Survey Methodology, June 1990
Vol. 16, No. 1, pp. 117–129
Statistics Canada

Sample Design of the 1988 National Farm Survey

C. JULIEN and F. MARANDA[1]

ABSTRACT

The National Farm Survey is a sample survey which produces annual estimates on a variety of subjects related to agriculture in Canada. The 1988 survey was conducted using a new sample design. This design involved multiple sampling frames and multivariate sampling techniques different from those of the previous design. This article first describes the strategy and methods used to develop the new sample design, then gives details on factors affecting the precision of the estimates. Finally, the performance of the new design is assessed using the 1988 survey results.

KEY WORDS: Multi-purpose sampling; Multiple frame; Area frame; Multivariate stratification.

1. INTRODUCTION

The National Farm Survey (NFS) is a probability-based sample survey focussing on several subjects related to agriculture in Canada. It is conducted annually in June and July in all provinces except Newfoundland, where a separate survey is carried out.

The previous NFS sample design, dating from 1983, was based on the results of the 1981 Census of Agriculture. A description of it may be found in Ingram and Davidson (1983). However, since 1981 the farm population has changed significantly, reducing the effectiveness of this design. Furthermore, the requirements of the survey have changed somewhat over the years, resulting in the need to update the samples.

A new sample design was therefore developed based on the results of the 1986 Census of Agriculture, and became operational in the summer of 1988.

2. OBJECTIVES OF THE SURVEY

The primary objective of the survey is to provide timely, reliable estimates of levels and annual trends for over 100 agriculture variables. Essentially, these variables may be divided into three categories: cropland areas for the current year; livestock numbers on July 1; and receipts and operating expenses for the previous calendar year. In terms of reliability, the objective of the survey is to obtain coefficients of variation (CV) below 5% at the provincial level for the major parameters.

Survey data are normally summarized to the provincial level. However, primarily for analysis purposes, results for sub-provincial regions are also produced using domain estimation methods.

Another important objective of the survey is to obtain a master sample from which subsamples are chosen for use in other farm surveys conducted by Statistics Canada.

[1] C. Julien is a methodologist with the Census Data Quality and Analysis Section, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6; F. Maranda is chief of the Agriculture Survey Methods Section, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

3. TARGET POPULATION AND SURVEY POPULATION

The target population includes all farms in the provinces surveyed which received $250 or more from the sale of agricultural products during the 12 months preceding the survey. Also included are farms which do not meet the $250 criterion at the time of the survey, but which expect to earn at least this sum during the 12 months following the survey. Such farms, which either began operating just prior to the survey or are temporarily inactive, are relatively few in number.

The survey population, or the group from which the sample is selected, excludes farms operated by institutions as well as those located on Indian reserves or settlements. The terms institution, Indian reserve and Indian settlement are defined in Statistics Canada (1987, pp. 115-117, 145, 152). The cost-benefit ratio associated with collecting data on these types of farms is very high. Because of this, they are excluded in order to enable more efficient use of the resources available for the survey. The contribution of such exclusions to national agricultural production is small and is estimated using adjustment factors which are based on Census data.

4. SAMPLING FRAMES AND THEIR USE

In theory, the survey population is divided into two groups, the first of which includes the farms enumerated in the Census and the second all other farms. These include the undercoverage from the Census and so-called new farms, that is, those which began operating after the Census.

The first group is covered all or in part, depending on the province, by one or two list frames created from the list of census farms. To complement the list frames and ensure complete coverage of the survey population, an area frame, created from the agricultural enumeration areas (EAs), is used. An enumeration area is the geographical region enumerated by a census representative. Furthermore, an EA is said to be agricultural if it contains at least one census farm. An area frame is needed to compensate for the shortcomings of the list frames, particularly their difficulty to identify new farms.

The estimation requirements of the survey and the characteristics of agriculture in Canada vary by region. To better account for these variations, the territory covered by the survey is divided into three regions and a different sample design is used in each one. The three regions involved are: the Prairie provinces and the Peace River district in British Columbia; Quebec and Ontario; and, finally, the Maritime provinces and the rest of British Columbia. The first of these regions is called the Canadian Wheat Board (CWB) region, since the entire region comes under the jurisdiction of this organization.

The total sample size in each of the three regions is essentially based on the overall budget available for data collection. Within each region, sample allocation among the various provinces and, where applicable, among the various frames, depends on several factors. The primary ones are the square root rule applied to the size of the survey population, historical allocations in the survey, and the results of various analyses centred on the expected precision of the estimates.

4.1 The Canadian Wheat Board Region

In this part of Canada, two list frames and one area frame are used in each province.

The first list frame (L1 list) essentially includes the large and medium-sized census farms in relation to key crop, livestock and expense variables. This list is obtained using an iterative process which consists in establishing a threshold for each key variable and including in the

list all farms that exceed at least one of these thresholds. Each threshold is adjusted separately upward or downward so that the L1 list, once completed, includes approximately 35% of the survey population's farms and accounts for 50% to 90% of the total agricultural activity, depending on the key variable in question. These percentages are used because experience has shown that the resulting list is composed of farms which, individually, are more stable over time than the rest of the farms in the survey population. This stability leads to the creation of strata which remain homogeneous over the years, which is a factor in maintaining the efficiency of the sample design.

In each province, the L1 list is then stratified within sub-provincial regions based on nine key variables. A sample of farms is selected and used to obtain data on crops and livestock. Because data on expenses are more difficult and costly to collect, only a sub-sample, called the core sample, is used to obtain this information.

The second list frame (L2 list) includes all census farms with more than 20 acres of cropland which were not included in the L1 list. The L2 list is stratified within crop districts based on a single key variable, namely, cropland area at the time of the Census. The L2 list is used to complement the L1 list for preliminary crop data. These data must be collected within very tight deadlines which, for operational reasons, cannot be met using the area frame.

The area frame includes all agricultural enumeration areas, except those on Indian reserves and in the so-called marginal agricultural regions, that is regions with little agricultural activity. Marginal regions are found mostly in the northern parts of the provinces and in urban fringes. The few census farms located in marginal regions are added to the L1 list, since it is the only list used to collect data on all survey variables.

The area frame is stratified using the same sub-provincial regions and key variables as the L1 list. It ultimately produces a sample of segments which are delineated on topographic maps. The identity of the farmers operating land in one of these segments is obtained through on-site enumeration. Manual matching of names and addresses then enables detection of segment farms overlapping one of the list frames. This detection is essential because each time the area frame is used to complement a list frame, only those segment farms that do not overlap the list in question are used, thus ensuring that the list and area frames represent mutually exclusive domains.

Complete information is required on all segment farms except those overlapping the L1 list, as the data for this list are obtained from the sample selected from it.

4.2 Quebec and Ontario

In each of these provinces, a single list frame, called L1, and an area frame are used.

The list frame is composed of all census farms in the survey population. The methodology used in sampling from this list is similar to that used for the CWB region L1 list, apart from two differences. First, incorporated farms, or farms founded as business corporations, are separated from the other farms, and strata are created independently within the two groups. This preliminary separation is performed because only incorporated farms are required to report their expenses in the survey, since the expenses of the non-incorporated farms are obtained from Revenue Canada tax records. It should be noted that the confidentiality of these records is completely protected under the Statistics Act. Second, sub-sampling for expenses is unnecessary because less than 25% of the farms in the survey population are incorporated.

The area frame and its sample design have not been modified following the last Census, due to a lack of resources. Only the marginal regions were updated, resulting in their enlargement.

4.3 Maritime Provinces and the Rest of British Columbia

In each province of this region, the sample design includes only one list frame, again called L1, which is made up of all census farms in the survey population. Given that a list frame tends to deteriorate with time and that there is no area frame to supplement it, it becomes more difficult to completely cover the survey population. However, because of the relatively small number of farms, under 30 000 in these provinces, more complex procedures were implemented to keep the list up-to-date. Notably, farms which were missed in the Census or which began operating following it may be detected through these procedures. Thus, for all practical purposes, the list frame is considered to ensure full coverage of the survey population.

In each province of this region, the list is stratified and a sample of farms is selected using the same approach as in Quebec and Ontario. All the estimates required are produced from this sample.

5. LIST SAMPLING TECHNIQUES

Samples are taken from the list frames using a one stage, statified sample design where the farms constitute the sampling units. The strategy and methods used to develop this design are essentially the same, regardless of the province and list involved. However, the combination of methods and key variables used may vary from case to case.

The first step consists in identifying the farms with distinct characteristics and in automatically including them in the sample. There are essentially two kinds of these so-called self-representative or take-all farms. The first group includes those with a unique operating structure such as community pastures and multiholding corporations, while the second group contains the farms which clearly stand out from the majority because of their very large contributions to key crop, livestock and expense variables. Due to the skewness (to the right) of the distributions involved, complete enumeration of these farms is an efficient way to reduce sampling variance.

Farms with very large contributions are identified through an intuitively-based rule which produced good results in the previous sample design. This rule, called the sigma-gap rule, is applied separately to each key variable using all farms having a non-zero value for the variable in question. Farms with a sufficiently high contribution to one of the key variables, as determined by this rule, are said to be take-all.

The sigma-gap rule, as adapted to the survey, functions as follows. Given a univariate distribution of points x_i, $i = 1, 2, \ldots, N$, $x_i > 0$ for all i, and given σ as its standard deviation, the points are arranged in increasing order $x_1 \leq x_2 \leq \ldots \leq x_N$; for the half of the distribution to the right of the median, the distance between each successive pair of points $d_i = x_i - x_{i-1}$ is determined; given i_o, the smallest i for which $d_i \geq \sigma$, all points $i \geq i_o$ correspond to take-all farms. If $d_i < \sigma$ for all i, no point in this distribution distinguishes itself sufficiently from the others to be declared a take-all farm.

The second step consists in dividing the rest of the farms in the list into take-some strata. In most cases, the strata are formed within sub-provincial regions according to nine key variables representing the usual three categories: crops, livestock and operating expenses. The number of variables in each category is one, six and two respectively.

The underlying principle to the stratification is as follows. Each farm is characterized by nine variables, and neighbouring farms, defined in terms of Euclidian distance, are grouped together. Two multivariate clustering algorithms are used for this purpose. These algorithms are called FASTCLUS and CLUSTER, since they are available in the procedures of the same name in the SAS statistical analysis software package (version 5).

The FASTCLUS algorithm divides a set of observations into a predetermined number of mutually exclusive clusters. First, the algorithm chooses observations which serve as initial cluster seeds. Each observation is then assigned to the nearest seed, and once this is completed, the cluster seeds are updated by the means of the clusters thus formed. The process is repeated until the changes in the seeds become minimal. The FASTCLUS algorithm is based on work by Hartigan (1975) and MacQueen (1967).

The CLUSTER algorithm groups a set of observations into mutually exclusive clusters in a hierarchical structure. Initially, each observation forms a cluster in itself. Based on a technique inspired by Ward (1963), the two most similar clusters are combined into one, which subsequently replaces them. The process is repeated until only one cluster remains. Massart and Kaufman (1983) provide an introduction to this type of classification. Thus, the set of observations is broken down into as many partitions as there were observations to begin with, and each partition corresponds to a stratification.

These algorithms are used successively as follows. FASTCLUS is used first to group the farms into 250 clusters, which are then progressively combined to form the strata using CLUSTER. Initial classification is performed with FASTCLUS, since using CLUSTER directly with a high number of records would require excessive computer time.

Each of the three categories of variables must contribute equally to strata formation. To ensure this, the initial stratification variables are transformed so that the sum of the transformed variables in each category has a mean 0 and a predetermined variance, usually 1. The crop category with its single variable may be standardized in the usual manner by subtracting its mean and dividing by its standard deviation. In each of the other two categories, two successive transformations are performed independently. Given X_i, the initial variables of a given category C, a principal components analysis was performed to obtain transformed variables Y_i. These new variables, with mean μ_i and variance σ_i^2, are linear combinations of the former ones and mutually independent. The Y_i are then standardized to obtain final stratification variables Z_i as follows:

$$Z_i = \frac{Y_i - \mu_i}{\left(\sum_{i \in C} \sigma_i^2\right)^{1/2}}.$$

Thus, the mean and variance for $\sum_{i \in C} Z_i$ are 0 and 1 respectively.

An empirical approach is used to determine the number of strata. Several stratifications and allocations are performed by varying the number of strata. Then, the coefficient of variation curve is drawn as a function of the number of strata for all key variables and many others. These curves generally resemble Figure 1. Stratification gains are considered to have been virtually fully attained at the point where the majority of curves are practically horizontal. The number of strata chosen is a compromise between this point and the desire to avoid forming too many strata so as to attenuate the effects of incorrect initial classification and stratum jumpers over time, two major causes of outliers or influential observations.

Sample allocation is multivariate and is generally carried out using the same key variables used for stratification. The allocation algorithm consists in minimizing a linear combination of the square of the coefficients of variation of the key variables, within the constraint of a fixed total sample size. Given c_i, coefficient of variation for a key variable, $a_i > 0$ as constant and n_o total sample size, $\sum a_i c_i^2 = f(n)$ must be minimized within the constraint $n = n_o$. The algorithm used is described in Bethel (1986). Adjustments are then made to obtain a minimum sample size of 4 and a maximum weighting factor of 50 in each stratum.

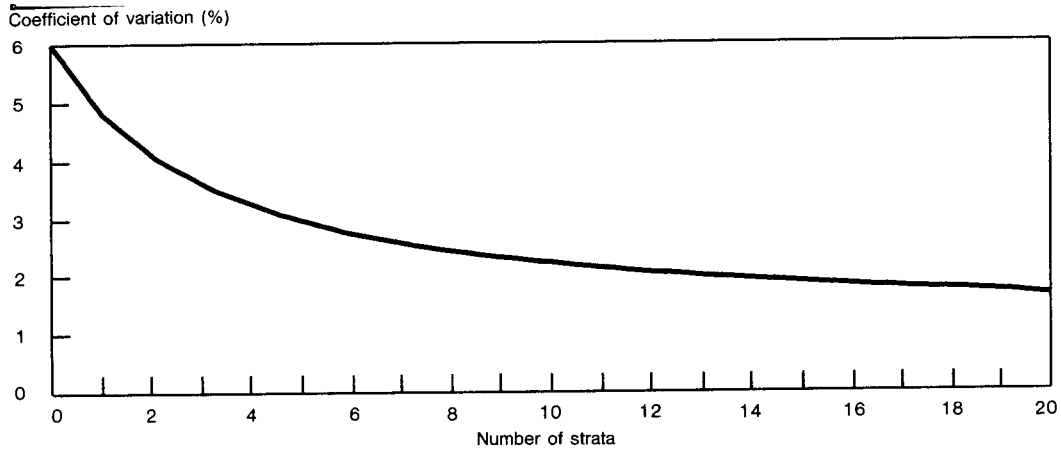

Figure 1. General Curve of the Coefficient of Variation as a Function of the Number of Strata

Finally, once allocation has been completed, the farms are sorted within each stratum by sub-provincial region and total operating expenses and a sample is selected using circular systematic sampling. For the L1 list in the CWB region, the complete sample is chosen first; the core sub-sample is then selected from it using circular systematic sampling.

6. AREA SAMPLING TECHNIQUES

Area samples are selected according to a two-stage stratified sample design. The Census enumeration areas and segments represent the primary and secondary sampling units respectively.

Given that the area sample design has not been modified for Quebec and Ontario, the following paragraphs apply only to the CWB region.

The first step consists in measuring the agricultural activity in each of the frame's EAs by summarizing to the EA level the data for the census farms not included on the L1 list. Excluding the L1 list farms from the summarization process produces EA distributions which accurately reflect the characteristics of small farms. Subsequent use of these distributions enables an area sample complementing the L1 list with respect to small farms to be selected with greater efficiency.

Once the summarization process has been completed, each EA is treated as a farm for sampling purposes. The EA selection strategy and methods are very similar to those applied

to the CWB region L1 list. First, take-all EAs are determined using the sigma-gap rule. The remaining EAs are then allocated to take-some strata within sub-provincial regions using the CLUSTER multivariate clustering algorithm. Preliminary classification with FASTCLUS is unnecessary in this case due to the relatively low number of EAs, never more than 3000 per province, to be processed. Furthermore, the usual standardizations suffice for transforming the key variables. A principal components analysis was not used because the area frame's contribution to provincial estimates does not justify such an approach.

Allocation to strata is performed with the same algorithm used for the list, and the minimum sample size is again established at 4. The sample size is then divided by four in each stratum, and four separate replicates are selected using circular systematic sampling. Replicates facilitate variance calculation, as a single secondary unit is often chosen per primary unit.

Once the EAs have been selected, their boundaries are traced on topographic maps and they are divided into segments of approximately 7.5 km^2 (3 mi^2). Natural boundaries such as roads and rivers are used as much as possible to facilitate the work of field interviewers. Simple random sampling without replacement of the segments is performed at a minimum rate of 1 out of 30 in each selected EA. There are, however, some exceptions to the rule: additional segments are taken so that the overall weighting factor does not exceed 180; a minimum of two segments are selected in each EA belonging to the strata subjected to first-stage complete enumeration; and, finally, when the same EA appears in more than one replicate, measures are taken to avoid selecting the same segment more than once. Nevertheless, these exceptions are rare.

7. RESULTS OF THE SAMPLE DESIGN

Table 1 contains the results of the list frame sample design. The following items are included: the number of farms in the list (N); the number of strata (H); the number of farms in the sample (n); and, finally, the number of farms in the core sub-sample (n-core) in those provinces where it applies.

Table 1

Results of the List Frame Sample Design

Province	L1 List				L2 List		
	N	H	n	n-core	N	H	n
P.E.I.	2,830	26	451				
N.S.	4,273	35	550				
N.B.	3,544	39	498				
Quebec	41,380	80	6,096				
Ontario	72,598	78	8,401				
Manitoba	6,712	48	1,364	490	18,058	29	2,267
Saskatchewan	15,668	48	3,625	1,106	45,798	41	4,573
Alberta	13,928	63	2,981	909	38,504	25	2,973
B.C. (Peace)[a]	494	25	190	190	1,187	6	170
B.C. (rest)[b]	17,042	41	1,999				
Total	178,469	479	26,155	2,695	103,547	101	9,983

[a] Peace River district in British Columbia.
[b] British Columbia minus the Peace River district.

Table 2 contains the results of the area sample design in those provinces where such a design is used. The following items are indicated: the number of EAs in the frame (N); the number of strata (H); the total number of EAs sampled (n); the number of EAs sampled where each EA is counted only once when it appears in more than one replicate (n-once); and, finally, the number of segments chosen (m).

8. FACTORS AFFECTING THE PRECISION OF THE ESTIMATES

To better appreciate the results obtained from the 1988 survey, three factors affecting the reliability of the estimates must be discussed. These factors are the sample size, the treatment of the total non-response and the estimation methodology.

First, the sample size for the L1 list in the CWB region was reduced by 10% in relation to that of the corresponding list used in the previous sample design. This reduction was prompted mainly by the desire to lower costs.

Second, the methodology used to treat total non-response was modified in 1988. Previously, when a farm failed to respond to the survey, its data were imputed using the data from another farm in the same stratum. These imputed data enabled the sample to be completed to its original size. However, in 1988, the cases of total non-response were not imputed; instead only the respondent sample was used and the weighting factors adjusted upward. The actual sample is therefore reduced in relation to the former method.

In the 1988 survey, the total non-response rate varied between 2% and 13%, depending on the province. The national rate was 10%. Non-response rates are presented in detail in Table 3.

Table 2
Results of the Area Sample Design

Province	N	H	n	n-once	m
Quebec	2,065	43	191	182	230
Ontario	2,687	49	195	185	259
Manitoba	794	21	277	264	305
Saskatchewan	1,496	26	328	308	477
Alberta	1,623	32	328	319	434
B.C. (Peace)[a]	54	7	36	32	58
Total	8,719	178	1,355	1,290	1,763

[a] Peace River district in British Columbia.

Table 3
Total Non-response Rate (%) by Province

Province	Refusals	No Contact	Total
P.E.I.	0.00	3.55	3.55
N.S.	0.00	2.18	2.18
N.B.	0.00	1.61	1.61
Quebec	1.71	6.56	8.27
Ontario	2.27	11.11	13.38
Manitoba	3.45	4.03	7.48
Saskatchewan	4.06	6.46	10.52
Alberta	2.68	7.95	10.63
B.C.	1.78	10.28	12.06
Total	2.32	8.11	10.43

The last factor to be discussed is the estimation methodology. The usual estimators corresponding to a stratified simple random sample are used for list frames. For area frames, an estimator described in Wolter (1986 pp. 19-26) and corresponding to a sample design with independent replicates is used. Provincial estimates are obtained by adding the contribution of the list and area frames since, as previously mentioned, these two frames are independent and represent mutually exclusive domains. Details on the estimation methodology are found in Lynch (1988).

9. ASSESSING THE PERFORMANCE OF THE NEW DESIGN

To assess the performance of the new design, the precision of the estimates obtained in 1988 is compared first to that of the 1987 survey, then to the precision anticipated during the development of the sample design.

9.1 The 1988 and 1987 Surveys Compared

Two opposite tendencies are in effect in a comparison of the precision of the estimate in the 1988 and 1987 surveys. The 1988 estimates should be more precise because the 1987 sample design was already four years old. However, the two sample size reduction factors described in section 8 would indicate less precise estimates for 1988.

Precision is compared using the coefficient of variation of the provincial estimates obtained by combining the L1 list and area frames. The estimates used are those for several key variables whose coefficient of variation in 1987 did not exceed 20%.

The precision of 234 estimates is compared in the charts in Figure 2, where each square represents the CV achieved in 1987 on the x-axis and achieved in 1988 on the y-axis for a given estimate. The frequency (as a percentage) of the key variables located within each zone delineated by the straight lines $Y = X/2$, $Y = X$ and $Y = 2X$ is also presented.

Nearly 60% of crop estimates were more precise in 1988 than in 1987. The majority of those that were less precise were so to a small degree only. Close to 95% of livestock estimates were more precise in 1988 than the previous year; in fact, 32% of the estimates were even twice as precise. Finally, over 60% of operating expense estimates were more precise in 1988. Some of the 1987 estimates were a good deal less precise, and 7% were even two times less precise. The latter are from Quebec and Ontario, where data on operating expenses are collected from incorporated farms only. Further more, the legal status of a farm in these provinces is difficult to identify, both in the Census and the survey.

Despite the reduction in the effective sample due to total non-response and cutbacks during the sample design development stage, the 1988 survey generally provided more precise estimates for each category of variables.

9.2 Precision Obtained Versus Precision Anticipated

The precision obtained is expected to be inferior to the precision anticipated for two reasons. First, when the weighting factors are adjusted to account for the total non-response, the variance increases slightly. Second, the data used to create the sampling frame were taken from the 1986 Census of Agriculture. These data are subject to error and the sampling frame deteriorates with changes in agricultural activity.

Precision is compared using the coefficient of variation of L1 list frame provincial estimates only. These estimates are for several key variables whose anticipated CV did not exceed 20%.

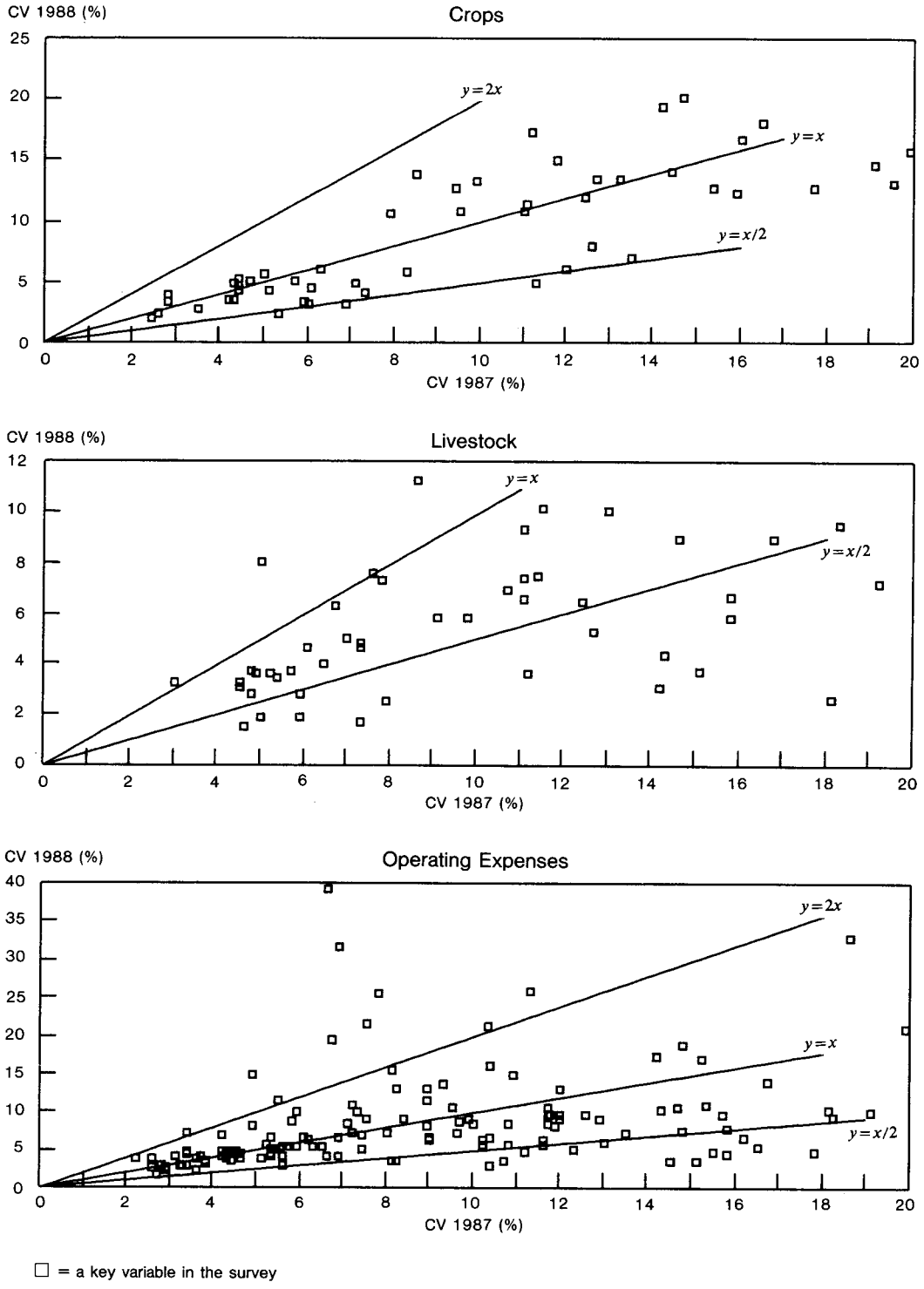

Figure 2. Comparaison of the Precision of Key Variable Estimates in the 1987 and 1988 Surveys by Category of Questions.

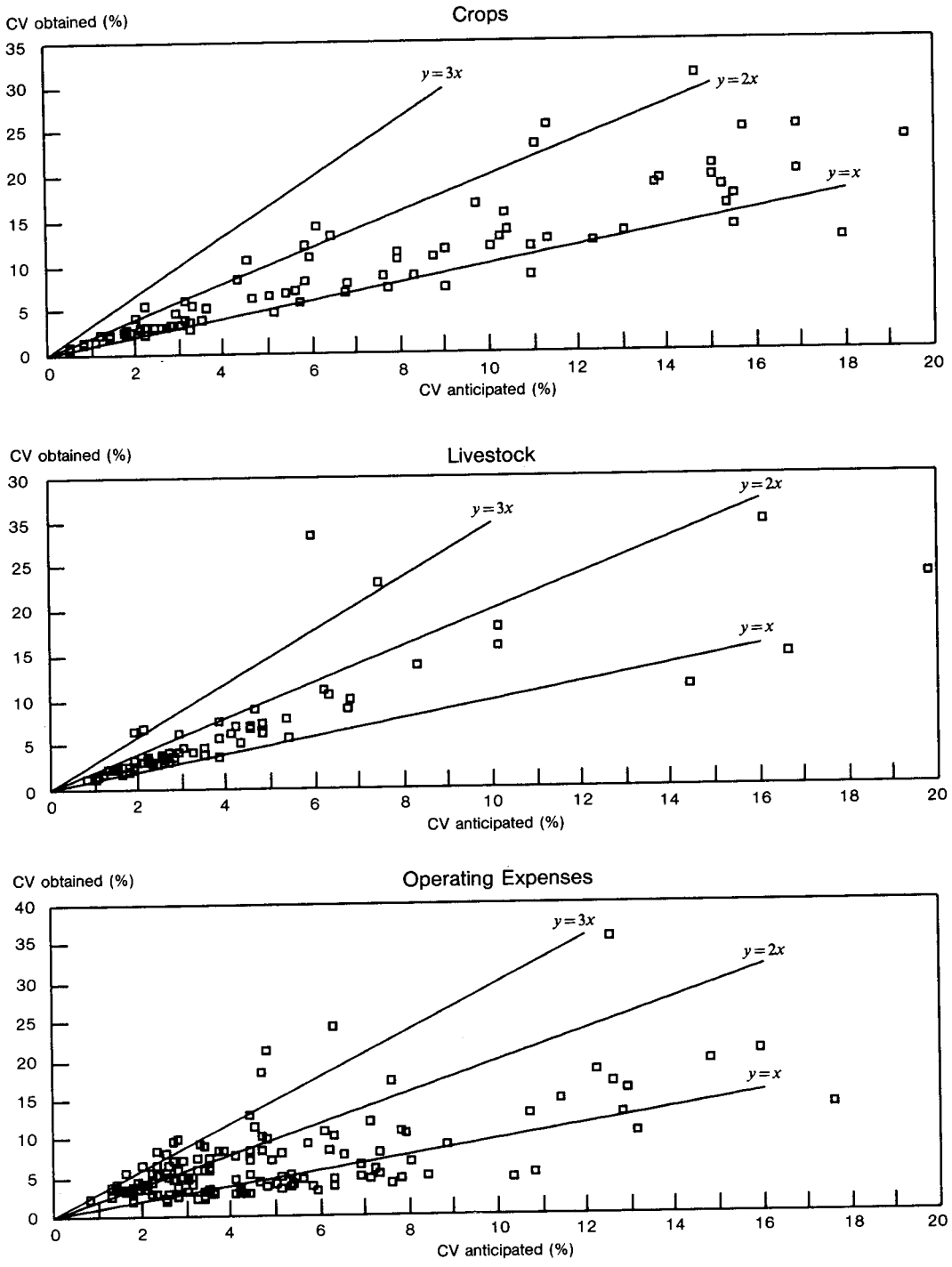

Figure 3. Comparison of the Precision of Key Variable Estimates Obtained in the 1988 Survey and the Precision Anticipated during Development of the Sample Design.

A comparison of the precision of 288 estimates is presented in chart form in Figure 3. In these charts, each square represents the anticipated CV on the x-axis and the obtained CV in 1988 on the y-axis for a given estimate. The frequency (as a percentage) of the key variables located within each zone delineated by the straight lines $Y = X$, $Y = 2X$ and $Y = 3X$ is shown in the charts.

For the crop and livestock categories, approximately 90% of the estimates are sufficiently precise, given the non-response rate, as most of the key variables are located closer to straight line $Y = X$ than to straight line $Y = 2X$. Two tendencies can be seen for the operating expense estimates. Surprisingly, the CV obtained is lower than the anticipated CV in 28% of the cases, the vast majority of which are found in the CWB region. However, 31% of all estimates are more than two times less precise than anticipated. These cases are found in Quebec and Ontario for the reasons given in section 9.1.

Finally, a complementary study was conducted in which the precision obtained was compared to the anticipated precision based on the size of the sample actually observed. This study revealed that the frequency of estimates at least two times less precise than anticipated dropped from 12% to 5% for crops, from 9% to 5% for livestock and from 31% to 7% for operating expenses.

These studies show that in general the precision obtained is acceptable and differs from the anticipated precision mainly because of the treatment for total non-response. This indicates that the sample design is therefore sound and the L1 list frame is adequate. On the other hand, less precise estimates were obtained for operating expenses due to a problem in identifying incorporated farms in Quebec and Ontario in the Census and in the survey. Finally, the list frame, which was two years old at the time of the survey, was observed to have deteriorated somewhat due mostly to bankruptcies and farm sales.

10. CONCLUSION

In general, survey results were substantially improved following implementation of the new sample design. Moreover, the reduction in sample sizes led to cost savings and a considerable reduction in the response burden on the farmers surveyed. Difficulties remain, however, especially regarding the operating expense variables for incorporated farms in Quebec and Ontario. Further studies to resolve these difficulties are being envisaged.

ACKNOWLEDGEMENTS

The authors would like to thank the editor of the journal and the referees, whose valuable comments helped to improve this article.

REFERENCES

BETHEL, J. (1986). An optimum allocation algorithm for multivariate surveys. Technical report of the United States Department of Agriculture, Statistical Reporting Service, Statistical Research Division, No. SF and SRB-89.

GERMAIN, M.-F., DOLSON, D., and MARANDA, F. (1989). Redesign of the 1988 National Farm Survey. Internal working document, Business Survey Methods Division, Agriculture Section, Statistics Canada.

HARTIGAN, J.A. (1975). *Clustering Algorithms*. New York: John Wiley and Sons.

INGRAM, S., and DAVIDSON, G. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 220-225.

LYNCH, J. (1988). Cas spéciaux d'estimation dans l'enquête nationale des fermes de 1988. Internal working document, Business Survey Methods Division, Agriculture Section, Statistics Canada.

MacQUEEN, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability*, 1, 281-297.

MASSART, D.L., and KAUFMAN, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley and Sons.

SAS Institute Inc. (1985). *SAS User's Guide: Statistics*, Version 5 Edition. Cary, NC: SAS institute.

STATISTICS CANADA (1987). 1986 Census Dictionary. Catalogue 99-101E, Statistics Canada.

WARD, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.