Survey Methodology, December 1984
Vol. 10, No. 2, pp. 165–179
Statistics Canada

The Use of Matching in the Evaluation of Non-Sampling Errors in the 1981 Canadian Census of Agriculture

J. COULTER[1]

ABSTRACT

This paper discusses the use of matching between files of comparable data in the evaluation of non-sampling error. As an example of the technique, the data quality evaluation of the 1981 Canadian Census of Agriculture is described and some results presented.

KEY WORDS: Non-sampling error; Coverage; Response error; Matching; Record Linkage; Census of Agriculture.

1. INTRODUCTION

As the use of probability sampling in data collection has evolved, the evaluation and control of sampling errors has been a constant concern. Extensive research has been devoted to the design of sampling schemes which would reduce sampling error and facilitate its measurement. In many situations, however, major portions of the survey error arise not from sampling, but from the effects of other components of the data collection operation. In censuses particularly, in which data are obtained through 100 percent enumeration of the population of interest, sampling error is nonexistent. Instead survey error is due entirely to the influences of respondents, interviewers, coders, keyers, and others during the collection, capture, and processing stages of the survey operation. As the impact of these non-sampling errors on data quality has become more fully understood, the development of techniques to control and measure them has gained in importance.

2. MODELS FOR SURVEY ERROR

Early papers on total survey error, such as that by Deming (1944), outlined the potential sources of error and discussed the need to consider their varying effects when planning data collection operations. As the study of survey error developed, general models were propose by Hansen et al. (1951), Sukhatme and Seth (1952), Hansen, Hurwitz, and Bershad (1961), and others to describe the components of sampling and non-sampling error. Studies were conducted on the correlations between errors which result from influences such as interviewers or coders, and methods were developed for measuring their effects. Fellegi (1964) presented a detailed model which included correlations between numerous error sources.

Other models have followed which consider both single and correlated non-sampling errors and propose methods for evaluating them. Some examples include the U.S. Bureau of the Census survey error model described by Nisselson and Bailar (1976), the discussion of measure-

[1] J. Coulter, Census Operations Division, Statistics Canada, 2nd Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

ment errors by Cochran (1977), and the model of survey error presented by Andersen et al. (1979) which was based on an earlier model by Kish (1965). In a recent paper Hartley (1981) described a model with terms for interviewer, coder, and respondent errors, and proposed a sample design to facilitate estimation of these errors.

Throughout the literature the components of non-sampling error have been categorized in a variety of ways. In this paper we will divide non-sampling error into two elements, coverage error and response error. A coverage error will be defined to have occurred when a unit which satisfies the definition of the universe of interest is missed or counted more than once, or a unit not belonging to the desired universe is included. Coverage errors cause the invalid inclusion or exclusion of all data for the incorrectly enumerated unit. As a result they may influence the estimates for any or all data items.

Response errors will be defined as affecting the values of individual items within the data for units which have been correctly included in the enumeration. They may have arisen at the initial collection of the data or during subsequent processing stages; potential sources include misinterpretation of a question by the respondent, total or partial non-response, the influence of the interviewer, and data capture or coding error.

3. EVALUATING NON-SAMPLING ERROR

In most survey error models, error is defined as the difference between the true value and the collected value for the particular data item. Thus in order to evaluate non-sampling error, one would in theory propose to compare the data collected by the survey with the true values for the item of interest. In practice, however, the true values are seldom known, if in fact they even exist. Instead the survey data must be compared to estimates from an alternate source which is believed to provide the closest available approximations to the true values.

In determining the data source most suitable to represent the unknown true values, a number of factors must be considered. The alternate data must be collected independently from those of the survey of interest. Optimally, the definitions and concepts employed in the collection of the data, and the reference periods to which the data applied, would be equivalent for the two sources. The universe covered by the alternate data would be the same as that of the survey, or comparable subuniverses would be identifiable. The purpose and methods of collection of the alternate data, and of any subsequent processing or updating stages, would be fully understood. Perhaps most important of all, the data would be of high enough quality to act as a standard against which the survey data could be compared.

In practice, of course, all of these conditions are seldom satisfied by a single alternate data source. In some cases one source may be best for a particular subpopulation while another is preferable for the remainder of the universe. Adjustments may be possible to remove the effects of differences in reference dates or definitions of variables between the two data sources. For the majority of cases, however, even the best estimator of the true values will involve major failures of some of these conditions, and their influence on the comparison may not be measurable or even identifiable.

Approximations to the true values may be obtained from a variety of sources. Estimates from one census or survey may be compared to those from another. Examples include the comparisons of the Current Population Survey and the U.S. Census of Population, the Labour Force Survey and the Canadian Census of Population, and the Agriculture Enumerative Survey and the Canadian Census of Agriculture (Statistics Canada 1979). Demographic projections have also been employed to approximate the true values, such as in the evaluations of the Labour Force Survey and the Canadian Census of Population described by Fellegi (1973).

Administrative data are also being used more and more in evaluation studies as respondent burden beccomes an increasing concern. Estimates from income tax, family allowance, motor vehicle licence, agricultural marketing board, and other files may provide approximations to the required true values. Income tax files, for example, are currently being studied at Statistics Canada for use in the collection and evaluation of farm income data.

As these few examples imply, a wide range of alternate data sources have been used as standards of comparison for survey estimates. However, while macro-level comparisons provide indications of the total error in the survey results, they cannot identify or measure the components of the error. The effects of coverage and response errors cannot be identified. Hence additional methods are required which facilitate more detailed investigations into the total error observed.

One technique, which can provide the analyst with a wealth of information on errors and their sources, links the survey results at the individual record level to a file of comparable data. Using the matched records, one-to-one comparisons can be made between the values reported for the survey and those from the alternate data source, which are assumed to represent the true values. Cross-classifications over other data items provide insight into the characteristics of units displaying certain types of inconsistencies. As well, the study of records which could not be matched can indicate areas of potential over or undercoverage by the survey.

In this paper we will consider the use of matching in the evaluation of survey non-sampling error. The strengths and weaknesses of the technique, and the types of studies which it makes possible, will be discussed. As an example of the use of matching, the data quality evaluation of the 1981 Canadian Census of Agriculture will be outlined and some results presented.

4. THE USE OF MATCHING

Study of the literature on the evaluation of total survey error reveals many studies which have made use of matching. As with the macro-level comparisons, a wide variety of comparable data sources have been employed. Post enumeration surveys, aimed at collecting data of a higher quality and in more detail than the original survey, have been conducted in a number of situations. Examples include reinterview by appraisers in a study on the reporting of the market value of homes (Kish and Lansing 1954), the post enumeration survey used to evaluate the U.S. Census of Agriculture (U.S. Bureau of the Census 1982), the Labour Force Survey reinterview program (Tremblay, Singh, and Clavel 1976), and the Vacancy Check Operation of the Canadian Census of Population and Housing (Statistics Canada 1980).

Independently-collected censuses and surveys have also been linked in order to evaluate data quality. Matches have been performed, for example, between the Labour Force Survey and the Canadian Census of Population (Krotki 1980), the Current Population Survey and the U.S. Census of Population (U.S. Bureau of the Census 1964), and the Agriculture Enumerative Survey and the Canadian Census of Agriculture (Statistics Canada 1979).

With the concern for reducing respondent burden, administrative data have been used increasingly in evaluation studies. Records of doctors and hospitals have been matched to health survey results by Andersen et al. (1979) and Horvitz (1981). Immigration and birth records have been employed in the Reverse Record Check Content Study of the Canadian Census of Population (Krotki 1980), and tax records of the IRS have been used to study response errors in the U.S. Census of Population (U.S. Bureau of the Census 1970). Other examples of linkage with administrative data include the evaluation of agricultural survey results by Faulkenberry and Tortora (1981) and the reporting of sensitive topics by Marquis, Marquis, and Polich (1981).

In general, the quality of an evaluation study based on a record linkage operation depends on two major factors:

1) the quality of the data on the alternate source to which the survey is matched, and
2) the uniqueness of the identifiers used for the match, and the accuracy of the match technique itself.

Firstly, by definition of the study objectives, the alternate data are to act as approximations to the true values. Random errors in the data which tend to cancel one another out over all records do not noticeably affect macro-level comparisons. However such errors can have a serious effect on studies conducted at the individual record level.

Some data bases chosen to act as standards for comparison may be assumed to be free from error. Birth records, for example, provide very accurate data on place of birth and age. Other data sets may be known to contain certain response or coverage errors, but if the errors are measurable they can be taken into account in the analysis, and the assumption of no error remains valid. In many cases, however, the comparable data are subject to errors which cannot be completely identified or measured. In these situations, the best approximation to the true values is provided by the data set which is least affected by error. Data which have been collected using more accurate methods or better trained staff than the survey of interest may be assumed to contain less error, and hence can provide a reasonable basis for comparison.

The second major factor affecting the quality of the study is a function of the match operation itself. In some situations, each member of the population will have been assigned a unique identification number which has been accurately stored on both files. Linking the records would then be a straightforward process of matching on these unique identifiers. At the other end of the scale, the best available identifiers may be non-unique characteristics, such as name, which are prone to the introduction of error during data collection or capture.

The linkage algorithm itself can also have an impact on the quality of the match, particularly when the keys or identifiers are less than perfect. The algorithm may tend to allow invalid matches between records with similar keys, or may prevent valid matches when the keys differ due to minor errors or omissions. The extent to which such errors occur can have a significant effect on the composition of the files of matched and unmatched records.

Other factors which affect the comparability of the two sets of data, as for macro-level studies, include differences in collection date and method, concepts and definitions, and reference period. Due to the greater detail of investigation and cross-classification over related variables which is involved in micro-levels studies, such differences can have a much greater impact on the analysis than for the macro-level comparisons.

In order to study the use of matching in the analysis of non-sampling error, we will now consider the example of the data quality evaluation of the 1981 Canadian Census of Agriculture. For this study, independently-collected agricultural data for macro and micro-level comparisons were provided by the Agriculture Enumerative Survey (AES) and the Farm Enumerative Survey (FES), annual probability surveys conducted by Statistics Canada.

5. COMPARING THE CENSUS AND SURVEYS

The Canadian Census of Agriculture was conducted on June 3, 1981, sharing field operations with the quinquennial Census of Population and Housing. Data were to be collected for every census farm in Canada, defined as any farm, ranch or other agricultural operation which received $250 or more from the sale of agricultural products during the twelve months prior to census day, or which had the potential to produce that value in the next twelve months. During drop-off of the population and housing questionnaire, the census representative was to ask at each household whether any member operated a farm or other holding which satisfied the above definition. If so, a Census of Agriculture form was left to be completed by the operator.

In order to improve coverage, results of the 1976 Census and subsequent agricultural surveys were used to identify farms which were major producers of one or more specified agricultural commodities. The census representatives then had to account for each of these "specified farms" located in the area to be enumerated.

The questionnaire, delivered prior to June 3 to the operator of each census farm, was to be completed by self-enumeration on census day. Items covered in the census included crops, livestock, land use, sales, expenses, and other areas of interest to the public and private sectors. (Further details on the methodology and content of the 1981 Census of Agriculture may be obtained from the publication Statistics Canada (1982).)

The Agriculture Enumerative Survey (AES) and Farm Enumerative Survey (FES) together covered the majority of Canada's agricultural land. The FES enumerated the Prairie provinces of Manitoba, Saskatchewan, and Alberta plus the Peace River district of British Columbia, and the AES covered the remainder of British Columbia and the provinces of Prince Edward Island, Nova Scotia, New Brunswick, Quebec, and Ontario. The survey universe consisted of agricultural holdings which satisfied the census farm definition described above. However it excluded types of organization which were of marginal economic influence, such as institutional farms, and areas which contain little or no agricultural activity, such as urban cores. In order to provide comparable universes for the evaluation, the census file was adjusted by removing operations of these types. The deletions consisted of only 2.8 percent of the farms and 1.8 percent of the total farm area from the complete census file.

The probability surveys collected data on the same major agricultural variables as the census, such as crops, livestock, land use, and operating expenses, and used similar concepts and definitions. Some differences existed in wording and format, and in the instructions on what to include or exclude, for particular questions. As will be indicated in the discussion of the results, the effect of these inconsistencies had to be taken into consideration when comparing data from the two sources.

The AES and FES were conducted on July 1, 1981, approximately one month after the June 3 census date. Some data, such as farm expenses for the previous year, were expected to be relatively unaffected by the difference in reference data. However, other items were more likely to change between June 3 and July 1. The effect was expected to be particularly significant for livestock items, due to the constant fluctuations in inventories caused by birth, deaths, purchases, transfers, etc. As a result, operators responding to the survey were asked to indicate the changes in numbers of cattle and pigs between June 3 and July 1. Evaluation indicated that, while the data obtained were of some use in reconciling the differences due to reference date, they were subject to high non-response and questionable accuracy. Hence, the comparison of these and any other variables which would tend to be influenced by date of response had to take into consideration the difference in reference dates between the census and survey.

The samples for the AES and FES were selected from an area frame of agricultural enumeration areas, supplemented by a list frame of farms which were major producers of certain important commodities. Data collection was performed by trained enumerators during a personal interview with the operator of each selected farm. Following the necessary processing stages, an estimation procedure was applied to scale the counts up to the level of estimates for the entire population of interest. (Further details on the sample design are found in Statistics Canada (1984) and Phillips (1978).)

The survey estimates were subject to the same types of non-sampling errors as those from the census. However, due to the concentration on a smaller number of holdings, and the improved control of operations which was thus possible, it was expected that these types of errors would have a lesser impact on the surveys. Hence the surveys provided acceptable approximations to the true data values. On the other hand, the survey estimates were affected by samp--ling error, which had to be taken into account when making comparisons with macro-level estimates obtained from the census.

5.1 Macro-level Comparisons

Prior to the evaluation using the matched file, estimates from the complete census and survey data files were studied. Since the two vehicles covered comparable universes, these macro-level comparisons for provinces and regions provided initial indications of census coverage. By comparing census point estimates with survey 95 percent confidence intervals for totals of livestock, crop acreages, and other items, areas of potential over or underestimation were identified. Further investigation of the macro-level differences was then initiated to determine if they were confined to particular categories of the items of interest. The macro-level studies, in addition, provided the experience and familiarity with the two sets of data which were required for the detailed analysis which followed.

As an example of the results of the macro-level comparisons, Table 1 presents estimates for Canada for the number of farms, total farm area, and land use. A significant difference between census and survey estimates was observed for total farm area in Canada, yet the size and direction of differences varried greatly among the component land use categories. Census estimates for classes of improved land differed from the survey estimate by as much as 25.5 ± 7.7 percent for other improved land to as little as -3.4 ± 2.2 percent for cropland. The

Table 1

Comparison of Census and AES-FES Estimates for Number of Farms, Area and Land Use (in thousands of acres), 1981, Canada[a]

Item	Census Estimate[b,c]	Survey Estimate[c]	Percent Difference[d]
Total number of farms	309,410*[e]	319,476	− 3.2 ± 2.6
Total area of farms	159,866*	175,543	− 8.9 ± 2.4
Improved land	112,390	114,610	− 1.9 ± 2.3
Cropland	75,532*	78,211	− 3.4 ± 2.2
Improved pasture	10,523*	9,460	11.2 ± 7.3
Summerfallow	23,827*	24,939	− 4.5 ± 3.7
Other improved land	2,509*	1,999	25.5 ± 7.7
Unimproved land	47,477*	60,933	−22.1 ± 4.3
Woodland	8,211*	17,751	−53.7 ± 3.9
Other unimproved land	39,265*	43,182	− 9.1 ± 6.5

[a] Excluding Newfoundland,Yukon and Northwest Territories.

[b] Excluding specified marginal areas and farms not belonging to the survey universe.

[c] Census and survey totals may not equal the sum of the components due to rounding. Survey estimates for Canada are based on a sample of 18,327 farms.

[d] Percent Difference $= \frac{(\text{Census Estimate} - \text{Survey Estimate})}{\text{Survey Estimate}} \times 100$; the percent difference may not be consistant with the totals represented due to rounding. The indicated confidence interval, resulting from the sampling error in the survey, is equal to

$$\pm 2 \times (\text{survey coefficient of variation}) \times \frac{\text{census estimate}}{\text{survey estimate}}.$$

[e] An asterisk, identifying a significant difference between estimates, is indicated when the census estimate lies outside the survey 95 percent confidence interval.

major discrepancies in land area, however, were concentrated in the categories of unimproved land, particularly woodland. Further analysis into the reporting of woodland, which was prompted by these results, is discussed in section 5.5.

Macro-level comparisons also included the study of estimated frequency distributions prepared from the census and survey files. Distributions of the estimated number of farms over variables such as type of organization, land area, area of cropland, and sales were compared. Differences in the distributions identified possible over or undercounting of farms with particular characteristics.

Table 2 presents the census and survey frequency distributions by type of organization for the estimated number of farms in Canada. No significant differences were observed between the estimates for individual or family farms or corporations. However further study was initiated into the coverage of partnerships on the basis of the discrepancies noted for this category.

The limitation of the macro-level comparisons for evaluation of coverage was the inability to separate the effects of response errors from the effects of coverage errors. For example, the differences between census and survey estimates for improved land categories, shown in Table 1, seemed to exhibit too much variation in direction and magnitude to be the result of coverage errors alone. The discrepancies for woodland might also have been caused by factors other than coverage. Perhaps differences in field procedures or questionnaire format had resulted in inconsistencies between the census and surveys in the inclusion or exclusion of land of questionable agricultural value, or the classification of certain categories of land use. The micro-level match provided the needed mechanism for investigating these types of issues.

Table 2

Comparison of Census and AES-FES Estimates for Number of Farms by Type of Organization, 1981, Canada[a]

Type of Organization	Census Estimate[b]	Survey Estimate[c]	Percent Difference[d]
Total number of farms	309,410*[e]	319,476	− 3.2 ± 2.6
Individual or family farm	268,199	267,396	0.3 ± 3.0
Partnership			
- with a written agreement	11,160*	15,908	−29.8 ± 16.7
- with no written agreement	17,646*	22,855	−22.8 ± 10.8
Corporation	11,744	12,160	− 3.4 ± 10.4
Other type of organization	661*	1,142	−42.1 ± 13.0

For footnotes, see Table 1.

5.2 The Micro-level Match

Past experience with other agricultural censuses and surveys has indicated that even the most careful attention to quality cannot entirely prevent response errors. Despite all attempts to provide clear, unambiguous questions, problems such as differing interpretations of certain agricultural terms across regions of Canada, or a lack of consensus on the appropriate classification for certain types of land use, influence the data collected. Misinterpretation is particularly common for items which are of marginal economic or agricultural value,or which do not apply to most respondents. The micro or record level match with the AES-FES files provided the means to evaluate the impact of response errors on the 1981 Census of Agriculture.

The match between the Census of Agriculture and the AES-FES was based on the operator name, address, telephone number, and postal code for each holding. The link was performed in thirteen stages, each requiring a match on a different combination of the identifiers or their components. At each stage of the procedure survey records which has not yet been matched were identified, and the census file was searched by computer to locate the corresponding records. For each survey holding, the specified matching variables or keys were compared character by character with those of the census records which had not yet been linked. A match was identified if all characters of the matching variables were equal. At the Canada level, a computer match rate of 75.7 percent was achieved for the 18,327 survey records.

It was inevitable that, for a certain number of survey records, no census farm would be identified by the computer. Discrepancies in spelling of names and addresses, which had arisen during collection or capture of the census or survey data, prevented links in many cases. For example, J. Smith might have been reported on the census as opposed to J. Smyth on the survey, James Smith as opposed to Jim Smith, or St. Catherines rather than St. Catharines. Partnerships or corporations for which one operator had responded on the census but a different partner or manager had been interviewed by the survey could not be matched by a computer link on operators. Similarly, records for holdings which had changed operators between the census and survey collection dates could not be linked by computer.

In order to improve the match rate, and eliminate the possible biases in the matched file which might have resulted from those operations which could not be linked by computer, a manual resolution process was initiated. Using additional data from the questionnaires, such as corporate or farm name, addresses and names of partners, land description of the holding, and comments, clerical staff attempted to identify the corresponding census farm for each unmatched survey record. Of the 4,459 unmatched survey farms which remained following the computer link, 3,228 were matched during the manual resolution process. At its completion, 93.3 percent of the total 18,327 AES and FES records for Canada had been linked to census operations.

With further input of time and resources, it may have been possible to link some of the remaining 6.7 percent of the survey records to the census data base. However, in many cases the needed identifiers has not been collected on either the census or survey, and would have required investigation of administrative records or contact with the operators themselves. It was felt that the possible benefits were not sufficient to warrant the expenditures required, and no further manual resolution was attempted.

The studies which were facilitated by the record linkage can be grouped into two main types, those based on the unmatched survey records and those using the matched census-survey pair.

5.3 Studies of the Unmatched Records

In order to study the characteristics of census undercoverage, the unmatched records were assumed to be representative of the farms which should have been enumerated by the census but were missed. It was known that the unmatched records overestimated the number of missed farms, due to certain conditions of the data sources and the matching algorithm. For example, it was probable that some records on the survey file had been covered in the census, but could not be matched by either computer or manual means due to missing or invalid name or address data.

Because of the resulting potential for overestimation of the land and commodities missed by the census, one had to proceed with caution in using the estimates produced from the unmatched records.Nonetheless the Canada level estimates were most valuable as initial indicators of the characteristics of the farms which were underenumerated by the census.

In the first stage of the study, sample expansion factors were applied to the unmatched records to produce commodity estimates for the "missed farms". These were then compared to the commodity estimates for the entire survey universe, and the fraction of the total estimate which was accounted for by the "missed farms" was calculated. This fraction was then compared with the fraction which the missed farms comprised of the total estimated number of farms. For example, Table 3 shows that the unmatched file contained only 4.4 percent of the estimated total farm area and 3.9 percent of the cropland, whereas it was responsible for almost 9.7 percent of the total estimate of farms. These results provided an initial indication that the "missed farms" were not representative of the complete universe, but were smaller than average in terms of land area and other characteristics. This implied that the extent of undercoverage could not be measured by the number of farms missed alone. Instead, the characteristics of the missed farms over particular commodities had to be considered.

To provide further insight into the characteristics of undercoverage, frequency distributions of the estimated number of missed farms were prepared over classes of land area, sales, livestock, and other commodities. Comparison with similar frequency distributions for the entire survey universe indicated that the missed farms had a higher proportion of holdings with small acreages and low sales. It can be seen from Table 4, for example, that 42.3 percent of the estimated missed farms reported less than 70 acres of total farm area, as compared with 15.8 percent of the complete survey population. Sales of less than $1,200 were reported by an estimated 27.7 percent of the missed farms, but only 7.3 percent of the survey universe.

The frequency distributions were also compared by considering the ratio of the unmatched estimate to the estimate for the entire survey universe, that is, the fraction of the total survey estimate accounted for by the unmatched farms. As shown in Table 4, the unmatched file contained 36.5 percent of the estimated farms with less than 10 acres of land, the smallest size range presented as comapred with only 4.3 percent of those in the largest range of 760 acres or more. Similarly 36.8 percent of the estimated farms with sales less than $1,200 were obtained

Table 3

Comparison of AES-FES Estimates for Total Farms and Unmatched Farms, Area and Land Use (in thousands of acres), 1981, Canada[a]

Item	Estimate from Total AES-FES File[b]	Estimate from Unmatched AES-FES Records[c]	Percent of the Total Estimate Accounted for by the Unmatched Farms
Number of Farms	319,476	30,975	9.7
Total Farm Area	175,543	7,768	4.4
Total Improved Land	114,610	4,502	3.9
Cropland	78,211	2,792	3.6
Improved Pasture	9,460	603	6.4
Summerfallow	24,939	1,004	4.0
Other Improved Land	1,999	104	5.2
Total Unimproved Land	60,933	3,266	5.4
Woodland	17,751	1,325	7.5
Other Unimproved Land	43,182	1,941	4.5

[a] Excluding Newfoundland, Yukon and Northwest Territories.
[b] Survey estimates are based on a sample of 18,327 farms.
[c] The unmatched file contained 1,231 farms.

Table 4

Percentage Distribution of AES-FES Estimates for Total Farms and Unmatched Farms by Total Farm Area and Total Value of Agricultural Products Sold During 1980, 1981, Canada[a]

Item	AES-FES Estimate of Total Number of Farms[b]	AES-FES Estimate of Number of Unmatched Farms[c]	Percent of the Total Estimate Accounted for by the Unmatched Farms
	Cumulative Percent	Cumulative Percent	Percent
Total Farm Area			
Under 10 acres	3.5	13.2	36.5
10 - 69 acres	15.8	42.3	22.9
70 - 399 acres	62.8	85.0	8.8
400 - 759 acres	78.4	90.5	3.4
760 acres and over	100.0	100.0	4.3
Total Value of Agricultural Products Sold			
Under $1,199	7.3	27.7	36.8
$ 1,200 - $ 2,499	12.3	41.2	26.5
2,500 - 9,999	29.6	65.4	13.5
10,000 - 49,999	67.8	88.3	5.8
50,000 and over	100.0	100.0	3.5

[a] Excluding Newfoundland, Yukon and Northwest Territories.
[b] Survey estimates are based on a sample of 18,327 farms.
[c] The unmatched file contained 1,231 farms.

from the unmatched file, compared to 3.5 percent of those with sales of $50,000 or more.

In summary, the results of the study of unmatched records were able to provide concrete evidence that the holdings missed by the census tended to be smaller than average in terms of agricultural production and value, a theory which had been widely held but not proven.

5.4 Studies Using the Matched File

The matched file was composed of agricultural holdings for which census and survey records could be linked by the computer and manual processes described. Since these census and survey values were assumed to have been collected from the same set of holdings, the effects of coverage differences were removed. In addition, the influence of imputation was lessened by excluding records for which census or survey data had been entirely imputed due to non-response. Hence the nature and extent of potential response differences between the two data collection vehicles could be studied. Prior to discussing the results of the matched record studies, however, a number of limitations which existed within the matched file, and which influenced the evaluation process, should be described.

Although every effort was taken to lessen the chance of spurious matches, it is possible that a small number of survey farms may have been linked to the wrong census holding due to similarities in name or address. In the case of extremely large agricultural operations which made a significant contribution to provincial commodity totals or land areas, linkage to the wrong census operation could noticeably skew the results. A detailed study of a sample of matched records, undertaken to determine the quality of the computer link, had not been completed at the time of the evaluation. As a result, the potential influence of spurious matches had to be considered when studying results from the matched file.

The second limitation on the matched file analysis affected the comparison of total counts from the census and survey data. The matched records consisted of a subset of the non-self-weighting sample of the AES and FES, since they contained only the survey farms which could be linked to census records.It would have been preferable to apply sample expansion factors to produce weighted estimates for the matched file. However, the expansion factors for the area frame were calculated using reported land use values from the survey, and it was as yet undetermined whether the factors were valid when applied to census data from the matched file. Census-related expansion factors could not be calculated using the census data, since one component of the factor, the farm area inside the selected segment of land, was collected only by the survey. As a result of this uncertainty regarding the application of survey expansion factors to census matched data, it was decided to restrict the analysis to unweighted census and survey totals from the matched file. (Study of the use of weighted estimates from the matched file was underway at the time of writing, but no conclusions had yet been reached.)

Despite the limitations of a non-self-weighting sample and the possible existence of spurious matches, the matched file proved to be a valuable evaluation tool. When matched totals identified discrepancies between census and survey values, further detailed investigations were undertaken into the possible causes of the observed differences.

As an example of the use of the unweighted matched totals, Table 5 presents counts for total farm area and categories of land use at the Canada level. The results indicate that less land was reported on the census than the survey for all land use items except improved pasture and other improved land. Relative differences between census and survey totals were smallest for items such as cropland, which are of major economic value and hence are clearly defined and seldom misunderstood by farm operators. Items of more marginal agricultural and economic value, however, tended to display greater discrepancies. The largest relative differences were observed for the category of woodland. In order to demonstrate some of the detailed evaluation techniques made possible by the matched file, further results of the study of data on woodland will be discussed.

Table 5

Comparison of Census and AES-FES Totals for Matched Farms, Land Use (thousand of acres), 1981, Canada[a]

Item	Census Total[b]	Survey Total[b]	Percent Difference[c]
Total area of farms	13,059	14,091	−7.3
Improved land	8,798	8,801	− −
Cropland	6,046	6,167	−1.9
Improved pasture	804	682	18.0
Summerfallow	1,777	1,816	−2.1
Other improved land	170	137	24.3
Unimproved land	4,261	5,291	−19.5
Woodland	523	1,102	−52.5
Other unimproved land	3,737	4,189	−10.8

[a] Excluding Newfoundland, Yukon and Northwest Territories.

[b] Records for which census or survey data were entirely imputed have been excluded leaving 16,388 matched farms. Census and survey totals may not equal the sum of the components due to rounding.

[c] Percent Difference $= \frac{\text{Census total - Survey total}}{\text{Survey Total}} \times 100$

5.5 Detailed Comparisons of the Matched Records

While the matched totals provided a measure of the overall biases in reporting, they masked detailed information on where and why the differences occurred. For instance, was the difference in woodland caused by large discrepancies in only a handful of holdings, or was it consistent across all records? Did the response differences vary in magnitude or direction across types of operations or regions of the country? By comparing census and survey responses at the individual record level, the characteristics of the reporting differences were studied in detail.

Table 6 presents an example of the type of investigation facilitated by the matched evaluation file. Only those holdings for which a non-zero value of woodland was reported on either the census or survey are included. The operations are classified by the size and direction of the difference between the census and survey values for the item, and cross-classified by the amount of woodland reported on the census.

The table indicates that less woodland was reported on the census than on the survey for the majority of holdings. Differences tended to be small, clustered in the 1 to 50 acres and 51 to 150 acres ranges even for operations with large amounts of woodland on the census. Of note also were the 3,177 holdings, or 34.3 percent of the universe of interest, for which the census value of woodland was zero but the survey value was greater than zero. This is in contrast to the 807 holdings, or 8.7 percent of the universe, in which the opposite case of zero acres of woodland on the survey but greater than zero acres on the census was observed. Examination of these results suggested that the census and surveys may not have been obtaining measures of the same quantity, and prompted further study into their collection methodologies and questionnaire formats.

On the census questionnaire the respondent was asked to report the area of woodland, with further instructions in the census representative's manual indicating that only land "with seedlings or trees which had or would have value as timber, fuelwood, or Chirstmas trees" be included. In contrast the AES and FES interviewers instructed respondents to "include woodlots, cut-over land, etc." with no additional instructions that the land be of present or future commercial value. As well, woodland was requested twice on the surveys, once immediately after the reporting of total area, and later as a component of land use, and thus received greater emphasis.

Table 6

Comparison of Census and AES-FES Responses for Matched Records, Difference in Woodland by Census Value of Woodland, Canada[a], 1981

Difference: Census Woodland - Survey Woodland (acres)	Census Value of Woodland (acres)								
	0	1 to 2	3 to 9	10 to 69	70 to 239	240 to 399	400 or more	Total	Percent of Reporting Farms
	Number of Reporting Farms[b]								
less than −500	241	0	3	16	16	6	11	293	3.2
−500 to −251	248	0	4	24	21	12	7	316	3.4
−250 to −151	268	4	1	27	44	9	4	357	3.9
−150 to −51	715	5	30	167	145	24	19	1,105	11.9
−50 to −1	1,705	59	277	1,053	370	57	28	3,549	38.3
0	-	26	165	481	151	21	21	865	9.3
1 to 50	-	55	229	1,288	492	53	30	2,147	23.2
51 to 150	-	-	-	50	294	45	32	421	4.5
151 to 250	-	-	-	-	65	25	17	107	1.2
251 to 500	-	-	-	-	-	33	42	75	0.8
greater than 500	-	-	-	-	-	-	36	36	0.4
Total	3,177	149	709	3,106	1,598	285	247	9,271	100.0

[a] Excluding Newfoundland, Yukon, and Northwest Territories.

[b] Including all operations which reported woodland on either the census or survey. Excluding records for which either the census or survey data were totally imputed due to non-response.

As a result of these differences, it was believed that certain areas of woodland of questionable commercial value, which were reported on the survey, may have been excluded from the census. As well, some areas may have been reported on the census under different categories of land use, such as other unimproved land. Study continues into these and other hypotheses, using the matched file to investigate possible causes of the observed response differences.

When summary tabulations from the matched file failed to suggest causes for observed biases, the study of individual records which displayed large discrepancies between census and survey values for the items of interest was often informative. By comparing census and survey responses for other related items, it was sometimes possible to identify misclassification between categories or other causes of reporting differences.

Table 7 shows a number of variables for a record with one of the largest differences between census and survey values for total farm area. In this case, the discrepancy in total area was

Table 7

Census and Survey Recorded Values for a Particular Matched Record

	Total Farm Area (acres)	Land Use (acres)						Total Cattle and Calves
		Cropland	Improved Pasture	Summerfallow	Other Improved Land	Woodland	Other Unimproved Land	
Census Values	2,640	1,035	0	1,240	15	0	350	920
Survey Values	17,000	1,010	0	970	20	0	15,000	815

concentrated in the category of other unimproved land; only minor differences existed between responses for cropland, summerfallow, and other improved land. It appears that most of the land which was reported as other unimproved on the survey was excluded from the census response. Similar studies of other individual records identified other potential discrepancies in the reporting of unimproved land. Theories developed from these studies of individual cases were then tested on the entire file to determine if they might apply in general.

6. CONCLUSION

The link between the 1981 census and survey files was a powerful tool for the evaluation of census coverage and response errors. Errors were known to exist in the survey data which were being used as approximations to the true values, and limitations of the linkage operation were known to have caused spurious matches and unmatches. However, the matched file was still a valuable source of data for investigation of quality concerns. Studies based on the record level matches broadened the initial results obtained from the macro-comparison of census and survey estimates. In addition they brought individual problems into focus by allowing detailed investigation of particular aspects.

The evaluation produced valuable results on census undercoverage. Studies based on the unmatched survey records showed that the holdings missed by the census were, in general, smaller than average in terms of total land area, livestock, and value of agricultural products sold. Thus concrete evidence was provided to support the widely-held theory that the census tended to miss holdings of marginal economic and agricultural value.

Studies of response differences for land use identified categories in which discrepancies were concentrated, tendencies for confusion between certain classes, and variations in differences among regions of the country. Possible revisions to questionnaire format and wording, or collection methods, have been considered as a result of the study. Some of the variations are known to have been caused by difficulties in defining certain land use categories, due to the lack of clarity in the concepts themselves. Problems such as these, that result from confusion in the minds of the respondents as to which land should be reported under which category of usage, may never be completely solved. However, the recognition of the existence of a problem, and the study of the characteristics of its occurrence and its effect on the data, are very valuable contributions to future planning.

The 1981 data quality evaluation had far-reaching effects for the census. In response to its main goal, the study identified quality concerns in the 1981 census data. A publication (Statistics Canada 1984) was prepared to provide users with an indication of data quality, and to advise them with respect to particular problems which had a noticeable impact on the data. Looking further ahead, the evaluation has served as input to the planning of 1986 census procedures. By identifying items for which coverage or response errors occurred in 1981, the study has provided a list of areas requiring further consideration of collection and processing methods.

The impact of the data quality evaluation was not restricted to the Census of Agriculture alone. The comparison of census and survey responses also identified problem areas in the other data collection vehicles. Improvements to the National Farm Survey, the annual probability survey which has replaced the AES and FES, may result from the census study.

A further benefit of the study is the knowledge gained on the use of record linkage for evaluation purposes. The experience in matching data at the individual record level, using both computer and manual means, could provide valuable input to other linkage projects. In particular, knowledge of the problems encountered, their causes, characteristics, and possible solutions, could result in improved procedures for other studies.

In summary, the 1981 Census of Agriculture Data Quality Evaluation project has provided further evidence of the power of matching in the study of non-sampling error. The investigations using matched and unmatched records, and macro and micro-level comparisons, have produced measures of quality of the 1981 census data, and identified items for which errors have impacted significantly upon the results. As an extension of the original project mandate, input was provided to the planning of the 1986 and subsequent censuses, by indicating areas in which further research into possible changes in methodology was required. The evaluation also identified potential problems in the surveys used for comparison, thereby contributing to the planning of future vehicles for the collection of agricultural data. Finally, the study provided valuable experience and insight into the application of record linkage techniques for data quality evaluation.

ACKNOWLEDGEMENTS

The author would like to thank D. Royce, M.R. Dibbs, B.N. Chinnappa, K. Thatcher, the reviewer, and the other members of Statistics Canada whose support and valuable comments were most appreciated during the writing of this paper.

REFERENCES

ANDERSEN, R., KASPER, J., FRANKEL, M.R., and associates (1979). *Total Survey Error.* San Francisco: Jossey-Bass Publishers.

COCHRAN, W.G. (1977). *Sampling Techniques,* third edition. New York: John Wiley & Sons.

DEMING,W.E. (1944). On Errors in Surveys. *American Sociological Review,* 9, 359-369.

FAULKENBERRY, D., and TORTORA, R.D. (1981). Non-sampling Errors in an Agriculture Survey. *1981 Proceedings of the Section on Survey Research Methods, of the American Statistical Association,* 493-495.

FELLEGI, I.P. (1964). Response Variance and its Estimation. *Journal of the American Statistical Association,* 59, 1016-1041.

FELLEGI, I.P. (1973). The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences. *International Statistical Review,* 41, 1-14.

GOSSELIN, J.-F., CHINNAPPA, B.N., GHANGURDE, P.D., and TOURIGNY, J. (1978). *A Compendium of Methods of Error Evaluation in Censuses and Surveys* (Catalogue 13-564). Ottawa, Canada: Statistics Canada.

HANSEN, M.H., HURWITZ, W.N., MARKS, E.S., and MAULDIN, W.P. (1951). Response Errors in Surveys. *Journal of the American Statistical Association,* 46, 147-190.

HANSEN, M.H., HURWITZ, W.N., and BERSHAD, M. (1961). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute,* 38, 359-374.

HARTLEY, H.O. (1981). Estimation and Design for Non-sampling Errors of Surveys. In *Current Topics in Survey Sampling,* ed. D. Krewski, R. Platek, and J.N.K. Rao. New York: Academic Press.

HORVITZ, D.G. (1981). Response Error Research Issues in Health Surveys. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 397-399.

KIBLER, W.E. (1978). Controlling Non-sampling Errors in Surveys. Summary Report of the 29th Federal-Provincial Committee on Agricultural Statistics, Statistics Canada.

KISH, L. (1965). *Survey Sampling.* New York: John Wiley & Sons.

KISH, L., and LANSING, J.B. (1954). Response Errors in Estimating the Value of Homes. *Journal of the American Statistical Association,* 49, 520-538.

KROTKI, K. (1980). *Response Error in the 1976 Census of Population and Housing.* Working Paper, Ottawa, Canada: Minister of Supply and Services Canada.

MARQUIS, K.H., MARQUIS, M.S., and POLICH, J.M. (1981). Survey Responses to Sensitive Topics. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 339-341.

NISSELSON, H., and BAILAR, B.A. (1976). Measurement, Analysis, and Reporting of Nonsampling Errors in Surveys. *Proceedings of the 9th International Biometric Conference,* 2, 201-322.

PHILLIPS, J. (1978). 1979 Farm Expenditure Survey Design and Estimation Procedures. Working Paper, Institutional and Agriculture Survey Methods Division, Statistics Canada.

STATISTICS CANADA (1979). *1976 Census of Canada - Agriculture - Evaluation of Data Quality* (Catalogue 96-872). Ottawa, Canada: Minister of Supply and Services Canada.

STATISTICS CANADA (1980). *1976 Census of Canada - Quality of Data - Series I: Sources of Error - Coverage* (Catalogue 99-840). Ottawa, Canada: Minister of Supply and Services Canada.

STATISTICS CANADA (1982). *1981 Census of Canada - Agriculture* (Catalogue 96-901). Ottawa, Canada: Minister of Supply and Services Canada.

STATISTICS CANADA (1984). *1981 Census of Canada - Agriculture - Evaluation of Data Quality* (Catalogue 96-918). Ottawa, Canada: Minister of Supply and Services Canada.

SUKHATME, P.V., and SETH, G.R. (1952). Non-sampling Errors in Surveys. *Journal of the Indian Society of Agriculture Statistics*, 4, 5-41.

TREMBLAY, V., SINGH, M.P., and CLAVEL, L. (1976). Methodology of the Labour Force Survey Re-interview Program. *Survey Methodology Journal*, 2, 43-62.

U.S. BUREAU OF THE CENSUS (1964). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by the CPS - Census Match.* Series ER60, No. 5, Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1970). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check Study of the Accuracy of Income Reporting*. Series EDR60, No. 8, Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1982). *1978 Census of Agriculture Volume 5 Special Reports - Part 3 Coverage Evaluation*, (AC78-SR-3). Washington, D.C.: U.S. Government Printing Office.