GENERALIZED RECORD LINKAGE AT STATISTICS CANADA

Connie Nuyens, Statistics Canada
14-O R. H. Coats Bldg., Ottawa, Canada K1A 0T6

KEY WORDS: Record Linkage, Generalized Software

1.0 Introduction

GRLS (Generalized **R**ecord **L**inkage **S**ystem) is a probabilistic record linkage system developed at Statistics Canada. It detects records which refer to the same *entity* (person, business, farm or other unit) in files which do *not* contain *unique identifiers*. For example, one file contains work history information, another file contains education data. In order to perform studies on the relationship between education and occupation, the two files must be linked to garner information from each file. If common fields exist between the two files (eg. NAME, ADDRESS, AGE etc.) **GRLS** could be used to perform the linkage.

2.0 Probabilistic Record Linkage

GRLS was built to utilize probabilistic linkage methods developed by Newcombe and his associates [5] and formalized by Felligi and Sunter [3]. Probabilistic linkage is based on the assumption that any particular **outcome** for a pair of records (eg. **agreement** on NAME, ADDRESS, AGE) can argue for or against linkage. The **probability** of *a particular outcome* occurring in record pairs which are truly linked divided by the **probability** of *the same outcome* occurring in record pairs truly unlinked is referred to as an **odds ratio**. The *higher the odds, the more likely the pair is a true link.*

The odds ratio formula shows the probability (P) of an outcome (O) occurring in a linked set of records (L) divided by the probability (P) of an outcome (O) occurring in an unlinked set of records (U).

$$\frac{(P\ (O)\ \in\ L)}{(P\ (O)\ \in\ U)}$$

3.0 GRLS

There are three important stages when using **GRLS** to link records: the Search, Decide and Group stages.

3.1 Search Stage

This is where linkage parameters are defined in order to create record pairs. Comparison **rules** must be specified listing the fields to be compared and all the possible comparison **outcomes** to be evaluated when records are compared.

Sample Rule for AGE

```
If difference in AGE = 0
OUTCOME = Full Agreement
Else
        If difference in AGE <= 5
        OUTCOME = Partial Agreement
                Else (difference in AGE > 5)
                OUTCOME = Disagreement
```

Initial probabilities for the **outcomes** need to be determined (either by observation or through sampling from prior linkages) so the odds ratio can be calculated for each possible outcome. Note that if the odds ratio is greater than one it tends to argue for linkage while if it is less than one it tends to argue against linkage.

Sample Probabilities for AGE

	Linked Set	Unlinked Set	Odds Ratio
Full Agreement	.45	.15	3.0
Partial Agreement	.50	.20	2.5
Disagreement	.05	.65	.07

Outcome weights are calculated from the **probabilities** and specified to the system; **GRLS** provides *automated methods to assist the user* with weight calculations. Also, **pockets** ("blocks" of records) can be specified for efficiency, determining which records will be compared to each other. Records are then compared and result in *record pairs*. For each pair, for all outcomes which occurred, the appropriate outcome weights are summed to a **total weight**. The **total weight** is tested against

thresholds (specified by the user) to determine whether the pair is considered **possibly** or **definitely linked.** The higher the total weight, the more likely the record pair is truly linked.

By sampling various pockets and creating sub-sets of record pairs, the user can review the pairs and refine linkage parameters before linking large files.

3.2 Decide Stage

It is very important to try to minimize the *two types of error* which can occur: overlooking record pairs which should have been created but are not and creating record pairs which are deemed linked but are not. *Existing* record pairs can be re-classified manually when selected pairs are reviewed or automatically by updating selected record pairs or by changing thresholds.

Total weights for *existing* record pairs can be revised by applying **value-specific** frequency **weights.** These weights are based on frequency counts of the values agreed upon between records; more common values are associated with lower weights and less common values are associated with higher weights. For example, agreement on GIVEN NAME may be significant but agreement on the value *JOHN* is not as significant as agreement on the value *ZACHARIAH* and thus should be weighted appropriately.

3.3 Group Stage

At this point, records which link to each other are *grouped together.* For two-file linkages, groups can be re-structured so that only one-to-one, many-to-one or one-to-many linkages between records exist. This can be done automatically or manually.

4.0 GRLS Version 1

4.1 Technology

GRLS V1 was developed at Statistics Canada and has been used there and at other sites since the early 1980's. It was developed in PL/1 to run on an IBM mainframe (or compatable platforms) and uses a transposed file manager called RAPID, developed at Statistics Canada.

4.2 Features

GRLS V1 allows both one-file and two-file linkages. **Rules** can specify fields described as character or numeric, can be conditionally executed, and can be ordered to maximum efficiency in creating pairs. **Rules** written by the *user* can be included when fields are not independent or special processing is required. When comparing numeric data, "range" differences between values can be specified eg. consider the fields to agree if the difference between BIRTH-YEAR is within 5 years. Cross-comparisons of fields are allowed so that related fields eg. FIRST NAME and SECOND NAME can be compared. Also, user-written code can be inserted to exclude records from comparison based on certain conditions.

Note that direct matching (deterministic linkage) can be accomplished using **GRLS.** If only one component of an outcome weight is specified and if the lower threshold is set to the maximum sum of the outcome weights (and the upper threshold set to the same value as the lower threshold), record pairs can be considered *directly matched.* The next step could be to eliminate the records involved in "direct" matches and *with the remaining unmatched records* continue with a probabilistic linkage. Alternatively, "linked set" outcome weights could be calculated in **GRLS** based on the directly matched pairs and the linkage could be repeated using both unlinked and linked set weight components (probabilistic linkage). With the latter method, *all* records which link to each other will be found and grouped together. Using **GRLS** for direct matching may not be the optimum strategy, however it could be a convenient mechanism for some linkage projects.

4.3 Experiences

In Statistics Canada **GRLS V1** was used to create a Residential Address Register for urban areas in Canada. Using this software, duplicate address register records were identified (and eliminated, reducing 10 million records to 6.7 million). Then **GRLS V1** was used to link geographic identifiers to address register records. Originally, automatic direct linking with clerical clean-up was planned but based upon test runs, only 80% of matched records directly linked; the other 20% needed to be resolved manually. By using **GRLS V1,** 97% of matched records were linked automatically; only 3% were resolved manually. **GRLS V1** provided stability and high match rates for this project.

The Canadian Centre for Health Information at Statistics Canada has used **GRLS V1** extensively over the last 10

The Canadian Centre for Health Information at Statistics Canada has used **GRLS V1** extensively over the last 10 years for numerous linkage projects, for their own studies and for other health agencies. Many of these projects would not have been done because of large data volumes or because linkage criteria were so varied that custom software for each application was not feasible. **GRLS V1** is currently being used as a tool in developing the Canadian Cancer Registry. The personnel in the Canadian Centre for Health Information have developed considerable expertise in all aspects of **GRLS V1**, especially in weight calculation and writing rules. Some of their techniques have been incorporated into the second version of **GRLS**.

In the Agriculture division at Statistics Canada, approximately 350,000 Census of Agriculture farm records were recently linked to two tax files reporting non-corporate and corporate farm income. Using **GRLS V1**, an exact match was performed first and matched records were eliminated from the farm file and the tax files. Then **GRLS V1** was used to do probabilistic linkage between the reduced files. In prior years, this linkage was done with SAS using a hierarchical approach and a large number of records required manual review. Using **GRLS V1** record pairs were sorted in descending order on **total weight** (indicative of the quality of the link) thus reducing the number of linked records requiring manual review.

Statistics Canada has also established a Record Linkage Resource Centre to assist users with linkage projects. This group of mathematical statisticians has notably contributed to the success of various **GRLS** linkages.

GRLS V1 has been used for numerous other studies and surveys at Statistics Canada. It has also been released to external sites such as the University of Western Australia, WESTAT Maryland, U.S., Bureau of Labor Statistics, Washington, D.C. and the New York State Department of Health.

These experiences demonstrate that **GRLS** can effectively be used for varied types and sizes of linkages.

5.0 GRLS Version 2

5.1 Technology

GRLS V2 is an enhanced version of **GRLS V1** and is capable of running on UNIX platforms using ORACLE as the data base management system. To make the system as portable as possible, it is implemented in the "C" programming language with imbedded SQL (Structured Query Language). The new system takes advantage of technology advances such as on-line database access allowing concurrent queries and updates (many operators can review and update record pairs or groups simultaneously). It can operate in a distributed environment whereby one file can exist on a personal computer, the other on a mid- range computer or mainframe and the two files can be accessed transparently.

5.2 Features

GRLS V2 has many new features. Firstly, it has a user-friendly interface for entering linkage criteria and reviewing linkage results. Detailed documentation is provided with the system [7]. A concise overview and training (including a tutorial) are available and a Strategy Guide provides step-by-step linkage instructions including examples of linkage criteria and reports. Extensive *on-line* help is also available.

A new type of linkage called **interactive** record linkage has been added whereby a "search" record is entered manually or is retrieved from a file and then linked *on-line* to a master file using specified linkage criteria. If record pairs are found, the master file can be **updated** with information from the search record.
Multiple pocket values can be processed without duplicating or triplicating input files. For example, if two pockets such as SALES-REGION and SALARY are specified, records not compared by the first pocket could be compared by specifying the second pocket.

Systematic sampling and random sampling have been incorporated into the system, providing the ability to sample pockets when testing and also in production (different thresholds can be specified for different pocket values). Random sampling is built-in so that a set of randomly created pairs can be created, unlinked set weights calculated based upon those pairs and then those weights used when creating the linked set of pairs.

To determine whether weights and thresholds are satisfactory and whether you have selected effective comparison outcomes for your rules, numerous reports and features are available. For instance, associated probabilities for weights are displayed if you manually enter values for outcome weights. This helps the user determine whether the specified weights are reasonable.

The need for "special" rules to do sophisticated types of

comparisons has been *greatly* reduced because of new comparison outcomes **built-in to GRLS V2:**

- **Percent** differences for numeric or date values allow specification of a difference of 2% (for example) between values.

- **Specific** differences for numeric or date values allow specification of an exact difference between values.

- **Transposition** of character, numeric or date values eg. **FRED & FERD** or **778 & 787** or **1929 & 1992.** Also allows specification of *date component* transposition eg. **30-01-93 & 93-01-30.**

- **Mismatch** of character, numeric or date values eg. **FRED & FREG** or **777 & 778** or **1992 & 2992.**

- **Extra character** for character and numeric values eg. **FRED & FREED** or **778 & 7778.**

- **Date** comparisons (left-to-right comparison).

- **Alternate name** comparison for character values currently for comparing *given names* using a look-up table of "nick-names" associated with various "formal" names such as PEG, MEG etc. for MARGARET. This outcome is based on research performed by Newcombe, Fair and Lalonde [6].

Other new outcomes are the following string comparator functions which return values that indicate how closely 2 character strings resemble each other.

- a **Jaro** outcome is based on Jaro's string comparator formula (published by Winkler [9]). The formula checks for length of strings and partially accounts for typographical and transcription errors.

- a **Winkler** outcome, an extension of the Jaro outcome to account for agreement amongst the first 3 characters (see Budzinski [2]).

- a **Bickel** outcome based on Bickel's information theoretic likeness measure [1]. This involves computing a likeness value from comparing strings letter by letter and assigning varying weights for each letter.

- a **PF474** outcome based on computing a proximity value for a pair of strings as described by Taylor [8]. Basically, this involves comparing sets of letters forwards and then backwards.

These are just some of the new features that are provided with **GRLS V2.**

5.3 Experiences

Experience with **GRLS V2** is limited at this point. It has recently been made available to the Cancer Treatment and Research Foundation of Nova Scotia and to the Ontario Cancer Treatment and Research Foundation. Both sites are in the process of evaluating the software and providing suggestions for future enhancements. In Statistics Canada, the Address Register has just started an evaluation of **GRLS V2** and the Agriculture Division is also looking into using **GRLS V2** on their UNIX platform.

6.0 Future of GRLS

6.1 Future Enhancements

GRLS V2 will be augmented as long as there is justification in improving the product. The current list of planned enhancements include extending or adding comparison outcomes such as the use of alternate names with *any* character or numeric field. This comparison would be useful for equivalent street names or hospital codes or business names to name just a few. Also, distance comparisons could be very valuable when trying to determine whether related fields are similar. For example, comparing associated location coordinates for WORK-PLACE and PLACE-OF-DEATH could determine that the places are *physically* close to each other while comparison on the place names themselves might not produce any level of agreement.

Some experimentation with the E/M algorithm for weight calculation as documented by Jaro [4] is planned.

6.2 Technology

GRLS V2 will continue to be upgraded within the current technology (ORACLE, "C", SQL) on the current UNIX platform. A second release of **GRLS V2** is due in 1993 using the newest version of the ORACLE data base (Version 7.0). This data base should provide improved performance for multiple users.

As computing environments evolve, it is likely that the **GRLS** software will evolve too. There is potential for the eventual development of **GRLS V3** to be

implemented with a graphical user interface and to be capable of running on UNIX as well as other platforms such as WINDOWS/NT.

ACKNOWLEDGEMENTS

The author would like to thank Ted Hill and Dan Ducharme for their assistance with this paper.

REFERENCES

1.BICKEL, M. A. (1987), "Automatic Correction to Misspelled Names: A Fourth-Generation Language Approach", Computing Practises, Volume 30 No. 3, March 1987

2.BUDZINSKY, C. D. (1991), "Automated Spelling Correction", Statistics Canada

3.FELLIGI, I. and SUNTER, A. (1969), "A Theory of Record Linkage", Journal of the American Statistical Association, 64 1183-1210

4.JARO, M. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", Journal of the American Statistical Association, Volume 84, No. 406, June 1989

5.NEWCOMBE, H. B. (1988), Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business, Oxford: Oxford University Press

6.NEWCOMBE, H. B., FAIR, M. E., LALONDE, P. (1992), "The Use of Names for Linking Personal Records", American Statistical Association Journal of the American Statistical Association, Vol. 87, No. 420

7.STATISTICS CANADA (1993), GRLS V2 Concepts, GRLS V2 Strategy Guide, GRLS V2 Tutorial, GRLS V2 Installation Guide

8.TAYLOR, D. (1986), "Wordz that Almost Match", Computer Language, November

9.WINKLER, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Felligi-Sunter Model of Record Linkage", Proceedings of the ASA Section on Survey Research Methods

AUTOMATED CODING AT STATISTICS CANADA

Dianne Miller, Statistics Canada
14P R.H. Coats Bldg., Ottawa, K1A 0T6

KEY WORDS: Automated Coding, Text Recognition

1.0 INTRODUCTION

ACTR is a generalized automated coding package and may be used for any coding application. It employs word standardization techniques to match input respondent text to an existing reference file of phrases in order to return a code. As part of its generalized implementation, **ACTR** provides: a parsing mechanism to reduce text to a standard format, functions to create and maintain reference files of phrases and codes, and a searching and matching algorithm to perform the coding. All that is required of the user is a reference file of phrases or descriptions and their associated codes.

ACTR can be used in batch or conversational mode with the existing end-user interface or embedded in a user-developed application in program-callable mode.

Reference File Maintenance

ACTR supports the loading and updating of reference files for a coding application. The use and the content of these files is entirely at the user's discretion and will most likely contain standard text and codes from existing reference manuals and responses as they are received from the survey participants, complete with spelling, grammatical and syntax errors. As experience is gained with the format of responses obtained, the reference file(s) and corresponding parsing strategy can be updated in order to improve the quality of the coding.

Parsing

The parsing function is capable of handling problems such as rearranged words, plural vs. singular, missing words, extraneous words, spelling variations, synonyms, abbreviations, inconsistently hyphenated words and variable punctuation and syntax. Ideally, the resulting form should be such that any two phrases with the same words will be identical in their **ACTR** representation regardless of their syntactical and grammatical differences. **ACTR** provides the user with default parsing rules and allows the user to control how, if at all, each step is executed. Examples of the application of parsing rules include: deleting strings which contain *'S* or *'T* or replacing strings such as *'T.V.'* with *'TELEVISION'*; removing trivial words such as *I, AM, A*; removing suffixes and prefixes or standardizing words which are synonymous such as *CAR* and *AUTOMOBILE.*

The set of words resulting from the parse of the supplied text is then examined for the presence of duplicates and the duplicate is removed. The words are then sorted.

Matching

Any text input to **ACTR** for matching is first parsed according to the parsing strategy defined. If, after this step, **ACTR** is able to locate a matching phrase on the reference file with all of its words in common with all of the words in the input text, then the match found is referred to as a ***Direct Match***. If a direct match cannot be found, **ACTR** can continue to search the reference file for the closest match. This is called an ***Indirect Match***.

The ***direct match*** is implemented with the use of a key which is generated and stored for all phrases in the reference file and is also generated for each input phrase to be coded. If the key of the input phrase matches an existing key on the reference file then the associated code is returned.

Indirect matching begins with the set of words resulting from the parsing process. The reference file is searched to retrieve and evaluate all the phrases which contain the words in the input text.

The nearest matching phrase is determined by calculating a score for each of the possible matches. The score is based on the weights of the words which are in common between the reference file and input phrases. Generally, the less frequently a word occurs in the reference file, the higher its weight and it will contribute more to the score. Of all the database phrases evaluated in this manner, the highest scoring phrase is the one which is considered to be the closest match.

Both direct and/or indirect matching can be performed in a coding run.

Readers interested in the details of **ACTR** are referred to the **ACTR Version 1** and **ACTR Version 2** documentation sets.

2.0 DEVELOPMENT HISTORY

Two versions of **ACTR** have been developed at Statistics Canada, with a third implementation currently being planned.

2.1 ACTR Version 1

ACTR Version 1 was developed for use in a mainframe only environment (IBM, MVS/TSO) using an ISPF user-interface, the PL/1 programming language and RAPID (a relational DBMS developed at Statistics Canada). The searching and matching algorithms are based on a methodology developed at the U.S. Bureau of the Census (Hellerman, 1982).

This version has been used successfully in a number of applications. These include:

> The ***1991 Census of Population*** used ACTR to code 13 variables, totalling approximately 17 million responses, over a 2 1/2 month timeframe. An overall match rate of 92% was achieved and savings of over $3 million were realized compared to a manual coding operation.
>
> The ***1991 Census of Agriculture*** used ACTR to code approximately 28,000 write-in responses over a 4 month timeframe. Match rates of 80-90% were achieved at one-third the cost of a manual coding operation.
>
> ***Transportation Division*** uses ACTR to code 300,000-500,000 commodity descriptions per year for a Trucking survey. The descriptions are captured from bills of lading and are lengthy, verbose and vary greatly in content. Without the use of ACTR, the division would have to reduce the volume of respondent data captured as there are limited resources to code it manually.
>
> ***The Canadian Centre for Health Information*** used ACTR to code the occupation data from 138,000 death certificates for a "Mortality by Occupation" study. Match rates of 85% were achieved at a saving of $100,000 over a manual coding operation.
>
> ***Education, Culture and Tourism*** uses ACTR to code 200,000 geographic responses per year for four International Travel surveys. Match rates of 90% are being achieved.

Many users have taken advantage of automated coding because the use of **ACTR** eliminated software development costs. However, **ACTR** has been easily integrated into application-specific programs. All users have cited cost-savings, quality and consistency of coding as major benefits of **ACTR**.

No further enhancements are planned for this version.

2.2 ACTR Version 2

ACTR Version 2 was a development effort to prototype new technology and new methods for weighting and scoring.

This version was developed within the guideline of a development strategy put in place at Statistics Canada which provided clear direction for any future development of general systems (Outrata, Doucet, 1989). Briefly, it states that all software which falls within the GSFD (Generalized Survey Function Design) concept must use 32-Bit technology, be driven by a relational database, adhere to international standards such as **SQL** for data management and **OSI** (Open Systems Interconnect) for interprocess communication and be portable to other computer environments.

ACTR V2 was developed with a UNIX operating system, ORACLE Version 6 RDBMS with SQL*FORMS 2.3 and an ANSI 'C' compiler.

Major changes were made to the weighting and scoring methods.

ACTR Version 1 was developed with the flexibility to use both direct and indirect matching to 'automatically' assign a code. In reality, most applications of **ACTR** use only the direct match feature to assign codes automatically, while the indirect match has generally been used for computer-assisted coding.

The word weight and phrase scoring methods used in **Version 1** require statistical information on the universe of responses which is often not known until after the coding operation is complete and also produce score values which are largely non-intuitive. These methods have proven inappropriate for computer-assisted coding and make it difficult to determine applicable matching parameters. **ACTR Version 2** employs methods which produce scores that are more intuitive and provide a relative measure of 'how close' a phrase matches in comparison with other phrases (Wenzowski, 1988).

A more flexible parsing strategy is also provided which allows the user more control over the order of the parsing components. The interface to modify the parsing order and contents is much easier to use than that of

Version 1.

Beta versions have been sent to a number of external users and we have installed the software on a central UNIX environment which can be accessed by interested clients at Statistics Canada.

There is great interest in this version as it overcomes some constraints of Version 1. However, most potential users cannot afford to acquire the technology to use this version or would prefer to have it available in a PC Windows environment. ACTR V2 also requires expertise with the operating system (UNIX) and the data management software (ORACLE).

No further development is planned.

2.3 ACTR Version 3

ACTR Version 3 is currently in the planning stage. The main objectives are to produce an automated coding system that:

- can be applied in a variety of subject matter areas having widely varying sizes of reference files and volumes of responses to code
- can be used in several computing environments including those where the reference files reside on a different computer than that of the data to be coded (i.e. client-server models), and
- will insulate the user from the data management function.

This version will provide the functionality of Version 1 with the new weighting and scoring methods and parsing changes of Version 2. Other features which will be implemented include new parsing options of retaining duplicate words and maintaining word order and the searching of multiple reference files in the coding process.

Retain Duplicate Words

ACTR Version 1 removes duplicate words in the resulting parsed text. For example, in the Census of Population Mobility question, the response of "**QUEBEC CITY QUEBEC**" currently results in parsed text of "**QUEBEC**" with the removal of "**CITY**" as a trivial word and the reduction of duplicate words, "**QUEBEC QUEBEC**", to a single occurrence. The search for reference file phrases containing this text resulted in over 15,000 cases being found, as every location (towns, cities, villages, etc.) in the province of Quebec has "**QUEBEC**" as part of its text. **ACTR Version 3** will permit the removal or retention of duplicate words.

Maintain Original Word Order

ACTR Version 1 sorts the resulting words from the parsing step in the order of the collating sequence in effect i.e. "**PROGRAM COMPUTER**" will become "**COMPUTER PROGRAM**". In some coding applications there is a requirement to maintain the original order specified. This was quite evident in the Mobility and Major Field of Study variables in the Census of Population. The Mobility coding consists of geographic data and there were many instances where validly different locations resulted in the same parsed phrase e.g. " **SPRING HILL**" and "**HILL SPRING**" appeared equal in **ACTR Version 1.**

The Major Field of Study response often contains a list of courses taken in order of decreasing importance. With the resultant parsed phrase being sorted, this information is lost. The users have been able to circumvent these problems by identifying the specific cases, but this is a tedious process and some cases may be missed.

ACTR Version 3 will permit the maintenance of word order for a given coding application.

Multiple Reference Files

ACTR Version 1 allows access to only one set of reference files in a coding run, although English and/or French reference files may be used. As sources for reference files may vary and require different parsing strategies, it may become necessary to create more than one reference file for a variable. For example, the coding of the Industry variable for the Census could use Company Name data, Kind of Business data and possibly Geographic data to assign a code. These sources have many different characteristics and it would be difficult to implement them in one reference file, applying one set of parsing rules to all data.

ACTR Version 3 will allow the searching of multiple reference files.

3.0 FUTURE OF ACTR

The immediate plan is to produce a first release of ACTR Version 3 for March 1994. This version will be assessed by the 1996 Census of Population for coding

the Industry, Occupation, Relationship to Person 1 and Place of Work responses.

The longer term plans will include enhancements to the computer-assisted coding function. This will round out the applicability of **ACTR** and make it useful in coding applications where the 'automated' coding aspects do not provide the appropriate results. This will include experimenting with the use of phonetics and partial word match algorithms to enhance the computer-assisted coding features.

The feasibility of using **ACTR** to code data at the point of capture, such as in the Regional Offices, will be studied.

The continued work in the above areas promises to increase the potential of **ACTR** in addressing a wider variety of coding needs and environments at Statistics Canada.

BIBLIOGRAPHY

Hellerman, E.
Overview of the Hellerman I&O Coding System
December 1, 1982.

Wenzowski, M.J.
Research Issues in Automated Coding
November 1988.

Research and General Systems Informatics Services and Development
Division
A C T R Version 1.06
Automated Coding by Text Recognition
Documentation Set
March 1989.

Outrata, E., Doucet, J.E.
Strategies for the Use of Newer Technologies in Statistical Organizations
-the Experience of Statistics Canada-
March 1989.

Berljawsky, A.
MARINE/TRADE SYSTEM
*The Use of **ACTR** Automated Coding Software to Capture Vessel Name*
Preliminary Report and Evaluation
June 30, 1989.

Wilkins, R., Ratnasingham, G.
Progress Report
Mortality By Occupation Study
Canada, 1986
September, 1989.

Lavigne, C.
Evaluating the use of ACTR for the 1991 Census of Agriculture
A Summary Report
September, 1990.

Research and General Systems Informatics Services and Development
Division
A C T R Version 2
Automated Coding by Text Recognition
Documentation Set
August 1991.

METHODOLOGY OF AGRICULTURAL SURVEYS IN INDIA

Prem Narain, Indian Agricultural Research Institute
A.K. Srivastava, Indian Agricultural Statistics Research Institute
Prem Narain, I.A.R.I., Pusa, New Delhi – 110 012, India

KEY WORDS: Sample surveys, agriculture, methodology

Introduction

India is mainly an agricultural country with nearly 80 percent of its population residing in villages. Indian agriculture has undergone significant changes during the last three decades. The agricultural statistics system and the related infrastructure available in the country are the result of a sustained effort over a long period. Sample surveys have played a major role in the development of reliable agricultural statistics in the country. With diversity in the nature of various agricultural crops/commodities and with a variety of data needs from planners and users, the pressure on the methodological aspects of agricultural surveys can well be understood. A methodology should be flexible enough to take into account both natural and introduced (such as technological) changes over time as well as the changing data needs for policy formulations. It is in this context that methodological investigations and research have received special attention in the agricultural statistics system of the country. Agriculture is being considered here in its broad sense including crops, livestock and fisheries. There is some amount of commonness in the approach of various agricultural surveys. This commonness is mainly due to similarity in the infrastructural set up for various crops/commodities. This paper deals with the basic approaches of methodologies for estimation of area and production of field crops, minor crops like fruits and vegetable, livestock numbers, livestock products and fish products.

The infrastructural set-up

The infra-structural set up for crop estimation surveys is spread over states as well central agencies. At the state level the different activities relating to planning of surveys, training of field staff, organisation of field work and tabulation of data are done by State Agricultural Statistics Authorities (SASA's) in each state/union territory. These are either the Director of Agriculture or the Director of Land Records depending upon the administrative set up of the State Government. The agencies involved in the collection of crop estimates are the State Department of Revenue, Agriculture, Community Development, Statistics etc., singly or in combination with their normal duties. The designated SASA's coordinate the work to build the crop estimates at the state level and report the same to Central Agency.

At the central level the National Sample Survey Organisation has overall responsibility for assisting all the states in planning and organising the work on crop estimation surveys.

Supervision of field work and improvement of the system of collection of agricultural statistics falls under the purview of their normal functions. The Directorate of Economics and Statistics in the Ministry of Agriculture of the Central Government is the apex organisation for collection of estimates from the State and for preparation and issue of all India crop estimates of area and production.

Surveys for estimation of area and yield of field crops

Production is normally estimated as a product of area and yield rate. This is mainly due to the fact that estimation of area is comparatively simple and reliable due to accurately measured fields and a reliable reporting system whereas estimation of average yield requires carefully planned experimentation on a smaller number of fields. The rationale behind carefully planned crop cutting experiments is to reduce the measurement errors in observing the average yield of crop. The method based on enquiry from farmers has been carefully discussed and rejected in order to avoid possible response errors (both deliberate as well as unintentional).

Area Estimation

From the point of view of area statistics, the entire country may be considered as divided into four categories :

(i) Cadastrally surveyed and possessing primary reporting agency.
(ii) Cadastrally surveyed but not possessing such reporting agency.
(iii) Unsurveyed but possessing primary reporting agency.
(iv) Unsurveyed and also without reporting agency.

For category (i), area statistics are based on complete enumeration through field to field inspection by the primary reporting agency. For category (ii) estimates of area are based on sample surveys using stratified multistage sampling designs. This area constitutes nearly 9.1 per cent of the total reported area. The categories (iii) and (iv) taken together constitute

nearly 9.2 per cent of the reported area and estimates are based on eye estimates. It may be mentioned that acreage statistics of nearly 93.3 percent of the total area is reported. The quality of area statistics varies, based on different methods of collection, but is fairly accurate because the majority is based on complete enumeration.

Estimation of yield rates

The traditional method of estimating yield per unit area, before the advent of crop cutting experiment approach, was to multiply normal yield by the reported condition factor of the crop during the season. The normal yield was defined as yield on average soil in a year of average character. There were set procedures for obtaining normal yield as well as condition factors. However, as subjectivity could not be eliminated, this method was unreliable with unknown amount of margin of error.

The random sampling approach based on crop cutting experiments was based on a series of experimentation spread over several decades. Several questions such as "how to locate a random plot", "what should be the size and shape of the plots", "what should be the agency to conduct the surveys on a large scale and on a routine basis" etc. were raised and tackled successfully. At present the field work is handled by the regular departmental staff of various departments of State Government such as revenue, agriculture etc. The sampling design is a stratified multistage sampling with Revenue Inspector Circles as strata, villages as primary sampling units, fields as second stage units and standard plots (rectangular and square plots of size 10 ×5 and 5×5 sq. meters) as the ultimate sampling units. The number of fields selected in each village is two. Crop cuts are taken when the crops are ready for harvest. The harvested produce from the plot is weighed and a correction factor for driage is obtained on the basis of a small portion (one kg) of the produce. This approach for estimation of average yield is being successfully followed in the entire country on a routine basis.

Surveys for estimation of area and yield of fruits and vegetables

Surveys on fruit crops

The sampling approach developed for the estimation of area and production of various field crops can not be directly applied to trees due to inherent differences in various aspects like sowing, growth period, cultivation practices, harvesting etc. Based on a series of pilot sample surveys, a methodology for estimation of extent of cultivation and production of fruit crops has been developed. However, an initial version of the approach which was suitable for conducting the surveys for a single fruit crop at district level is being presented here to give the essential features of the methodology.

The sampling design is a stratified three stage sampling design with taluks/tehsils (taluks/tehsils are geographical sub–divisions of districts) as strata, villages the primary sampling units, orchards as the second stage units and clusters of trees as the ultimate units of selection. Within each tehsil, villages are classified into two categories according to the latest information available with the primary reporting agency :

Category (i) Villages reporting area under fruit crop under study.

Category (ii) Villages not reporting area under the fruits crop.

Nearly 100 to 150 villages are allocated to the strata roughly in proportion to area under fruits and the selected villages are completely enumerated for fruit trees. A fraction (about 2/5th) of the villages is further selected at random for the purpose of collecting data on yield. In each of the selected villages, a sample of five orchards is selected at random and within each orchard three clusters of four trees each of bearing age are selected for the purpose of yield estimation. A suitable sample from the non–reporting group of villages category (ii) is also selected to determine the extent of fruit cultivation, if any, in this group. The selected trees are observed for multiple harvests as and when the harvesting is done on these trees. The actual observation of the produce from the selected trees becomes inconvenient, sometimes, from the field work point of view, and therefore some understanding with the cultivator is needed for the produce to be recorded accurately. Wherever, a better recording system is available, the approach has got scope for further improvement.

Surveys on vegetable crops

For these crops also while the broad approach of objectivity through random sampling is applicable, the special features of the crops pose some operational difficulties in the conduct of such surveys. Some of the features are the multiple pickings, short duration of the crops and multiple crops.

The sampling design for vegetable surveys at the district level may be described as stratified multistage random sampling. For estimation of area the design is unistage with taluks as strata, and clusters of three villages each forming the sampling units. For estimation of yield rate and production, clusters of villages are psu's, fields growing vegetables are ssu's and plots of size 5 ×5 sq. meters are the ultimate sampling units. For estimation of production of vegetables, a year is divided into 4–6 periods of equal duration. Fields are selected in the beginning of each period from those fields which

are to be harvested during that period. Selected fields are observed for the pickings as and when the picking takes place. This involves considerable time and effort and also restricts the movements of enumerators. Efforts are being made to develop a method for estimation of production of vegetables on the basis of partial harvests only.

Surveys for estimation of livestock numbers and livestock products

One of the main sources for data on livestock numbers is the quinquennial livestock census where the data is collected on the number of different categories of livestock. However, for intercensal periods, the only source is sample surveys conducted in a well planned manner. Moreover, sample surveys are also needed for estimation of livestock products such as milk, wool, meat and eggs. Investigations on these aspects have resulted in methodologies for conducting sample surveys for estimating the livestock numbers and products.

The estimation of a livestock product in a year involves the simultaneous estimation of number of animals/layers as also the average yield per animal/layer per day in the year. The estimate of total annual production can then be obtained as a product of the two, further multiplied by 365. The estimation of each of the first two factors further requires data on number of animals as also the yield spread over the period of entire year as also the entire population. In view of the uncertainties of data obtained by enquiry, it is desirable to obtain data on the above aspects by physical measurements or actual observations on the spot, as far as possible.

The sampling design for estimation of livestock numbers and products is broadly stratified multistage random sampling with groups of districts forming the strata, tehsils/taluks as psu's, and clusters of 2 to 3 adjoining villages as ssu's. In the case of milk and eggs a cluster of 2 to 5 households is the third stage unit. In the case of milk further selection of animals is done within each selected cluster of households whereas in the case of eggs no further selection of layers is done. In the case of wool, all flocks in the selected village are observed.

The design described above is suitable when individual surveys are to be conducted for each product. When interest lies in the simultaneous estimation of all the products in different years, an approach of integrated surveys is adopted. The design in these surveys is more or less the same as described for individual surveys. However, a fraction of primary sampling units in the sample is matched over the season in the three consecutive years. The main livestock product is studied on an intensive scale in a year whereas the other products are covered on a smaller scale so as to provide indices of changes over seasons. Successive sampling procedure *viz.* retention of some psu's and replacement of others over seasons within a year is used when a livestock product (*viz.* milk, egg, wool or meat) is the main product under investigation. Double sampling is used when the product is covered on a reduced scale in a year. Information on more than one product is obtained from the same sample of psu's. Currently, livestock products are estimated through the integrated approach on a routine basis with a simplified version of the sample design described above.

Surveys for estimation of fish catch

Efforts for the development of methodology for estimation of fish catch started with a pilot study for estimation of marine fish catch during 1950–51. Based on pilot studies a methodology was developed and further improvements were subsequently made at the Central Marine Fisheries Research Institute which is responsible for research on marine fisheries. Presently, these surveys are conducted by the respective State Governments. The sampling design used in these surveys is as follows :

The entire coast line is divided into a total of 54 zones or strata which vary from State to State. Each zone is divided into two sub-strata based on intensity of fishing. Three landing centres are selected from each sub-stratum. The time stratum in the survey is a month, which is further divided into three 10 day–periods. For the first time strip of 10 days, 6 consecutive days are selected with a random start and with an interval of 10 days, six days are selected systematically from each of the remaining two time strips. Of the three landing centres selected in a sub–stratum, the first one is observed on the first two days, the second one on the next days and the third on the last two days. A working day (6 a.m. to 6 p.m.) is divided into four intervals of three hours each. Information pertaining to catch and count is recorded during two time intervals on the first day and the remaining two time intervals on the second day. A pre-determined fraction of total number of boats landing in a time interval is selected systematically for the collection of data on composition of fish catch.

Methodologies have also been developed for estimation of fish catch from inland resources including fresh water, brackish water and riverine resources.

Conclusions

The paper provides a broad overview of the methodologies of sample surveys of agriculture in India. It may be observed that although the methodologies differ because of the specific problems of individual surveys, they are similar in the basic ap-

proach. The emphasis is on the objective approach of measurements on items for which collection of data through other methods (like enquiry) may lead to sizeable response errors. Another important aspect of these methodologies is that the available infrastructure of data collection and analysis have been taken into account while developing the methodologies. This is essential for viability and adaptability of the methodologies. An important aspect of the methodological research is that it is a continuous process in which the theoretical and other technological changes must be taken into account simultaneously. A continuous interaction amongst the producers of the data, users of the data and research workers is therefore a basic requirement for methodological research in sample surveys.

Selected References

Narain, P., Goel, B.B.P.S. and Garg, J.N., 1985. Handbook on methodology of sample surveys for estimation of livestock numbers and products. I.A. S.R. I. (ICAR), New Delhi.

Narain, P., Kathuria, O.P. and Srivastava, A.K., 1985. Estimation of extent of cultivation and production of fruits and vegetables in India. Proc. 45th Session International Statistical Institute. Amsterdam Vol. 51 Book 2, 477-478.

Panse, V.G., 1963. Thirty years of statistics in Agriculture in India. ICAR review series No. 27.

Panse, V. G. and Sukhatme, P.V., 1948. Crop surveys in India-I, Jour. Indian Soc. Agril. Statistics 1 (1), 34-58.

Singh, D., Manwani, A.H. and Srivastava, A.K., 1976. Survey on fresh fruits in Tamil Nadu. ICAR, New Delhi.

Sukhatme, P.V., Panse, V.G. and Sastry, K.V.R., 1958. Sampling technique for estimating catch of fish in India. Biometrics 14. 78-96.

QUALITATIVE TESTING OF THE FARM FINANCIAL SURVEY QUESTIONNAIRE

David Lawrence and Frances Laffey, Statistics Canada
David Lawrence, Questionnaire Design Resource Centre, Statistics Canada, Ottawa, Tel. (613) 951-9003

KEY WORDS: Focus groups, questionnaire testing, measurement error

1. Introduction

Statistics Canada conducted a research project using qualitative techniques in the development and evaluation of the questionnaire for the 1993 *Farm Financial Survey (FFS)*. This paper provides an overview of the project methodology, including highlights of the findings and recommendations for improving the survey questionnaire. The research findings provide useful insights into the questionnaire design and the application of qualitative techniques in the development and testing of establishment surveys.

2. Background

The Farm Financial Survey is an annual survey conducted jointly by Agriculture Canada and Statistics Canada. The main purpose of the survey is to collect up-to-date information on the financial situation of Canadian farm operators. The data are used to identify new opportunities to provide credit to farm operators and to analyze the financial situation of Canadian farmers. The information also is used to estimate probable costs of various policy alternatives.

The survey population consists of approximately 260,000 farming operations. A sample of about 12,000 farms is selected nationally. The selected sample is stratified by province, farm-type, and farm-size (as defined by sales). Survey estimates are produced by the stratification characteristics for the variables of interest.

Key variables of interest on the questionnaire include: Assets, Liabilities, Sales, Income, Capital Invested and Capital Borrowed. Some general profile information about the farm and farm operator(s) also is collected. Data collection takes place from January to March. Selected farm operators are contacted by telephone to arrange a personal interview.

A similar survey, the *Farm Credit Corporation Survey (FCC)*, had been conducted on an ad hoc basis five times between 1980 and 1992. The *FCC Survey* had experienced high non-response rates (approaching 20 percent) compared to other farm surveys. The *Farm Financial Survey* team was concerned about the non-response (refusal) rates and the associated respondent burden of changing the survey to an annual collection. As part of the development of the 1993 *Farm Financial Survey*, personal interviews and focus groups were conducted with farm operators from three regions of Canada. The goals of the research were to evaluate the existing questionnaire (*Farm Survey, 1992*) and to review the data collection procedures.

The Questionnaire Design Resource Centre (QDRC) of Statistics Canada conducted the research. The fieldwork was carried out during July and August 1992. The interviews and focus groups provided insights into questionnaire and data collection issues as well as suggestions for improving procedures. The study identified sources of measurement error and provided ways to improve the "respondent-friendliness" and "interviewer-friendliness" of the questionnaire.

3. Study Overview and Methodology

The objective of the research study was to determine farm operators' perceptions of the *Farm Financial Survey* and to discuss ways of addressing their concerns. The research was designed to assess respondents' ability to report the desired information, to evaluate respondents' understanding of the questions, to test the general flow of the questionnaire, and to assess respondents' concerns regarding the sensitivity and confidentiality associated with reporting financial information.

Specifically, the interviews and focus groups investigated the following issues: availability and ease of reporting information as requested, sensitivity of financial information, ability to report information by calendar year *versus* crop year, clarity and respondent comprehension of questions, understanding of survey terms and concepts, questionnaire format and layout, incentives to encourage respondent participation, timing of the survey, concern about confidentiality, and respondent burden.

The first step involved a critical review of the questionnaire by the QDRC. The review provided insights on issues that may be problematic for respondents. The review also allowed the QDRC to become familiar with the subject matter.

The field work for the study was conducted at three locations across Canada to reflect the opinions of farmers in different regions. The three centres selected were Moose Jaw, Saskatchewan; St. Anselme,

Québec; and Kentville-Wolfville, Nova Scotia.

While coverage could have been improved by extending the research to other regions of the country, time and cost constraints did not permit this. The research study used qualitative techniques. Although the results do not reflect a statistically representative sample, the study produced important findings regarding the questionnaire.

The study used two qualitative techniques: one-on-one interviews and focus groups. Participants were contacted by telephone and first invited to participate in a personal interview. Respondents were then asked to attend a group discussion to be held a day or two later. A letter outlining the study and its details was mailed to operators who agreed to participate.

One-on-one interviews

The purpose of the one-on-one interview was to familiarize the respondent with the 1992 questionnaire before the group discussion. Personal interviews were conducted with farmers at their homes or farm operations. Respondents were asked to respond to the survey questions as they normally would.

The one-on-one interviewing technique allowed the researchers to administer the questionnaire in the field and to directly observe any difficulties or problems experienced by the respondents while answering the questions. The interviews for this study usually took between one hour and one and a half hours to complete.

In several cases during the study the researchers encountered operations where the farm was jointly run, usually by a husband and wife. In these situations, a paired interview was carried out. The paired interview involved both the farm operator and spouse (or partner) participating in the interview. During the paired interview the respondents frequently consulted each other before providing a response. Sometimes this involved recalling a value from memory, other times computer records or personal files were used. Respondents conferred with each other to decide what a survey question was asking, as well as to help recall items such as sales and loans incurred during the reference period. Paired interviews were useful in observing the process that respondents go through to formulate answers for the farm survey. Twenty-six one-on-one interviews were conducted for the study. In six cases participants were available only for personal interviews and not for the focus group discussions. In these cases, an **in-depth interview** was conducted with the respondent. After administering and observing the respondent complete the questionnaire, the researcher used a discussion guide to solicit information about the respondent's perceptions and impressions about the questionnaire. The follow-up discussion covered any difficulties or problems experienced by the respondent regarding such issues as: the availability of information, sensitivity and confidentiality of financial data, the ability to report information, and the clarity and respondent comprehension of questions. The in-depth interviews usually took about two hours.

Focus groups

Focus groups are a qualitative technique that bring a group of people together to discuss selected topics. The group is led by a moderator who promotes discussion among the participants on the subject of interest. The moderator uses a discussion guide of general topic areas. Typically the respondents do not know each other, but share common attributes that relate to the topic of interest. Discussion is interactive as participants influence each other by responding to ideas presented from the group.

Three focus groups were conducted for the study. For each group, the moderator reviewed the *Farm Survey* questionnaire, question by question. The participants were asked to share their thoughts on completing the form and discuss areas of difficulty or ease. The focus groups provided several suggestions and recommendations for improving the questionnaire. The one-on-one interviews and focus group sessions were observed by Statistics Canada and Agriculture Canada representatives. The focus groups were audio-taped. Each focus group lasted about two hours.

4. Recruiting Participants

Recruiting farm operators was a key component of the study design. It was felt that farmers may be reluctant to participate due to the potentially sensitive nature of the information and because the study would involve consulting them twice. Several steps were taken to encourage participation:

- Respondents were told of the importance of their participation toward the collection of quality agriculture information.
- Respondents were told that their help would directly influence the content of the questionnaire.
- The confidentiality of the discussions was stressed.
- An honorarium of $50 as well as several agricultural publications were given to the participants.
- The personal interviews were conducted in the participants' homes or at their farming operation.

5. Highlights of the Study

This section of the paper presents some of the study highlights. These items reflect the perceptions and comments expressed by the participants during the interviews and group discussions. Observations regarding the implementation of the questionnaire also are included.

In general, the *Farm Survey* questionnaire was well received by participants. The farm operators did not appear to have any difficulties with the survey terms or concepts. Respondents found the questionnaire straightforward and easy to answer compared to many of the forms they receive. Several participants noted that they had enjoyed the group discussions and interviews. Team members felt the study helped promote the "personality" of Statistics Canada and Agriculture Canada from a public relations perspective.

Response Burden

Participants did not feel that the questionnaire was difficult to complete. However, they did emphasize their concerns regarding the burden placed on respondents to complete farm-related questionnaires and forms.

The burden arises not only from federal and provincial government surveys and forms, but also from requests for information from private companies and financial institutions. There is a perception among farmers that a lot of the requested information is being asked time and time again. The participants felt that much of the required information is already available from some government department.

Understanding Survey Terms and Definitions

In general, participants experienced no problems with the terminology on the questionnaire. Participants felt the terminology was quite clear compared to other surveys. However, they felt that Revenue Canada and Statistics Canada used a lot of different definitions for financial information.

In St. Anselme, some of the French terms on the questionnaire were ambiguous, or presented interpretation difficulties for several participants.

For example, the term *bestiaux* is used to mean "livestock". Many of the participants felt bestiaux was related specifically to "beef", as in "beef cow". The term *animaux* was thought to refer more generally to livestock.

Availability of Requested Information

Most of the reporting difficulties experienced by participants occurred in the section of the questionnaire on assets of the farming operation:

Question 1:
What was the value of the following assets of this farm operation as of December 31, 1991?

- ***Report** at current market value and to the nearest $1000*

a) Farmland and buildings
b) Breeding and replacement stock
c) Machinery and equipment
d) Quota
e) Accounts receivable
f) Inputs
g) Crops for sale
h) Market livestock
i) Cash, bonds, savings, stocks, shares, RRSP's
j) Other farm assets
k) Non-farm assets

Several participants did not have the survey reference date *(December 31)* as a year-end. Many found it difficult to relate to imposed survey reference dates. Participants said they reply for their personal year-end, regardless of what the questionnaire requests. Respondents also found it difficult to report *Assets* according to the current market value. Many farmers said that they know the depreciated value, but not the current market value. Most agreed that they were reporting the depreciated value or the insured value. Participants had difficulties assessing the value of machinery and equipment: much of the machinery was old, but still functional. The respondents thought the market value of such items was negligible; however the equipment would be extremely expensive to replace. They wondered what the appropriate answer should be.

Interviewers frequently had to clarify the meaning of *"quota"* with respondents to determine whether the question referred to dairy or poultry quota. Also, some provinces include the quota in the value of land and buildings. Some other provinces report it separately.

One participant asked: "What if the car is used two-thirds of the time for farm business and one-third of the time for personal use?" If the operator has a vehicle strictly for family use, is this considered a non-farm asset?

The operating arrangements of some farms provide interesting findings regarding the ability to report some required information: a three-way partnership was encountered where the participant said that he could answer the operation-related questions accurately. The respondent noted that he

also could report the "non-farm" information for only his portion of the operation. He could not provide any estimate of non-farm information for his two partners. Furthermore, he indicated that his partners would refuse to answer the entire questionnaire.

Issues Influencing Participation in the Survey

Participants felt that when they answer a survey, some form of remuneration is deserved. Most are very interested in receiving survey results. The participants prefer to receive this information directly from the collecting agency, as opposed to reading an excerpt of the results in a local newspaper. Information on their farm or farm-type compared to similar farms in the province or country is useful to them.

Respondents felt very strongly that the survey be conducted in the "off season". The first quarter of the calendar year is the best time to contact respondents. The books are usually finalized by the latter part of January or early February. Respondents have a growing concern about telephone interviews. Several participants provided examples of scams and bad experiences they have had with telemarketers and telephone solicitations. Respondents appreciate a "warm" first contact such as a telephone call to inform them of the survey. Participants prefer the personal interview for actual administration of the questionnaire. Most stated they would not report any financial information over the telephone.

The personal interviews usually lasted between one hour and two hours. Most participants did not mind the time it took to complete the questionnaire. As one respondent stated: *"It felt more like a social visit ... ."*

The operators emphasized the importance of not personally knowing the interviewer. There were two reasons for this opinion: first, from a sensitivity and confidentiality perspective, participants felt uncomfortable about providing detailed information to interviewers that they personally knew; second, there was a feeling that, even if interviewers were bound by strict confidentiality guidelines, they still know personal information about them that the general public does not have. It was suggested that this might allow the interviewers to use the information to their own advantage.

The professionalism and friendliness of the interviewer has a definite impact on the respondent. The decision to participate is made based on the first contact with the respondent. The participants were agreeable to completing the survey annually for the proposed four-year period when a sampled operation would be in the survey. They could see the importance of collecting information over time.

Privacy and Confidentiality Issues

Participants were skeptical regarding the confidentiality of their reported information. However, respondents felt that Statistics Canada treats data more confidentially than the banks and other financial institutions. This perception is largely associated with the relative "remoteness" of the agency, as opposed to the practices of Statistics Canada to protect the confidentiality of respondents. Participants are uneasy about claims of confidentiality.

Participants were reticent to provide information that they felt did not apply to the farming operation. They wondered what their personal Registered Retirement Savings Plans (RRSP's) and inheritances had to do with operating the farm.

Literacy

Two situations occurred where the participant was illiterate or had difficulty reading or understanding the survey material. There is no indication of how prevalent this condition is among farm operators on the Farm Financial Survey sampling frame. It should be noted that such conditions exist and may play a part in a respondent's reluctance to participate in a survey.

Administration of the Questionnaire

There were several difficulties encountered by the researchers while administering the questionnaire in the field. Some of the more salient points are described below.

General Observations

The questionnaire should contain more "interviewer aids" to improve its interviewer-friendliness. Questions could be clarified with more examples or with more explicit "**Include**" or "**Exclude**" statements. Instructions intended for the *interviewer only* and instructions intended to be *read aloud* to the respondent should be distinguished from the text of the questions. This might be done with different fonts or by enclosing the instructions in shaded areas on the questionnaire.

The existing interviewer procedure guide provided little assistance. The guide presented difficulties for the research team in the field due to ambiguous and incorrect information. The manual should be redesigned to include items such as: an overview of the survey and the role of the interviewer, special

concerns or situations the interviewer may encounter, and how to respond to typical questions asked by the respondent.

The existing questionnaire is administered by a personal interview, but the document is designed like a business form. The document does not help the interviewer or respondent with the transition from one section to the next. A brief introduction for each section would help focus the interviewer and respondent on the information to be next asked. It is preferable to provide interviewers with the complete questionnaire with the exact wording. This helps control the consistency of wording and reduces the chance of interviewers improvising their own words.

Several questions seem appropriate as they are printed on the questionnaire, but when the questions are administered in a personal interview, do not flow in a clear and orderly fashion.

For example, in *Section G: Income & Expenses*, Question 2 asks: *In 1991, what was the total gross farm revenue of this operation before expenses?* Question 3 follows, asking: *Of the amount in question 2, what was the amount from ...* The respondent does not understand the association between the amount previously reported and the fact that the question was Question 2.

Section B. Physical Characteristics Of This Operation

Question 2:
For the total and cultivated land area of this operation at December 31, 1991, how much was:
a) Owned by this operation
b) Rented from others
c) Owned but leased to others

The question is double-barrelled (area of total land and area of cultivated land), and is difficult to implement, if asked verbatim. The interviewer frequently had to backtrack and re-ask each component. A different question structure would help alleviate this problem. Questions 2(a) and 2(c) are not mutually exclusive. A respondent may find it difficult to report the area of cultivated land for another operator. Even if the respondent could answer the question, it is not relevant to ask the amount of cultivated land for another operation.

Section C. Capital Investments and Sales

Question 1:
During 1991, did this farm operation invest any money in capital items or improvements, receive any money from the sale of capital items, or receive any capital through gifts or inheritances?

This question is actually three questions rolled into one. Respondents were completely baffled by the question. The question should be divided into three questions. Also, it is important to provide the respondent with examples of types of capital information requested. Several respondents asked if computers are considered capital investments.

The respondents did not understand why questions about "non-farm" finances were being asked. These questions caused some concern with all participants.

The location of these questions (at the end of Sections C, D, and E) tended to disrupt the flow of the questionnaire. Questions regarding cash, bonds, inheritances, etc. were often thought **not** to be related to the farm. Participants preferred not to answer these questions, or gave a quick "None" as a response. It was recommended that the non-farm questions be grouped together and placed in a section at the end of the questionnaire. The questions should then be sequenced from the least sensitive question (non-farm income) to the perceived most-sensitive question. The rationale for asking these questions should be explained in detail to all interviewers and covered in the interviewer training procedures. Interviewers should be prepared to probe respondents to determine the precise answer to these questions. Respondents are reluctant to report "non-farm" data; they regard this information as personal and not related to the farm operation.

Section E. Liabilities Outstanding and *Section F. Capital Borrowed*

Questions in this section focused primarily on how much money was owed by the operation, which lenders were owed the monies, and the term of the loan(s). The question was structured as a matrix of cells making the question appear cluttered and intimidating. The question was cumbersome to administer since usually only one or two cells were completed for each respondent. The main question in this section pertained to the amount of long-term credit borrowed and from which lender. When administering the questionnaire, these questions seemed repetitive. It was suggested to merge questions from each section in order to improve the flow of the document and reduce the burden on the respondent.

There was some problem distinguishing between provincial and federal loan agencies in Québec. The Québec provincial government does not lend money, it only guarantees loans.

Section G. Income and Expenses

Respondents often required clarification for two questions concerning payments to and withdrawals from support programs. Respondents frequently reported information for payments due for the crop year. The questions were intended to collect information for the calendar year.

6. Conclusion

The qualitative study provided the researchers and observers with in-depth knowledge on the perceptions and attitudes of the farm operators regarding the Farm Financial Survey. Through direct observation, the research provided insights into the response process of respondents and into areas of the questionnaire that may contribute to measurement error. Issues such as respondent burden, availability of information, privacy and confidentiality, and problems encountered administering the questionnaire are areas where steps can be taken to reduce potential sources of error.

Several of the study's recommendations had a direct impact on the survey design. Several modifications were incorporated directly on the 1993 *Farm Financial Survey* questionnaire. For example, all questions regarding "non-farm" finances were grouped together and placed in a new section at the end of the questionnaire; the triple-barrelled question regarding Capital investments and sales was split into three separate questions; and the section regarding Capital Borrowed was reduced in size and merged with the section on Liabilities Outstanding.

Changes were implemented in the layout of the questionnaire to help improve the interviewer-friendliness and the flow of the document. Several new instructions have been written for the interviewers as well as interviewer edits for reported data for areas of land use and capital borrowed and total liabilities. Other recommendations from the study are being considered for improving the questionnaire in 1994.

Acknowledgements

The authors wish to acknowledge the participation and contributions of Suzelle Giroux, Guy Laflamme, Laurent Roy, Phil Stevens, and Patricia Whitridge, Statistics Canada, and John Caldwell and Alfred Cho-Chung-Hing, Agriculture Canada.

Reference

Statistics Canada (1992), *Farm Financial Survey, Final Report on Focus Groups and Personal Interviews with Farm Operators*, Questionnaire Design Resource Centre.

USDA'S ANNUAL FARM COSTS AND RETURNS SURVEY: IMPROVING DATA QUALITY

by Bob Milton and Doug Kleweno, Estimates Division, National Agricultural Statistics Service
Rm. 5912-South Building, 14th & Independenc Ave., S.W., Washington, D.C. 20250

KEY WORDS: finance survey, data quality, data sharing, respondent burden

Survey Description and Use of Data

The Farm Costs and Returns Survey (FCRS) is a comprehensive farm finance survey conducted annually by the National Agricultural Statistics Service (NASS) for the U.S. Department of Agriculture (USDA). In total, some 1,300 data items are collected when all questionnaire versions of the FCRS are considered. Information on crop and livestock production, farm expenses, income, debt, assets, and socio-economic and demographic data are collected.

Information from the survey is the basis for USDA estimates of farm expenditures, income, cash flow, wealth, costs of production, and productivity. The FCRS is an integrated survey that provides information on the farm sector, household, business, and enterprise (for major farm commodities). Information from the survey is provided at the U.S. and regional levels and by type and size of farm. Size of farm is determined by value of annual sales. Much of this information is published routinely by USDA's Economic Research Service in its series Economic Indicators of the Farm Sector and in their Situation and Outlook reports. NASS also publishes detailed expenditure data annually from the FCRS.

The FCRS provides the only annual data set at the U.S. level for farm financial, production, and related information. The FCRS data base is used by ERS in analyzing numerous farm program and policy issues annually for USDA and other policy makers.

Survey Design

The FCRS is a multiple frame, probability survey of U.S. farms. The sample size over the past 5 years has averaged about 24,000 farms, just over one percent of all farms. A farm is defined as any establishment from which $1,000 or more of agricultural products are sold or could be sold during the year. Types of establishments included in the survey are those listed in the Federal Government Standard Industrial Code (SIC) for agricultural production of crops and livestock - major group codes 01 and 02.

Samples are selected from two sources. The first source is a list of operators of farms and ranches. Control data on type of farm and size are used to stratify the list. The list frame represents the larger, more specialized operations. The second source is an area frame where the continental United States is divided into small area sampling units, each with a known probability of selection. The area frame sample focuses on collecting data on smaller operations, less than $20,000 in annual sales, plus larger operations that are not on the list. Data for the area frame operators not on the list are used to measure the incompleteness of the list.

The survey is designed to provide reliable data at the regional level which represents 10 geographic groups of States with similar production practices. At the U.S. level, the coefficient of variation (C.V.) is about 2-5 percent for major expenditure and income items. C.V.'s at the regional level are generally in the range of 10 to 20 percent. The extent of nonsampling errors is not known. To minimize nonsampling errors, data collection procedures are uniform and consistent across the Nation by using extensive training and field supervision of data collectors.

The FCRS is designed to provide estimates of several types of information. Accordingly, several versions of the FCRS questionnaire are used to collect the types of information. Depending upon the questionnaire version, additional data are collected on cost of production for specific commodities on a 4-5 year rotation, on socio-economic and demographic data, and on detailed expenditure and income data. All questionnaire versions have basic income and expenditure questions so that all questionnaires are additive to generate certain basic financial information. The different questionnaire versions provide additional independent estimates of specific information depending upon questionnaire purpose.

Survey Problems and Data Quality

The largest obstacle confronting the FCRS evolves around the large amount of detailed data collected from a shrinking population of farmers.

Some 1300 separate data items are collected annually on the FCRS. Many of the these items are related to the costs of production surveys where minute detail is needed in constructing costs of production budgets and models.

The more detail collected, the greater the respondent burden becomes. Average interview time for the 1990 survey was nearly 1 1/2 hours overall (Rutz and Cadwallader, 1991). The average interview time for the 1990 cow-calf costs of production questionnaire version was nearly 2 hours and interviews of 3-4 hours were common (Appendix Table 1).

The interview time requirements for the FCRS is a major reason the survey response rate is relatively low (10-20 percentage points lower) compared with other NASS surveys and continues to erode (Appendix Maps 1 and 2) (Rutz and Cadwallader, 1991). Over the past five years (1987-92) the response rate for the FCRS has fallen from 73 to 66 percent. In research conducted on reasons for nonresponse to the 1990 survey, one-fourth of all refusals indicated they would not take time to complete the survey (Appendix Table 2) (O'Connor 1992). The overall refusal rate for the 1991 survey was 25 percent, but was as high as 33 percent for the corn costs of production questionnaire version. In two States, the overall refusal rate was above 50 percent. The response rate is also lower among the large farms. The response rate for the largest farms sampled from the list frame, farms with annual sales over $500,000, for 1990 was 57 percent compared with 69 percent for all farms (Appendix Table 3) (Rutz and Cadwallader, 1991). Field offices have also indicated that large farms have a greater tendency to refuse in the future once having completed a lengthy interview.

The higher level of nonresponse for the large farms is particularly critical with regard to data adjustment for nonresponse. Data are adjusted for nonresponse at the strata level within State by the ratio of good responses plus inaccessible and refusal samples to good responses. In many cases this adjustment more than doubles the expansion factor for responses from the largest farms, annual sales of over $500,000. This strata of farms accounts for only two percent of all farm numbers but over two-fifths of total farm expenditures and gross income.

Beginning with resummarization of the 1991 data, the nonresponse adjustment was modified so that all refusal and inaccessible samples were assumed to have positive farm data (Turner, 1992). Field enumerators were instructed to verify that refusal and inaccessible samples had positive farm data, some type of crop or livestock production. The modified adjustment removed the count of operations without positive farm data, out of scope operations, from both the numerator and denominator. The resulting larger nonresponse adjustment factor increased the expansion of total U.S. expenditures and income by about 9 percent. The increase due to the change in the nonresponse adjustment was greater than what was assumed before research proved otherwise. The greatest increase occurred in the upper strata, or large farm classes, where it had been assumed that there were fewer screenouts or out-of-business operations.

The nature of the FCRS, to collect personal financial data, is another major contributing factor to the relatively lower response rate on the FCRS. Beyond no reason given, the nonresponse research indicated that the second most frequent reason for refusing to complete the survey questionnaire was that the information was too personal. Besides the 25 percent that refused the initial interview, refusals or "don't knows" to some questions accounted for as much as 15-16 percent of expanded data for some items, specifically value of farm assets and landlords' share of government payments (Appendix Table 4) (Morehart and Johnson, 1992). On average, expanded data for refusal items amounted to 1-2 percent. For refusal or "don't know" items, data are imputed by combining all U.S. data into one file and calculating average by type and size of farm for the missing items. This level of imputation occurs after the raw survey data are considered "clean".

A thorough clerical and machine edit is also run on the raw data as it is received, prior to the imputation edit. Research on this edit concluded that the edit has little effect on the final results and that the small effects are accounted for by a very few reports (Hoge and Willimack, 1991). Nearly half of the edits move respondent data to the proper cell with little or not effect on data expansions. The same is true for detailed editing for incomplete allocation of aggregate reported data.

Once the machine edit is completed, the data are summarized and an outlier review takes place. An outlier is defined as a report whose expanded data account for 5 percent or more of the regional total or one-half of one percent of the U.S. total for major data items (Statistical Methods Branch, 1992). The

outlier adjustment process moves the report to the largest operator stratum where all large operations have the same expansion factor. If it is further determined by the outlier review board that the extreme operation is unique in itself or is similar to only a few operations, the expansion factor is further reduced.

One additional adjustment is made to the FCRS data to ensure complete farm coverage. Data are adjusted by sales class at the regional level by the ratio of FCRS expanded number of farms to estimated USDA number of farms (Statistical Methods Branch, 1992). The area frame expansion for the FCRS has been historically based upon a sample of resident farm operators. This expansion of farm numbers is generally about 15 percent below the official estimates. The reason for the incompleteness of the farm coverage from the area frame is largely due to the inability to pick up farm operators, especially small operators, in the urban and suburban land units, segments. Adjustment for undercoverage of farms was initiated with the resummarization of the 1991 data and added about 3 percent to total expenditure expansions.

Future Direction on the FCRS

Future direction on the FCRS should focus on increasing response rates. Of utmost importance to increasing response rates is reducing interview time. Preliminary plans are to expand the use of the aggregate expenditure questionnaire version that eliminates the detail or breakout of component expenditures from the group total and collects no commodity costs of production data. The interview length was reduced by about one-fourth hour for the aggregate questionnaire compared with the detailed expenditure questionnaire during 1992 tests. Expenditure data for the farm operation that is part of the cost of production questionnaire version will also be collected at only the aggregate level.

Expanded use of a global short version expenditure questionnaire to the operational level also fits within future plans. This questionnaire of 16 pages is even more abbreviated in length than the aggregate version. A global short version questionnaire was tested in the farm finance follow-up to USDA's Chemical Use Survey in 1992. Preliminary review of data from this global short version is promising with regard to collecting data at the aggregate rather than component level. Response rates for the global version were significantly higher than for the FCRS in the two States conducting the farm finance follow-up survey. In Louisiana, the response to the global short version was 76 percent compared with 65 percent for all versions of the FCRS and in Minnesota the response was 62 percent compared with 56 percent for the FCRS. Much of this increase in response rates is however attributable to the screening out of refusal and out-of-business operations before arriving at the sample size for the farm finance survey.

The high level of respondent burden on the larger farms due to frequent contacts for a variety of surveys causes a need to concentrate on sampling schemes that will reduce the number of contacts. Sampling plans are being considered that integrate the needs of several surveys with one sample selection using basically nonreplacement sampling of strata that meet the needs of all the surveys. Preliminary post-survey research analysis covering four major surveys in three States during 1991-92 indicates a potential reduction in individual respondent burden, or number of multiple contacts, of 60 percent (Preliminary research by NASS researchers Dr. Charles Perry and Jim Burt).

Current list building activities should enhance the sampling work. List building activities this year concentrated on trying to improve coverage on farms with annual sales of $100,000 or more. In 1992, list coverage at the U.S. level for farms with annual sales of over $100,000 was 89.3 percent (Geuder, 1992). The goal for 1993 is to improve the coverage of these larger farms to 95 percent. List concentration on adding large farms and improving their control data should enhance sampling and improve data accuracy due to better overall stratification and coverage.

Farm coverage for the FCRS should improve for the 1993 survey due to the switch to a weighted area estimator. Since all area tracts (separate operations within the land segment) and not just resident operator tracts will be eligible for selection, the sampling universe will be larger, reducing respondent burden for resident farm operators and possibly improving response rates. Data for selected area tracts will be expanded based upon the ratio of land within the tract to land in the entire operation. This weighted estimator reduces the undercoverage bias due to missed area frame farms, especially farm operators living near or in urban and suburban areas, because data are associated with the location of the farm rather than the location of the operator's residence. Starting with the 1993 survey, the current

procedure of adjusting data for farm coverage by the ratio of estimated number of farms by sales class to survey expanded number will be reevaluated.

An important factor in improving response rates that needs more consideration is the perception of the survey by the field enumerators conducting the face-to-face interviews. Enumerators play an important, if not the most important, role in obtaining survey response. Most respondent decisions to participate are heuristically based (Groves, Cialdini, and Couper, 1992). Enumerator experiences and expectations affect their ability and motivation to maintain interaction with the respondent. If the FCRS is presumed to be too much of a respondent burden, the questions too personal or too difficult in nature, and data of marginal value to users, response to the survey will suffer (Allen 1993). This situation can be addressed by getting the questionnaire length to a manageable level, providing additional training to enumerators, and "selling" the survey to the enumerators and public.

A task group has been formed within NASS to investigate the low response rate on the FCRS. The task group believes that field enumerators need additional, more specific, training to better handle potential refusal and inaccessible (by respondent choice) respondents. Role playing and special case situations need to be a basic part of training. Enumerators need more training on interviewing techniques, scheduling, and on the purpose and need for the survey.

Above all, field enumerators need to be convinced of the importance of the survey in order to "sell" it. NASS management in Headquarters and the States need to make additional efforts to demonstrate the importance of the FCRS to enumerators. This starts with more public relations work on the FCRS. Studies have shown that public relations more focused to gain the support of groups identified with and respected by the target population are helpful (Slocum, Emply, and Swanson, 1956). Historically, FCRS response rates for sugarbeet growers have been higher than other commodity groups because the industry visibly endorsed and encouraged cooperation. States need to work more with the industries, producers, and media throughout the year on the importance of the FCRS.

NASS is also researching incentives as inducement to improve response rates. Pocket calculators were given out on a trial basis to a portion of the FCRS sample in four States during the 1992 survey. An evaluation of the incentive research has not been completed to date; however, initial results suggest some improvement in response rates. Concerns over the effects on participation in other voluntary surveys have been raised.

Data Sharing

Another issue that is related to survey response is the confidentiality of the survey data relative to its use. Recently, NASS received a ruling from USDA's Office of General Counsel (OGC) on interpretation of the statutes governing sharing of individual record data such as that provided by the FCRS. OGC's interpretation of the statutes allows data sharing to other agencies, universities, and private entities as long as it enhances the mission of USDA and is through a contract, cooperative agreement, cost reimbursement agreement, or Memorandum of Understanding. Such entities or individuals receiving the data are also bound by the statutes restricting unlawful use and disclosure of the data.

It will be NASS policy that data sharing will occur on a case by case basis as needed to address an approved, specified USDA or public need. NASS and ERS have the responsibility to assure data providers that use of the data will be for public good only. NASS will explore opportunities to broaden the use of cooperative agreements with universities and other government agencies. Access to each data set provided to the cooperative party will need to be properly certified as to the confidential aspects of that data set and regulations. Data sets shared by NASS will be used on-site in USDA facilities and will also be returned or destroyed after meeting the specified need. To improve data access, NASS plans to make the FCRS data available to qualifying entities at two of its State offices on a trial basis in 1993.

Summary

The FCRS is a probability farm finance survey that produces the only annual comprehensive U.S. data set available that combines farm financial, production, and related information. The survey is the basis for USDA estimates for farm expenditures, income, cash flow, costs of production, and productivity. The detailed and personal nature of the survey is the major reason for the relatively low response rate.

During the past year, data adjustments for nonresponse and undercoverage have been modified to improve quality of expanded data. Nonresponse and respondent burden problems are more concentrated among the large farms who account for the majority of expanded data. In order to improve response rates, future efforts will focus on sampling schemes that reduce the reporting burden on large farms, shortening the length of the questionnaire to lessen respondent burden, providing more training to field enumerators in handling reluctant respondents, and publicizing the survey more to gain public acceptance. In order to improve access to the FCRS data set, NASS will make the data available to qualifying entities at two State office sites during 1993 on a trial basis.

REFERENCES

1. Allen J. Donald. (1993). "The Interviewer and the Interviewing Process", NASS Staff Report SMB-93-02.

2. Geuder, Jeff. (1992). "1992 NASS List Frame Evaluation", NASS Staff Report SSB-92-02.

3. Groves, R. M., Cialdini, R. B., and Couper, M. P. (1992). "Understanding the Decision to Participate in a Survey", Public Opinion Quarterly, 56:475-495.

4. Hoge, Stanley J. and Willimack, Diane K. (1991). "Analysis of Item Nonresponse, Imputation and Editing in the 1989 Farm Costs and Returns Survey for Iowa and North Carolina", NASS Staff Report SRB-91-09.

5. Morehart, Mitchell J., Johnson, James D., and Banker, David E. (1992). "Financial Performance of U.S. Farm Businesses, 1987-90", ERS Agricultural Economic Report No. 661.

6. O'Connor, Terry P. (1992). "Identifying and Classifying Reasons for Nonresponse on the 1991 Farm Costs and Returns Survey", NASS Staff Report SRB-92-10.

7. Rutz, Jack L. and Cadwallader, Chris L. (1991). "1990 Farm Costs and Returns Survey, Survey Administration Analysis", NASS Staff Report SMB-91-04.

8. Slocum, W. L., Emply, L. T., and Swanson, H. S. (1956). "Increasing Response to Questionnaires and Structured Interviews", American Sociological Review 21:221-225.

9. Statistical Methods Branch. (1992). "1992 Farm Costs and Returns Survey Specifications".

10. Turner, Kay. (1992). "Modification of FCRS Nonresponse Adjustment Procedures", NASS Staff Report SRB-92-08.