ou primaire) de ramification du tronc, avec probabilité proportionnelle à la surface de la coupe transversale, et à suivre la branche choisie jusqu'au prochain point de ramification. On répète cette opération jusqu'à ce que la surface de la coupe transversale d'une branche choisie subséquentement soit égale à au moins cinq pour cent et à au plus quinze pour cent de la surface totale cumulative des coupes transversales des branches primaires. Comme il n'est pas toujours possible de choisir une telle branche, il faut, dans certains cas, recenser deux branches. La branche choisie dans chaque arbre-échantillon est marquée pour identification future, étant donné que la même unité est recensée d'une année à l'autre.

les pêches et les griottes, l'agent recenseur demande au producteur de l'aider à énumérer tous ses vergers et à établir la taille (c'est-à-dire, le nombre d'arbres) de chacun. Dans le cas de l'enquête sur les raisins, il lui demande dresser une liste similaire pour chacune des trois variétés de raisins qui nous intéressent, nommément Concord, DeChaunac et "Autres". (Il est à noter que, certaines variétés étant cultivées ensemble, un vignoble peut figurer sur plusieurs listes. Cependant, la taille déclarée du vignoble pour une variété particulière est fonction du nombre de vignes de cette variété seulement.)

Dans le cas des enquêtes sur les pêches et les griottes, on choisit, dans chaque ferme-échantillon, un verger avec probabilité proportionnelle à la taille. Dans le cas de l'enquête sur les raisins, on choisit de façon indépendante un vignoble pour chacune des trois variétés effectivement cultivées sur la ferme, également avec probabilité proportionnelle à la taille. Soulignons à quel point il est important de suivre ces méthodes avec soin afin de ne pas compromettre la validité des estimations. Il faut surveiller de près l'étape de la sélection afin de s'assurer qu'aucun biais n'est introduit en faveur des petits vergers ou des vignobles à variété unique, bien entendu, plus faciles à recenser.

Afin d'éviter un chevauchement des vergers ou des vignobles faisant partie des fermes tirées dans les deux échantillons répétés ou tirées plus d'une fois dans le même échantillon répété, on choisit en même temps tous les vergers ou les vignobles pour une ferme donnée, en se servant d'une méthode d'échantillonnage systématique, avec probabilité proportionnelle à la taille. Suite à cette sélection, on procède, de façon aléatoire, à l'inclusion des vergers ou des vignobles dans les échantillons répétés. (Il est à noter que, dans l'enquête sur les raisins, on choisit, pour les fermes figurant dans les deux échantillons répétés, deux vignobles de chaque variété cultivée.)

2.5 Plan d'échantillonnage - 3^e degré

Une fois la sélection d'un verger ou d'un vignoble terminée, on procède au comptage des vignes ou des arbres producteurs, et on tire un échantillon aléatoire simple, sans remplacement, de quatre arbres producteurs ou de cinq vignes productrices. Les vignes ou les arbres choisis sont marqués pour

2.3 Plan d'échantillonnage - 1^{er} degré

Pour chaque enquête, nous avons tiré systématiquement, dans chaque région, deux échantillons répétés et indépendants de fermes (afin d'obtenir un échantillon représentatif), avec probabilité proportionnelle au nombre total d'arbres ou de vignes déclaré par l'exploitant dans le recensement de 1981. Les effectifs totaux des deux échantillons répétés, par région, figurent dans le tableau 1. Étant donné que les deux échantillons répétés sont tirés de façon indépendante et que les fermes plus importantes ont plus de chance d'être choisies, on s'attend à un certain chevauchement entre les échantillons répétés. En fait, certaines fermes sont tellement grosses qu'elles ont non seulement la garantie de faire partie de l'échantillon mais également la possibilité de figurer plus d'une fois dans le même échantillon répété [4]. Chaque inclusion dans l'échantillon est considérée comme un événement distinct, et chaque verger ou vignoble est tiré sans remplacement chaque fois qu'une ferme est choisie. Par conséquent, le nombre réel de fermes distinctes faisant partie de l'échantillon se trouve diminué, comme on peut le voir au tableau 2.

TABLEAU 2: Nombre total de fermes distinctes faisant partie de l'échantillon, par région

	RÉGION 1	RÉGION 2	RÉGION 3	RÉGION 4	TOTAL
PÊCHES	3	22	3	27	55
GRIOTTES	4	17	15	10	46
RAISINS	4	30	4	20	58

2.4 Plan d'échantillonnage - 2^e degré

À partir du deuxième degré, une certaine partie du travail d'échantillonnage se fait sur le terrain. Une fois que le contact initial avec le fruiticulteur est établi (au printemps de 1983), il est important de consentir tous les efforts nécessaires pour obtenir sa coopération. Dans le cas des enquêtes sur

Région 3: Villes de Pelham et de Thorold dans la municipalité régionale de Niagara (Townships 11 et 12, dans le comté 29)

Région 4: Villes de Niagara Falls et de Niagara-on-the-Lake, dans la municipalité régionale de Niagara (Townships 3 et 10, dans le comté 29)

Étant donné la demande croissante d'estimations de la production fruitière par aire géographique, nous avons tiré un échantillon distinct de fermes pour chacune des quatre régions et cherché à répartir les effectifs de l'échantillon (c'est-à-dire, le nombre de fermes choisies) de façon optimale entre les régions. Cependant, étant donné l'effectif exceptionnellement réduit de la population dans certaines régions (voir le tableau 1), nous avons fait un compromis entre une répartition proportionnelle, une répartition optimale et une règle établissant un "minimum de 2 fermes par région, par échantillon répété". Cette dernière règle nous a semblé appropriée, puisqu'elle réduit la possibilité d'obtenir une non-réponse totale dans un échantillon répété donné (comme cela pourrait être le cas si on ne choisissait qu'une seule ferme par échantillon répété). Le nombre d'arbres ou de vignes, dans chaque ferme, a servi de variable pour mesurer la taille de la ferme, aux fins de la répartition de l'échantillon et de la sélection de celui-ci, avec probabilité proportionnelle à la taille, aux 1^{er} et 2^e degrés. Des résultats antérieurs [6] indiquent que d'autres variables de substitution (telles que la superficie cultivée) ne sont probablement pas meilleures que la variable fondée sur le nombre d'arbres pour mesurer la taille de la ferme.

TABEAU 1 : Effectifs de la population et de l'échantillon, par région, pour les enquêtes sur les fruits tendres

	RÉGION 1		RÉGION 2		RÉGION 3		RÉGION 4		TOTAL	
	POPU- LATION	ÉCHAN- TILLON	POPU- LATION	ÉCHAN- TILLON	POPU- LATION	ÉCHAN- TILLON	POPU- LATION	ÉCHAN- TILLON	POPU- LATION	ÉCHAN- TILLON
PÊCHES	15	4	198	22	15	4	195	30	423	60
GRIOTTES	20	4	55	20	30	20	40	12	145	56
RAISINS	67	4	275	32	46	4	164	22	552	62

2.1 Population cible, bases de sondage et effectif total de l'échantillon

La population cible des trois enquêtes sur le rendement prévu comprend tous les fruiticulteurs commerciaux de la péninsule de Niagara. Le MAAO a défini les fruiticulteurs commerciaux comme étant ceux qui ont déclaré plus de 200 - pêcheurs, 200 griottiers ou 5,000 vignes lors du recensement des fruiticulteurs (arbres fruitiers et vignes) de 1981. À partir de cette définition, nous avons établi une base de sondage distincte pour chacune des trois enquêtes. Les listes servant aux enquêtes sur le rendement en pêches, en griottes et en raisins comptent respectivement 423, 145 et 552 fruiticulteurs commerciaux. Les contraintes budgétaires du MAAO ont servi à déterminer l'effectif total de l'échantillon (nombre de vergers et de vignobles à recenser) pour chaque enquête, soit environ 60 vergers de pêches, 55 vergers de griottiers et 155 vignobles. Dans le cas de l'enquête sur les raisins, comme toutes les variétés de raisins qui nous intéressent doivent être échantillonnées sur une ferme choisie, il est impossible de connaître d'avance l'effectif final de l'échantillon. Cependant, en nous fondant sur le recensement des vignobles de 1981, nous estimons que 62 fermes fourniront un échantillon d'environ 155 - vignobles.

2.2 Stratification et répartition des effectifs de l'échantillon entre les régions

La péninsule de Niagara a été divisée en quatre régions, pour lesquelles il faut faire des estimations distinctes. Les quatre régions ont été définies comme suit (selon les limites établies lors du recensement de 1976) :

Région 1: Ville de Grimsby dans la municipalité régionale de Niagara et township de Saltfleet dans la municipalité régionale de Hamilton-Wentworth (Township 8, dans le comté 29, et township 4 dans le comté 17)

Région 2: Villes de St. Catharines et de Lincoln dans la municipalité régionale de Niagara (Townships 5 et 9, dans le comté 29)

Figure 1. Plan d'échantillonnage des enquêtes
sur le rendement prévu en fruits tendres

BASE
DE SONDAGE :

TOUS LES FRUITICULTEURS COMMERCIAUX
DES QUATRE RÉGIONS VISÉES PAR LE
RECENSEMENT DE 1981 DU MAAO

STRATIFICATION
ET RÉPARTITION :

STRATIFIER L'ÉCHANTILLON PAR RÉGION, ET
COMPENSER PAR UNE RÉPARTITION OPTIMALE DES
EFFECTIFS DE L'ÉCHANTILLON ENTRE LES QUATRE STRATES.

1er DEGRÉ :

DANS CHAQUE STRATE, CHOISIR LES FERMES EN
TIRANT UN ÉCHANTILLON RÉPÉTÉ, SYSTÉMATIQUE,
AVEC P.P.T., EN FONCTION DU NOMBRE D'ARBRES OU
DE VIGNES DÉCLARÉ DANS LE RECENSEMENT DE 1981

2e DEGRÉ :

POUR CHAQUE FERME-ÉCHANTILLON, CHOISIR UN
VERGER OU UN VIGNOBLE, AVEC P.P.T., EN FONCTION
DU NOMBRE D'ARBRES OU DE VIGNES (DANS LE CAS
DE L'ENQUÊTE SUR LES RAISINS, CHOISIR UN VIGNOBLE
POUR CHACUNE DES VARIÉTÉS DE RAISINS CULTIVÉES
SUR CETTE FERME)

3e DEGRÉ :

TIRER UN ÉCHANTILLON ALÉATOIRE DE 4 ARBRES
OU DE 5 VIGNES DANS CHACUN DES VERGERS OU DES
VIGNOBLES SÉLECTIONNÉS

4e DEGRÉ :

ENQUÊTE SUR
LES GRIOTTES

ENQUÊTE SUR
LES PÊCHES

ENQUÊTE SUR
LES RAISINS

CHOISIR, SUR CHAQUE ARBRE
SÉLECTIONNÉ UNE BRANCHE-
ÉCHANTILLON, AVEC PROBA-
BILITÉ PROPORTIONNELLE À
LA SURFACE DE LA COUPE
TRANSVERSALE

TIRER, SUR CHAQUE
VIGNE SÉLECTIONNÉE,
UN ÉCHANTILLON
ALÉATOIRE SIMPLE DE
5 GRAPPES DE RAISINS

COLLECTE
DES DONNÉES :

COMPTER TOUS LES FRUITS COMMERCIALISABLES SUR
TOUTES LES UNITÉS CHOISIES (BRANCHES DE
GRIOTTIERS, PÊCHERS, GRAPPES DE RAISINS)

étaient complètes et précises et qu'elles contenaient suffisamment de renseignements pour constituer la base de sondage des enquêtes sur les fruits tendres. Par conséquent, nous utiliserons dorénavant, pour les enquêtes sur le rendement en pêches, en griottes et en raisins, trois échantillons indépendants tirés des listes du recensement mené par le MAAO en 1981.

Les trois enquêtes ont pour objet de prévoir la quantité totale de fruits effectivement vendue (pour la vente fraîche ou la transformation). Nous établissons ces prévisions en estimant le rapport entre le nombre de fruits destinés à la vente pour l'année courante et le total correspondant pour l'année précédente et appliquons le ratio obtenu à la quantité de fruits, en tonnes, effectivement vendue l'année précédente (donnée fournie par le comité de la statistique sur les fruits et les légumes de l'Ontario). Ainsi, nous supposons qu'il y a une corrélation importante entre le poids des fruits et le nombre de fruits. Étant donné l'écart temporel entre les enquêtes et la récolte, nous supposons en outre que toute perte de fruits entre ces deux événements est constante d'une année à l'autre.

2. VUE GÉNÉRALE DU PLAN D'ÉCHANTILLONNAGE

Les échantillons des trois enquêtes sur le rendement prévu ont été tirés de façon indépendante, conformément à un plan d'échantillonnage stratifié (selon les aires géographiques), presque complètement autopondéré (tous les arbres et toutes les vignes ont à peu près la même chance d'être choisis), à plusieurs degrés, avec échantillons répétés et avec p.p.t. (les fermes de même que les vergers et les vignobles sont choisis selon une probabilité proportionnelle à leur taille). Un schéma du plan d'échantillonnage est présenté à la figure 1. Il est à noter que, comme les variables de pondération sont recueillies à des moments différents, le plan d'échantillonnage n'est pas parfaitement autopondéré.

REMANIEMENT DE L'ENQUÊTE SUR LE RENDEMENT PRÉVU EN FRUITS TENDRES DANS LA PÉNINSULE DE NIAGARA

J. Kovar¹

Des enquêtes sur le rendement prévu en pêches, en griottes (cerises acides) et en raisins sont menées chaque année, depuis 1964, dans la péninsule de Niagara; elles servent à prévoir l'amplitude des variations, d'une année à l'autre, dans la production de fruits destinés à la vente. Il est essentiel d'établir à temps les estimations afin de permettre à l'Office de commercialisation des fruits tendres de l'Ontario et à l'Office de commercialisation des raisins de l'Ontario de dresser leurs plans de commercialisation bien avant la récolte. Dans le présent rapport, nous expliquons les principales modifications apportées à l'enquête par suite du deuxième remaniement effectué en 1982. Nous examinons plus particulièrement le plan d'échantillonnage, la collecte des données et les formules d'estimation.

1. INTRODUCTION

Lors du premier remaniement, en 1974, il avait été décidé de changer la base de sondage de l'enquête, soit de se fonder sur des aires géographiques plutôt que sur une liste, principalement parce qu'on ne disposait d'aucune liste complète des fruiticulteurs commerciaux dans la péninsule de Niagara. Toutefois, en 1981, le ministère de l'Agriculture et de l'Alimentation de l'Ontario (MAAO) a fait le recensement des fruiticulteurs (arbres fruitiers et vignes). Avec les données ainsi obtenues, nous pouvons remanier l'enquête pour la deuxième fois afin de tenir compte des changements survenus dans l'industrie au cours des huit dernières années. Suite à des entretiens avec le MAAO, il a été décidé que les listes du recensement des fruiticulteurs

¹ J. Kovar, Division des méthodes d'enquêtes-entreprises, Statistique Canada. Ce travail a été fait lorsque l'auteur était à la Division des méthodes d'enquête-institutions et agriculture, Statistique Canada.

Tableau 1

		Mois Occupé						
a)		0	1	2	3	4	5	6
Mois où les données sont présentes	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					
		Mois en Chômage						
b)		0	1	2	3	4	5	6
Mois où les données sont présentes	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					

- [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] Statistique Canada, Guide d'utilisation des données de l'enquête sur la population active. No 71-528 au catalogue, hors série, juillet 1979.
- [11] Statistique Canada, Méthodologie de l'enquête sur la population active du Canada 1976. No 71-526 au catalogue. Hors série. Octobre 1977.
- [12] Wong, F. A Technique To Correct The Response Rate In the 4x4 Labour Force Gross Flow Matrix. Technical Report, Statistics Canada, 1983.

le but d'estimer des flux bruts entre les catégories de la population active serait certainement différente de l'enquête sur la population active. Ainsi, les données longitudinales de l'enquête ne sont pas idéales pour les estimations des flux bruts. Les données, cependant, sont disponibles, et si elles peuvent être utilisées pour fournir des estimations raisonnables des flux bruts, alors on produit des renseignements supplémentaires utiles à un coût relativement peu élevé.

RÉFÉRENCES

- [1] Bishop, Y.M.M., Fienberg, S.E. et Holland, P.W. Discrete Multivariate Analysis: Theory and Practice. The MIT Press, 1975.
- [2] Deming, W.E. et 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. et Tanur, J.M. The Design et 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 Presentation: 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. et Lawes, M. Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active du Canada. Techniques d'enquête, 1982, 8, 53-93.

n'utilise que les renseignements des individus qui ont répondu les deux mois. Il y a également des renseignements disponibles des individus qui n'ont répondu qu'un seul de ces deux mois. Stasny (1983) présente une méthode d'estimation des flux bruts d'un mois à l'autre qui emploie les renseignements disponibles des individus qui ont répondu seulement un de ces deux mois et qui peut être utilisée lorsque la non-réponse est reliée au temps ou au statut d'activité. Dans cette méthode, nous considérons les données de flux bruts observés comme le résultat final d'un processus à deux étapes. Dans la première étape du processus, que nous n'avons pas l'occasion d'observer, les individus sont distribués entre les 16 cellules de la matrice des flux bruts conformément à un schéma d'échantillonnage multinomial simple. Ensuite, dans la deuxième étape, chaque individu possède une probabilité quelconque de perdre son statut d'activité soit dans le mois $t-1$ ou dans le mois t . On peut établir le modèle de la probabilité de perdre le statut d'activité d'un mois selon le mois, ou le statut d'activité, ou les deux. Les estimations de probabilité maximale pour les paramètres de la distribution multinomiale de la première étape et les probabilités de perdre le statut d'un mois sont obtenues au moyen de méthodes itératives.

Lorsqu'on a appliqué ces modèles aux données de l'enquête sur la population active provenant d'un seul panel, Stasny (1983) a trouvé que le modèle où la probabilité de perdre le statut d'activité d'un mois dépend du statut d'activité correspond raisonnablement aux données pour toutes les matrices de flux bruts à l'exception de la matrice des mois 1-2. Pour les données du mois 1 au mois 2, la probabilité de la perte du statut d'activité d'un mois semble dépendre du mois. Ceci peut être dû au fait qu'il y a un taux de non-réponse plus élevé le premier mois qu'un panel est inséré dans l'enquête. Nous croyons qu'il vaudrait la peine d'appliquer ce genre de modèle à des données supplémentaires de l'enquête sur la population active pour vérifier si l'on obtient des résultats semblables avec d'autres panels.

Il est clair que le problème d'obtenir de bonnes estimations des flux bruts à partir des données de l'enquête sur la population active n'est pas un problème simple. L'enquête est organisée pour fournir des données pour la production des estimations mensuelles de l'activité sur le marché du travail, et non des estimations des flux bruts. Une enquête organisée spécifiquement dans

Considérons les probabilités sous-jacentes aux pourcentages observés présentés dans la partie a) du tableau 1. Soit

π_i = la probabilité qu'un individu soit Occupé pendant i de 6 mois pour $i = 0, 1, \dots, 6$.

En postulant que la non-réponse se produit au hasard, les probabilités correspondant à la première colonne de ce tableau peuvent s'écrire:

P (qu'on observe 0 mois Occupé sur 6-k mois de réponse)

$$= \sum_{j=1}^k \pi_j / \binom{6}{j}, \text{ pour } k = 0, 1, \dots, 5. \quad (13)$$

À noter que ces probabilités augmentent de la première à la dernière ligne de la colonne.

De même, on peut montrer que, si les données manquent au hasard, alors les probabilités sous-jacentes doivent augmenter de façon descendante dans chaque colonne dans les deux tableaux. La première colonne de chaque tableau dévie de ce comportement de façon très nette. Dans les deux cas, les pourcentages observés diminuent tout au long des quatre premières lignes du tableau et augmentent dans les deux dernières. Il ne semble pas probable que les variations d'échantillonnage par elles-mêmes puissent être responsables d'un tel comportement dans les deux tableaux. Par conséquent, il semble bien que la non-réponse ne se produise pas au hasard.

Bien entendu, cette analyse n'est basée que sur un seul panel des données de l'enquête sur la population active. Cependant, dans une étude plus considérable utilisant des données de 1980 et 1981, Paul et Lawes (1982) ont également trouvé des preuves d'une relation entre le statut d'activité et la non-réponse. Par conséquent, il est nécessaire d'envisager des méthodes d'estimation des flux bruts qui ne nécessitent pas le postulat que la non-réponse se produit au hasard.

La méthode que propose Statistique Canada pour l'estimation des flux bruts

et à l'extérieur de celle-ci. Une autre possibilité qui devrait être prise en considération est l'élimination des poids mensuels aux fins de l'estimation des flux bruts, et le calcul d'un poids longitudinal pour chaque individu inclus dans l'échantillon de l'enquête sur la population active dans l'un ou l'autre des deux mois.

En tant que statisticiens, nous acceptons facilement des estimations des flux bruts dont les totaux marginaux ne correspondent pas aux totaux de l'activité sur le marché du travail publiés mensuellement; cependant, nous sommes conscients des problèmes qui pourraient être soulevés si on publiait des estimations des flux bruts non convergentes avec les totaux mensuels. Néanmoins, on ne devrait pas postuler automatiquement que les estimations mensuelles sont justes et que le problème se situe uniquement dans les estimations des flux bruts. Comme nous l'avons noté dans la section 3.5, la matrice des flux bruts est corrigée pour tenir compte des erreurs de classification. Les estimations mensuelles, cependant, ne sont pas corrigées pour le biais d'erreur de classification. En conséquence, lorsque l'ajustement proportionnel itératif est utilisé pour corriger la matrice des flux bruts pour la faire correspondre aux totaux mensuels, on change la matrice pour la faire converger avec des valeurs biaisées. Nous croyons qu'il serait plus approprié de faire face au problème des erreurs de classification du statut d'activité dans les données mensuelles là où elles se produisent plutôt qu'uniquement dans les estimations des flux bruts.

5. NON-RÉPONSE ET ESTIMATION DES FLUX BRUTS

La méthode proposée par Statistique Canada d'estimation des flux bruts corrige la non-réponse en ajustant les poids des répondants en fonction de l'échantillon. Cette méthode de traitement de la non-réponse est correcte si les données manquantes sont distribuées au hasard (voir Rubin, 1976). Pour explorer le postulat de la répartition au hasard de la non-réponse, nous avons utilisé un fichier longitudinal pour un seul panel pour produire les données du tableau 1. Ce tableau indique les pourcentages non pondérés des individus interviewés occupés ou en chômage entre 0 à 6 mois selon le nombre de mois où ils ont répondu à l'enquête.

active n'est pas conçue pour l'estimation de nombres de personnes à l'extérieur de la population étudiée. Si nous voulons obtenir des estimations raisonnables pour les cellules d'entrée et de sortie de la matrice, il peut devenir nécessaire d'inclure des individus de l'extérieur de la population étudiée dans l'échantillon de l'enquête sur la population active ou d'utiliser un échantillon spécial supplémentaire.

Dans la section 4.3, nous avons vu que les surestimations des cellules d'entrée et de sortie pouvaient résulter de mouvements des individus entre une strate dont la population a augmenté et une autre dont la population a diminué. Le fait que ce sont les mouvements entre les strates qui ont causé les problèmes résulte des postulats simplificateurs qui ont été posés. Nous avons postulé que l'échantillon final a été choisi au hasard à l'intérieur de chaque strate. Par conséquent, les poids attribués aux individus inclus dans l'échantillon provenant d'une même strate étaient égaux. Si, à la place, nous avions supposé que les strates avaient été divisées en grappes et que les échantillons aléatoires d'individus avaient été choisis dans ces grappes, alors tous les individus inclus dans l'échantillon provenant d'une même grappe se seraient vu attribuer le même poids et la surestimation aurait été produite par les mouvements entre les grappes.

Pour corriger cette surestimation et les sous-estimations correspondantes, directement lorsque les échantillons finals sont choisis au hasard de l'intérieur des strates, nous aurions besoin d'estimations du nombre de sujets qui se sont déplacés entre chaque paire de strates lorsque la population d'une strate a augmenté et que celle d'une autre a diminué. Si les échantillons finals sont choisis au hasard à l'intérieur de grappes, des estimations semblables seraient nécessaires pour chaque paire de grappes. Ceci exige une quantité considérable de renseignements. Une complication supplémentaire est que, en pratique, les corrections proportionnelles appliquées aux poids rendent possible l'attribution de poids différents aux divers membres d'un même ménage.

Comme il a été suggéré plus tôt, si des individus à l'extérieur de la population étudiée étaient inclus dans l'échantillon, nous pourrions obtenir directement des estimations des mouvements à l'intérieur de la population étudiée

Finalement, nous notons que dans la matrice des flux bruts 2x2 présentée ci-haut la cellule Dans-la-population à Dans-la-population doit contenir une sous-estimation égale à la surestimation de la cellule de sortie. Quelle que soit la taille de cette surestimation, elle est étalée sur les neuf cellules Dans-la-population à Dans-la-population de la matrice des flux bruts 4x4. De plus, la taille de la surestimation est petite en comparaison avec la taille totale des neuf cellules Dans-la-population.

4.4 Commentaires sur la méthode d'estimation des flux bruts proposée

Les résultats décrits dans les deux sous-sections précédentes illustrent des problèmes de la méthode proposée pour le traitement des différences des poids entre les mois aux fins de l'estimation des flux bruts. Ces résultats ne constituent pas une surprise pour les chercheurs de Statistique Canada. Grâce à leur expérience des méthodes et des données de l'enquête sur la population active, ils savaient que les mouvements des individus à l'intérieur de la population pourraient expliquer une partie de la surestimation des cellules d'entrée et de sortie de la matrice des flux bruts. Les résultats obtenus en basant le processus sur un modèle confirment cette opinion et précisent dans quelle mesure les mouvements des individus influencent les estimations. De plus, l'établissement du modèle a fait surgir un problème dont Statistique Canada n'était pas conscient: la sous-estimation compensatoire étalée à travers les neuf cellules Dans-la-population à Dans-la-population de la matrice des flux bruts.

Dans la section 4.2, nous avons vu que les augmentations nettes des strates sont attribuées aux cellules d'entrée alors que les diminutions nettes sont attribuées aux cellules de sortie en fonction des fractions des individus observés classés comme Occupés, Chômeurs et Inactifs au cours du mois t et du mois $t-1$ respectivement. Cette technique d'estimation des entrées et des sorties n'est valide que si les individus qui entrent dans la population étudiée et qui en sortent constituent un échantillon aléatoire des individus et, par conséquent, sont "identiques" aux individus qui demeurent à l'intérieur de la population en question. Les individus inclus dans l'échantillon qui sont classés comme Hors-de-la-population étudiée apparaissent dans l'échantillon accidentellement plutôt qu'intentionnellement; l'enquête sur la population

$$\begin{aligned} \sum_{u \neq A} (m_{u,A} - m_{A,u}) + \sum_{u \neq B} (m_{u,B} - m_{B,u}) &= m_{O,A} - m_{A,O} + m_{B,A} - m_{A,B} + m_{C,A} - m_{A,C} \\ &+ m_{O,B} - m_{B,O} + m_{A,B} - m_{B,A} + m_{C,B} - m_{B,C} \\ &= m_{O,A} - m_{A,O} + m_{O,B} - m_{B,O} + m_{C,A} - m_{A,C} + m_{C,B} - m_{B,C}. \end{aligned}$$

À noter que les mouvements entre les strates A et B s'annulent, mais que les termes indiquant les mouvements entre les strates A et C et les strates B et C demeurent dans la somme.

En général, les cellules d'entrée contiennent des termes supplémentaires de forme $m_{v,u} - m_{u,v}$ pour chaque strate v qui a perdu des sujets alors que la strate u a gagné des sujets. De même, la cellule de sortie contient des termes supplémentaires de forme $m_{x,y} - m_{y,x}$ pour chaque strate y qui a augmenté alors que la strate x a perdu des sujets.

Dans la cellule d'entrée, la quantité $\sum_{u \neq s} (m_{u,s} - m_{s,u})$ pour chaque strate s qui a gagné des sujets entre les mois $t-1$ et t sera positive, bien que chaque terme distinct de la somme ne soit pas nécessairement positif. Si

$$\sum_{u \neq s} (m_{u,s} - m_{s,u}) > m_{O,s} - m_{s,O} \quad (12)$$

alors la contribution à la strate s est supérieure à l'entrée dans la strate s en provenance de l'extérieur de la population étudiée. Cet excédent provient des termes de forme $m_{u,v} - m_{v,u}$ comme décrit ci-haut. C'est-à-dire que la surestimation est due à des mouvements entre les strates à l'intérieur de la population. On obtient un résultat semblable pour la cellule Dans-la-population à Hors-de-la-population de la matrice.

Les analystes de Statistique Canada signalent que la méthode qu'ils ont proposée pour le traitement des différences de poids entre les mois semble produire des surestimations des cellules d'entrée et de sortie de la matrice des flux bruts. Bien qu'ils soient basés sur des postulats simplificateurs, nos résultats indiquent une explication possible pour cette surestimation, qui peut être imputable à des mouvements à l'intérieur de la population étudiée.

est trouvée de la même façon. Ainsi, la matrice des flux bruts 2x2 se présente comme suit

		Mois t		
		Dans-la-population	Hors-de-la-population	
Mois t-1	Dans-la- population	$\sum_{s=1}^S \{N_t^s + \min[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]\}$	$\sum_{s=1}^S \max[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]$	$\sum_{s=1}^S N_{t-1}^s$
	Hors-de-la- population	$\sum_{s=1}^S \max[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]$	0	

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

Considérons la quantité qui se trouve dans la cellule entrée de cette matrice des flux bruts. Cette cellule devrait contenir l'augmentation nette de la population qui de la population qui provient de l'extérieur de la population étudiée, $m_{0,s} - m_{s,0}$, pour chaque strate qui a reçu des sujets de l'extérieur de la population. Ce que cette cellule contient est $\sum_{u \neq s} (m_{u,s} - m_{s,u})$ pour chaque strate s qui a augmenté à la suite de mouvements entre les strates et en provenance de l'extérieur de la population étudiée. La somme $\sum_{u \neq s} (m_{u,s} - m_{s,u})$ inclut la quantité $m_{0,s} - m_{s,0}$ mais elle peut également contenir d'autres termes.

Par exemple, supposons que la population est constituée de trois strates appelées A, B et C. Si les strates A et B ont augmenté entre le mois t-1 et le mois t, et que la strate C a perdu des sujets, alors la cellule d'entrée contient

$$N_t^s = N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u}). \quad (7)$$

Les poids attribués aux individus dans la strate s dans les mois $t-1$ et t respectivement sont

$$W_{t-1}^s = N_{t-1}^s / r_{t-1,t}^s \text{ et } W_t^s = N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u}) / r_{t-1,t}^s. \quad (8)$$

Étant donné que nous nous attachons dans cette section au mouvement d'entrée dans la population étudiée et de sortie de celle-ci, il n'est pas nécessaire pour nous de diviser les membres de la population étudiée entre Occupés, Chômeurs et Inactifs. Ainsi, la matrice des flux bruts utilisée ici est une matrice 2×2 formée par la sommation des trois premières lignes et des trois premières colonnes de la matrice des flux bruts 4×4 utilisée dans la sous-section précédente.

L'entrée pour la strate s dans la cellule Dans-la-population à Dans-la-population est de

$$\begin{aligned} \min(W_{t-1}^s, W_t^s) r_{t-1,t}^s &= \min \{ N_{t-1}^s / r_{t-1,t}^s, [N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})] / r_{t-1,t}^s \} r_{t-1,t}^s \\ &= \min [N_{t-1}^s, N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})] \\ &= N_{t-1}^s + \min [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]. \end{aligned} \quad (9)$$

La valeur de la strate s dans la cellule Hors-de-la-population à Dans-la-population, ou entrée est

$$\begin{aligned} \max(0, W_t^s - W_{t-1}^s) r_{t-1,t}^s &= \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u}) / r_{t-1,t}^s] r_{t-1,t}^s \\ &= \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]. \end{aligned} \quad (10)$$

La valeur de la cellule Dans-la-population à Hors-de-la-population, ou sortie

taille des strates peut produire des estimations biaisées des cellules d'entrée et de sortie de la matrice des flux bruts.

4.3 Effets des mouvements entre les strates

Les poids utilisés aux fins de l'estimation des flux bruts figurant dans l'expression (3) sont déterminés par le nombre de répondants des deux mois $t-1$ et t , une quantité qui demeure constante pour les deux mois, et par la population des strates. La population d'une strate change si a) des individus y entrent venant de l'extérieur de la population étudiée, comme lorsque les gens atteignent leur quinzième année ou quittent les Forces armées à plein temps, b) des individus sortent de la population étudiée, comme lorsqu'ils entrent dans les Forces armées ou une institution, ou c) des individus de la population étudiée changent de strate. La présente sous-section décrit les effets de ces changements de la taille de la population sur les valeurs de la matrice des flux bruts.

Comme dans la sous-section précédente, nous supposons que la population étudiée est divisée en S strates. Encore une fois, les individus de l'échantillon sont tirés au hasard de chaque strate chaque mois, sont interviewés six mois consécutifs et sont ensuite retirés de l'échantillon. Soit $r_{t-1,t}^s$, comme auparavant, le nombre d'individus de la strate s qui sont interviewés dans les deux mois $t-1$ et t .

Ensuite, nous supposons qu'il y a N_{t-1}^s individus dans la strate s au cours du mois $t-1$. Indiquons les mouvements vers cette strate et hors de celle-ci entre les mois $t-1$ et t par:

$m_{u,v}$ = nombre d'individus qui passent de u à v , $u \neq v$, entre les entrevues pour les mois $t-1$, et t où u et v peuvent prendre les valeurs

s = strate s pour $s = 1, 2, \dots, S$ et

H = hors de la population étudiée.

En employant cette notation, la population de la strate s dans le mois t est

A noter que chaque terme de la somme des neuf cellules de sujets dans la population active (les cellules qui indiquent les flux bruts entre Occupés, Chômeurs, et Inactifs) est le produit de la taille nette de la strate et de la fraction observée des sujets qui avaient les diverses classifications de la population active dans les mois $t-1$ et t . La cellule Hors-de-la-population vers Occupés de la matrice des flux bruts contient la somme des termes de chaque strate qui a augmenté entre le mois $t-1$ et le mois t . Chaque terme est le produit de l'augmentation nette de la taille de la strate et de la fraction des sujets de la strate qui se sont déclarés Occupés le mois t . Les cellules Hors-de-la-population à Chômeurs et à Inactifs contiennent les sommes de termes semblables sauf que l'augmentation nette de la taille de chaque strate est multipliée par la fraction des sujets de la strate qui étaient Chômeurs ou Inactifs respectivement le mois t . En d'autres mots, l'augmentation nette de la taille de chaque strate est attribuée proportionnellement aux trois cellules d'entrée de la matrice des flux bruts sur la base des fractions observées des Occupés, des Chômeurs et des Inactifs au cours du mois t . De même, la diminution nette de la taille de chaque strate qui a diminué entre les mois $t-1$ et t est attribuée proportionnellement aux cellules de sortie de la matrice sur la base des fractions observées des Occupés, des Chômeurs et des Inactifs au cours du mois $t-1$.

Dans ce modèle, nous avons postulé que la seule façon pour des chiffres d'apparaître dans les cellules d'entrée et de sortie est une différence de poids. En pratique, un petit nombre d'individus qui entrent et sortent de la population étudiée apparaissent dans l'échantillon et les poids qui leur sont attribués sont ajoutés aux cellules d'entrée et de sortie appropriées. L'effet de ces individus sur les estimations est très petit.

Les fractions f_{+O}^S , f_{+C}^S , f_{+I}^S , f_{O+}^S , f_{C+}^S et f_{I+}^S sont estimées en utilisant les individus qui apparaissent dans l'échantillon au cours des deux mois. Presque tous les individus classés, par exemple comme H_0 , ne pourraient pas être répondants dans les deux mois parce qu'ils n'ont pas été choisis comme sujet à cause du plan ou parce qu'ils ont déménagé. Ainsi, les gens qui n'auraient pu être répondants les deux mois sont représentées par des individus qui ont été répondants ces deux mois. Dans la mesure où ces groupes sont différents, l'allocation proportionnelle des augmentations et des diminutions nettes de la

$$= \max(0, N_t^s - N_{t-1}^s) n_{00}^s / r_{t-1,t}^s.$$

Les différences des individus tombant dans les cellules CO et IO contribueront également à la cellule HO. Ainsi, la contribution totale de la cellule HO de la strate est :

$$\begin{aligned} & \max(0, N_t^s - N_{t-1}^s) \{ (n_{00}^s / r_{t-1,t}^s) + (n_{CO}^s / r_{t-1,t}^s) + (n_{IO}^s / r_{t-1,t}^s) \} \\ &= \max[0, N_t^s - N_{t-1}^s] n_{+0}^s / r_{t-1,t}^s \\ &= \max(0, N_t^s - N_{t-1}^s) f_{+0}^s \end{aligned} \quad (6)$$

où f_{+0}^s = fraction de tous les individus de la strate s, interviewés dans dans les deux mois t-1 et t, qui étaient occupés dans le mois t.

On obtient les totaux de toutes les cellules de la matrice des flux bruts de façon semblable. La matrice des flux bruts qui en résulte est la suivante :

Matrice des flux bruts - Mois t-1 à mois t

		Mois t			
		O	C	I	H
Mois t - 1	O	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{00}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{0C}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{0I}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{0+}^s$
	C	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{C0}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{CC}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{CI}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{C+}^s$
	I	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{I0}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{IC}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{II}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{I+}^s$
	H	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+0}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+C}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+I}^s$	0
		$\sum_{s=1}^S N_t^s f_{+0}^s$	$\sum_{s=1}^S N_t^s f_{+C}^s$	$\sum_{s=1}^S N_t^s f_{+I}^s$	

$$w_{t-1}^s = N_{t-1}^s / r_{t-1,t}^s \text{ et } w_t^s = N_t^s / r_{t-1,t}^s. \quad (3)$$

Aussi longtemps que ces mouvements entre les strates et la sélection des groupes constants sont des "processus aléatoires", ces poids représentent l'inverse de la probabilité qu'un individu à l'intérieur d'une strate soit interviewé dans les deux mois $t-1$ et t . Étant donné que tous les individus à l'intérieur d'une strate ont le même poids un mois donné, on peut utiliser des agrégats pour chaque strate. Par conséquent, nous définissons:

n_{ij}^s = nombre d'individus inclus dans l'échantillon de la strate s classée dans la catégorie d'activité i au mois $t-1$ et à la catégorie j au mois t pour $i, j = 0, C, I, H$.

La méthodologie proposée par Statistique Canada exige que le minimum des poids des mois $t-1$ et t de chaque individu soit ajouté à la cellule appropriée dans la matrice des flux bruts. Cette différence est ajoutée à la cellule d'entrée appropriée si le poids du mois t est supérieur au poids du mois $t-1$ et à la cellule de sortie appropriée dans le cas contraire. Ainsi, par exemple, l'entrée de la strate s dans la cellule 00 (occupés à occupés) de la matrice des flux bruts est:

$$\begin{aligned} \min(w_{t-1}^s, w_t^s) n_{00}^s &= \min [N_{t-1}^s / r_{t-1,t}^s, (N_t^s / r_{t-1,t}^s)] n_{00}^s \\ &= \min (N_{t-1}^s, N_t^s) N_{00}^s / r_{t-1,t}^s \\ &= \min (N_{t-1}^s, N_t^s) f_{00}^s \end{aligned} \quad (4)$$

où f_{00}^s = fraction de tous les individus de la strate s , interviewés les deux mois $t-1$ et t , qui étaient Occupés ces deux mois.

La contribution de la strate s à la cellule H0 de la matrice des individus occupés dans le mois t est:

$$\max(0, w_t^s - w_{t-1}^s) n_{00}^s = \max [0, (N_t^s - N_{t-1}^s / r_{t-1,t}^s)] n_{00}^s \quad (5)$$

Étant donné que le plan de l'enquête sur la population active est plutôt complexe, nous commençons avec une série de postulats simplificateurs. Dans notre modèle, nous postulons que:

1. on choisit un échantillon stratifié à un seul degré
2. il n'y a pas d'erreur de réponse et
3. les non-réponses ne se produisent que parce que des individus au hasard changent de strates ou à cause de leur insertion ou de leur retrait de l'échantillon.

4.2 Attribution des changements nets de population aux cellules d'entrée et de sortie

Supposons que la population étudiée est divisée en S strates identifiées par $s = 1, 2, \dots, S$. Soit

N_k^s = taille de la population dans la strate s au mois k .

Chaque mois, un échantillon aléatoire simple est choisi de chaque strate pour l'enquête et les individus choisis sont interviewés six mois consécutifs avant d'être retirés de l'échantillon. Notre but est d'estimer les flux bruts entre le mois $t-1$ et le mois t .

Aux fins de l'estimation des flux bruts, seuls les individus qui sont interviewés dans les deux mois $t-1$ et t seront utilisés. Ceci exclut les individus qui sont insérés dans l'échantillon ou retirés de celui-ci et les personnes qui passent d'une strate à l'autre. Soit

$r_{t-1,t}^s$ = le nombre d'individus inclus dans l'échantillon de la strate s interviewée dans les mois $t-1$ et t .

Chacun des $r_{t-1,t}^s$ répondants de la strate s reçoit les poids suivants dans les mois $t-1$ et t respectivement aux fins de l'estimation des flux bruts:

d'une étape à l'autre.

Des vérifications à Statistique Canada ont montré que les changements de cellules résultant de l'application de la remise à l'échelle proportionnelle itérative étaient petits en valeur absolue et en valeur relative et se situaient généralement à l'intérieur des limites de variation de l'échantillon associées aux cellules. Ceci laisse croire que la correction pour la convergence ne biaise pas sérieusement les estimations des flux bruts.

3.5 Correction du biais d'erreur de classification

La méthode proposée par Statistique Canada pour l'estimation des flux bruts comprend également une étape de correction du biais d'erreur de classification. Il s'agit du biais qui résulte de la mauvaise attribution du statut d'activité d'un individu. Une technique mise au point par Fred Wong (1983) à Statistique Canada utilise les données d'une nouvelle entrevue pour corriger le biais d'erreur de classification.

4. CONSÉQUENCE DE LA MÉTHODE PROPOSÉE PAR STATISTIQUE CANADA

4.1 Les modèles des flux bruts

Chaque étape de la méthode proposée par Statistique Canada décrite dans la section précédente constitue une tentative logique de corriger des problèmes concernant la production d'estimations valides des flux bruts. On ne connaît pas exactement, cependant, l'effet des diverses corrections sur les estimations finales de la matrice des flux bruts. Pour mieux comprendre la proposition de Statistique Canada de traiter les différences de poids comme étant attribuables aux entrées dans la population étudiée et aux sorties de celle-ci, nous exposerons dans la présente section un modèle du processus de flux bruts. Notre discussion se centrera sur les valeurs des cellules d'entrée et de sortie de la matrice des flux bruts estimés, étant donné que les problèmes affectant la méthode proposée par Statistique Canada semblent se produire dans ces cellules.

proportionnellement de façon que les entrées et les sorties totales indiquées dans la matrice des flux bruts soient égales aux estimations du recensement des entrées et des sorties respectivement.

Soit I l'estimation indépendante du recensement des entrées dans la population étudiée et F les estimations du recensement des sorties de cette population. Appelons la somme des entrées estimées $X_{H+} = X_{H0} + X_{HC} + X_{HI}$ et la somme des sorties estimées $X_{+H} = X_{0H} + X_{CH} + X_{IH}$. Les entrées ajustées proportionnellement sont

$$y_{Hj} = x_{Hj} I / X_{H+} \quad \text{pour } j = 0, C, I. \quad (1)$$

Les sorties ajustées proportionnellement sont

$$y_{jH} = x_{jH} F / X_{+H} \quad \text{pour } j = 0, C, I. \quad (2)$$

3.4 Convergence des estimations des flux bruts avec les totalisations mensuelles

Statistique Canada aimerait que les estimations des flux bruts convergent avec les estimations mensuelles publiées des taux d'activité totaux. Ainsi, les totaux des lignes de la matrice des flux bruts doivent être les estimations du taux d'activité du mois $t-1$ et les totaux des colonnes doivent être les estimations ponctuelles du mois t . Les totaux marginaux de la matrice des flux bruts établis comme on l'a décrit ci-haut ne convergent pas avec les totaux mensuels de la population active.

Statistique Canada a l'intention d'utiliser la méthode de remise à l'échelle proportionnelle itérative, originellement proposée par Deming et Stephan (1940) et décrite en détail par Bishop, Fienberg et Holland (1975), pour corriger la matrice des flux bruts en fonction des totaux mensuels de la population active. Lorsqu'elle est utilisée pour corriger la matrice des flux bruts, la remise à l'échelle proportionnelle itérative successivement 1) ajuste les lignes de la matrice à la somme des estimations $t-1$ et ensuite 2) ajuste les colonnes à la somme des estimations du mois t . On répète les étapes 1) et 2) jusqu'à ce que les données de la matrice ne changent pas

différence, $W_{t,i} - W_{t-1,i}$, est ajoutée à la cellule HO, étant donné qu'on estimerait que ces cinq personnes se trouvaient en dehors de la population étudiée au cours du mois $t-1$ et se sont ajoutées à la population comme personnes occupées au cours du mois t .

Si, par ailleurs, l'individu est Occupé les deux mois, mais que $W_{t-1,i} = 305$ et que $W_{t,i} = 300$, alors on ajoute encore une fois le 300 à la cellule OO mais le poids supplémentaire de 5 est ajouté à la cellule OH. Ici, la différence entre les poids représente 5 personnes qui étaient Occupées au cours du mois $t-1$ et qui sont sorties (Hors) de la population étudiée au cours du mois t .

Un individu classé comme Hors de la population étudiée dans le mois $t-1$ et qui est ensuite, disons, Occupé dans le mois t , aura un $W_{t-1,i} = 0$. Si $W_{t,i} = 300$ alors on ajoute 300 à la cellule HO. Les individus classés comme Hors-de-la population étudiée au cours du mois t sont traités de la même façon, en ajoutant $W_{t-1,i}$ à la cellule appropriée dans la dernière colonne de la matrice des flux bruts.

Parce que les personnes Hors de la population étudiée se voient attribuer un poids zéro, une personne classée ainsi dans les deux mois $t-1$ et t auraient donc $W_{t-1,i} = 0$ et $W_{t,i} = 0$. Par conséquent, x_{HH} , la valeur de la cellule HH de la matrice des flux bruts doit toujours être de zéro.

3.3 Correction des cellules d'entrée et de sortie

L'addition des différences des poids aux cellules d'entrée et de sortie de la matrice des flux bruts fournit une méthode de traitement des changements des poids basés sur l'échantillon d'un mois à l'autre et fournit des estimations des entrées dans la population étudiée et des sorties de celle-ci. Des estimations indépendantes des entrées et des sorties, disponibles à partir des données du recensement, laissent croire que cette méthode surestime la valeur réelle des mouvements dans et hors de la population étudiée. Ainsi, Statistique Canada a l'intention d'ajuster les x_{HO} , x_{HC} , x_{HI} , x_{OH} , x_{CH} , et x_{IH} dans la matrice des flux bruts. Ces cellules seront corrigées

non-réponse d'un mois à l'autre. Statistique Canada propose de repondérer les fichiers des individus qui ont répondu dans les deux mois $t-1$ et t pour compenser pour cette non-réponse supplémentaire.

Lorsque la repondération sera terminée, Statistique Canada aura en main un fichier de données uniques comprenant des renseignements provenant de toutes les personnes qui ont répondu deux mois consécutifs. Ce fichier contiendra des renseignements géographiques et démographiques pour chaque individu ainsi que le statut d'activité de l'individu et la pondération qui lui est attribuée pour les mois $t-1$ et t .

3.2 Différences de poids

Comme nous l'avons noté dans la section 2.3, le poids attribué à un individu peut changer d'un mois à l'autre à cause de la technique de rotation utilisée, de la non-réponse et des changements dans la taille de la population étudiée. Même lorsque le facteur de correction pour la non-réponse est calculé sur les données d'un mois à l'autre, il peut encore arriver, pour un individu quelconque i , que $W_{t-1,i} \neq W_{t,i}$. Si l'on veut utiliser les données pour estimer les flux bruts, il est essentiel d'avoir une méthode pour traiter cette différence de pondération.

Statistique Canada propose de résoudre ce dilemme en postulant que les différences entre les deux poids ne se produisent que comme résultat des entrées et des sorties de la population étudiée. Ainsi, les différences des poids sont ajoutées à la cellule appropriée soit dans la dernière rangée ou la dernière colonne de la matrice des flux bruts. Cette procédure est fortement dépendante de l'interprétation des poids suggérés à la section 2.3, c'est-à-dire que l'individu i de l'échantillon représente $W_{t,i}$ personnes de la population au cours du mois t .

Pour illustrer cette procédure, supposons qu'un individu soit classé comme Occupé dans les mois $t-1$ et t , mais que $W_{t-1,i} = 300$ et $W_{t,i} = 305$. Le poids minimum, 300, est ajouté à la cellule 00 du tableau des flux bruts. Cette

4x4 comme ci-dessous:

		Statut d'activité pour le mois t			
		O	C	I	H
Statut d'activité pour le mois t-1	O	X_{OO}	X_{OC}	X_{OI}	X_{OH}
	C	X_{CO}	X_{CC}	X_{CI}	X_{CH}
	I	X_{IO}	X_{IC}	X_{II}	X_{IH}
	H	X_{HO}	X_{HC}	X_{HI}	X_{HH}

où: O = Occupé
 C = Chômeur
 I = Inactif
 H = Hors de la population étudiée, et
 X_{ij} = nombre estimé de personnes ayant le statut d'activité i dans le mois t-1 et le statut j dans le mois t.

On peut utiliser les dossiers finals de l'enquête sur la population active pour deux mois consécutifs pour obtenir les données d'estimation des flux bruts de la matrice 4x4. Pour utiliser ces données pour la production des estimations des flux bruts, Statistique Canada doit apparier les dossiers individuels extraits des deux fichiers mensuels consécutifs à l'aide des numéros d'identification personnels attribués aux individus enquêtés pour la durée de leur passage dans l'enquête.

Un individu présent dans le fichier des données dans un mois donné peut être absent du fichier dans un autre mois à cause de son insertion dans l'échantillon ou de son retrait de celui-ci ou pour avoir déménagé, avoir été absent de la maison ou avoir refusé de répondre. La pondération de l'échantillon décrite dans la section 2.4 comprend une correction pour la non-réponse chaque mois. Dans le traitement des flux bruts, nous devons également considérer la

n'est pas nécessairement vrai que $W_{t-1,i} = W_{t,i}$.

2.4 Structure longitudinale et estimation des flux bruts

Bien que l'objectif principal de l'enquête sur la population active soit de produire des estimations ponctuelles de l'activité sur le marché du travail, la technique du panel de l'enquête produit une base de données longitudinales dans laquelle environ les cinq sixièmes des ménages enquêtés un mois donné se retrouvent dans l'échantillon le mois suivant. Évidemment, on rejoint moins des cinq sixièmes des individus ou des ménages de l'échantillon au cours de deux mois consécutifs à cause de la non-réponse et des déménagements. Cependant, Statistique Canada s'intéresse à la possibilité d'utiliser les renseignements des individus qui ont répondu deux mois consécutifs pour produire des estimations des flux bruts entre les catégories de la population active.

Les estimations des flux bruts servent à répondre à des questions telles que a) quelle proportion de l'augmentation du chômage est due à des pertes d'emploi et quelle proportion est due à des personnes non auparavant actives commençant à se chercher un emploi? ou b) combien de personnes en chômage se découragent et quittent la population active?

Nous discutons du problème de l'estimation des flux bruts entre les catégories de la population active dans les deux prochaines sections.

3. MÉTHODE PROPOSÉE PAR STATISTIQUE CANADA POUR L'ESTIMATION DES FLUX BRUTS

Dans cette section, nous décrivons la procédure à plusieurs degrés d'estimation des flux bruts mise au point par Statistique Canada (voir Macredie et Veevers, 1977; Wong, 1983). Notre description de la procédure comprend diverses interprétations de l'effet des divers degrés.

3.1 Données nécessaires à l'estimation des flux bruts

Statistique Canada a proposé d'estimer les flux bruts à l'aide d'une matrice

population active sont les ménages. On trouvera une description détaillée du plan d'échantillonnage de l'enquête dans Méthodologie de l'enquête sur la population active, 1976, Statistique Canada (1977).

Les ménages choisis pour l'enquête sur la population active font partie de l'enquête pendant six mois consécutifs et sont ensuite retirés de l'échantillon. Par exemple, les ménages intégrés à l'échantillon en janvier sont interviewés six mois de suite, et ensuite retirés de l'échantillon après l'entrevue de juin. Chaque groupe de ménages ainsi intégré à l'échantillon et retiré ensuite constitue un panel. Chaque mois, l'échantillon de l'enquête sur la population active comprend des sujets de six groupes constants différents.

2.3 Pondération basée sur l'échantillon

Les données ponctuelles, les renseignements d'un mois donné pour tous les sujets dans les six panels interviewés au cours de ce mois, sont utilisées pour produire les estimations mensuelles de l'activité sur le marché du travail. Les estimations mensuelles sont des moyennes pondérées des valeurs pour chaque individu de l'échantillon. On utilise une moyenne pondérée parce que chaque sujet est considéré comme "représentant" un certain nombre de personnes dans la population étudiée. Le poids attribué au dossier d'un individu correspond au nombre de personnes de la population que cet individu de l'échantillon représente.

Soit $W_{t,i}$ le poids attribué à l'individu i au mois t . Si l'individu i est classé comme à l'extérieur de la population étudiée au cours du mois t , alors $W_{t,i} = 0$. Sinon, les poids attribués sont fixés par la probabilité de sélection de la grappe, la probabilité de sélection du ménage à l'intérieur de la grappe, la non-réponse à l'intérieur du mois, les facteurs rural/urbain, les corrections de sous-échantillonnage pour les zones dont la croissance est rapide et l'ajustement des rapports pour les facteurs province, âge, sexe.

Le poids attribué à un individu peut changer d'un mois à l'autre à cause du fait que chaque mois, on remplace un sixième de l'échantillon, à cause également de la non-réponse et, dans une moindre mesure, à cause des changements de la taille de la population étudiée. Ainsi, pour un individu donné, i , il

2. DESCRIPTION DE L'ENQUÊTE SUR LA POPULATION ACTIVE

2.1 Champ d'application de l'enquête

Environ 56,000 ménages, choisis dans les dix provinces canadiennes, sont inclus chaque mois dans l'échantillon de l'enquête sur la population active. Un questionnaire est rempli pour chaque membre des ménages de l'échantillon âgé de 15 ans et plus, ne faisant pas partie des forces armées et ne vivant pas en institution. Les questions de l'enquête concernent principalement les activités du sujet reliées au travail au cours de la semaine de référence, qui est la semaine précédant la semaine de l'enquête et qui contient généralement le 15e jour du mois. Les réponses aux questions de l'enquête sont utilisées pour classer les sujets comme occupés, chômeurs ou inactifs. Pour une discussion sur la classification de l'activité, voir Guide d'utilisation des données de l'enquête sur la population active, Statistique Canada (1979).

2.2 Conception de l'enquête

L'enquête sur la population active a été conçue pour permettre l'estimation des niveaux et des taux de l'emploi et du chômage pour chacune des dix provinces individuellement. Ainsi, excepté en ce qui concerne les exigences pour la taille totale de l'échantillon, les échantillons de chaque province sont indépendants.

Les régions économiques (RE), zones de structure économique semblables, forment la strate fondamentale à l'intérieur des provinces. Les RE sont divisées en unités autoreprésentatives (UAR) et unités non autoreprésentatives (UNAR). Les UAR sont des centres urbains importants et les UNAR sont généralement composées d'un petit centre urbain et d'une zone rurale. L'échantillonnage est effectué séparément pour les UAR et les UNAR.

L'échantillon des UAR est un échantillon stratifié à deux degrés. L'échantillonnage des UNAR est effectué selon un schéma d'échantillonnage stratifié à plusieurs degrés. En plus des UAR et des UNAR, certaines unités d'échantillonnage sont tirées d'un univers des immeubles d'appartements et d'un univers des zones spéciales. Les unités d'échantillonnage finales de l'enquête sur la

multiples des sujets fournissent une base de données longitudinales additionnelle qui pourrait être exploitée pour fournir des estimations des changements dans le temps pour des coûts supplémentaires minimes (voir, par exemple, Kalachek, 1979, et Fienberg et Tanur, 1983).

On a fait quelques tentatives d'utiliser les données longitudinales tirées des enquêtes par panel. Par exemple, les données longitudinales de la Current Population Survey sont utilisées depuis 1948 pour produire des tableaux montrant les mouvements bruts des individus entre les diverses catégories de la population active d'un mois à l'autre. Bien que ces tableaux soient produits chaque mois, ils n'ont pas été publiés depuis 1952 à cause de problèmes statistiques. Smith et Vanski (1979) discutent la production de données des variations brutes en utilisant la structure longitudinale de la Current Population Survey.

Récemment, Statistique Canada a commencé une étude sur les utilisations possibles des données longitudinales disponibles comme sous-produits de l'enquête sur la population active. Eux aussi souhaiteraient trouver une méthode de production d'estimations fiables des mouvements bruts entre les catégories de la population active. Dans le présent document, nous discutons les méthodes à l'étude à Statistique Canada pour la production de ces données des variations brutes.

Dans la section 2, nous donnons une brève description du champ d'application et de la conception de l'enquête sur la population active, et nous décrivons la structure des données produites. À la section 3, nous décrivons les méthodes proposées d'estimation des flux bruts mises au point par Statistique Canada, qui nécessitent l'utilisation de poids fonction de l'échantillon, de corrections pour les entrées dans la population étudiée et les sorties de celle-ci, de corrections pour l'homogénéité et de corrections pour le biais causé par les erreurs de classification. En mettant au point certains modèles simples pour le processus des flux bruts, nous analysons dans la section 4 les conséquences de la méthode proposée par Statistique Canada. Finalement, dans la section 5, nous décrivons certains travaux sur le traitement de la non-réponse dans les estimations des flux bruts.

L'ESTIMATION DES FLUX BRUTS MENSUELS DE L'ACTIVITÉ SUR LE MARCHÉ DU TRAVAIL¹

Stephen E. Fienberg et Elizabeth A. Stasny²

L'enquête sur la population active au Canada est une enquête sur les ménages effectuée chaque mois dans le but de fournir des estimations ponctuelles du nombre de personnes occupées, en chômage ou inactives. L'enquête utilise une technique de groupe constant avec renouvellement dans laquelle tous les individus d'un ménage inclus dans l'échantillon sont interviewés chaque mois, six mois de suite. Dans le passé, cette structure longitudinale a été peu exploitée, bien qu'on ait montré beaucoup d'intérêt pour les flux bruts d'un mois à l'autre (transitions) entre ces diverses catégories d'activité. Dans le présent texte, nous discutons des méthodes de production des estimations des flux bruts à l'étude à Statistique Canada, mais à partir du point de vue de la production d'un modèle.

1. INTRODUCTION

L'enquête sur la population active au Canada est une enquête mensuelle sur les ménages utilisée pour produire des estimations ponctuelles, ou transversales, de l'activité sur le marché du travail. Cette enquête, cependant, comme la Current Population Survey aux États-Unis et beaucoup d'autres enquêtes par échantillonnage d'envergure, utilise la technique du panel où les sujets sont interviewés plusieurs fois avant d'être retirés de l'échantillon. Bien que le principal but de l'enquête soit l'obtention d'estimations ponctuelles, on a depuis longtemps constaté que les renseignements provenant de ces interviews

¹ Cette recherche a été financée en partie par un contrat avec Statistique Canada. Les auteurs désirent remercier Murray Lawes, Larry Swain et Richard Veevers pour leurs explications de la méthodologie de l'enquête sur la population active et de ses données ainsi que l'éditeur et un évaluateur pour leurs commentaires et suggestions utiles.

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

Tableau 7 (suite)

**Incidence du cancer
1969-1978
Cancers primitifs multiples de chaque siège**
(Canada à l'exclusion de l'Ontario)

Siège du cancer ICDA	Cancers primitifs multiples			
	Nombre	% d'in- cidence	% dans le même registre	% ayant le même 4 ^e chiffre du code ICDA
180 Col de l'utérus	133	1.4	60.9	S/0
181 Chorionépipithéliome	1	1.0	0.0	S/0
182 Autre: de l'utérus	149	1.1	73.8	91.3
183 Ovaire, etc.	126	1.5	80.2	97.6
184 Org. gén. fém., autres	31	1.5	71.0	83.9
185 Prostate	64	1.6	69.2	S/0
186 Testicule	27	1.5	44.4	S/0
187 Org. gén. masc., autres	9	1.4	88.9	100.0
188 Vessie	274	1.6	77.7	S/0
189 Org. urin. autres et sans autre précision	119	1.5	72.3	79.8
190 Œil	13	1.1	46.2	S/0
191 Cerveau	93	1.6	46.2	S/0
192 Autre système nerveux	6	0.4	33.3	66.7
193 Glande thyroïde	49	1.5	55.1	S/0
194 Autres glandes endocrines	5	0.6	80.0	100.0
195 Sièges mal définis	4	0.6	100.0	100.0
196 Ganglions lymphatiques sec. et non spécifiés	5	0.3	100.0	100.0
197 Secondaires; Resp. et Digestif	9	0.3	88.9	77.8
198 Autres secondaires	1	0.1	100.0	100.0
199 Sans précision du siège	12	0.3	75.0	83.3
200 Lymphosarcome, etc.	68	1.1	60.3	97.1
201 Maladie de Hodgkin	85	2.2	55.3	S/0
202 Autre, du tissu lymphoïde	22	0.8	90.9	72.7
203 Myélome multiple	45	1.3	64.4	S/0
204 Leucémie lymphatique	64	1.5	62.5	84.4
205 Leucémie de la moelle	32	1.0	53.1	90.6
206 Leucémie monocyttaire	4	1.0	75.0	100.0
207 Leucémie, autre et non spécifié	11	0.7	81.8	90.9
208 Polyglobulie essentielle	1	0.2	100.0	S/0
209 Myélofibrose	1	0.4	100.0	S/0

Tableau 7

**Incidence du cancer
1969-1978
Cancers primitifs multiples de chaque siège
(Canada à l'exclusion de l'Ontario)**

Siège du cancer ICDA	Cancers primitifs multiples			
	Nombre	% d'in- cidence	% dans le même registre	% ayant le même 4 ^e chiffre du code ICDA
Total des sièges (excepté la peau, 173)	6,113	1.7	76.8	88.4
140 Lèvre	87	1.4	90.8	83.9
141 Langue	30	1.6	63.3	63.3
142 Glande salivaire	8	0.6	12.5	75.0
143 Gencive	8	1.5	100.0	62.5
144 Plancher buccal	18	1.8	94.4	S/0
145 Bouche, autre et non spécifié	16	1.4	75.0	62.5
146 Arrière-gorge	14	1.0	92.9	78.6
147 Rhinopharynx	8	1.1	62.5	S/0
148 Laryngopharynx	3	0.5	100.0	33.3
149 Pharynx, non spécifié	3	1.3	66.7	S/0
150 Oesophage	37	1.1	75.7	S/0
151 Estomac	178	1.0	71.9	74.2
152 Petit intestin	12	1.2	91.7	91.7
153 Gros intestin, à l'exclusion du rectum	623	1.9	82.3	44.0
154 Rectum	256	1.4	77.0	74.6
155 Foie	11	0.6	63.6	90.9
156 Vésicule biliaire	22	0.7	77.3	77.3
157 Pancréas	96	1.0	70.8	54.2
158 Péritoine	6	0.7	100.0	100.0
159 Organes digestifs non spécifiés	1	0.3	100.0	S/0
160 Nez, etc.	5	0.7	80.0	80.0
161 Larynx	94	2.0	77.7	57.4
162 Trachée, bronches, poumon	795	1.9	75.0	99.5
163 Org. resp. autres et sans autre précision	7	0.7	85.7	85.7
170 Os	28	2.0	78.6	82.1
171 Tissus conjonctifs	32	1.3	81.3	75.0
172 Mélanome de la peau	31	1.4	78.7	54.1
174 Sein	1,971	3.6	84.4	S/0

Tableau 6

Mortalité du cancer
Disponibilité et intégralité des données élémentaires

Canada (à l'exclusion de l'Ontario) et les provinces comptant les pourcentages
les plus élevés et les moins élevés d'intégralité des données

DONNÉES ÉLÉMENTAIRES	ANNÉE DE DÉCÈS	CANADA	POURCENTAGE LE PLUS ÉLEVÉ	POURCENTAGE LE MOINS ÉLEVÉ
DATE DE NAISSANCE				
Jour	1969 - 1973	96.6	99.4	0.0
	1974 - 1978	97.8	99.7	43.9
	1969 - 1978	97.2	99.6	23.0
Mois	1969 - 1973	97.0	99.7	0.0
	1974 - 1978	98.1	99.9	44.2
	1969 - 1978	97.6	99.8	23.2
Année	1969 - 1973	100.0	100.0	99.2
	1974 - 1978	99.9	100.0	99.6
	1969 - 1978	99.9	100.0	99.4
Date de naissance complète	1969 - 1973	96.6	99.4	0.0
	1974 - 1978	97.8	99.7	43.9
	1969 - 1978	97.2	99.6	23.0
ÂGE	1969 - 1973	100.0	100.0	100.0
	1974 - 1978	99.9	100.0	99.6
	1969 - 1978	100.0	100.0	99.8
LIEU DE NAISSANCE (Pays ou 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
RÉSIDENTE Division du recen- sement	1969 - 1973	99.8	100.0	99.6
	1974 - 1978	99.7	100.0	99.8
	1969 - 1978	99.8	100.0	99.7
Subdivision du recensement	1969 - 1973	92.2	99.9	51.9
	1974 - 1978	99.7	99.5	99.4
	1969 - 1978	96.2	99.7	77.1

Tableau 5

Incidence du cancer
Disponibilité et intégralité des données élémentaires

Canada (à l'exclusion de l'Ontario) et les provinces comptant les pourcentages
les plus élevés et les moins élevés d'intégralité des données

DONNÉES ÉLÉMENTAIRES	ANNÉE DE DIAGNOSTIC	CANADA	POURCENTAGE LE PLUS ÉLEVÉ	POURCENTAGE LE MOINS ÉLEVÉ
DATE DE NAISSANCE				
Jour	1969 - 1973	63.7	98.8	0.0
	1974 - 1978	71.6	99.8	4.0
	1969 - 1978	68.0	99.3	2.2
Mois	1969 - 1973	65.6	98.8	0.0
	1974 - 1978	73.3	99.8	4.1
	1969 - 1978	69.8	99.4	2.2
Année	1969 - 1973	92.9	100.0	15.9
	1964 - 1978	96.3	100.0	21.9
	1969 - 1978	94.7	100.0	19.1
Date de naissance complète	1969 - 1973	63.6	98.7	0.0
	1974 - 1978	71.6	99.8	4.0
	1969 - 1978	68.0	99.3	2.2
ÂGE	1969 - 1973	99.4	100.0	98.6
	1974 - 1978	100.0	100.0	100.0
	1969 - 1978	99.7	100.0	99.4
LIEU DE NAISSANCE (Pays ou 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
RÉSIDENTE Division du recen- sement	1969 - 1973	89.6	100.0	4.1
	1974 - 1978	89.6	100.0	82.8
	1969 - 1978	89.6	100.0	49.0
Subdivision du recensement	1969 - 1973	16.6	43.2	0.0
	1974 - 1978	32.3	76.4	0.0
	1969 - 1978	25.2	61.6	0.0

Tableau 4

Appariement des dossiers incidence aux dossiers mortalité

Classification des causes de décès pour un échantillon de cas de cancer du poumon
tirés du fichier incidence du cancer (Distribution de pourcentage)

Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux d'appariement les
plus élevés ou les moins élevés de cas de cancer du poumon parmi les appariements

ICDA-8	CLASSIFICATION SELON LE FICHIER MORTALITÉ	CANADA	TAUX LE + ÉLEVÉ	TAUX LE - ÉLEVÉ
	1. CANCER DE L'APPAREIL RESPIRATOIRE	92.0	100.0	85.3
162.1	(a) POU MON; primitif	91.4	100.0	82.9
160-163	(b) AUTRES PARTIES DE L'APPAREIL RESPIRATOIRE; primitif	0.3		0.0
197.0-197.3	(c) APPAREIL RESPIRATOIRE; secondaire	0.3		2.4
	2. AUTRES CANCERS	4.3		9.8
174	(a) SEIN	0.3		0.0
200-209	(b) SYSTÈME LYMPHATIQUE ET HÉMATOPOÏÉTIQUE; Primitif	0.3		0.0
196	Secondaire	-		0.0
	(c) AUTRE SIÈGE PRIMITIF SPÉCIFIÉ	3.1		9.8
195,199	(d) SIÈGE MAL DÉFINI OU NON DÉFINI	0.6		0.0
	3. PAS DE CANCER	3.7		4.9
	TOTAL	100.0	100.0	100.0

TAILLE DE L'ÉCHANTILLON POUR LES APPARIEMENTS	350	45	41
TAUX DE RÉUSSITE DES APPARIEMENTS (%)	69.4	80.4	73.2

Tableau 3

Appariement des dossiers mortalité aux dossiers incidence

Classification des maladies selon le fichier mortalité cancer pour un échantillon de décès dus au cancer (Distribution de pourcentage)

Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux d'appariement les plus élevés ou les moins élevés de cas de cancer du poumon parmi les appariements

ICDA-8	CLASSIFICATION SELON LE FICHIER MORTALITÉ	CANADA	TAUX LE + ÉLEVÉ	TAUX LE - ÉLEVÉ
	1. CANCER DE L'APPAREIL RESPIRATOIRE	95.2	100.0	83.7
162.1	(a) POUMON; primitif	92.5	100.0	74.4
160-163	(b) AUTRES PARTIES DE L'APPAREIL RESPIRATOIRE; primitif	1.0		4.7
197.0-197.3	(c) APPAREIL RESPIRATOIRE; secondaire	1.7		4.7
	2. AUTRES CANCERS			16.3
		4.8		
174	(a) SEIN	0.7		2.3
	(b) SYSTÈME LYMPHATIQUE ET HÉMATOPOÏÉTIQUE; Primitif	0.7		0.0
200-209				
196	Secondaire	1.0		4.6
	(c) AUTRE SIÈGE PRIMITIF SPÉCIFIÉ	1.9		9.3
195,199	(d) SIÈGE MAL DÉFINI OU NON DÉFINI	0.5		0.0
	TOTAL	100.0	100.0	100.0

TAILLE DE L'ÉCHANTILLON POUR LES APPARIEMENTS	415	54	43
TAUX DE RÉUSSITE DES APPARIEMENTS (%)	82.3	96.4	76.8

Tableau 2
Ratios mortalité-incidence du cancer

Décès au cours de la période (mortalité) en pourcentage des
nouveaux cas enregistrés (incidence)
Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux
les plus élevés ou les moins élevés.

Femmes		Principaux sièges selon le groupe d'âge et le sexe					
Siège du cancer	Âge	1969 - 1973			1974 - 1978		
		Canada	taux le plus élevé	taux le moins élevé	Canada	taux le plus élevé	taux le moins élevé
Sein (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
Côlon (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
Utérus (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
Col de l'utérus (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
Ovaire (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
Tous les cancers (140-209) à l'exclusion du cancer de la peau (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

- Soit la mortalité, soit l'incidence soit les deux sont nulles.

* Taux basés sur moins de 10 cas tant pour la mortalité que pour l'incidence

Tableau 1
Ratios mortalité-incidence du cancer

Décès au cours de la période (mortalité) en pourcentage des
nouveaux cas enregistrés (incidence)
Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux
les plus élevés ou les moins élevés.

Hommes	Principaux sièges selon le groupe d'âge et le sexe						
Siège du cancer	Âge	1969 - 1973			1974 - 1978		
		Canada	taux le plus élevé	taux le moins élevé	Canada	taux le plus élevé	taux le moins élevé
Poumon (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
Côlon (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
Vessie (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
Estomac (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
Tous les cancers (140-209) à l'exclusion du cancer de la peau (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

- Soit la mortalité, soit l'incidence soit les deux sont nulles.

* Taux basés sur moins de 10 cas tant pour la mortalité que pour l'incidence

BIBLIOGRAPHIE

- [1] Waterhouse, J., Muir, C. et al. CANCER INCIDENCE IN FIVE CONTINENTS, Volume III, Centre international de recherche sur le cancer, OMS, Lyon, 1976, p. 3.
- [2] MANUEL DE L'OMS: registres normalisés du cancer (registres hospitaliers), OMS, Genève 1976.
- [3] MANUEL FOR CANCER RECORDS OFFICERS, Institut national du cancer du Canada.
- [4] Op. cit. en [1], pp 45-51.
- [5] Nagnur, D.N., S.G. Currie et B. Heath, ÉVALUATION DE LA QUALITÉ DES STATISTIQUES DE L'ÉTAT CIVIL: étude pilote. Statistique Canada. Division de la santé.
- [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, JNC1, Vol. 66, N° 6, Juillet 1981.
- [8] Percy C., COMPARISON OF THE CODING OF DEATH CERTIFICATES RELATED TO CANCER IN SEVEN COUNTRIES, Public Health Reports, Vol. 93, N° 4, juillet-août 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, JNC1, Vol. 67, N° 5, Novembre 1981, p. 1019.

fet, les règles actuelles sur la déclaration des cancers primitifs multiples exigent des rapports distincts sur les cancers affectant les deux côtés de la plupart des organes bilatéraux.

Dans l'ensemble, nous croyons que même s'il existe une certaine inconsistance provoquée par la déclaration des cancers primitifs multiples et par les déclarations en double, le problème est très négligeable en regard de celui qui est provoqué par la sous-représentation.

4. SOMMAIRE

Les techniques décrites dans cet article ont permis d'identifier les niveaux différents de qualité des données sur l'incidence et la mortalité du cancer. Nous avons trouvé que les ratios mortalité-incidence en particulier peuvent servir à évaluer les erreurs de représentation qui constituent l'une des principales préoccupations des responsables d'un système de bonne qualité pour déterminer l'incidence du cancer. La qualité des données touchant ceux qui sont inscrits dans le système de déclaration des cas de cancer est suffisamment bonne pour qu'il soit possible d'évaluer la qualité de la classification des causes de décès dans le système de mortalité, par le biais d'un appariement de micro-données. En fait, un appariement informatisé de micro-données pourrait servir à évaluer la sous-représentation parce que l'on pourrait déterminer le nombre de décès dus au cancer non inscrits auparavant au Système national de déclaration des cas de cancer.

REMERCIEMENTS

Les auteurs tiennent à remercier J. Gorman, D. Lawrence, K. McClean, S. Moore et P. Walsh qui les ont aidés à préparer les données destinées à cet article. Ils sont également reconnaissants à M. J. Silins pour ses commentaires. Ils remercient enfin les membres du comité de lecture pour leurs commentaires.

Les règles sur la déclaration des cancers primitifs multiples sont difficiles à interpréter; aussi doit-on s'attendre à certaines différences à cet égard parmi les provinces.

L'inscription involontaire en double du même cas de cancer peut survenir si un bureau provincial ne peut déterminer si le même cas a été inscrit auparavant (peut-être parce que l'information sur l'identité est insuffisante) ou si le même cas est déclaré par deux bureaux différents⁷. Nous avons limité notre recherche sur les cancers primitifs multiples aux entrées multiples décrivant le même siège du cancer (au niveau à 3 chiffres du code ICDA).

Nous n'avons pas tenté de séparer les inscriptions en double des cas de cancers primitifs multiples, bien que l'on puisse présumer que la majorité des cas déclarés par deux bureaux différents font double emploi alors que les cas déclarés par le même bureau sont davantage susceptibles d'être des cancers primitifs multiples, surtout dans le cas de ceux dont le 4e chiffre du code de diagnostic est différent.

En nous servant de critères très stricts d'appariement et en excluant les cancers de la peau (autres que le mélanome de la peau) nous avons déterminé que 1.7 % (6113) des dossiers formant le fichier sur l'incidence du cancer entre 1969 et 1978 constituaient des inscriptions multiples.

Ce taux variait de 0.5 % à 1.9 % selon la province. Seulement 1.4 % des dossiers multiples avaient été déclarés par deux bureaux différents. Dans 88.4 % des cas, il y avait concordance jusqu'au niveau du 4e chiffre du code de diagnostic.

Quant au siège du cancer, si l'on ne tient compte que des sièges figurant dans plus de cinquante dossiers, le taux des cancers primitifs variait entre 1.0 % pour le cancer de l'estomac et du pancréas et 3.6 % pour le cancer du sein. Le taux élevé en ce qui concerne le cancer du sein n'est pas étonnant; en ef-

⁷ Le Système national de déclaration des cas de cancer ne procède pas à des vérifications régulières de ce genre de double emploi.

et 1974-1978), on relève une certaine amélioration au cours de la période plus récente.

Dans les dossiers sur la mortalité du cancer pour la même période, une date de naissance complète existe dans plus de 95 % des cas. Toutefois au moins une partie de ce taux élevé est due au fait que le système de mortalité impute une date de naissance tirée de l'âge et de la date de décès lorsque la date de naissance n'est pas déclarée. En 1976, le taux d'imputation était de 11.5 % (5). On ne procède pas à des imputations de ce genre dans le cadre du système de déclaration des cas de cancer⁶.

Les analyses de petites régions où il se produit des cas de cancer nécessitent une information détaillée sur le lieu de résidence. Les données sur la mortalité liée au cancer sont bien plus utiles à ces fins parce que les codes de résidence dans la division de recensement (comté) figurent dans 99.8 % des dossiers et les codes de subdivision de recensement (cité, ville, village) dans 96.2 % des dossiers. Par contre, dans le fichier sur l'incidence du cancer, les codes de division figurent dans 89.6 % des dossiers et les codes de subdivision dans seulement 25.2 % des dossiers. Dans le fichier en question, il y a eu au cours de la deuxième période une amélioration quant à l'information fournie sur la subdivision de recensement.

3.2.4 Inscriptions multiples dans le système de déclaration des cas de cancer (Tableau 7)

La possibilité de comparer les données sur les cas de cancer est affectée s'il existe des différences dans la déclaration des cancers primitifs multiples chez la même personne et dans une inscription involontaire faisant double emploi.

⁶ Les imputations peuvent être utiles, à des fins statistiques mais elles sont trompeuses dans les études de post-observation, à moins qu'il ne soit précisé que l'information est basée sur une imputation.

Pour ce qui est des dossiers dont l'appariement a réussi, il y avait concordance du diagnostic d'un cancer primitif du poumon dans 91.4 % des cas, taux très semblable à celui que nous avons trouvé dans la comparaison inverse. Les échantillons destinés aux deux comparaisons avaient été choisis de façon indépendante; il est donc rassurant de constater la concordance des résultats en ce qui concerne le diagnostic. Parmi les cas restants, 4.3 % présentaient la cause sous-jacente du décès attribuée à des sièges de cancer autres que l'appareil respiratoire et 3.4 % une cause qui n'était pas le cancer. Dans le cas de ce dernier groupe, il est possible que le certificat de décès ait mentionné le cancer comme une cause ayant contribué au décès. Cette analyse est faisable, mais nous ne l'avons pas entreprise.

3.2.3 Disponibilité et intégralité des données (Tableaux 5 et 6)

Une mesure de la qualité et de l'utilité des fichiers de données est la fréquence d'une information valide pour des données élémentaires spécifiques. Cette mesure est brute parce que le terme "valide", selon la définition que nous vous en donnons ici, signifie valide conformément aux contrôles d'édition informatisés et n'exclut pas l'invalidation de l'information par l'imputation de l'information manquante ou des erreurs de définition ou de classification de la donnée élémentaire.

Sous réserve de ce qui précède, la mesure en question peut se révéler utile en faisant ressortir s'il y a eu des améliorations dans le temps dans la déclaration des données élémentaires et où ces améliorations se sont produites et s'il est possible ou non de soumettre les données à des analyses particulières. Par exemple, l'information sur la date de naissance revêt de l'importance dans les analyses purement statistiques (spécifiques à l'âge) des données ainsi que dans les études post-observation qui dépendent de l'exactitude de l'information sur l'identité des personnes.

Dans le fichier incidence du cancer, une date de naissance complète (jour, mois et année) ne figure en moyenne que dans 68 % des dossiers relatifs à la période 1969-1978. Si l'on considère séparément les deux périodes (1969-1973

des fichiers (ou peut-être le fait qu'un cancer du poumon ait été diagnostiqué pour la première fois avant 1969, suivi par l'inscription ultérieure d'un autre cancer primitif) explique le manque de concordance. Parmi les provinces, le taux de concordance des diagnostics variait entre 74.4% et 100%.

Il est intéressant de noter que pour la province présentant une concordance totale des diagnostics, nous avons aussi enregistré le taux de succès le plus élevé (96.4 %) pour repérer les dossiers correspondants dans les deux fichiers. Cela pourrait éventuellement indiquer une liaison étroite entre le bureau provincial de l'état civil et le bureau d'enregistrement des cas de cancer dans la province. Lors d'un appariement inverse d'un échantillon de dossiers sur l'incidence du cancer avec des dossiers sur la mortalité (décrit à la partie 3.2.2), c'est la même province qui présentait le taux le plus élevé d'appariements réussis ainsi qu'une concordance totale en matière de diagnostic.

3.3.2 Recherche dans le fichier incidence/mortalité

La recherche inverse, basée sur le fichier incidence pour les années 1969 à 1978, comme point de départ, et visant à repérer un dossier correspondant dans le fichier complet de mortalité pour la même période, a produit un taux de succès de 69.4 % de l'échantillon de 504 cas étudiés où le diagnostic indiquait un cancer primitif du poumon. Le taux d'appariements manqués était donc plus élevé que dans le cas des appariements mortalité/incidence pour toutes les provinces et il variait de 19.6 % à 44.6 %. Les raisons possibles de ce manque d'appariement sont a) soit que le patient était toujours en vie à la fin de l'année 1978, b) soit que l'information sur son identité n'est pas suffisante pour permettre l'appariement. En règle générale, il est moins probable que l'on trouve un dossier correspondant lorsqu'on recherche dans le fichier incidence/mortalité, car certaines personnes chez qui on a diagnostiqué le cancer du poumon survivent effectivement, alors que toutes les personnes qui meurent de ce cancer devraient être inscrites comme des cas nouveaux soit avant le décès soit au moment du décès.

en entreprenant par exemple des recherches sur la précision du codage sur le terrain de la cause du décès et du diagnostic, comme le décrivent deux rapports américains (8) et (9), cela faciliterait l'interprétation des résultats des recherches épidémiologiques.

3.2.1 Recherche dans le fichier mortalité/incidence

Pour l'échantillon de 504 enregistrements de décès indiquant le cancer du poumon comme cause sous-jacente de décès entre 1969 et 1978, nous avons trouvé 415 enregistrements correspondants (82 %) dans le fichier incidence du cancer pour les années de diagnostic 1969-1978⁵. Le taux des appariements manqués variait entre 3.6 % et 26.8 % parmi les neuf provinces. Ce taux est influencé par quatre facteurs principaux:

a) le fait que le cancer ait été diagnostiqué pour la première fois avant 1969, b) le fait qu'il reflète une sous-représentation des nouveaux cas, c) le fait que l'information sur l'identité ne suffisait pas pour permettre l'appariement des dossiers, ou d) le fait que le code de la cause du décès indiquait par erreur le cancer. Nous ne disposons pas de données suffisantes pour évaluer le poids relatif de chacun de ces facteurs.

Parmi les 415 enregistrements de décès pour lesquels nous avons trouvé des dossiers correspondants dans le fichier incidence, il y avait concordance du diagnostic (cancer primitif du poumon) dans 92.5 % des cas. Il y avait une concordance de 95.2 % sur la présence d'un cancer de l'appareil respiratoire. Les quelques autres dossiers du fichier "incidence" indiquaient des diagnostics de sièges autres que l'appareil respiratoire. Il est généralement reconnu que l'information sur le diagnostic qui figure dans les registres du cancer est plus précise que l'information sur la cause du décès figurant aux certificats de décès. Toutefois, étant donné l'envergure de cette étude, il n'était pas possible de déterminer si une classification erronée dans l'un ou l'autre

⁵ Dans six cas, il existait pour la même personne plus d'un enregistrement correspondant dans le fichier mortalité. Nous avons retenu un seul de ces dossiers comme un appariement réussi.

les patients plus âgés, la sous-représentation de la population plus âgée est plus marquée dans les bureaux qui disposent généralement de systèmes d'inscription moins complets.

Les données recueillies au Canada vont donc dans le sens des recommandations de l'Union internationale contre le cancer (1970) et du Centre international de recherche sur le cancer (1976) et de la réitération de la recommandation faite dans un récent article de Doll et Peto (7), à savoir "que l'on n'obtient des comparaisons suffisamment dignes de foi (de l'incidence du cancer) que si ces comparaisons se limitent aux hommes et aux femmes de l'âge moyen".

3.2 Appariement des dossiers mortalité et incidence (Tableaux 3 et 4)

Comme Statistique Canada est responsable de la gestion des fichiers de données tant sur l'incidence que sur la mortalité attribuable au cancer, il est possible de comparer les inscriptions sur les personnes qui figurent dans les deux registres distincts pour vérifier l'information qui y est consignée.

Pour cette partie de l'étude sur la qualité des données, la précision de la détermination du diagnostic et de la cause du décès présentait un intérêt particulier. Dans le cadre de l'étude, on ne peut que décrire les résultats - les raisons qui expliquent les divergences découvertes demeurent inconnues. Toutefois, nous estimons que les résultats sont révélateurs et qu'ils indiquent que, dans le cas du cancer particulier choisi aux fins d'analyse (cancer du poumon), la concordance de diagnostics entre les deux fichiers est généralement élevée (plus de 90 %).

Notre étude indique aussi qu'il serait possible de procéder à un appariement sur une plus grande échelle pour évaluer la précision des codes de diagnostic et de cause de décès. Évidemment, si une étude sur une plus grande échelle était basée sur un échantillon, il serait préférable de stratifier celui-ci selon l'année de diagnostic ou l'année de décès. En théorie, si l'on se sert de techniques informatisées d'appariement, il est possible d'appliquer ce type d'analyse à tous les sièges de cancer. Si l'on voulait compléter cette étude

Parmi les principaux sièges du cancer qui ont été examinés pour les hommes, les cancers du poumon et de l'estomac allaient de pair avec les ratios mortalité-incidence les plus élevés dans tous les bureaux, mais la variation interprovinciale des ratios était la plus élevée pour les cancers du côlon (rectum exclu), de la prostate et de la vessie. Dans les études visant à mettre en relief les différences de risque du cancer selon la région géographique (province), l'emploi des données sur l'incidence du cancer constituerait donc une méthode plus fiable dans le cas des deux premiers sièges de cancer cités.

Pour les femmes, parmi les principaux sièges examinés, les cancers du côlon et de l'ovaire correspondaient aux ratios les plus élevés dans tous les bureaux. La variation interprovinciale des ratios était la plus élevée pour le cancer de l'utérus et du col de l'utérus ainsi que pour le cancer du côlon. Aux fins de comparaisons interprovinciales, les données sur l'incidence du cancer du sein et du cancer de l'ovaire seraient donc plus fiables.

Dans le cas des sièges du cancer de l'utérus (autre que celui du col) et du cancer du col, les ratios indiquent de grandes variations interprovinciales si l'on tient compte des sièges séparément. La variation se réduit considérablement si l'on combine les deux sièges, ce qui laisse entendre qu'il existe des différences dans la précision du diagnostic et la détermination de la cause du décès dans le cas de ces deux sièges.

Pour les principaux groupes d'âge, nous avons examiné les ratios mortalité-incidence spécifiques au siège. Les taux les plus élevés se sont invariablement manifestés dans le cas des personnes âgées (65 ans et plus) dans tous les bureaux et pour tous les sièges indiqués. Il s'agit là d'un phénomène normal puisque le risque de décès augmente avec l'âge; proportionnellement, il y a donc plus de décès que de nouveaux cas de cancer parmi les personnes âgées. Il est aussi reconnu qu'il est généralement plus difficile de diagnostiquer et de déclarer les cancers chez les personnes âgées. Toutefois, l'augmentation relative des ratios dans le cas de ces personnes est bien plus élevée pour les bureaux qui ont, au départ, les ratios moyens les plus élevés. Cela indique que même si tous les bureaux peuvent éprouver certaines difficultés à inscrire

général, les bureaux qui se servent d'un plus grand nombre de sources différentes d'inscription n'ont pas plus d'inscriptions multiples pour le même siège que les bureaux qui se servent d'un nombre moins élevé de sources.

Les ratios mortalité-incidence du cancer dans les six autres bureaux se ressemblaient davantage. Dans ces bureaux, nous n'avons pas relevé une tendance systématique selon laquelle un bureau présentait des rapports plus élevés ou moins élevés pour tous les sièges et pendant toutes les deux périodes à l'étude.

Il existe de nombreux facteurs qui peuvent influencer les variations des ratios observés selon le site du cancer. Parmi les facteurs qui tendent à produire une inscription moins complète des nombreux cas - et par conséquent, des ratios mortalité-incidence plus élevés - il faut citer la difficulté de diagnostiquer le cancer (par exemple dans les organes profondément enfouis) et le manque d'accès à des sources spécifiques de données (tels que les rapports d'hématologie confirmant un diagnostic de leucémie).

Les facteurs qui peuvent provoquer la surinscription des nouveaux cas et des rapports moins élevés sont les programmes de dépistage de masse (qui peuvent aboutir à l'inclusion des cas répandus, surtout dans le cas des tumeurs à croissance lente), les inscriptions qui se chevauchent, l'inclusion de cancers in situ ainsi que l'inclusion de cancers latents découverts seulement au moment de l'autopsie (tel est notamment le cas du cancer de la prostate). Par ailleurs, les différences de précision dans la détermination du diagnostic ou de la cause du décès peuvent mener à des différences artificielles. Par exemple, les certificats de décès peuvent porter l'inscription "cancer de l'utérus, non spécifié" ou "leucémie, non spécifiée" comme étant la cause du décès alors qu'un bureau d'inscription du cancer disposera souvent d'une information plus précise et affectera des codes plus précis (6). C'est pourquoi une analyse des ratios mortalité-incidence au niveau du diagnostic plus détaillé pourrait révéler des divergences importantes.

siège, un sexe ou un groupe d'âge donné, il faut se douter qu'il y a des différences sur le plan de l'intégralité des inscriptions de nouveaux cas de cancer. Un rapport plus élevé, qui signifie une proportion plus élevée de décès en regard des nouveaux cas pendant la même période peut indiquer une inscription moins élevée de ces cas.

Pendant toutes les deux périodes, il y avait deux bureaux où l'on relevait constamment les taux les plus élevés pour tous les sièges réunis et pour la plupart des principaux sièges indiqués. Il n'y a guère de doute que ces taux élevés sont dus à une sous-représentation de nouveaux cas; en effet, ces bureaux sont les seuls qui ne se servent pas de certificats de décès comme sources d'inscription. En outre, l'un de ces bureaux ne se sert que d'une seule source de données, à savoir les rapports des hôpitaux, pour inscrire les cas de cancer. Le bureau en question avait auparavant présenté les conclusions d'une étude spéciale qui montrait que ce bureau ne recevait des avis que sur un pourcentage estimé à 70 % des victimes du cancer admises dans les hôpitaux jusqu'à la fin de 1976. À la suite de cette étude, d'importants changements ont été apportés en vue d'améliorer le système de déclaration. Depuis 1977, ce bureau déclare un nombre plus élevé de cas de cancer, ce qui se traduit par une réduction marquée des ratios mortalité-incidence.

Tous les autres bureaux d'inscription du cancer au Canada se servent des sources multiples de déclaration, ce qui est considéré essentiel pour couvrir un vaste champ d'observation et ce qui pourrait avoir une incidence positive sur l'intégralité et la qualité des données élémentaires. Pour mener notre étude, nous avons examiné l'intégralité des données élémentaires déclarées (voir sous-section 2.3). Toutefois, il se trouve que le seul bureau qui a recours à une seule source de données occupe en fait un rang passablement élevé sur le plan de l'intégralité de l'information touchant de nombreuses données élémentaires.

Le recours à des sources multiples d'inscription comporte un risque: le double emploi. Toutefois, l'analyse des inscriptions multiples concernant la même personne et le même siège de cancer ne confirme pas ce risque. En

Parmi ces dossiers, nous avons relevé les inscriptions multiples comme suit:

- l'année, le mois et le jour de naissance étaient indiqués et concordaient, ou
- le jour de naissance n'était pas indiqué sur au moins un dossier mais les mois de naissance concordaient.

Nous avons aussi vérifié à la main tous les groupes comprenant au moins trois personnes où l'information sur le mois et le jour ne concordait pas et tous les groupes de deux personnes où l'information sur le mois ou le jour manquait dans au moins un dossier.

En tout, nous avons fini par relever 6,113 dossiers susceptibles de faire double emploi. Ces dossiers correspondent à 5,947 personnes. (À noter que certaines personnes faisaient double emploi plus d'une fois). Nous n'avons pas jugé s'il s'agissait là d'inscriptions multiples valables ou effectivement de doubles emplois.

Pour chaque valeur ICDA-8 de 3 chiffres, nous présentons dans le tableau 7 la ventilation de ces doubles emplois possibles selon qu'il s'agit de dossiers provenant du même bureau ou de bureaux différents et s'il s'agit du nombre de doubles emplois possibles qui ont le même quatrième chiffre de la classification ICDA-8.

3. DISCUSSION

3.1 Ratios mortalité-incidence (tableaux 1 et 2)

Les ratios mortalité-incidence du cancer peuvent fournir une indication sur l'intégralité des inscriptions. Dans tous les bureaux, ces taux varient selon le siège du cancer (les taux les plus élevés correspondant aux taux de survie les moins élevés), et selon l'âge et le sexe. Toutefois si une comparaison des ratios des différents bureaux indique de grandes différences pour un

provinces qui ont la déviation la plus élevée par rapport à la moyenne nationale. Pour le dossier mortalité, nous indiquons les résultats uniquement pour les décès dus au cancer dans les neuf provinces autres que l'Ontario afin que la comparaison avec le SNDCC soit plus valable.

2.4 Inscriptions multiples au système de déclaration des cas de cancer

Le principe du Système national de déclaration des cas de cancer est de consigner tous les cas nouveaux de néoplasmes malins. Une personne devrait être inscrite plus d'une fois lorsqu'il se développe chez elle des néoplasmes malins multiples. Afin d'éviter la double inscription du même cas ou la double déclaration des patients inscrits dans plus d'une province, tous les responsables des registres provinciaux de cas de cancer suivent des procédures réglementaires. Il peut arriver malgré cela que le même cas de cancer soit déclaré deux fois. Dans le but d'évaluer l'ampleur des doubles emplois, nous avons recherché les dossiers susceptibles de faire double emploi. Notre recherche a été loin d'être exhaustive, de sorte que le nombre de doubles possibles que nous avons trouvés est en deçà du chiffre réel. Parmi les 457,158 dossiers, nous avons éliminé ceux qui indiquaient des noms ou des années de diagnostic invalides. Pour les dossiers où il manquait l'année de naissance, nous avons déterminé l'année en question à partir de l'âge si celui-ci était connu. Nous avons aussi éliminé les dossiers sur le cancer de la peau (ICDA-8: 173) puisque l'on sait que ce type de cancer apparaît à de multiples occasions.

Parmi les dossiers restants, nous avons retenu ceux qui remplissaient toutes les conditions suivantes:

- Concordance parfaite des années de naissance ou des années de naissance calculées
- concordance parfaite des noms
- exactitude des quatre premières lettres du prénom
- conformité du code de trois chiffres ICDA-8.

Le taux moyen d'appariements réussis a été de 69.4 %. Les taux variaient entre 55.4 % et 80.4 % parmi les neuf provinces. Sur les appariements réussis, 92.0 % avaient la même classification à 4 chiffres (ICDA-8). Ces taux variaient entre 74.4 % et 100.0 %. Pour les provinces ayant les taux d'appariements les plus élevés ou les moins élevés comportant la même classification ICDA-8, nous indiquons au tableau 4 la ventilation des causes observées des décès classifiés.

2.3 Disponibilité et intégralité des données

La fréquence relative de données valides pour des éléments d'information spécifiques offre une mesure simple de la qualité des fichiers de données. Dans le cas du Système national de déclaration des cas de cancer et dans le cas des décès dus au cancer consignés dans le fichier mortalité, nous nous concentrons sur les éléments d'information suivants:

- date de naissance (jour, mois, année)
- âge
- lieu de naissance
- comté et subdivision de résidence

Nous choisissons ces éléments pour souligner à quel point il serait facile ou difficile d'apparier des dossiers provenant d'autres fichiers (ex. partie 2.2) ou de créer des tabulations spéciales telles que des statistiques sur de petites régions. Pour chaque élément d'information, nous classifions les données en données valides et en données invalides. Outre les valeurs nulles, des données invalides apparaissent lorsque l'on trouve des caractères alphabétiques dans une variable numérique ou que la valeur numérique dépasse l'intervalle de variation. Nous avons groupé les fréquences relatives en deux ensembles de cinq ans (1969-1973 et 1974-1978) afin de pouvoir observer si la qualité a changé de manière appréciable au cours des années ultérieures.

Dans les tableaux 5 et 6, nous présentons les moyennes nationales pour les deux fichiers centraux; nous indiquons également les valeurs pour les

2.2.2 Rapport incidence-mortalité

Nous avons aussi choisi un échantillon de dossiers tirés du SNDCC et les avons appariés au fichier mortalité. Nous avons choisi au hasard dans chacun des bureaux de l'état civil 56 dossiers indiquant des néoplasmes malins du poumon et des bronches (ICDA-8: 162.1) et rassemblé ainsi un échantillonnage de 504 dossiers. Nous avons ensuite vérifié les dossiers appariés pour déterminer la cause sous-jacente du décès d'après le fichier complet mortalité de 1969 à 1978. Pour cette étude, nous n'avons pas vérifié la cause du décès figurant sur le certificat original de décès, mais on pourrait procéder à cette vérification pour les prochaines études.

Les résultats tirés de cet appariement manuel peuvent être classés comme suit:

- (a) on ne trouve pas de dossier apparié,
- (b) on trouve un appariement, mais la cause du décès est autre que le cancer,
- (c) on trouve un appariement où le cancer est la cause du décès mais non le cancer du poumon ou des bronches,
- (d) on trouve un appariement indiquant la même cause de décès.

Si l'on ne trouve pas d'appariement, cela indique l'insuffisance du processus d'appariement, à moins que le décès n'ait eu lieu après 1978 ou à l'extérieur du Canada ou encore que la personne soit toujours en vie.

Si l'on trouve un appariement où la cause sous-jacente du décès est différente, cela indique l'une des possibilités suivantes:

- (a) un risque concurrent a prévalu,
- (b) le cancer a été une cause contribuant au décès, mais non la cause sous-jacente, ou encore
- (c) la cause sous-jacente du décès indiquée par le fichier central mortalité était inexacte, ou bien le siège du cancer indiqué par le SNDCC était inexact (ce dernier cas étant censé être moins probable).

- (a) on ne trouve aucun dossier apparié
- (b) on trouve un appariement mais le dossier indique un siège différent pour le cancer
- (c) on trouve un appariement et le dossier indique le même siège pour le cancer.

Si on ne trouve aucun appariement cela indique qu'il y a une sous-représentation, ou que le cancer a été diagnostiqué pour la première fois en Ontario ou à l'extérieur du Canada ou bien avant 1969, ou encore que le décès n'était pas réellement attribuable au cancer. Il est possible aussi comme on l'a déjà indiqué, que le processus d'appariement proprement dit ne soit pas parfait. Si l'on trouve un appariement, mais que les dossiers indiquent des sièges différents pour le cancer, cela peut indiquer une erreur soit dans le fichier mortalité, soit dans le SNDCC. Comme nous l'avons déjà mentionné, on croit généralement que le dossier mortalité offre une classification plus précise des maladies en raison du taux élevé de confirmations histologiques.

Nous avons entrepris une étude à échelle réduite pour mesurer ce phénomène. Dans chacune des provinces, à l'exception de l'Ontario, nous avons choisi un échantillon aléatoire de 56 dossiers indiquant un néoplasme malin du poumon ou des bronches comme la cause sous-jacente du décès (ICDA-8: 162.1) et rassemblé ainsi un échantillonnage de 504 dossiers. Nous avons retenu seulement les décès survenus entre 1969 et 1978.

À l'échelle nationale, le taux d'appariements réussis était de 82.3 %. Les taux variaient entre 73.2 % et 96.4 % parmi les neuf provinces. Sur les appariements réussis, 92.5 % avaient la même classification à 4 chiffres (ICDA-8). Ces taux variaient entre 74.4 % et 100 %. Pour les provinces ayant les taux d'appariements les plus élevés ou les moins élevés comportant la même classification ICDA-8, nous indiquons au tableau 3 la ventilation des classifications de maladies observées.

Comme points de départ, nous avons choisi dans les deux fichiers les dossiers faisant état de néoplasmes malins du poumon ou des bronches (ICDA-8⁴: 162.1) pour les années 1969 à 1978. Ce choix a été basé sur des considérations d'incidence élevée, de mortalité élevée et de courtes périodes de survie; les conditions étaient donc favorables pour trouver des dossiers appariés dans les deux fichiers. Malgré cela, en raison du temps écoulé entre le diagnostic et décès il est vrai que les décès provoqués par le cancer au cours des premières années et les cancers nouvellement diagnostiqués au cours des années ultérieures sont moins susceptibles de faire l'objet d'un dossier correspondant dans le fichier apparié. L'analyse des résultats serait améliorée dans les prochaines études si la conception de l'échantillon permettait de contrôler l'année du décès et l'année du diagnostic.

Nous avons choisi deux échantillons indépendants: l'un du fichier de la mortalité et l'autre du fichier du Système national de déclaration des cas de cancer (SNDCC).

2.2.1 Rapport mortalité-incidence

Tous les décès dus au cancer entre 1969 et 1978 devraient avoir, du moins en théorie, un dossier correspondant dans le fichier du SNDCC. Les principales exceptions à cette règle sont les cas où le cancer a été diagnostiqué pour la première fois en Ontario ou à l'extérieur du Canada ou bien les cas où le cancer a été diagnostiqué pour la première fois avant 1969. Observation faite de ces exceptions, l'absence d'un dossier correspondant dans le SNDCC indique une sous-représentation. On suppose, bien entendu, que la cause sous-jacente du décès qui figure dans le fichier mortalité est dépourvue d'erreur.

Par conséquent, si l'on choisit un échantillon de décès dus au cancer et si on les apparie au SNDCC, on est en présence d'un certain nombre de possibilités:

⁴ International Classification of Diseases (Classification internationale des maladies) adaptée pour les États-Unis, huitième révision.

Si les ratios mortalité-incidence sont trop différents pour la plupart des sièges, on pourra négliger de tenir compte des différences de taux en a) et b). Quant aux taux d'erreur (c), les études sur l'erreur de codage de la cause sous-jacente du décès ont fait ressortir des taux d'erreur de moins de 10 % (5). Dans un rapport non publié, on a trouvé que ces taux d'erreur pourraient varier entre 3 % et 18 % parmi les bureaux de l'état civil. Toutefois, les différences observées parmi les ratios mortalité-incidence (voir tableaux 1 et 2) ne peuvent entièrement s'expliquer par ces taux d'erreur. Par ailleurs, comme presque 90 % des inscriptions de nouveaux cas de cancer sont confirmées sur le plan histologique, le taux d'erreur en d) est négligeable. Aussi, ces ratios mortalité-incidence donnent-ils une indication sur l'erreur de représentation dans le Système national de déclaration des cas de cancer.

En nous concentrant sur les sièges les plus fréquemment diagnostiqués (à l'exception du cancer de la peau) d'après le fichier SNDCC, nous présentons aux tableaux 1 et 2, pour chaque sexe, les ratios mortalité-incidence à l'échelle nationale ainsi que les provinces comptant les ratios les plus élevés ou les moins élevés, pour deux périodes de cinq ans, ainsi qu'une ventilation par groupes d'âge. Nous excluons l'île-du-Prince-Édouard de notre étude parce que le nombre de cas observés dans cette province est trop faible pour permettre une comparaison valable.

2.2 Appariement des dossiers mortalité et incidence

Pour établir la possibilité d'évaluer les erreurs de classification des causes de décès, ou les erreurs de classification des sièges du cancer, on peut tirer un échantillon des inscriptions provenant soit du fichier mortalité, soit du fichier incidence, et ensuite rechercher dans l'autre fichier des dossiers appariés. La recherche manuelle ne garantit pas que l'on trouvera tous les appariements valables et, en fait, le taux d'appariement peut être différent d'un bureau de l'état civil à l'autre, parce que le niveau de détail des variables d'appariement peut varier d'un bureau à l'autre. (Voir à la partie 2.3 une étude sur la disponibilité des données).

2. DESCRIPTION DES MESURES

Dans cette partie, nous décrivons certaines méthodes pour étudier la qualité des données contenues dans les registres sur la mortalité et l'incidence du cancer.

2.1 Ratios mortalité-incidence

En vue d'étudier les taux relatifs de sous-représentation parmi les inscriptions de l'incidence du cancer, nous considérons les ratios entre les décès et les cas de cancer. Étant donné qu'au Canada tous les décès sont enregistrés selon la cause du décès et que suivant le "Système national de déclaration des cas de cancer" (SNDCC), tous les nouveaux cas de néoplasme malin sont enregistrés³, si les deux systèmes de consignation étaient de la même qualité dans tous les bureaux de l'état civil on devrait s'attendre à ce que la proportion entre les taux de mortalité et les taux d'incidence relatifs à un siège donné soit passablement uniforme au sein d'une population d'un âge et d'un sexe donnés et au cours de périodes suffisamment longues. (Nous calculons ces proportions de décès et de cas pour des périodes de cinq ans et de dix ans afin de réduire l'effet du décalage dans le temps entre la déclaration d'un cas de cancer et le décès). Cette proportion manquerait de cohérence si l'un ou l'autre des taux ci-dessous différait d'un bureau de l'état civil à l'autre.

- (a) taux de survie ou accroissement soudain du taux d'incidence,
- (b) taux de mortalité associée à d'autres risques concurrents,
- (c) taux d'erreur de codage de la cause sous-jacente de décès
- (d) taux d'erreur dans la classification du siège du cancer dans le cas des nouveaux cancers
- (e) taux de sous-représentation ou de sur-représentation des cas de cancer

³ Exceptionnellement, on enregistre les cancers métastatiques lorsque le siège primaire est inconnu.

du cancer et le sous-ensemble de dossiers tirés du fichier sur la mortalité, lequel fait apparaître le cancer comme cause sous-jacente du décès.

Voici les aspects de la qualité des données que nous avons retenus pour notre étude:

1. L'intégralité des inscriptions de nouveaux cas de cancer par le biais d'une comparaison entre la mortalité attribuable au cancer et l'incidence du cancer pendant la même période. Cette comparaison constitue un indicateur très approximatif (mais couramment utilisé) de l'intégralité des inscriptions.
2. La compatibilité des codes de détermination du diagnostic et de la cause du décès par le biais d'un appariement des dossiers individuels formant les deux fichiers.
3. La disponibilité et l'intégralité des données élémentaires par le biais d'une analyse de la fréquence à laquelle des valeurs valides apparaissent dans les fichiers.
4. L'inscription des cas de cancer primitifs multiples par le biais d'un appariement des dossiers formant le registre.

Notre étude couvre la période comprise entre 1969 et 1978 pour laquelle on dispose des données sur l'incidence du cancer. Nous avons exclu l'Ontario du champ de toutes nos recherches étant donné que le Système national de déclaration des cas de cancer comprend des données sur cette province uniquement pour la période qui s'étend de 1969 à 1971².

La 2^e partie de cet article expose la méthode adoptée pour les recherches et la 3^e partie les conclusions de l'étude.

² L'Ontario a créé un système passif d'inscription qui fait appel aux rapports sur les victimes du cancer, préparés à d'autres fins. Cette province prépare actuellement les données relatives aux dernières années.

d'autres aspects de la qualité des données telles que le détail de l'information sociodémographique et géographique qui est fournie. Par ailleurs, les données sur l'incidence du cancer sont sensibles à des facteurs tels que les programmes de dépistage de la masse qui mènent à l'inclusion de cas répandus non diagnostiqués auparavant.

Les rédacteurs de "Cancer Incidence in Five Continents" utilisent et expliquent plusieurs indices qui peuvent servir à déterminer si les données consignées sont complètes et fiables (4). Parmi ces indices, il faut citer les ratios mortalité-incidence du cancer en tant qu'indicateur de l'intégralité des inscriptions (voir parties 2.1 et 3.1 de cet article).

La déclaration des décès est une obligation légale dans la plupart des pays développés; on suppose donc que les erreurs d'enregistrement y sont minimales. Aux fins de la recherche épidémiologique sur le cancer, les limitations connues ou présumées des données sur la mortalité comprennent un manque d'information sur les cancers non mortels, une information moins précise sur le diagnostic et des erreurs de classification plus fréquentes attribuables à l'explication et au codage de la cause sous-jacente des décès, d'où une précision moindre en comparaison de l'information sur le diagnostic dans les statistiques sur l'incidence du cancer.

Une évaluation de la qualité des statistiques de l'état civil entreprise sous forme d'étude pilote fournit certaines indications sur la qualité des données concernant la mortalité au Canada (toutes les causes), notamment sur les taux d'erreur dans le codage de la cause sous-jacente du décès. Ce taux d'erreur s'élevait à 7.2 % en 1976. Près des deux tiers des erreurs portaient sur le premier et le deuxième chiffres du code de la cause du décès, lequel comprend 4 chiffres. La variation du taux d'erreur selon les causes spécifiques du décès n'a pas été étudiée.

Dans le présent article, nous évaluerons la possibilité de mesurer certains aspects touchant la qualité des deux fichiers de données de Statistique Canada qui servent aux études épidémiologiques, à savoir le fichier sur l'incidence

2. Données sur l'incidence du cancer à l'échelle nationale basées sur les déclarations des registres provinciaux des cas de cancer. Cette série de données a été établie en 1921.

La bonne qualité des statistiques fournissant des renseignements fiables sur les différentielles de risque dépend de l'intégralité et de la précision des cas de cancer enregistrés et de la comparabilité des données entre les différents bureaux de l'état civil et les différentes périodes. En tant qu'indice du risque réel lié au cancer, chacune des données sur l'incidence du cancer et la mortalité due à cette maladie présente de l'intérêt tout en ayant des limites qui lui sont propres.

Les données sur l'incidence du cancer se prêtent particulièrement bien à l'étude épidémiologique de cette maladie parce qu'elles fournissent des renseignements sur tous les cancers et non uniquement sur ceux qui sont mortels, parce qu'elles permettent de prévoir les problèmes naissants et parce que l'information sur le diagnostic est généralement détaillée et de très bonne qualité. Par exemple, la publication "Cancer Incidence in Five Continents (1)" souligne le rôle que jouent les comparaisons de cette incidence à l'échelle internationale en fournissant des indices sur les causes du cancer en dépit de certaines limitations bien connues des données. Parmi ces restrictions, citons la difficulté d'obtenir la déclaration de tous les nouveaux cas de cancer et les différences entre les bureaux de l'état civil lorsque les nouveaux cas sont bien déclarés. Les principaux facteurs qui influent sur le relevé effectif des cas sont les suivants: quels sont le nombre et les types de sources de données utilisées, dans quelle mesure on procède à une recherche active des cas, depuis combien de temps le bureau de l'état civil existe et si la déclaration des cas de cancer constitue ou non une obligation légale dans la région où est situé le bureau. Au Canada, les bureaux d'enregistrement des cas de cancer emploient une méthodologie passablement hétérogène pour la collecte de leurs données, mais ils essaient tous de suivre les directives tant internationales (2) que nationales (3) pour la consignation normalisée des données sur l'incidence du cancer. Les différences de sources et de techniques influencent non seulement le champ d'observation mais aussi

CERTAINS ASPECTS DE LA QUALITÉ DES STATISTIQUES SUR LA MORTALITÉ DUE AU CANCER ET L'INCIDENCE DE CETTE MALADIE

D. Binder et A. Malhotra¹

Depuis 1921, Statistique Canada, organisme central de regroupement de statistiques au Canada, compile des données sur la mortalité à l'échelle nationale, notamment celles qui touchent la mortalité due au cancer. Il dispose également de données sur l'incidence du cancer qui remontent à 1969.

On peut évaluer de diverses façons la qualité des données de ces fichiers. Les rapports entre la mortalité due au cancer et l'incidence de cette maladie donnent certaines indications sur les erreurs de représentation. L'appariement des micro-données entre les fichiers "incidence" et "mortalité" donnent un aperçu des erreurs de classification. De même, les inscriptions multiples de l'incidence du cancer posent le problème du double emploi. Par ailleurs, l'intégralité et la disponibilité des données élémentaires revêtent de l'importance dans le cas d'études spéciales.

Dans cet article, nous étudions la possibilité de nous servir de ces mesures de la qualité des données et les conséquences que peuvent avoir les mesures en question.

1. INTRODUCTION

Les statistiques démographiques du cancer servent de base aux recherches épidémiologiques sur la distribution et les déterminants de cette maladie ainsi qu'aux programmes de santé visant la prévention, le diagnostic et le traitement du cancer. Statistique Canada, organisme central de regroupement de statistiques, compile deux types de statistiques de ce genre sur le cancer:

1. Données sur la mortalité à l'échelle nationale basées sur les rapports provenant des registres provinciaux de l'état civil. Ces données remontent à 1921.

¹ D. Binder, Division des méthodes d'enquête-institutions et agriculture, Statistique Canada et A. Malhotra, Division de la santé, Statistique Canada.

BIBLIOGRAPHIE

- [1] Burchell, J.M.: "International Air Passenger Origin/Destination Project"; exposé présenté à la Division des transports et des communications de Statistique Canada, mars 1981.
- [2] "International Air Passenger O & D Statistics System - User Requirements Specifications"; document du Centre des statistiques de l'aviation (CSA) de la Division des transports et des communications de Statistique Canada, revu en mars 1981.
- [3] "International Air Passenger O & D Statistics System - Feasibility Study Report"; document du CSA, janvier 1981.
- [4] "International Air Passenger O & D Statistics System - Requirements for Functional Specifications (Concepts)"; document du CSA, revu en septembre 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"; rapport confidentiel établi pour la Direction générale des études économiques du transport de passagers et de l'aviation, Direction de la recherche, Commission canadienne des transports, février 1977.
- [6] Carpenter, R.: "Specifications for Estimates of the Reliability for the Air Scheduled International Passenger Origin and Destination Estimates"; document de la Division des méthodes d'enquêtes-entreprises de Statistique Canada, revu en janvier 1982.

d'échantillonnage est inférieure à 10% quand il est nécessaire d'inclure des données imputées dans l'estimation du nombre global de passagers dans un marché. Ainsi, l'inclusion de f_j au lieu de 0.10 dans l'expression de la variance d'une estimation produit un coefficient ajusté qui s'applique aux estimations globales relatives à un marché.

Cette méthode tient compte du fait que des données d'échantillonnage sont utilisées dans la technique d'attribution, mais elle est fondée, comme le procédé qui détermine les estimations elles-mêmes, sur l'hypothèse que les voyages DOD tronqués représentent bien le trafic sans correspondance.

8. PROJETS À L'ÉTUDE

Le besoin de disposer de renseignements sur les marchés des transports aériens pour les négociations bilatérales sur les services internationaux a été expliqué dans une des sections précédentes. L'échange de statistiques constitue un des éléments négociables prévus dans les accords sur le transport aérien, et des ententes permettant de tels échanges existent déjà entre le Canada et plusieurs pays. Les notions que ces pays appliquent dans leurs enquêtes ainsi que la qualité de leurs données nous laissent croire que ces chiffres pourraient être utilisés dans le système d'estimation de l'ASIPOD. De telles données fourniraient un échantillonnage des trajets non dénombrés dans l'univers des billets émis pour les vols internationaux. Des études de faisabilité actuellement en cours tentent de déterminer si le nombre de pays qui pourraient échanger ces statistiques contribuerait à améliorer la précision d'un nombre d'estimations suffisant pour justifier l'expansion du système d'estimation de l'ASIPOD en vue d'inclure des "données échangées".

REMERCIEMENTS

Les auteurs remercient leurs collègues de la Division des transports et des communications et de la Division des méthodes d'enquêtes-entreprises de Statistique Canada pour les nombreuses remarques pertinentes faites pendant la rédaction de ce document.

7.2 Variance des estimations globales relatives à un marché

La méthode de calcul des coefficients de variation pour l'ensemble d'un marché doit assurer que les données imputées sont une fonction des données de l'échantillon. Autrement dit, des échantillons différents produisent des valeurs différentes des données imputées et, en conséquence, des grandeurs différentes des estimations globales relatives à un marché. La réutilisation de certaines parties de l'échantillon a un effet sur la distribution d'échantillonnage des estimations.

La méthode que nous proposons corrige la fraction d'échantillonnage de 10% pour qu'elle corresponde au pourcentage de passagers réellement échantillonnés dans chaque domaine d'intérêt. Comme on prend pour acquis l'application de la technique d'attribution, on suppose également que les voyages DOD tronqués représentent bien le trafic sans correspondance. Notre mesure de fiabilité constitue une mesure de précision seulement si cette hypothèse est valable. Il est donc raisonnable de rectifier les poids des voyages DOD échantillonnés afin de prendre en considération le trafic sans correspondance. Ainsi, la fraction d'échantillonnage nécessaire pour l'estimation du j^e domaine d'intérêt serait f_j , comme définie dans l'équation (1) plus haut. Dans l'équation (2), f_j remplacerait f pour le calcul de $\hat{\text{var}}(\hat{d})$, ce qui donne la formule pour la variance des estimations globales relatives à un marché:

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

Ensuite, le coefficient de variation des estimations globales relatives à un marché serait:

$$\hat{c}v(\hat{d}_j^I) = \sqrt{\hat{\text{var}}(\hat{d}_j^I)} / \hat{d}_j^I .$$

Signalons que $f_j < 0.10$ lorsque $a_j > 0$. Ce qui signifie que la fraction

L'enquête sur l'origine et la destination des passagers payants utilise effectivement un échantillon aléatoire simple de 10%, puisque

- (i) le choix des volets dont le numéro d'ordre se termine par "0" produit un échantillonnage systématique, et
- (ii) la distribution des billets ne comporte aucun cycle qui causerait une relation entre le nombre de passagers estimé par l'enquête et le dernier chiffre du numéro d'ordre.

L'estimation de la variance peut donc s'écrire:

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

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

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

et le coefficient de variation peut s'exprimer sous la forme:

$$c\hat{v}(\hat{d}) = \sqrt{\hat{\text{var}}(\hat{d})/\hat{d}} .$$

A noter que $\hat{\text{var}}(\hat{d})$ peut s'écrire également:

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

où n est supposé suffisamment grand pour que $n/(n-1)$ égale approximativement 1.

Prenons les estimations du nombre total de passagers sortants et entrants, sans distinguer la direction du voyage. Parmi les diverses paires d'origines et de destinations, le billet en question ajouterait des passagers aux domaines suivants:

<u>Domaine d'intérêt</u>	<u>Nombre de passagers</u>
Winnipeg - Londres	1
Toronto - Londres	1.
Canada - Londres	2
Canada - Europe	2
Est du Canada - Europe	1
Ouest du Canada - Europe	1

Notons que le nombre de passagers par billet dépend du niveau de répartition géographique et, par conséquent, de l'origine et de la destination (c'est-à-dire des domaines d'intérêt) pour lesquelles des estimations du trafic sont calculées.

On peut trouver $\hat{d}$, l'estimation du nombre de passagers qui voyagent entre une origine et une destination en particulier, d'après l'enquête sur l'origine et la destination des passagers payants, à l'aide de la formule suivante:

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

où x_i est le nombre de voyages simples appartenant à une paire d'origines et de destinations pour le i^e billet.

n est le nombre total de billets dans l'échantillon, et

N est le nombre total de billets dans la population de l'enquête sur l'origine et la destination des passagers payants.

f est la fraction d'échantillonnage, c'est-à-dire, $n/N = 0.1$.

7. ESTIMATION DE LA VARIANCE

L'application de la technique d'attribution à l'estimation des marchés internationaux représente une solution nouvelle pour un problème de taille. Elle ne règle pas complètement les difficultés causées par la non-exhaustivité du dénombrement, mais elle constitue un grand pas en avant, comme d'ailleurs la production de statistiques sur la variance des estimations. Les chiffres sur la variance devraient à la fois tenir compte de la technique d'attribution et de ses hypothèses, et fournir une mesure claire de la fiabilité des estimations relatives aux DOD.

L'estimateur de la variance des estimations du trafic international par paires d'origines et de destinations est une modification simple de l'estimateur de la variance utilisé dans l'enquête sur l'origine et la destination des passagers payants.

7.1 Variance de l'estimation du trafic avec correspondance

Le calcul de l'estimateur de la variance des estimations de l'origine et de la destination des passagers payants dépend du nombre de passagers que chaque billet attribue aux paires d'origines et de destinations (c'est-à-dire au domaines d'intérêt). Rappelons que chaque billet est choisi avec une probabilité de 0.1 et que l'itinéraire peut être divisé en plusieurs segments ou voyages DOD. Un billet peut faire inscrire 0, 1, 2, etc. passagers dans un domaine d'intérêt donné. Par exemple, l'itinéraire

YWG - AC - LON - BA - YYZ

serait coupé, selon le programme qui calcule les points de division, en deux voyages simples:

YWG - AC - LON
et
LON - BA - YYZ

croient que de telles erreurs dans cette enquête sont petites. Toutefois, des consultations sont actuellement en cours avec les grands transporteurs aériens canadiens pour trouver des façons de réduire les exigences des organismes gouvernementaux en matière de déclarations. Au terme de ces discussions, l'activité aéroportuaire pourrait devenir l'objet d'une enquête-échantillon. Lorsque viendra le moment d'établir le plan de sondage, il faudra tenir compte de l'importance d'obtenir des données précises sur le volume des passagers aux points d'entrée ou de sortie.

La technique d'attribution suppose que tout le trafic des passagers non-dénombrés est pris en compte dans le trafic sans correspondance. Une façon intéressante de vérifier cette hypothèse serait de comparer, pour une même période de référence, le nombre de passagers d'un transporteur canadien déclaré dans l'enquête sur l'activité aéroportuaire et une estimation faite à partir de l'échantillon de l'enquête sur l'origine et la destination des voyageurs. Cette estimation correspondrait au nombre total de passagers dont les voyages DOD comportent le même transporteur pour le trajet international et le même aéroport canadien comme point d'entrée ou de sortie. Il faudrait pouvoir dire si les écarts entre le chiffre déclaré et les estimations proviennent de différences entre les notions appliquées par ces enquêtes et, dans l'affirmative, vérifier si ces différences sont prises en considération dans le système d'estimation de l'ASIPOD. Si le système n'en tient pas compte, cela pourrait confirmer que l'hypothèse selon laquelle tout le trafic des passagers non-dénombrés est compris dans le trafic sans correspondance soulève des problèmes.

L'étude d'un bon nombre de ces lacunes devrait donc être poussée. Le système actuel comporte de sérieuses lacunes, mais on n'a pas encore réussi à trouver une solution de rechange à la technique d'attribution, étant donné les contraintes de temps et de coût. Néanmoins, les améliorations apportées aux estimations du trafic pour chaque marché par le système d'estimation de l'ASIPOD seront appréciables.

La technique d'attribution est fondée sur le postulat que les voyages DOD tronqués représentent bien le trafic sans correspondance. Examinons l'exemple hypothétique que nous venons d'illustrer afin de voir si cette hypothèse est, par intuition, raisonnable. La technique d'attribution suppose que les passagers qui partent de Montréal à bord d'un vol d'un transporteur étranger particulier auront la même destination finale que les passagers dont l'origine est Winnipeg ou Toronto et qui passent par Montréal. Cela revient à utiliser le comportement des voyageurs des communautés ethniques, comme les groupes polonais de Toronto et allemands de Winnipeg, pour tracer le comportement des communautés à prédominance française de Montréal. En outre, le caractère douteux du postulat de similitude entre les voyages avec et sans correspondance a été démontré empiriquement dans une étude-pilote (Rosen et Conroy (1977)) effectuée par la Commission canadienne des transports, en 1977. Bref, il existe des arguments non seulement intuitifs, mais aussi empiriques qui réfutent le postulat de base de cette technique.

La précision des estimations du nombre de passagers par paire d'origines et de destinations est mise en cause par toute exception à l'hypothèse que les voyages DOD tronqués représentent bien le trafic sans correspondance. Or, comme dans le cas du tableau 1, de grands volumes de passagers sont attribués à des paires d'origines et de destinations à partir d'un échantillon "réel" de petite taille. Par exemple, la fraction d'échantillonnage "réelle" du marché formé entre le Canada et la région no 7 n'est pas 10%, comme dans l'enquête sur le trafic intérieur, mais 5.7% (c'est-à-dire $(100\% - 42.7\%) \times 10\%$), à cause de l'exclusion du trafic sans correspondance de l'univers. Par conséquent, un échantillon de moins de 10% sert à répartir un grand volume de passagers. L'existence d'exceptions à l'hypothèse en question pourrait donc causer de grands biais dans les estimations.

Comme les chiffres tirés de l'enquête sur l'activité aéroportuaire constituent des repères sur le volume du trafic, leur précision est très importante. Bien qu'aucune évaluation n'ait été entreprise concernant l'ampleur du biais attribuable à des erreurs d'observation dans les chiffres sur l'activité aéroportuaire, les économistes qui analysent les statistiques sur l'aviation

<u>Nombre de passagers dans l'échantillon</u>	<u>Estimation du nombre de passagers</u>	<u>DOD</u>
1	10	YWG-AC-YMX-BA-LON-LO-WAW
2	20	YYZ-CP-YMX-BA-LON-LH-HAM

où les codes doivent être interprétés comme suit:

<u>Code</u>	<u>Désignation</u>
YWG	Winnipeg
YMX	Montréal (Mirabel)
LON	Londres
WAW	Varsovie
YYZ	Toronto
HAM	Hambourg
AC	Air Canada
BA	British Airways
LO	LOT
CP	CP Air
LH	Lufthansa

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

Les voyages DOD tronqués, la proportion de ces voyages inclus dans D_i , le nombre de passagers à répartir et les estimations totales pour le marché sont donc:

<u>Voyages DOD tronqués</u>	<u>Proportions</u>	<u>Passagers à répartir (a_j)</u>
YMX - BA - LON - LO - WAW	1/3	30
YMX - BA - LON - LH - HAM	2/3	60

6.4 Lacunes de la technique d'attribution

Cette technique comporte toutefois certaines lacunes.

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

et où f_j est la fraction d'échantillonnage corrigée pour la j^{e} paire d'origines et de destinations dans l'enquête sur l'origine et la destination des passagers payants;

d_j est le nombre de passagers des vols internationaux pour la j^{e} paire d'origines et de destinations dans l'échantillon de l'enquête sur l'origine et la destination des passagers payants, et

a_j est le nombre de passagers attribués à la j^{e} paire d'origines et de destinations.

Il convient de souligner que, conformément au point (iii) ci-dessus, a_j peut s'exprimer:

$$a_j = \frac{d_j}{D_i} \times A_i, \quad j \in i$$

$$\text{où} \quad \sum_{j \in i} d_j = D_i$$

$$\text{et} \quad \sum_{j \in i} a_j = A_i$$

6.3 Un exemple d'application de la technique d'attribution

Prenons un exemple simple pour illustrer comment fonctionne la technique d'attribution. Les données sur l'activité aéroportuaire fournies par British Airways révèlent que 120 passagers sont embarqués à Montréal à destination de Londres. Dans ce cas, $C_i = 120$. Supposons que l'enquête sur l'origine et la destination des voyageurs payants n'indique que deux suites de voyages DOD:

dénombrée de la population cible, sont appelées des voyages DOD sans correspondance, puisqu'il s'agit de paires d'origines et de destinations de voyages simples assurés par des transporteurs étrangers qui n'ont pas de correspondances avec un transporteur participant canadien. Selon la technique d'attribution, on impute les voyages DOD sans correspondance au i^e groupe de la façon suivante:

- (i) Une liste de tous les voyages DOD inclus dans D_i est dressée (c'est-à-dire au même aéroport, par le même transporteur étranger et dans la même direction).
- (ii) Les trajets intérieurs sont éliminés (c'est-à-dire les trajets entre un point canadien et un point d'entrée ou de sortie canadien). (Ces trajets étant faits par un transporteur canadien, ils sont pris en compte dans l'enquête sur l'origine et la destination des passagers payants. Les voyages ainsi "tronqués" sont alors sans correspondance.)
- (iii) Le volume des passagers non dénombrés, A_i , est réparti parmi les voyages DOD ainsi "échantillonnés", selon la proportion de ces voyages dans D_i .

De nouveaux totaux de passagers classés par origine et destination de voyages simples sont produits, et il s'agit alors de "passagers répartis".

La technique d'attribution repose sur l'hypothèse que les voyages DOD tronqués représentent bien le trafic sans correspondance, ce qui donne plus de poids à certaines paires d'origines et de destinations que dans l'enquête qui porte seulement sur l'origine et la destination des passagers payants. On peut donc trouver l'estimateur $\hat{d}_j^T$, qui correspond au nombre total de passagers observés sur le marché pour la j^e paire d'origines et de destinations, en corrigeant la fraction d'échantillonnage de la manière suivante:

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

g est le nombre d'aéroports canadiens qui servent de point d'entrée ou de sortie, et

n_i est le nombre de transporteurs étrangers (non américains) détenant un permis d'atterrissage, qui desservent le i^e aéroport canadien de point d'entrée ou de sortie.

L'enquête sur l'origine et la destination des passagers payants permet, par recoupement des transporteurs et des aéroports canadiens, d'obtenir un classement semblable des passagers entrants et sortants par origine et destination du voyage simple. Ainsi, l'échantillon de cette enquête fournit également b comptages.

A la première étape de la technique d'attribution, le volume de passagers non dénombrés, A_i , inclus dans le groupe d'attribution i , peut être estimé au moyen de l'équation suivante:

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

où C_i est le volume total des passagers dans le groupe i , selon l'enquête sur l'activité aéroportuaire;

f est la fraction d'échantillonnage de l'enquête sur l'origine et la destination des passagers payants (c'est-à-dire $1/10$);

D_i est le nombre échantillonné de passagers des vols internationaux dans le groupe i ;

A_i correspond ainsi à l'estimation du nombre de passagers déplacés par un transporteur étranger et qui tombent dans le i^e des b groupes d'attribution, mais pour lesquels il manque des renseignements sur l'origine et la destination.

A l'étape suivante, on attribue le volume des passagers non dénombrés à des paires d'origines et de destinations. Ces paires reliées à la partie non

monde, on obtiendrait de plus hauts pourcentages de non-dénombrement pour chaque région du monde. Donc, il semble que, de façon générale, plus les régions géographiques sont désagrégées, plus le pourcentage de non-dénombrement est élevé. Pour citer un exemple, le pourcentage de non-dénombrement des passagers qui ont voyagé entre l'Est du Canada et la région no. 7 est de 55% contre 42.7% pour l'ensemble du Canada (tableau 1). Afin de clarifier l'affirmation qu'un haut degré d'agrégation géographique dans le choix des paires d'origines et de destinations produit généralement un faible pourcentage de non-dénombrement, il faut souligner que le non-dénombrement existe seulement dans le cas des vols des transporteurs étrangers qui prennent fin à un point d'entrée au Canada. Le dénombrement du trafic de passagers est exhaustif lorsque la partie canadienne des paires d'origines et de destinations n'est ni un point d'entrée, ni un point de sortie. Ainsi, lorsque le degré d'agrégation géographique augmente, on inclut plus de vols avec correspondance, ce qui fait diminuer le pourcentage de non-dénombrement.

6.2 La technique d'attribution

La technique d'attribution consiste d'abord à estimer le volume de passagers non dénombrés, pour ensuite répartir ce volume dans des paires d'origines et de destinations.

A partir de l'enquête sur l'activité aéroportuaire, on effectue b comptages repères des passagers entrants et sortants aux aéroports canadiens, par transporteur. La valeur de b , le nombre de catégories auxquelles on peut attribuer un volume de voyageurs, se définit de la façon suivante:

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

où le "2" s'explique par le fait qu'il y a deux comptages, un pour les passagers qui entrent au Canada et un pour ceux qui sortent du pays,

problème pour le système d'estimation de l'ASIPOD à cause de l'absence de données sur l'origine et la destination de la partie non dénombrée de la population cible.

Le tableau suivant donne un indice du volume de passagers qui se déplacent entre le Canada et neuf régions du monde. (Ces chiffres sont des estimations annuelles pour 1979. Les neuf régions du monde ne sont pas nommées parce que ces données sont confidentielles.)

Tableau 1 - Passagers des vols internationaux - estimations, 1979

Entre le Canada et...	Origine et destination des passagers payants	Non-dénombrement (le pourcentage du total est entre parenthèses)		Total
la région no 1	116,050	495	(0.4)	116,545
la région no 2	325,200	2,020	(0.6)	327,220
la région no 3	99,010	12,628	(11.3)	111,638
la région no 4	67,200	14,558	(17.8)	81,758
la région no 5	575,010	126,076	(18.2)	703,086
la région no 6	205,180	101,464	(33.1)	306,644
la région no 7	1,221,040	908,876	(42.7)	2,129,916
la région no 8	54,410	56,807	(51.1)	111,217
la région no 9	45,330	57,267	(55.8)	102,597
Total mondial	2,708,430	1,282,191	(32.1)	3,990,621

D'après les pourcentages de non-dénombrement, il est évident que ce problème est important.

Si ce tableau était fait pour l'Est du Canada et les neuf mêmes régions du

contrôle de ces erreurs d'observation, erreurs peu faciles à corriger, n'est pas un objectif de ce programme de révision.

- (ii) Mettre au point un système de calcul simple, facile à utiliser et qui fournit des messages indicateurs sommaires.

Quelque 625,000 enregistrements d'origine et de destination, et quelque 820,000 enregistrements concernant l'activité aéroportuaire doivent être traités annuellement avec un minimum d'intervention manuelle. Etant donné qu'une multitude de passagers est répartie parmi un nombre considérable de paires d'origines et de destinations, les messages imprimés à chaque étape du système doivent résumer le traitement jusqu'à ce point, tout en indiquant les problèmes éventuels.

- (iii) Fournir des tableaux des estimations du nombre de passagers sur les vols internationaux, soit régulièrement ou au besoin, de manière à simplifier les analyses que les utilisateurs doivent effectuer.
- (iv) Produire des estimations, dont la fiabilité est mesurable, sur l'origine et la destination des passagers à bord des vols internationaux à horaire fixe.

Bien qu'on ait cru dans le passé que la fiabilité de ces statistiques était variable, elle était de fait inconnue jusqu'à présent. L'estimation de la fiabilité de ces données permettra de mieux connaître la force des conclusions qu'on peut tirer de leur analyse. Les déductions faites sans une mesure de la fiabilité des données peuvent être en fait fort trompeuses.

6. RÉOLUTION DU PROBLÈME DE NON-EXHAUSTIVITÉ

6.1 Importance du problème

Comme dans d'autres enquêtes, la non-exhaustivité du dénombrement pose un

Supposons également que deux suites de voyages simples sont possibles:

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

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

Le point de sortie canadien dans cet exemple serait Toronto, puisque c'est là que les passagers empruntent le transporteur étranger.

5. OBJECTIFS DE LA RÉVISION

Les quatre principaux objectifs visés par le projet de révision, dont le quatrième sera analysé en détail dans ce document, sont les suivants:

- (i) Éliminer les problèmes signalés dans le système actuel.

Comme le système actuel ne prévoit pas d'imputation dans le cas des codes de transporteur inscrits de façon illisible sur le volet des billets d'avion, la catégorie "transporteurs inconnus" devient la troisième en importance dans les totalisations. En outre, il n'y a aucune vérification des codes qui permette de déterminer si un transporteur vole en direction ou à partir d'aéroports où il est effectivement autorisé à atterrir. Compte tenu des besoins des totalisations actuelles, il faut donc prévoir la vérification et l'imputation des données illisibles ou fausses inscrites sur les volets internationaux, en plus des corrections nécessaires aux coupons de vol exclusivement intérieurs.

D'autres erreurs d'observation ont été relevées, mais il est difficile de les corriger à l'aide d'un système d'estimation. Ce type d'erreurs inclut les cas où les transporteurs interprètent mal les instructions relatives aux coupons de vol; des erreurs fréquentes dans les numéros d'ordre des billets, qui sont utilisés pour le choix de l'échantillon; des erreurs dans les systèmes de traitement des transporteurs, etc. Le

américains) qui exploitent des services internationaux à horaire fixe à destination ou en provenance du Canada. Depuis le 1er janvier 1982, 10 transporteurs américains et 21 transporteurs étrangers remplissent une déclaration pour chaque aéroport canadien qu'ils desservent. Tout nouveau transporteur étranger qui obtient un permis d'exploitation de services à horaire fixe au Canada est automatiquement inclus dans le système de collecte des données sur l'activité aéroportuaire.

Les données sur le flux de passagers dans l'ensemble du trafic aérien sont extraites des renseignements sur l'activité aéroportuaire. Par "flux de passagers", on entend le nombre de personnes qui, pendant une période déterminée, voyagent à bord d'un transporteur donné entre un aéroport canadien déclarant et un point adjacent, c'est-à-dire le dernier aéroport visité ou le prochain point prévu à l'itinéraire. La technique d'attribution ne tient compte que des déclarations fournies par les transporteurs étrangers (non américains). Les éléments d'information extraits de cette enquête, qui sont utilisés pour déterminer les marchés formés par des paires d'origines et de destinations, sont le nombre de passagers payants embarqués ou débarqués au Canada, le transporteur au point d'entrée ou de sortie canadien et le nom du point d'entrée ou de sortie canadien. Par "point d'entrée ou de sortie canadien", nous entendons un aéroport qui participe à l'enquête et où un transporteur étranger (non américain) arrive au Canada ou en sort.

Cependant, d'après l'enquête sur l'origine et la destination des passagers payants, le point d'entrée/de sortie canadien pour les transporteurs canadiens et américains est le premier/dernier point canadien dans un itinéraire pour les vols à destination/en provenance du Canada. Quant aux transporteurs étrangers, le point d'entrée ou de sortie correspond au point à l'intérieur du Canada où un passager emprunte les services d'un transporteur étranger ou s'en sépare. Prenons l'exemple suivant:

Supposons qu'Air France assure un service Toronto - Montréal - Paris, et qu'il y a embarquement de certains passagers à Toronto et d'autres à Montréal.

défini comme "une suite de points de départ et de destination énumérés dans un ordre qui indique la direction suivie dans l'itinéraire". Les voyages simples (DOD) sont les parties d'un itinéraire, lequel est divisé de façon que chaque voyage simple corresponde à une seule direction. Cette méthode exige qu'on décompose les itinéraires semi-circulaires ("open-jaw"), et les itinéraires de retour comme les itinéraires symétriques et circulaires, en voyages simples à sens unique. Pour obtenir les voyages DOD, un programme qui calcule des points de division est appliqué aux données sur l'origine et la destination des voyages complets. Ce procédé est automatisé dans le système qui regroupe les statistiques sur l'origine et la destination des passagers aériens, et il comporte le calcul de la distance entre divers points d'un itinéraire, suivi de comparaisons entre ces distances et la distance totale de l'itinéraire. En règle générale, on divise les itinéraires à partir du point le plus éloigné du lieu d'origine. Chaque voyage DOD ainsi formé est présenté de nouveau au programme jusqu'à ce qu'aucun nouveau point de division ne puisse être trouvé.

Les données sur l'activité aéroportuaire proviennent de l'état 32. Les caractéristiques relevées pour chaque vol incluent:

- le transporteur participant;
- l'aéroport déclarant;
- le point d'origine et la destination finale prévue pour le vol;
- le nom du dernier aéroport visité, dans le cas des arrivées, ou le prochain aéroport prévu à l'itinéraire, dans le cas des départs;
- le nombre de débarquements et d'embarquements de passagers payants.

Les données sur l'activité aéroportuaire sont envoyées tous les mois par les transporteurs transcontinentaux (Air Canada et CP Air) et régionaux (Eastern Provincial, Québecair, Nordair et Pacific Western Airlines) du Canada, par Norcanair et par tous les transporteurs étrangers (y compris les transporteurs

transports aériens de la Commission canadienne des transports, en collaboration avec le ministère des Transports. Les données sont recueillies auprès des transporteurs aériens, pour le compte du Comité des transports aériens, par le Centre des statistiques de l'aviation (CSA) de Statistique Canada. En vertu des Règlements sur les transports aériens de la Loi sur l'aéronautique, les transporteurs sont obligés de remplir les états (questionnaires) du CSA.

Les statistiques sur l'origine et la destination (O & D) des passagers payants parviennent au CSA grâce à l'état 35. Parmi les renseignements recueillis, on trouve entre autres:

- l'origine et la destination inscrites sur le billet;
- les points où un passager s'arrête pour prendre une correspondance sur la même ligne ou sur une ligne différente;
- le transporteur inscrit sur chaque volet.

Les renseignements sur l'origine et la destination des passagers payants sont fournis chaque mois par les principaux transporteurs canadiens qui émettent des billets à tarif unitaire et exploitent des services à horaire fixe. Depuis le 1er janvier 1982, les sept transporteurs canadiens qui participent à cette enquête sont Air Canada, CP Air, Eastern Provincial Airways, Nordair, Pacific Western Airlines, Air Ontario et Québecair. Les données américaines sont recueillies par le Civil Aeronautics Board auprès de tous les transporteurs autorisés, à l'exception de ceux qui assurent les transports par hélicoptère et les transports intérieurs en Alaska. Les chiffres pour chaque trimestre sont regroupés et les renseignements en double supprimés de façon à constituer un fichier des itinéraires complets, d'après l'origine et la destination du voyage complet ("TOD", Ticket Origine and Destination).

Par contre, les statistiques sur l'origine et la destination des passagers aériens sont traitées selon le concept de l'origine et de la destination du voyage simple ("DOD", Directional Origine and Destination), qui peut être

sans correspondance qui le conduit de Paris à Montréal par Air France et de nouveau à Paris avec ce même transporteur, ce voyage ne figurera pas dans l'enquête sur l'origine et la destination des passagers payants. Toutefois, ce genre d'itinéraire fait effectivement partie de la population cible de l'étude des voyages internationaux.

Ce dénombrement incomplet de la population cible constitue en fait un problème de non-exhaustivité pour le système d'estimation de l'ASIPOD. Il s'agit d'un problème de "non-dénombrement" plutôt que de "sous-dénombrement" parce qu'il est impossible d'inclure une grande partie de l'univers dans le cadre de l'enquête sur le trafic international.

Dans le système actuel, on utilise les données des enquêtes sur l'origine et la destination des passagers payants et sur l'activité aéroportuaire, et on applique une méthode appelée la technique d'attribution, afin de produire des estimations globales pour chaque marché.

L'enquête sur l'activité aéroportuaire procède à un dénombrement complet des passagers à bord des vols internationaux à destination ou en provenance de chaque aéroport canadien. Tous les transporteurs canadiens, américains et étrangers qui effectuent des vols à horaire fixe sont inclus, mais sans considération du premier point de départ ou de la destination finale des passagers.

Ainsi, l'enquête sur l'activité aéroportuaire donne un dénombrement du volume total des passagers de tous les transporteurs qui desservent la population cible. La technique d'attribution consiste à estimer le volume de passagers non dénombrés par l'enquête sur l'origine et la destination des passagers payants et à les classer dans des paires d'origines et de destinations. Avant d'expliquer plus en détail cette technique, il convient d'abord de décrire ces deux enquêtes sur le trafic aérien de façon un peu plus complète.

4. PRINCIPAUX ASPECTS DES DEUX ENQUÊTES

L'organisation de ces deux programmes d'enquête est confiée au Comité des

destination des passagers entre le Canada et les États-Unis, dans lequel les États-Unis fournissent au Canada des renseignements sur les itinéraires complets où:

- (i) un point situé au Canada et un point situé aux États-Unis figurent sur le billet, ou
- (ii) un transporteur des États-Unis s'est rendu en un point situé au Canada ou a quitté un point canadien, ou
- (iii) un transporteur canadien s'est rendu en un point situé aux États-Unis ou a quitté un point américain.

A cause de cet accord, nous employons les expressions "étranger" ou "étranger (non américain)" pour désigner ce qui est "ni canadien, ni américain".

Ce type de billets est également l'univers de l'enquête sur l'origine et la destination des passagers payants, mais celle-ci vise uniquement les principaux transporteurs canadiens. Chaque transporteur participant prend un coupon des billets dont le numéro d'ordre se termine par "0", tenant compte seulement des volets dont il est le premier à prélever. Ainsi, l'enquête porte sur un échantillon de 10% des coupons de vol pour des itinéraires qui incluent les services d'au moins un des transporteurs canadiens ou américains participants.

Quelques renseignements sur les marchés des transporteurs étrangers (non américains) peuvent être tirés de l'enquête sur l'origine et la destination des passagers payants. Par exemple, si un passager voyage par Air Canada d'Ottawa à Montréal, et de là à Paris par Air France, l'enquête sur l'origine et la destination des passagers payants comptera ce trajet international parce qu'un transporteur canadien intervient quelque part dans l'itinéraire. Le transporteur canadien doit déclarer tous les détails sur les transporteurs et les trajets, y compris la partie concernant Air France.

L'enquête sur l'origine et la destination des passagers payants ne tient pas compte cependant des volets où des transporteurs étrangers sont inscrits du début à la fin d'un itinéraire. Par exemple, si un passager suit un itinéraire

- les transferts de fonds;
- l'exemption d'impôts sur le revenu.

Pour discuter de ces sujets et surtout pour échanger des routes, les négociateurs doivent savoir où se trouvent les marchés du transport aérien et s'ils sont en expansion. En conséquence, ils doivent disposer des résultats d'analyses de coûts et de bénéfices, tant pour le Canada que pour les autres pays, qui découlent de l'attribution de diverses routes internationales à des transporteurs canadiens et étrangers. Afin de permettre au Canada de négocier une juste part du marché, il est essentiel de posséder des estimations de l'origine et de la destination des passagers à bord des vols internationaux. A titre indicatif de l'importance de ces données pour l'économie canadienne, mentionnons que les recettes³ de l'ensemble des services aériens internationaux à destination ou en provenance du Canada ont atteint près de 2.3 milliards de dollars canadiens en 1980.

3. NATURE DU PROBLÈME DE NON-EXHAUSTIVITÉ

Le remaniement effectué dans le cadre de ce projet repose sur le principe qu'il faut conserver la méthodologie fondamentale appliquée dans le système actuel. Par conséquent, la plupart de ces travaux ont visé à chercher des améliorations souhaitables, à les classer selon leur degré d'utilité et à déterminer lesquelles pouvaient être adoptées compte tenu des délais et du budget. Le maintien de la méthodologie fondamentale a fourni les grandes lignes du calcul des estimations du nombre de passagers des vols internationaux, mais non celui des estimations de la fiabilité de ces chiffres.

La population cible de ce système est l'ensemble des billets émis pour un voyage international (c'est-à-dire entre un pays étranger et le Canada ou les Etats-Unis). Il existe un programme d'échange de données sur l'origine et la

³ Obtenues à partir de totalisations internes de la Commission canadienne des transports.

genre de service s'appelle un service à tarif unitaire.

Avant qu'un service commercial international à horaire fixe à destination ou en provenance du Canada puisse être assuré, il doit exister un accord officiel entre le Gouvernement du Canada et le gouvernement du deuxième pays. Cet accord peut prendre la forme d'un échange provisoire de notes diplomatiques ou d'une entente globale sur le transport aérien conclue par voie de négociations.

Des négociations bilatérales sur les transports aériens nécessitent la participation de fonctionnaires des ministères canadiens des Affaires extérieures, des Transports et de l'Industrie et du Commerce, ainsi que de la Commission canadienne des transports. Les pourparlers peuvent durer quelques mois ou quelques années.

Les routes aériennes pour les services à horaire fixe sont normalement l'objet principal des négociations, mais non le seul. Les articles contenus dans une entente sur le transport aérien peuvent concerner:

- le droit de survoler un certain territoire, ou d'y atterrir pour des raisons non liées au service normal;
- la désignation du transporteur qui peut desservir chaque route;
- le respect des lois et des règlements de chaque pays en matière d'entrée, d'autorisation, d'immigration, de passeports, de douanes et de quarantaine;
- la navigabilité, les certificats de compétence et les permis;
- l'échange de statistiques;
- les tarifs;

estimations du nombre de passagers sur les vols internationaux à horaire fixe, entre le Canada et les marchés étrangers, en fonction de diverses paires d'origines et de destinations. La première source, l'enquête sur l'origine et la destination des passagers payants, offre un échantillon de données sur les voyages internationaux où au moins un parcours d'un itinéraire est assuré par un transporteur canadien. Cette enquête, toutefois, présente des lacunes au niveau du champ d'observation, car elle ne fournit pas de données sur les voyages internationaux qui comportent les services d'un transporteur étranger pour l'ensemble de l'itinéraire. La deuxième source, l'enquête sur l'activité aéroportuaire, dénombre tous les passagers qui sortent du Canada ou qui y entrent d'un autre pays à bord d'un transporteur canadien ou étranger qui offre des services à horaire fixe, sans tenir compte de l'origine ou de la destination des passagers.

Ce document présente d'abord un aperçu des besoins des utilisateurs en matière de données sur l'origine et la destination des passagers sur les vols internationaux. On examine ensuite les principaux aspects des deux sources de données sur le trafic aérien, ainsi que le problème de la non-exhaustivité du dénombrement. Enfin, ce texte explique de quelle façon le système d'estimation de l'ASIPOD permet de calculer des estimations du nombre de passagers, et les coefficients de variation correspondants, pour diverses paires d'origines et de destinations, que ces villes soient situées ou non dans les parties du marché international visées par l'enquête sur les passagers payants.

2. BESOIN DES UTILISATEURS

Des estimations par origine et destination des passagers des vols commerciaux internationaux à horaire fixe sont nécessaires aux utilisateurs pour les négociations bilatérales.

Un service aérien commercial international à horaire fixe est défini comme l'exploitation de vols entre des points situés au Canada et des points à l'extérieur pour le transport de passagers, de biens ou de courrier par aéronef selon un horaire fixe et moyennant un taux de transport unitaire. Ce

MÉTHODOLOGIE DU SYSTÈME CANADIEN D'ESTIMATION DE L'ORIGINE ET DE LA DESTINATION DES PASSAGERS DES VOLS INTERNATIONAUX À HORAIRE FIXE¹

Greg Hunter et Lisa DiPiéto²

Le système d'estimation de l'origine et de la destination des passagers des vols internationaux à horaire fixe, communément appelé L'ASIPOD (Air Scheduled International Passenger Origin and Destination) est fondé sur les données de deux enquêtes sur le trafic aérien et permet de produire des estimations du nombre de passagers à bord des vols internationaux en fonction de l'origine et de la destination. La technique d'attribution offre une solution au problème du non-dénombrement du trafic sans correspondance. Les hypothèses sur lesquelles reposent cette technique soulèvent assez de doutes pour qu'une évaluation du biais des estimations mérite d'être entreprise. Cependant, le nouveau système comportera des améliorations importantes qui diminuent le biais des estimations, tout en calculant un estimateur de la fiabilité des résultats. Par conséquent, les négociations bilatérales sur le transport aérien international nous permettront de mieux déterminer la valeur des conclusions tirées au sujet des marchés des transports aériens à partir de ces estimations.

1. INTRODUCTION

En 1979, Statistique Canada a entamé une révision du programme fédéral de la statistique de l'aviation en invitant le ministère des Transports et la Commission canadienne des transports (les deux "organismes utilisateurs") à former une équipe interministérielle. Le remaniement du système d'estimation de l'ASIPOD est un des divers projets de révision du programme.

Selon le système d'estimation de l'origine et de la destination des passagers des vols internationaux à horaire fixe, communément appelé l'ASIPOD, on puise les données de deux enquêtes sur le trafic aérien pour produire des

¹ Présenté aux réunions conjointes de l'American Statistical Association à Cincinnati, août 1982.

² Greg Hunter, Division des méthodes d'enquêtes-entreprises, Statistique Canada. Lisa DiPiéto, Division des transports et des communications, Statistique Canada.

- [3] "Report of the Interdepartmental Working Group on For-hire Trucking - Phase II Review"; rapport rédigé par la Division des transports et des communications de Statistique Canada, avril 1979.

- [4] Statistique Canada, Entreprises de camionnage et de déménagement, no 53-222 au catalogue, annuel.

- [5] Lussier, R. (1981), "For-hire Trucking Survey: Survey Design", Techniques d'enquête, Statistique Canada, vol. 7, no 1, pp. 74-92.

faut procéder à des analyses pour évaluer les renseignements fournis. En ce qui a trait à l'année de référence 1981, seulement une bande sera utilisée; elle porte sur un transporteur qui a effectué environ cinq millions d'expéditions en 1981. Au cours des prochaines années, plus d'entreprises fourniront des données sur bande, et des accords sont sur le point d'être conclus avec cinq autres transporteurs pour l'année de référence 1982.

Toutefois, lorsqu'on obtient finalement une bande de données qui satisfait aux exigences de Statistique Canada, on doit tout de même faire beaucoup de traitement informatique pour exploiter la bande. En outre, des interventions manuelles sont nécessaires pour régler le problème des données qui ne figurent pas sur les diverses listes. Il faudra donc probablement échantillonner les dossiers contenus sur chaque bande selon le plan de sondage du deuxième degré appliqué aux documents d'expédition des entreprises des classes 1 et 2.

Une autre possibilité qui devrait être explorée serait de demander aux entreprises elles-mêmes d'échantillonner leurs documents. Par exemple, au moment où les bordereaux d'envoi sont établis, un transporteur pourrait photocopier ceux dont le numéro de série se termine par un certain chiffre et faire parvenir ces photocopies à Statistique Canada tous les mois.

Enfin, des travaux seront consacrés à une évaluation complète des diverses parties de l'enquête et à la formulation de propositions pour l'améliorer. On espère pouvoir appliquer ces propositions à l'enquête portant sur l'année de référence 1982.

BIBLIOGRAPHIE

- [1] Statistique Canada, L'enquête sur le transport routier de marchandises pour compte d'autrui, no 53-224 au catalogue, annuel.

- [2] "Report on the Findings and Recommendations of the Working Group on the For-hire Trucking Survey Phase I Review"; rapport rédigé par la Division des transports et des communications de Statistique Canada, en date du 7 juillet 1978.

Une utilisation particulière de l'enquête a été de définir les caractéristiques des marchés du camionnage à l'aide de variables telles que la marchandise transportée, la longueur moyenne des parcours et le poids des expéditions. Une autre étude a été réalisée sur divers aspects des activités des transporteurs dans des domaines réglementés et non réglementés. Elle contient une analyse du comportement des transporteurs face aux coûts à partir de caractéristiques du trafic, comme la taille des expéditions et la longueur moyenne des parcours.

Dans le passé, les demandes spéciales visant des estimations provenaient de ministères intéressés par le commerce, d'agences fédérales et provinciales de réglementation des transports, de transporteurs, d'experts-conseils d'universités, d'associations industrielles et d'un bon nombre d'autres organismes et particuliers qui s'intéressent au secteur des transports.

14.2 Méthodes de diffusion

L'enquête remaniée diffusera des renseignements de trois façons, comme le faisait l'ancienne enquête.

Premièrement, la publication relative à l'enquête présentera les estimations produites par le système régulier, de même que des mesures d'erreur comme les coefficients de variation. Deuxièmement, on répondra aux demandes individuelles, mais sous réserve des contraintes de coût et de fiabilité des renseignements. Enfin, la base des données sur les expéditions qui est constituée pour cette enquête pourrait être mise à la disposition de certains utilisateurs sur bandes magnétiques, à condition que les normes de confidentialité soient respectées.

15. PROJETS

Comme il est mentionné dans une section antérieure, l'enquête exploite trois sources de données, dont les bandes magnétiques fournies par certains répondants. Ce genre de données s'est avéré difficile à traiter et, malgré les travaux déjà commencés sur ce problème, les résultats sont décevants jusqu'à présent. Il faut entreprendre des négociations complexes avec les entreprises afin d'obtenir les données nécessaires sur bande et ensuite il

L'ensemble des données sur les entreprises des classes 1 et 2 est épuré par l'élimination des expéditions hors du champ de l'enquête. Ces dernières incluent les expéditions à destination ou en provenance des Etats-Unis, les expéditions qui totalisent 15 milles ou moins de l'origine à la destination, les expéditions faites hors des routes publiques, les expéditions qui seraient comptées deux fois parce qu'elles sont des expéditions de transfert ou parce qu'elles ont été déclarées par des déménageurs qui sont des agents de transporteurs sur longue distance et aussi par les transporteurs sur longue distance eux-mêmes, les expéditions qui n'ont pas produit de recettes au titre du transport interurbain, et les enregistrements concernant les services autres que le transport, comme l'entreposage, l'emballage, la location de matériel, le chargement et le déchargement.

Les estimations des recettes et du nombre de tonnes et de tonnes-kilomètres, qui doivent être publiées, sont enfin produites par la sommation des données pondérées dans les domaines d'intérêt correspondants. Des mesures d'erreur, comme les coefficients de variation, sont fournies avec les estimations. Les coefficients de variation sont calculés à l'aide de la formule issue du plan de sondage, mais on fait aussi l'hypothèse que l'échantillon systématique d'expéditions est un échantillon aléatoire simple.

14. UTILISATION DES DONNÉES ET MÉTHODES DE DIFFUSION

14.1 Utilisation des données

Les demandes d'estimations faites à partir de l'ancienne enquête émanent d'un grand nombre de sources. De plus, la nature de ces demandes est très variée. Il est prévu que la demande d'estimations faites à partir de la nouvelle enquête sera semblable à ce qu'elle était dans le passé.

Ces estimations sont devenues des instruments très répandus pour satisfaire cinq besoins principaux, à savoir mesurer le volume des marchandises du commerce intérieur transportées dans les provinces et entre celles-ci par les transporteurs interurbains pour compte d'autrui, déterminer le taux de croissance du transport interurbain de marchandises, fournir des données sur l'expansion régionale, aider à la réalisation d'études sur le transport et étayer la présentation de mémoires, de communications et de demandes aux organes et commissions de réglementation.

transcrits et à comparer ces chiffres avec le total du nombre d'expéditions, des recettes et du nombre de tonnes que le représentant du LDD a déclaré au cours de l'interview. Des vérifications semblables sont faites pour les entreprises de la classe 3. Dans tous les cas où on trouve un écart, on effectue un suivi.

13. TECHNIQUES D'ESTIMATION

Pour l'estimation, on a décidé de considérer l'échantillonnage systématique au deuxième degré dans les entreprises des classes 1 et 2 comme un échantillonnage aléatoire simple sans remise (EASSR). On a pris cette décision parce que, en premier lieu, on a considéré les documents comme étant en ordre aléatoire et, en deuxième lieu, un EASSR permet de calculer une estimation de la variance échantillonnale.

La première étape de l'estimation consiste à calculer des coefficients de pondération. Il y a des coefficients de premier et deuxième degré pour les expéditions des entreprises des classes 1 et 2, mais seulement des coefficients de premier degré en ce qui concerne les entreprises de la classe 3. Généralement, un coefficient de premier degré correspond à l'inverse de la probabilité qu'un LDD soit sélectionné dans sa strate, alors qu'un coefficient de deuxième degré représente l'inverse de la probabilité qu'une expédition d'un LDD soit choisie supposant qu'on effectue un EASSR. L'ordinateur corrige les coefficients de premier degré en fonction du nombre d'entreprises pour laquelle la collecte des données a échoué, mais aucune correction n'est faite pour les fermetures de LDD ou d'entreprises, ni pour les sorties d'entreprises du champ de l'enquête, étant donné qu'on les considère comme n'ayant fait aucune expédition. Un coefficient de pondération final est inscrit sur chaque enregistrement dans l'ensemble de données des entreprises des classes 1 et 2 et dans l'ensemble de données des entreprises de la classe 3.

Des rapports détaillés sont produits sous la forme de tables contenant diverses totalisations des données. Ces chiffres sont très utiles à l'analyse des données et aux dernières vérifications de la qualité des données.

Lorsqu'une paire ne figure pas sur la liste, on calcule une distance aérienne (X) à l'aide de la latitude et de la longitude des points d'origine et de destination. Ensuite, X est convertie en distance routière Y au moyen du modèle de régression linéaire simple

$$Y = a X + b$$

où a et b varient en fonction de 12 régions d'origine et de 12 régions de destination. La distance routière est ajoutée aux autres données.

Il y a aussi imputation des renseignements qui manquent concernant les expéditions partiellement transcrites des entreprises des classes 1 et 2. La technique d'imputation qu'on applique varie d'après la variable ou la paire de variables manquantes. Les principales imputations se font à l'aide de relations fixes entre les chiffres transcrits, de facteurs de conversion des unités de poids et de tables de distribution au prorata. Voici un exemple d'une relation fixe entre les renseignements déclarés:

$$\text{poids} = \frac{\text{recettes} \times 100}{\text{tarif}}$$

Cette formule peut servir à imputer le poids lorsque les recettes et le tarif sont connus ou à imputer les recettes quand le poids et le tarif figurent parmi les données. Les facteurs de conversion des unités de poids sont des coefficients déterminés par type d'unité (par exemple, caisse, sac, litre, etc.) et par code de la CTP. Sachant l'unité et le code de la CTP, on peut appliquer le bon facteur de conversion au nombre d'unités pour calculer le poids. Enfin, les tables de distribution au prorata indiquent le tarif par catégorie de marchandise, par bloc numérique de distance parcourue et par groupe de recettes ou de poids. Ces tables sont fondées sur les données de l'année précédente, mais sont mises à jour à mesure qu'on traite les données valides recueillies pour la période courante. Au moyen de ces tables, on calcule le poids lorsque les recettes sont connues, ou les recettes quand le poids est connu.

Dans les cas où trop de caractéristiques doivent être imputées, l'expédition est classée comme inutilisable.

Par la suite, des vérifications d'extrapolation sont effectuées. Dans le cas des entreprises de classe 1 ou 2, ce type de vérification consiste à pondérer approximativement le nombre d'expéditions, les recettes et le nombre de tonnes

Les autres traitements incluent la conversion du poids en unités métriques et du tarif en \$/100 kilogrammes. On procède également à une comparaison des noms des origines et des destinations (c'est-à-dire les villages, les villes, etc.) avec une liste de municipalités dans le but d'obtenir un code de la Classification géographique type (CGT), ainsi que la latitude et la longitude. Lorsque le nom de l'origine ou de la destination ne figure pas sur cette liste, l'opérateur doit utiliser un synonyme. De même, le nom de chaque marchandise est assorti à une liste de marchandises pour trouver un code (3 chiffres) de la Classification type des produits (CTP). Si la marchandise ne figure pas sur la liste, l'opérateur se sert d'un synonyme ou inscrit le mot "unknown". Il y a donc toujours un code de la CTP inscrit pour chaque expédition. De plus, le mini-ordinateur produit le nombre requis d'enregistrements de transcription pour chaque type d'expédition à partir des données tirées des profils des LDD des classes 1 et 2.

Enfin, les données sont extraites du mini-ordinateur et deux ensembles de données sont créés: un ensemble pour les expéditions des entreprises des classes 1 et 2 et un autre pour les renseignements concernant les types d'expéditions des entreprises de la classe 3. La principale différence entre ces deux ensembles de données est que le premier porte sur des expéditions distinctes tandis que le deuxième contient des données globales. Il convient aussi de noter que le premier compte plus de variables (par exemple, le tarif, le lieu d'origine plutôt que la province d'origine, etc.) que le deuxième ensemble.

12.3 Contrôles du système principal et imputations

La distance routière entre l'origine et la destination de chaque expédition des entreprises des classes 1 et 2, qui est visée par l'enquête, doit être connue afin d'estimer le nombre de tonnes-kilomètres parcourus par les transporteurs des classes 1 et 2. On vérifie donc chaque paire d'origines et de destinations, codées selon la CGT, en fonction d'une liste de distances pour obtenir la distance routière en kilomètres entre les deux points.

référence. À titre d'exemples, une entreprise pour laquelle on constate que 100 % de ses recettes proviennent d'expéditions locales est considérée comme hors du champ de l'enquête, tandis qu'un transporteur qui possède un seul LDD et qui refuse de collaborer ou est touché par une grève est considéré comme un cas d'échec.

Deuxièmement, il y a une vérification des données recueillies sur les profils des LDD des classes 1 et 2 afin de déterminer le nombre d'expéditions qui auraient dû être transcrites dans chaque catégorie d'expédition déclarée, si tous les documents avaient été dans les dossiers. On obtient ce nombre en faisant des calculs avec le nombre total d'expéditions comprises dans un profil, et avec l'origine choisie au hasard et l'intervalle de sondage qui auraient été utilisés si les documents avaient été disponibles. Le résultat est ensuite codé pour que l'ordinateur puisse produire le nombre requis d'enregistrements de transcriptions dans chaque catégorie d'expédition, comme si on avait obtenu originalement toutes les transcriptions.

12.2 Saisie des données

Après le traitement manuel, les formules arrivent au stade de la saisie des données. Cette opération est faite au moyen d'un mini-ordinateur qui permet d'effectuer des vérifications et d'autres traitements en ligne.

Diverses vérifications sont faites sur le mini-ordinateur. Certaines produisent des messages d'erreur et requièrent des corrections, mais d'autres font apparaître des avertissements qui demandent de vérifier les données introduites en ordinateur et d'apporter des corrections seulement s'il y a lieu. Certaines vérifications concernent la validité de chaque réponse individuellement tandis que d'autres examinent les rapports entre les caractéristiques valides d'une même expédition. Les opérateurs du mini-ordinateur sont censés connaître l'enquête en détail de façon à pouvoir apporter des corrections en ligne. Des imputations manuelles sont effectuées au besoin, étant donné qu'il n'y a pas d'imputation automatique pour les LDD de la classe 3.

11.5 "Profils" des LDD des classes 1 et 2

Il arrive parfois qu'un LDD de la classe 1 ou 2 ne peut pas fournir de documents, ne conserve pas de documents utilisables pour l'échantillonnage ou ne peut pas fournir une partie de ses documents d'expédition et qu'il est impossible de représenter cette partie au moyen des documents disponibles. Ce dernier cas peut se produire, par exemple, lorsque les documents qui manquent correspondent à des contrats particuliers qui ont été enlevés des dossiers pour des besoins de vérification comptable. Dans ces cas, l'interviewer doit faire un "profil" des documents qui manquent, c'est-à-dire qu'il doit demander à un représentant d'un LDD de décrire les types d'expédition consignés sur les documents qui ne peuvent être consultés. Les profils ressemblent aux descriptions des types d'expédition établies pour les entreprises de la classe 3, sauf qu'il faut indiquer l'origine et la destination précises de chaque expédition (c'est-à-dire le village, la municipalité, la ville, etc.).

Dresser un profil peut prendre beaucoup de temps dans certains LDD parce que leurs activités peuvent être assez complexes. Il faut une bonne collaboration de la part des représentants des LDD.

12. TRAITEMENT DES DONNÉES

12.1 Traitement manuel

Les formules remplies sont envoyées au bureau central de Statistique Canada, à Ottawa, où elles sont enregistrées et où on vérifie les numéros d'identification. Deux petites tâches sont également accomplies avant l'étape suivante.

Premièrement, un bref examen est effectué afin de repérer et de coder les fermetures de LDD ou d'entreprises, la sortie d'une entreprise du champ de l'enquête et l'échec de la collecte des données auprès d'un transporteur. Une entreprise qui tombe hors du champ de l'enquête est un transporteur actif dont les recettes comprises dans le champ de l'enquête sont nulles pour l'année de référence. Une entreprise pour laquelle on dit que la collecte des données a échoué est un transporteur actif pour lequel aucune information n'a été recueillie bien que les interviewers savaient que ce transporteur avait des recettes non nulles comprises dans le champ de l'enquête pour l'année de

Une fois qu'une expédition est choisie, l'interviewer transcrit les caractéristiques requises. Ce travail est souvent difficile parce qu'il n'est pas toujours facile de comprendre les divers documents et les codes qui figurent sur certains. Ce problème s'applique surtout aux noms des marchandises, et l'interviewer doit éviter d'inscrire des marques de commerce, des noms propres et des noms qui ont plus d'une signification. Il faut souvent interpréter les renseignements consignés dans un document et transcrire les données sur des feuilles de codage sous une forme exploitable par l'ordinateur.

11.4 Description générale de la collecte de données dans les entreprises de classe 3

L'interviewer envoie une lettre de présentation à l'entreprise deux ou trois semaines avant de téléphoner. Par la suite, l'interviewer communique avec le représentant de l'entreprise qui est le mieux placé pour fournir les renseignements requis, ce qui peut nécessiter plusieurs appels téléphoniques. Cela fait, l'interviewer procède à une interview téléphonique.

Les questions posées pendant l'interview sont semblables à celles posées dans le cas des LDD des classes 1 et 2, mais il y a une différence importante: aucune question ne porte sur le genre de documents utilisés ou sur les méthodes de classement appliquées par l'entreprise. Lorsque cette première partie de l'interview est terminée, l'interviewer demande au répondant de décrire les types d'expédition dont s'occupe son entreprise. Pour chaque catégorie d'expédition, la description doit spécifier la province d'origine, la province de destination et le nom de la marchandise transportée. On demande ensuite au représentant de donner une estimation du nombre d'expéditions, du poids moyen et des recettes moyennes provenant de chaque catégorie d'expédition.

Les experts-économistes en statistiques du transport considèrent généralement que les activités de n'importe quelle entreprise de la classe 3 sont assez homogènes. Chaque transporteur a donc seulement un petit nombre de types d'expédition à déclarer. On croit aussi que la couverture obtenue par cette méthode est acceptable du point de vue des utilisateurs. Toutefois, aucun test n'a été appliqué à cette hypothèse concernant les transporteurs de la classe 3.

Les types de système de classement comprennent par ordre numérique absolu, par ordre numérique avec interruption, par ordre chronologique, par ordre alphabétique (par exemple suivant le nom du client), par terminus, par genre de marchandise, ou sans aucun ordre. Les documents peuvent même être recoupés, par exemple, suivant le numéro d'ordre et d'après le nom du client. Dans un même système de classement, les documents peuvent être conservés dans des tiroirs de classeurs, dans des relieurs ou des fichiers "Shannon", sur des rayons, dans des tiroirs ou même dans des livres.

Les données globales sur les activités du LDD couvrent plusieurs variables, comme le total des recettes provenant du transport, le nombre total de tonnes transportées, le nombre total d'expéditions, le pourcentage de ces trois derniers éléments qui correspond aux expéditions interurbaines et aux expéditions internationales, les types de marchandises transportées et le pourcentage que chacun de ces types représente dans le total des recettes au titre du transport.

Il arrive souvent que l'interviewer puisse choisir parmi quelques méthodes de classement pour trouver les documents qui fournissent les renseignements requis. L'interviewer évalue à quel point les divers systèmes de classement sont complets par rapport aux cinq principales caractéristiques et l'année de référence, et il choisit le système qui présente le moins de sous-dénombrement. Toutefois, si le sous-dénombrement (le cas échéant) a la même ampleur dans deux systèmes ou plus, l'interviewer prend celui qui comprend le plus petit nombre de documents hors du champ d'enquête ou celui qui permet d'enlever ces documents du fichier ou de ne pas les compter.

L'interviewer sélectionne ensuite l'échantillon d'expéditions de la façon suivante. A partir du nombre d'expéditions déclaré par le représentant du LDD, l'interviewer obtient d'une table un intervalle de sondage et une origine choisie au hasard. Dans certains cas, l'intervalle et l'origine peuvent avoir été déterminés préalablement par le bureau central de Statistique Canada.

Puis, lorsque les documents sont classés en ordre numérique, on additionne l'origine choisie au hasard ou l'intervalle de sondage aux numéros d'ordre des documents pour obtenir les expéditions faisant partie de l'échantillon. Autrement, il faut compter un nombre de documents égal à l'origine choisie au hasard ou à l'intervalle pour sélectionner les expéditions.

d'exercices présentés par le chargé d'enquête et un ou des spécialistes en méthodes d'enquête. Pendant la formation sur le terrain, des groupes de trois ou quatre personnes se rendent à un LDD afin d'appliquer les connaissances acquises en classe et afin d'en discuter.

11.2 Planification des travaux de collecte

Après leur formation, les chargés de projet des opérations régionales recrutent les interviewers et leurs donnent à leur tour une formation complète. Par la suite, les interviewers, suivant les conseils de leur chargé de projet, se fixent un horaire de travail et planifient leurs itinéraires de visites aux LDD des entreprises des classes 1 et 2. Les itinéraires sont élaborés de façon à éviter les déplacements inutiles et à assurer le meilleur rendement possible. Les interviewers envoient aux représentants des LDD une lettre de présentation qui décrit brièvement la nature de l'enquête. Plus tard, les interviewers communiquent par téléphone avec eux pour fixer un rendez-vous. La collecte des données a lieu entre mai et septembre pour l'enquête qui porte sur l'année civile précédente.

11.3 Description générale de la collecte des données dans les LDD des entreprises des classes 1 et 2

Au moment du rendez-vous, l'interviewer mène une interview avec les représentants des LDD, leur expliquant l'enquête et à quoi servent les données. L'interviewer évalue le temps nécessaire pour effectuer le travail et demande des renseignements au sujet de l'entreprise. Ces questions concernent principalement les modifications à apporter aux noms et aux adresses, les changements de propriétaire, le(s) type(s) de documents et le système de classement utilisés, et des données globales sur les activités du LDD pendant l'année de référence.

Les types de documents d'expédition les plus communs sont les bordereaux d'envoi, les connaissements, les manifestes, les feuilles de route et les factures. Une entreprise peut se servir de n'importe quelle combinaison de ces documents.

dangereuses, ou d'autres types d'entreprises pour lesquelles le chargé d'enquête pourrait vouloir élargir la base des données. Au cours des prochaines années, il sera peut-être aussi nécessaire de faire un ajustement pour les entreprises qui exercent leurs activités dans un domaine où la fiabilité des estimations obtenues l'année précédente atteignait plus ou moins le niveau souhaité.

Pour les LDD des entreprises de la classe 3, il n'y a pas de deuxième degré d'échantillonnage. On n'échantillonne pas des expéditions distinctes dans les dossiers des LDD, mais on recueille des données globales au niveau des LDD.

11. OPÉRATIONS SUR LE TERRAIN

Les opérations sur le terrain sont différentes pour les entreprises des classes 1 et 2 et celles de la classe 3. Dans le cas des deux premières classes, le travail consiste à tirer des documents d'expédition des dossiers des LDD et à transcrire les caractéristiques de ces expéditions sur des feuilles de codage. Pour les entreprises de la classe 3, on recueille par téléphone des renseignements globaux sur les activités du transporteur.

La présente section porte sur les tâches du personnel des opérations régionales de Statistique Canada, notamment la formation des chargés de projet pour les opérations régionales, la planification des travaux de collecte, la collecte de données aux LDD des entreprises des classes 1 et 2, la collecte de données auprès des transporteurs de la classe 3 et, enfin, les profils des LDD des classes 1 et 2.

11.1 Formation des chargés de projet pour les opérations régionales

Toutes les ans, les chargés de projet des opérations régionales de Statistique Canada reçoivent une formation concernant tous les aspects de l'enquête. Les séances de formation durent quatre jours et ont lieu au mois de mars. Ce programme est divisé en deux parties: une formation théorique et une formation sur le terrain. La formation théorique comprend une série d'exposés et

9. PLAN DE SONDAGE DU PREMIER DEGRÉ D'ÉCHANTILLONNAGE

Le premier degré d'échantillonnage consiste à sélectionner dans chaque strate un nombre d'entreprises qui correspond à n_h , le nombre déterminé lors de la répartition de l'échantillon.

Toutes les entreprises dont les recettes de transport étaient d'au moins \$2,700,000 ont été désignées unités take-all c'est-à-dire qu'elles ont été sélectionnées avec une probabilité de un dans l'enquête de 1981 sur le transport routier de marchandises pour compte d'autrui. La raison de cette décision est qu'on sait que cette classe d'entreprises est hétérogène en ce qui concerne les principales variables à estimer et que ces transporteurs représentent une grande proportion des recettes qu'il faut estimer.

Enfin, l'échantillon d'entreprises est converti en un échantillon de LDD par l'inclusion dans ce dernier de tous les LDD des transporteurs choisis.

10. PLAN DE SONDAGE DU DEUXIÈME DEGRÉ D'ÉCHANTILLONNAGE

La sélection d'un échantillon systématique d'expéditions parmi les dossiers de chaque LDD choisi constitue le deuxième degré du plan d'échantillonnage des LDD des entreprises des classes 1 et 2. Le choix des documents est effectué dans les LDD par les interviewers de la Division des opérations régionales de Statistique Canada. L'intervalle de sondage varie en fonction du nombre d'expéditions faites par une entreprise et, généralement, les interviewers le trouvent dans une table de distribution qui leur est fournie. Cette table indique des intervalles de nombres possibles de documents d'expédition d'une compagnie et l'intervalle de sondage pour chaque cas. Cet intervalle peut cependant être établi à l'avance, pour n'importe quelle entreprise, par un membre du personnel du bureau central de Statistique Canada. Ce cas peut se produire lorsqu'un transporteur a de nombreux LDD, parce qu'un interviewer dans un LDD donné pourrait ignorer le nombre d'expéditions faites par l'ensemble de l'entreprise. Un autre exemple serait les entreprises qui présentent certaines particularités, comme les transporteurs de marchandises

On fait ensuite l'addition de la taille initiale corrigée des échantillons de toutes les strates pour obtenir un total corrigé de la taille initiale de l'échantillon global.

A l'étape suivante, la taille des échantillons est encore modifiée pour que, dans chaque strate, la taille soit égale ou supérieure à celle qu'on aurait obtenue si on avait distribué le total corrigé de la taille initiale de l'échantillon global d'une classe d'entreprises selon le rapport entre la racine carrée du nombre d'entreprises dans chaque strate et le total des racines carrées du nombre d'entreprises dans toutes les strates à l'intérieur de cette classe. Bref:

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

où la somme est effectuée sur toutes les strates de la même classe que la strate h.

Enfin, le chargé d'enquête peut subjectivement ajuster la taille des échantillons à n_h .

On a retenu cette méthode de répartition de l'échantillon parce qu'il s'agit d'un algorithme qui a donné de bons résultats lors des essais et qui est fondé sur la seule variable qu'on peut mesurer pour toutes les entreprises, c'est-à-dire les recettes faisant partie du champ de l'enquête. Toutefois, il faut rappeler que ces recettes ne sont pas recueillies directement dans l'enquête sur le transport routier de marchandises pour compte d'autrui, mais qu'on transcrit les recettes d'un échantillon d'expéditions. La méthode décrite ci-dessus fait donc abstraction du deuxième degré d'échantillonnage.

D'abord, un programme informatique établit le nombre initial d'entreprises qui doivent être choisies dans chaque strate afin d'atteindre un coefficient de variation cible pour l'estimation des recettes qui font partie du champ de l'enquête à l'intérieur d'une strate. Ce coefficient de variation cible est celui qu'on voudrait obtenir si l'estimation était calculée à partir des recettes totales déclarées par un échantillon d'entreprises choisies, par échantillonnage aléatoire simple, dans une population d'entreprises dont la distribution des recettes comprises dans le champ de l'enquête était la même que celle observée l'année précédente dans l'enquête sur les entreprises de camionnage et de déménagement. La formule est la suivante:

$$n_1^h = \frac{N_h^2 S_h^2}{N_h S_h^2 + Y_h^2 (C.V._h)^2}$$

- où n_1^h : le nombre initial d'entreprises qui doivent être choisies parmi les unités non take-all dans la strate h;
- N_h : le nombre d'unités non take-all dans la strate h;
- Y_h : le total des recettes incluses dans le champ de l'enquête des unités non take-all dans la strate h;
- S_h^2 : la variance des recettes incluses dans le champ de l'enquête des unités non take-all dans la strate h;
- $C.V._h$: le coefficient de variation cible dans la strate h (la valeur utilisée est la même pour toutes les strates d'une classe donnée mais elle peut varier d'une classe à l'autre).

Deuxièmement, la taille initiale des échantillons est corrigée afin d'assurer qu'un nombre minimal d'entreprises soit tiré de chaque strate, c'est-à-dire:

$$n_2^h = \min \{ \max (m, n_1^h), N_h \}$$

- où n_2^h : le nombre initial corrigé des entreprises qui doivent être choisies parmi les unités non take-all dans la strate h;
- m : le nombre minimal d'entreprises à choisir dans la strate h, dans la mesure du possible⁵.

⁵ Pour l'année de référence 1981, ce nombre minimal a été fixé à 3 dans toutes les strates.

liquide, de déchargements, de produits forestiers, de matériaux de construction, de vrac sec et (ou) de liquides réfrigérés, de machines lourdes, de solides réfrigérés, de produits explosifs et (ou) d'autres marchandises dangereuses, de produits agricoles ou d'animaux. Les entreprises spécialisées dans les petites expéditions de marchandises générales sont les transporteurs de marchandises générales dont les recettes moyennes sont de moins de \$85.00 par expédition; les autres transporteurs de marchandises générales sont considérés comme ceux qui font de grandes expéditions. La région desservie indique si une entreprise assure des services de transport dans une province en particulier, dans les Territoires du Yukon ou du Nord-ouest ou dans une combinaison spécifique de ces endroits. Par exemple, la région desservie pourrait être le Nouveau-Brunswick, ce qui signifierait qu'une entreprise exploite des services seulement dans cette province. Un autre exemple serait "l'Atlantique", ce qui voudrait dire qu'un transporteur dessert au moins deux des quatre provinces atlantiques (Terre-Neuve, l'Île-du-Prince-Édouard, la Nouvelle-Écosse et le Nouveau-Brunswick) mais aucune autre province au Canada. Vingt zones d'exploitation ont été délimitées pour l'enquête.

Les valeurs limites fixées pour la stratification selon les recettes et le type d'activité (c'est-à-dire \$85, \$350,000 et \$2,700,000) sont souples et pourraient varier au cours des prochaines années en fonction des changements qui surviendront dans la population de l'enquête.

La stratification décrite ci-dessus crée 840 strates, dont 355 n'étaient pas vides dans l'enquête sur le transport routier de marchandises pour compte d'autrui de 1981.

Une fois que la base de l'enquête est stratifiée, des agents spécialisés peuvent désigner des entreprises comme unités take-all, c'est-à-dire des entreprises qu'ils souhaitent inclure dans l'échantillon avec une probabilité de un. Par la suite, un spécialiste en méthodes d'enquête détermine le nombre d'entreprises à échantillonner parmi les entreprises non take-all dans une strate. Pour obtenir ce nombre, il faut effectuer quelques calculs dont sont exclues les unités take-all.

Comme la collecte de données coûte très cher à cause des déplacements vers les régions éloignées, on a essayé de restreindre le nombre de LDD choisis pour l'échantillon dans la catégorie des transporteurs dont les recettes annuelles dépassent \$350,000. La limite a été fixée à 875 LDD par année, soit le nombre utilisé au cours des dix dernières années de l'ancienne enquête.

7.2 Nombre total maximum d'expéditions transcrites

La deuxième contrainte d'ordre administratif concerne le nombre total d'expéditions transcrites. Le budget actuel permet de relever au plus 418,000 expéditions, mais ce nombre peut varier d'une année à l'autre selon les résultats des négociations qui ont lieu entre Statistique Canada et les utilisateurs qui participent aussi au financement de l'enquête.

7.3 Nombre maximal d'expéditions transcrites par entreprise

Une autre contrainte administrative a été imposée concernant le nombre maximal d'expéditions transcrites pour chaque entreprise. Une limite implicite a été fixée relativement au nombre de jours qu'une équipe de collecte des données peut passer dans un endroit donné, afin que la présence de ces personnes n'in-dispose pas les répondants.

8. STRATIFICATION ET RÉPARTITION DE L'ÉCHANTILLON

A partir des résultats de l'enquête de l'année précédente sur les entreprises de camionnage et de déménagement, les entreprises sont stratifiées en fonction de leurs recettes faisant partie du champ de l'enquête, par type d'activité et par région desservie. Ces variables ont été choisies parce qu'elles illustrent bien le caractère hétérogène de l'industrie. Les recettes faisant partie du champ de l'enquête indiquent si l'entreprise appartient à la classe 1, 2 ou 3, c'est-à-dire si elle a tiré respectivement des recettes de \$2,700,000 ou plus, entre \$350,000 et \$2,700,000 ou entre \$100,000 et \$349,999 au titre du transport interurbain de marchandises au Canada, à l'exclusion des services de déménagement et de transport par véhicule blindé. Le genre d'activité révèle si une entreprise est spécialisée dans les grandes expéditions de marchandises générales, les petites expéditions de marchandises générales, le déménagement sur grande distance, le transport d'automobiles, de pétrole

lieu lorsqu'une expédition est assurée par un transporteur jusqu'à un point intermédiaire, d'où un autre transporteur continue l'expédition vers un autre point. Les données sur les expéditions de transfert servent à supprimer les doubles comptes d'une expédition.

Les renseignements secondaires qui doivent être consignés sont la date de l'expédition, la quantité de la marchandise et l'unité de mesure (par exemple, 5 pieds-planche, 20 gallons, 15 sacs), des données sur le poids transcrit pour l'expédition (par exemple, le poids minimum, le poids commode utilisé pour le calcul des recettes), le tarif exigé et le code des taux (par exemple, un code qui indique si le tarif est minimum, ou exprimé par 100 livres ou par heure) et le code des recettes (c'est-à-dire un code qui indique s'il est impossible d'obtenir les recettes exactes de transport ou si une expédition est exclue du champ de l'enquête).

Les données globales recueillies auprès des petits transporteurs décrivent l'expédition moyenne ou typique en fonction de la province d'origine, de la province de destination, de la marchandise, des recettes moyennes, du poids moyen et du nombre d'expéditions.

7. CONTRAINTES ADMINISTRATIVES

La quantité de ressources affectées à la collecte et au traitement des données, de même que l'objectif visant à faciliter la tâche imposée aux répondants ont limité le nombre d'entreprises choisies ainsi que le nombre d'expéditions échantillonnées et de renseignements transcrits.

7.1 Nombre maximal d'entreprises dans l'échantillon

La population de l'enquête sur le transport routier de marchandises pour compte d'autrui de 1981⁴ est composé de 2,711 entreprises, dont 1,288 ont des recettes annuelles de plus de \$350,000 et 1,423 font entre \$100,000 et \$350,000 annuellement.

⁴ L'enquête de 1981 sur le transport routier de marchandises pour compte d'autrui correspond à l'enquête menée en 1982 pour l'année de référence 1981.

Pour cette raison, la base de l'enquête est composée d'une liste de toutes les entreprises dont les recettes annuelles de transport interurbain intérieur dépassent \$100,000. Les entreprises peuvent être aussi divisées en LDD lorsque les documents d'expédition ne sont pas entreposés en un lieu central. La base de l'enquête est établie à partir du recensement annuel des transporteurs pour compte d'autrui effectué par Statistique Canada par le biais de l'enquête sur les entreprises de camionnage et de déménagement³.

5. UNITÉ D'ÉCHANTILLONNAGE FINALE

L'enquête exploite les données de trois sources, c'est-à-dire celles contenues sur les bandes magnétiques de certains transporteurs, les renseignements transcrits à partir des documents d'expédition échantillonnés et, finalement, des données globales provenant des transporteurs qui touchent des recettes annuelles de \$100,000 à \$350,000.

Les bandes renferment des données sur les expéditions, dont les caractéristiques sont les mêmes que celles sur les expéditions qui sont échantillonnées et transcrites manuellement. Ainsi, l'unité d'échantillonnage finale est l'expédition, que ce soit pour les entreprises qui fournissent leurs bandes magnétiques ou pour les transporteurs dont les documents d'expédition sont échantillonnés. Dans le cas des entrepreneurs en camionnage qui touchent des recettes entre \$100,000 et \$350,000, on recueille des données globales parce que, habituellement, ces transporteurs ne conservent pas les documents nécessaires sur les expéditions. Pour cette catégorie de transporteur, l'unité d'échantillonnage finale est l'entreprise.

6. INFORMATION RECUEILLIE

Les principales caractéristiques qu'il faut noter pour chaque expédition échantillonnée d'un transporteur déclarant des recettes annuelles de plus de \$350,000 au titre du transport interurbain intérieur sont l'origine réelle et la destination finale, la description de la ou des marchandises transportées, le poids et l'unité de poids, les recettes perçues au titre du transport et les renseignements concernant les expéditions de transfert. Un transfert a

³ L'enquête de Statistique Canada sur les entreprises de camionnage et de déménagement est un recensement annuel des entrepreneurs en transport routier. Elle vise à obtenir des données d'entrée et de sortie sur les établissements, comme les recettes, les dépenses, les renseignements relatifs aux bilans et le matériel exploité.

b) Limites du remaniement

Les principales contraintes imposées sur le remaniement ont été les suivantes: certains types d'entreprises de camionnage pour compte d'autrui devaient être exclus de la population de l'enquête, notamment les déménageurs pour compte propre et les transporteurs de pétrole; la stratification devait être améliorée afin de respecter la structure économique de cette industrie; trois types de données devaient être utilisés à savoir les bandes de certains répondants, les transcriptions provenant d'échantillons de documents d'expédition des LDD d'un échantillon d'entreprises déclarant des recettes annuelles de plus de \$350,000 au titre du transport interurbain intérieur, et les données globales d'un échantillon d'entreprises qui ont touché des recettes annuelles se situant entre \$100,000 et \$350,000 au titre du transport interurbain intérieur; enfin, l'enquête remaniée devait pouvoir être appliquée à l'année de référence 1981, la collecte de données commençant au printemps de 1982.

4. POPULATION ET BASE DE L'ENQUÊTE

La population de l'enquête regroupe toutes les expéditions faites pendant l'année de référence par les entreprises de transport qui satisfont à la définition établie pour le champ de l'enquête. Une expédition est définie comme une quantité de marchandises acheminées par un transporteur et envoyées par une personne ou par un organisme à une autre personne ou à un autre organisme. Les entreprises échantillonnées sont celles dont les recettes annuelles provenant du transport de marchandises entre villes dépassent \$100,000, dont le camionnage constitue l'activité principale et qui sont situées au Canada. Les éléments exclus de cette population sont les expéditions faites par certains types de transporteurs spécialisés comme les transporteurs de pétrole et les déménageurs à compte propre.

Comme cette population idéale n'est pas accessible, on utilise plutôt les entreprises comme grappes naturelles d'expéditions pour l'échantillonnage de premier degré.

des transporteurs. On a donc ajouté une troisième phase à l'étude de l'enquête. Son objectif était d'effectuer les analyses nécessaires afin de formuler et de suggérer des directives générales pour une enquête révisée. Les analyses devaient se conformer aux propositions issues de la phase II.

En juin 1980, le groupe de recherche travaillant sur la troisième phase a proposé que l'enquête soit remaniée de façon à intégrer quatre types de données: les bandes de certains répondants; les transcriptions provenant des échantillons de documents d'expédition tirés de chaque lieu de dépôt des documents (LDD) déclarant des recettes annuelles supérieures à \$1,500,000 au titre du transport interurbain intérieur; les transcriptions provenant des échantillons de documents d'expédition tirés d'un échantillon de LDD ayant des recettes annuelles de \$350,00 à \$1,500,000 au titre du transport interurbain intérieur; et les données globales recueillies auprès d'un échantillon de LDD dont les recettes annuelles au même titre atteignaient entre \$100,000 et \$350,000. La décision de recueillir des données globales auprès des petits transporteurs a été fondée sur le fait que ces entreprises ne possèdent pas les documents nécessaires à l'échantillonnage.

3.2 Remaniement de l'enquête

a) Objectif du remaniement

Après la troisième phase de l'examen de l'enquête, on a décidé de procéder à un remaniement complet de tous les aspects de l'enquête dans le but de recueillir des renseignements plus fiables et plus détaillés sur l'origine et la destination des marchandises acheminées par les transporteurs canadiens pour compte d'autrui. On s'attend à une amélioration non seulement de la fiabilité des données en comparaison avec l'"ancienne enquête", mais aussi de la quantité de renseignements recueillis au niveau régional et au niveau des marchandises.

b) Étapes de l'examen

Initialement, l'examen de l'enquête a été décomposé en deux phases.

La première phase avait pour objet d'esquisser des propositions visant à améliorer l'enquête, à l'intérieur du cadre existant, au moyen de seulement quelques ressources supplémentaires. Comme prévu, les propositions ont porté sur une redéfinition de la population de l'enquête, l'amélioration des variables de stratification et l'élargissement de la taille de l'échantillon des documents d'expédition. Les propositions ont été présentées dans un rapport [2].

La deuxième phase avait pour but d'évaluer les propositions formulées, en fonction des besoins des utilisateurs, d'indiquer le coût et les méthodes d'application possibles de chacune des propositions retenues et de poursuivre des analyses de l'enquête. Cette phase a produit une refonte de certaines propositions de la première phase et a présenté quelques nouvelles propositions visant une population d'entreprises plus petite que celle de l'enquête existante et mieux stratifiée en groupes homogènes. En plus de suggérer l'application de ces propositions, on a étudié quatre façons d'agrandir la taille de l'échantillon, à savoir le statu quo, une augmentation de 50 % de la taille de l'échantillon des documents d'expédition, une augmentation de 100 % de cet échantillon et, finalement, une augmentation de 25 % de l'échantillon plus le traitement des bandes de données pour une quarantaine d'entreprises de transport. Après une évaluation des avantages et des coûts de chacune de ces possibilités, la quatrième a été approuvée en principe, parce qu'elle permettait une augmentation considérable de la taille de l'échantillon au coût le plus bas et avec le fardeau de réponse le plus faible. Les recommandations et les détails de l'argumentation figurent dans un rapport [3].

Une première évaluation de la portée des propositions a révélé qu'il fallait effectuer des travaux supplémentaires, surtout pour déterminer le coût total de l'utilisation des bandes contenant les bordereaux d'envoi

3. RÉVISION DE L'ENQUÊTE SUR LE TRANSPORT ROUTIER DE MARCHANDISES POUR COMPTE D'AUTRUI

La révision s'est faite en deux étapes principales. En premier lieu, on a procédé à une évaluation détaillée de l'enquête existante. En deuxième lieu, selon les recommandations issues de cet examen, une refonte complète de l'enquête a été effectuée.

3.1 Examen de l'enquête

a) Motifs de cet examen

Au début de 1978, Statistique Canada a amorcé un examen de l'enquête sur le transport routier de marchandises pour compte d'autrui pour les raisons suivantes. Premièrement, la Division des transports et des communications de Statistique Canada prévoit toujours une étude périodique des enquêtes en cours. L'enquête sur le transport routier de marchandises pour compte d'autrui n'avait pas été évaluée depuis 1973. Deuxièmement, il était impossible de satisfaire aux besoins courants et prévus de renseignements plus détaillés sur l'origine et la destination des marchandises à l'intérieur des contraintes imposées par l'enquête. Troisièmement, l'expérience acquise par le biais de l'enquête sur le transport routier de marchandises pour compte d'autrui et d'autres enquêtes connexes a fourni des renseignements supplémentaires permettant d'améliorer la base de sondage, les variables de stratification et les techniques d'imputation. Quatrièmement, certaines innovations dans l'industrie du camionnage, comme la production de données sur l'origine et la destination sous forme lisible par une machine, laissaient entrevoir la possibilité d'utiliser des bandes magnétiques pour élargir la base des données et, en même temps, réduire le fardeau des unités déclarantes.

En outre, le caractère de plus en plus complexe des besoins des utilisateurs ont fait sentir la nécessité de perfectionner les techniques de diffusion des données, tandis que des progrès en informatique non seulement ont rendu le système de traitement des données existant désuet du point de vue technique, mais lui ont fait perdre sa rentabilité.

b) L'enquête sur le transport routier de marchandises pour compte d'autrui (1969-1979)

C'est en 1969 qu'ont été entrepris les premiers travaux sur une enquête visant à mesurer en fonction de l'origine et de la destination les mouvements de biens entre villes canadiennes effectués par l'ensemble des industries canadiennes de camionnage pour compte d'autrui. A cette époque, une étude a été menée pour examiner diverses méthodes de recueillir les renseignements sur l'origine et la destination des marchandises. Les résultats ont démontré qu'une enquête-échantillon des dossiers administratifs des transporteurs, notamment leurs documents d'expédition, était une méthode convenable de rassembler les données nécessaires.

En 1970, on a procédé à une enquête pilote afin d'évaluer l'efficacité de la méthode de sondage. Pour cette expérience, il fallait examiner les documents d'expédition de 187 entreprises de camionnage pour compte d'autrui dans l'ensemble du pays. La réaction favorable à cette enquête pilote ainsi que la facilité de trouver des renseignements sur l'origine, la destination, les marchandises, le poids et les recettes ont confirmé que la méthode de sondage était applicable.

Ainsi, l'enquête sur le transport routier des marchandises pour compte d'autrui a été menée pour les années de référence 1970 et 1971 avec les objectifs énumérés ci-dessus. Quant à l'année de référence 1972, on a modifié les objectifs de façon à limiter l'enquête aux transporteurs pour compte d'autrui situés au Canada dont les recettes annuelles provenant du transport interurbain étaient d'au moins \$100,000. Pour l'année de référence 1973, on a utilisé une base de sondage mise à jour et mieux définie, qui comprenait les transporteurs routiers réglementés, et on a élaboré une technique d'échantillonnage plus efficace. Depuis l'année de référence 1973, l'enquête se poursuit et les résultats sont publiés annuellement par la Division des transports et des communications de Statistique Canada [1] [5].

La combinaison de tous ces facteurs influe sur le plan d'enquête, surtout pour ce qui a trait à la stratification.

2.2 Historique des enquêtes sur l'origine et la destination des marchandises transportées par des entreprises de camionnage au Canada

a) Motor Transport Traffic Survey (1957-1963)

La première tentative de mesurer le volume de la circulation des camions au Canada a commencé en 1957 avec l'instauration de la Motor Transport Traffic Survey (MTTS - enquête sur le transport routier), une enquête-échantillon sur les véhicules automobiles utilisés pour transporter des marchandises. La base de l'enquête était une liste de véhicules automobiles immatriculés dressée à partir des dossiers des gouvernements des provinces et des territoires. Cette base était stratifiée selon le type d'activité et le poids brut du véhicule.

La taille de l'échantillon représentait à peu près 10 % de tous les véhicules immatriculés. L'échantillon était prélevé en quatre segments trimestriels dont chacun représentait environ un quart de l'échantillon total. Pour chaque échantillon trimestriel, le travail de collecte des données était réparti en trois semaines d'enquête, un tiers de cet échantillon étant visé pendant une période de sept jours tous les mois.

Comme l'enquête portait sur les véhicules, aucun renseignement n'était recueilli sur l'origine et la destination des marchandises transportées. C'était une enquête sur l'origine et la destination des camions et les marchandises étaient secondaires. On recueillait aussi des données comme la description du véhicule, le nombre de milles parcourus, la quantité de carburant consommé et les coûts d'exploitation du véhicule.

Cette enquête a été menée de 1957 à 1963 inclusivement, puis abandonnée en 1964 par suite de changements dans le système d'immatriculation des véhicules et de la réorganisation de l'industrie du camionnage mais, surtout, parce qu'on avait observé une diminution très importante des taux de réponse.

détriment des autres modes de transport. A la fin des années 1970, Statistique Canada a entrepris, en collaboration avec les principaux utilisateurs, une révision complète de l'enquête.

Le présent document a deux objets: d'abord, présenter quelques renseignements généraux sur l'enquête et décrire les étapes du processus de révision et, deuxièmement, expliquer la méthodologie de l'enquête remaniée qui sera menée pour l'année de référence 1981. Il faut souligner que les détails de la méthodologie de certaines phases n'ont pas encore été mis au point de façon définitive; ce document en décrit toutefois la portée générale.

2. RENSEIGNEMENTS GÉNÉRAUX

2.1 Bref aperçu de l'enquête sur le camionnage pour compte d'autrui

L'industrie canadienne du camionnage pour compte d'autrui se caractérise par un très grand nombre de petits transporteurs, et la grande diversité des marchandises transportées, des tailles des entreprises et des zones d'exploitation témoigne de son caractère hétérogène.

Les petits transporteurs, c'est-à-dire ceux dont les recettes annuelles sont moins de \$100,000, représentent numériquement 88 % de l'industrie, mais seulement 20 % des recettes d'exploitation. Etant donné l'instabilité de ces transporteurs et leur poids relativement faible dans le total des recettes, il a été décidé de les exclure de la population de l'enquête.

Les entreprises de camionnage assurent le transport d'une grande variété de marchandises, qui exige divers types d'équipement et différents modes d'exploitation. Les transporteurs (déménageurs, transporteurs de marchandises générales, de pétrole en vrac, etc.) se distinguent non seulement par les biens qu'ils acheminent, mais également par le volume des expéditions.

On peut aussi constater le caractère hétérogène de l'industrie du camionnage par l'étendue des régions desservies par les entreprises. Certains transporteurs ne font que des expéditions locales, d'autres des expéditions interprovinciales, tandis que certaines grandes entreprises desservent chaque province ainsi que le marché international.

REMANIEMENT D'UNE ENQUÊTE SUR L'ORIGINE ET LA DESTINATION DES MOUVEMENTS DE MARCHANDISES EFFECTUÉS PAR LES TRANSPORTEURS POUR COMPTE D'AUTRUI AU CANADA¹

Robert Lussier et Steven Mozes²

Ce document présente d'abord quelques renseignements généraux sur l'enquête sur le transport routier de marchandises pour compte d'autrui et décrit les étapes du processus de révision qui a conduit à la décision de remanier l'enquête. Le document explique dans un second temps la portée générale de la méthodologie de l'enquête remaniée qui sera menée pour l'année de référence 1981.

1. INTRODUCTION

L'enquête sur le transport routier de marchandises pour compte d'autrui a été entreprise pour la première fois par Statistique Canada en 1971 afin d'obtenir des renseignements sur l'origine et la destination des mouvements de biens par les soins de transporteurs pour compte d'autrui. Aux fins de cette enquête, l'industrie du transport pour compte d'autrui est définie comme étant l'ensemble des entreprises de camionnage qui, moyennant rétribution, assument le transport de marchandises. L'enquête était fondée sur un échantillonnage probabiliste des expéditions enregistrées sur les documents conservés par les transporteurs canadiens pour compte d'autrui. Or, depuis 1971, la demande de renseignements plus fiables et plus détaillés n'a pas cessé d'augmenter. On peut attribuer cette progression de la demande à de nombreux facteurs, comme la très forte croissance du camionnage depuis le début des années 1950, le caractère de plus en plus complexe des besoins des utilisateurs des statistiques sur le transport, l'intérêt porté à la question de la réglementation de ce secteur économique ou de la levée de normes et, enfin, l'augmentation de la part du marché du transport des marchandises enregistrée par le camionnage au

¹ Ce document est une version révisée d'un exposé présenté sur demande aux Joint Statistical Meetings of the American Statistical Association, the Biometric Society, ENAR and WNAR, and the Institute of Mathematical Statistics, Cincinnati, 16 au 19 août 1982.

² Robert Lussier, division des méthodes d'enquêtes - entreprises, et Steven Mozes, division des transports et des communications, deux divisions de Statistique Canada.

TECHNIQUES D'ENQUÊTE

1983

Vol. 9

No. 1

Préparé par Statistique Canada

Comité de rédaction:	G.J.C. Hole	
	C. Patrick	
	R. Platek	- Président
	M.P. Singh	- Rédacteur en chef
	P.F. Timmons	
	H. Lee	- Rédacteur adjoint

Politique de la rédaction:

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

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages Statistique Canada, 4e étage, Édifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi.

TECHNIQUES D'ENQUÊTE

1983

Vol. 9

No. 1

Préparé par Statistique Canada.

TABLE DES MATIÈRES

Remaniement d'une enquête sur l'origine et la destination des mouvements de marchandises effectués par les transporteurs pour compte d'autrui au Canada ROBERT LUSSIER et STEVEN MOZES.....	1
Méthodologie du système canadien d'estimation de l'origine et de la destination des passagers des vols internationaux à horaire fixe GREG HUNTER et LISA DIPIÉTRO.....	29
Certains aspects de la qualité des statistiques sur la mortalité due au cancer et l'incidence de cette maladie D. BINDER et A. MALHOTRA.....	54
L'estimation des flux bruts mensuels de l'activité sur le marché du travail STEPHEN E. FIENBERG et ELIZABETH A. STASNY.....	85
Remaniement de l'enquête sur le rendement prévu en fruits tendres dans la péninsule de Niagara J. KOVAR.....	111
Une estimation précise et rapide de la superficie plantée en pommes de terre grâce à Landsat: Résultats d'une démonstration R.A. RYERSON, J.-L. TAMBAY, R.J. BROWN, L.A. MURPHY et B. MCLAUGHLIN.....	129
Échantillonnage à deux reprises avec pptr G.H. CHOUDHRY et JACK E. GRAHAM.....	153

8-3200-501

No. de références

Z - 079

ISSN: 0714-0045



Statistique Canada Statistics Canada

TECHNIQUES D'ENQUÊTE

1983

volume 9

numéro 1

**Préparé par
Statistique Canada**

Canada