Techniques d'enquête, juin 1990
Vol. 16, n° 1, pp. 127–139
Statistique Canada



# Le plan de sondage de l'enquête nationale sur les fermes de 1988

C. JULIEN et F. MARANDA[1]

RÉSUMÉ

L'Enquête nationale sur les fermes est une enquête par échantillonnage qui produit des estimations annuelles sur une variété de sujets reliés à l'agriculture canadienne. En 1988, l'enquête a été dotée d'un nouveau plan de sondage. Ce nouveau plan fait intervenir des bases de sondage multiples et des méthodes d'échantillonnage multidimensionnelles qui sont différentes de celles du plan précédent. Dans cet article, on décrit d'abord la stratégie et les méthodes utilisées pour le nouveau plan de sondage. Ensuite, on apporte des précisions sur quelques facteurs qui affectent la précision des estimations. Enfin, on évalue la performance du nouveau plan suite à son utilisation.

MOTS CLÉS: Échantillonnage à buts multiples; base multiple; base aréolaire; stratification multivariée.

## 1. INTRODUCTION

L'Enquête nationale sur les fermes est une enquête par échantillonnage probabiliste portant sur plusieurs sujets reliés à l'agriculture canadienne. Elle est menée annuellement en juin et juillet dans toutes les provinces à l'exception de Terre-Neuve où une enquête séparée est effectuée.

L'ancien plan de sondage de l'enquête datait de 1983 et il était basé sur les résultats du Recensement de l'agriculture de 1981. On en retrouve une description dans Ingram et Davidson (1983). Or, depuis 1981, la population des exploitations agricoles a subi plusieurs changements qui ont entraîné une perte d'efficacité de ce plan. De plus, les exigences de l'enquête ont quelque peu changé au cours des ans et il était devenu impératif d'apporter des corrections aux échantillons.

Pour ces raisons, on a développé un nouveau plan de sondage. Ce nouveau plan s'appuie sur les résultats du Recensement de l'agriculture de 1986 et il est devenu opérationnel à l'été 1988.

## 2. OBJECTIFS DE L'ENQUÊTE

L'objectif premier de l'enquête consiste à produire des estimations actuelles et fiables reflétant les niveaux et les tendances annuelles de plus d'une centaine de variables agricoles. Ces variables se répartissent essentiellement en trois catégories: les superficies ensemencées de l'année en cours; la taille des cheptels au premier juillet; et enfin les recettes et dépenses d'exploitation pour l'année civile précédente. En termes de fiabilité, l'enquête vise des coefficients de variation inférieurs à 5 % pour les paramètres importants à l'échelle provinciale.

Les données de l'enquête sont normalement agrégées au niveau des provinces. Cependant, surtout pour des fins analytiques, l'enquête produit également des résultats pour des régions infraprovinciales au moyen de méthodes d'estimation par le domaine.

Un autre objectif important de l'enquête est de procurer un échantillon maître duquel on choisit des sous-échantillons servant à d'autres enquêtes agricoles effectuées par Statistique Canada.

[1] C. Julien est méthodologiste, Section de la qualité des données et de l'analyse du recensement, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6; F. Maranda est chef, Section des méthodes d'enquêtes agricoles, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6.

## 3. POPULATIONS CIBLE ET ENQUÊTÉE

La population cible comprend toutes les fermes des provinces enquêtées dont la vente de produits agricoles s'est chiffrée à 250 dollars ou plus au cours des 12 mois précédant le début de l'enquête. Font également partie de la population cible les fermes qui ne satisfont pas au critère du 250 dollars en date de l'enquête mais qui anticipent réaliser au moins cette somme au cours des 12 mois suivant l'enquête. Ces dernières sont relativement peu nombreuses; ce sont des fermes qui ont débuté leurs activités juste avant l'enquête ou qui sont temporairement inactives.

La population enquêtée, c'est-à-dire celle qui est effectivement échantillonnée, exclut les fermes exploitées par les institutions ainsi que les fermes situées dans les réserves ou établissements indiens. Les termes institution, réserve indienne et établissement indien sont définis dans Statistique Canada (1987, pp. 115-117,145,152). Le rapport coûts-bénéfices associé à la collecte des données pour ces types de fermes est très élevé. Ainsi, on les exclut afin de permettre une meilleure utilisation des ressources consacrées à l'enquête. La contribution des exclusions à la production agricole nationale est faible et on l'estime en utilisant des facteurs d'ajustements calculés à partir des données du recensement.

## 4. BASES DE SONDAGE ET LEUR UTILISATION

En théorie, la population enquêtée se répartit en deux groupes, le premier renfermant les fermes qui ont été dénombrées au recensement et le second toutes les autres fermes. Ces autres fermes correspondent au sousdénombrement du recensement et aux fermes dites nouvelles, c'est-à-dire celles dont l'exploitation a débuté après le recensement.

Le premier groupe est couvert, en tout ou en partie selon la province, par une ou deux bases de liste formées à même le registre des fermes du recensement. Pour compléter les bases de liste et assurer une couverture complète de la population enquêtée, on a recours à une base aréolaire qui est créée à partir des secteurs de dénombrement (SD) agricoles. Par secteur de dénombrement, on entend la région géographique qui est dénombrée par un agent recenseur; de plus, un secteur est dit agricole s'il renferme au moins une ferme au recensement. Le recours à une base aréolaire est nécessaire afin de pallier aux lacunes des bases de liste, notamment en ce qui concerne leurs difficultés à détecter les nouvelles fermes.

Les exigences de l'enquête en matière d'estimation et les caractéristiques de l'agriculture canadienne varient selon la région. Pour mieux tenir compte de ces variations, on divise le territoire couvert par l'enquête en trois régions et on utilise un plan de sondage distinct dans chacune d'entre elles. Les trois régions concernées sont les suivantes: les provinces des Prairies et le district de la rivière de la Paix en Colombie-Britannique; le Québec et l'Ontario; et enfin les provinces Maritimes et le reste de la Colombie-Britannique. La première de ces régions est appelée région de la Commission canadienne du blé (CCB) car c'est le territoire auquel s'étend la juridiction de cet organisme.

La taille totale des échantillons dans chacune des trois régions est établie essentiellement à partir du budget global disponible pour la collecte des données. À l'intérieur de chaque région, la répartition des échantillons entre les diverses provinces et, selon le cas, entre les diverses bases, dépend de plusieurs facteurs. Les principaux facteurs en jeu sont la règle de la racine carrée de la taille de la population enquêtée, les répartitions historiques de l'enquête et les résultats de diverses analyses portant sur la précision espérée des estimations.

### 4.1 Région de la Commission canadienne du blé

Dans cette partie du Canada, on utilise deux bases de liste et une base aréolaire dans chaque province.

La première base de liste, notée L1, comprend essentiellement les grandes et moyennes fermes du recensement relativement à des variables clés de culture, de bétail et de dépenses. Cette liste est obtenue à l'aide d'un processus itératif qui consiste à établir un seuil pour chaque variable clé et à inclure dans la liste toutes les fermes qui excèdent au moins un de ces seuils. On ajuste indépendamment à la hausse ou à la baisse chacun des seuils de façon à ce que la liste L1 représente, une fois complétée, environ 35 % des fermes de la population enquêtée et de 50 % à 90 % de l'activité agricole totale selon la variable clé considérée. Ces pourcentages sont retenus car l'expérience a démontré que la liste qui en résulte est composée de fermes qui sont individuellement plus stables dans le temps que le reste des fermes de la population enquêtée. Cette stabilité permet de créer des strates qui demeurent homogènes au fil des ans, ce qui est un facteur de conservation d'efficacité du plan d'échantillonnage.

Dans chaque province, la liste L1 est ensuite stratifiée à l'intérieur de régions infraprovinciales selon neuf variables clés. Un échantillon de fermes est sélectionné pour obtenir des données sur les cultures et le bétail. Les données sur les dépenses étant plus difficiles et dispendieuses à recueillir, seul un sous-échantillon, appelé noyau, est tenu de fournir ces renseignements.

La deuxième base de liste, notée L2, renferme les fermes du recensement de plus de 20 acres qui n'ont pas été retenues dans la liste L1. Sa stratification se fait à l'intérieur des districts agricoles selon une seule variable clé, soit la superficie cultivée au recensement. La liste L2 sert à compléter la liste L1 pour les données préliminaires sur les cultures. Ces données doivent être recueillis dans des délais très courts et la base aréolaire ne peut, pour des raisons opérationnelles, satisfaire ces délais.

La base aréolaire comprend tous les secteurs de dénombrement agricoles, sauf ceux qui correspondent aux réserves indiennes et aux régions dites marginales, c'est-à-dire là où il y a peu d'activité agricole. Ces régions marginales comprennent surtout le nord des provinces et les zones en bordure des villes. Les rares fermes du recensement qui sont situées dans les régions marginales sont ajoutées à la liste L1 car seule cette liste est utilisée pour recueillir des renseignements sur toutes les variables de l'enquête.

La base aréolaire est stratifiée en utilisant les mêmes régions infraprovinciales et variables clés que la liste L1. Elle engendre ultimement un échantillon de segments qui sont délimités sur des cartes topographiques. L'identité des fermiers qui exploitent des terrains dans ces segments est obtenue par un dénombrement sur place. Ensuite, des appariements manuels sur noms et adresses permettent de détecter les cas de chevauchement entre les fermes de segments et l'une ou l'autre des bases de liste. Cette détection est essentielle car chaque fois que la base aréolaire est appelée à compléter une base de liste, seules les fermes de segments qui ne chevauchent pas la liste en question sont utilisées. Ainsi, on s'assure que les bases de listes et aréolaire représentent des domaines mutuellement exclusifs.

Des renseignements complets sont exigés de toutes les fermes de segments, sauf celles qui chevauchent la liste L1 car les données pour la liste L1 proviennent de l'échantillon tiré de cette liste.

### 4.2 Québec et Ontario

Dans chacune de ces deux provinces, on a recours à une seule base de liste, appelée L1, et à une base aréolaire.

La base de liste est composée de l'ensemble des fermes du recensement qui appartiennent à la population enquêtée. La méthodologie employée pour échantillonner cette liste est similaire à celle de la liste L1 de la région de la CCB, à deux différences près. La première différence consiste à séparer des autres les fermes incorporées, c'est-à-dire constituées en sociétés par actions, puis à former indépendamment des strates dans chacun de ces deux groupes. Cette séparation préliminaire est effectuée car seules les fermes incorporées doivent rapporter leurs dépenses dans l'enquête, les dépenses des autres fermes étant obtenues à partir des dossiers fiscaux de Revenu Canada. Il convient de noter que la confidentialité de ces dossiers est

entièrement sauvegardée par la Loi sur la statistique. La seconde différence est qu'il n'est pas nécessaire de sous-échantillonner pour les dépenses car moins de 25 % des fermes de la population enquêtée sont incorporées.

La base aréolaire et son plan d'échantillonnage n'ont pratiquement pas été modifiés, faute de ressources, à partir des résultats du dernier recensement. Seules les régions marginales ont été mises à jour, ce qui a résulté en leur agrandissement.

### 4.3 Provinces Maritimes et le reste de la Colombie-Britannique

Dans chaque province de cette région, le plan de sondage ne comprend qu'une seule base de liste appelée L1. L'ensemble des fermes du recensement qui font partie de la population enquêtée constitue cette liste L1. Étant donné qu'une base de liste a tendance à se détériorer avec le temps et qu'il n'y a pas de base aréolaire pour la compléter, il devient alors plus difficile de couvrir entièrement la population enquêtée. Cependant, en raison du nombre relativement peu élevé de fermes dans les provinces concernées, soit moins de 30,000, des procédures plus poussées pour tenir à jour la liste ont été mises en place. Ces procédures permettent notamment de détecter des fermes qui ont été oubliées au recensement ou qui ont commencé leurs activités depuis lors. On estime ainsi que la base de liste assure, à toutes fins pratiques, une couverture exhaustive de la population enquêtée.

Dans chaque province, on utilise la même approche qu'au Québec et en Ontario pour stratifier la liste et sélectionner un échantillon de fermes. Cet échantillon est utilisé pour produire toutes les estimations requises.

## 5. MÉTHODES D'ÉCHANTILLONNAGE DES LISTES

Le prélèvement des échantillons des bases de liste s'appuie sur un plan d'échantillonnage stratifié à un degré où les fermes constituent les unités d'échantillonnage. La stratégie et les méthodes qui sont employées pour le développement de ce plan sont essentiellement les mêmes quelles que soient la province et la liste considérées. Par contre, l'agencement des méthodes et les variables clés en jeu peuvent varier d'un cas à l'autre.

La première étape consiste à identifier les fermes qui ont des caractéristiques distinctes et à procéder à un tirage complet de ces dernières. Ces fermes, dites autoreprésentatives, sont essentiellement de deux types. D'abord, on retrouve celles qui ont une structure d'exploitation unique, soit les pâturages communautaires et les corporations à opérations multiples. En second lieu, il y a les fermes qui se démarquent nettement de la majorité en raison de leur très forte contribution à des variables clés de culture, de bétail et de dépenses. Le dénombrement complet de ces dernières constitue, en raison de l'asymétrie (vers la droite) des distributions traitées, une façon efficace de réduire la variance échantillonnale.

L'identification des fermes à très forte contribution se fait au moyen d'une règle dont les fondements sont intuitifs et qui a donné de bons résultats dans l'ancien plan de sondage de l'enquête. Cette règle, dite de l'écart sigma, est appliquée indépendamment à chaque variable clé sur l'ensemble des fermes ayant une contribution non nulle à la variable en question. Ensuite, sont déclarées autoreprésentatives toutes les fermes dont la contribution est jugée suffisamment élevée selon la règle à l'une ou l'autre des variables clés.

La règle de l'écart sigma adaptée à l'enquête fonctionne de la façon suivante. Soit une distribution unidimensionnelle de points $x_i$, $i = 1, 2, \ldots, N$, $x_i > 0$ pour tout $i$, et soit $\sigma$ son écart type; on ordonne la distribution en ordre croissant $x_1 \leq x_2 \leq \ldots \leq x_N$; on détermine, pour la moitié de la distribution se trouvant à droite de la médiane, la distance entre chaque couple de points successifs $d_i = x_i - x_{i-1}$; soit $i_o$ le plus petit $i$ pour lequel $d_i \geq \sigma$, alors tous les points $i \geq i_o$ donnent lieu à des fermes autoreprésentatives. Si $d_i < \sigma$ pour tout $i$, alors aucun point de cette distribution ne se distingue suffisamment des autres pour déclarer une ferme autoreprésentative.

La seconde étape consiste à répartir le reste des fermes de la liste en strates à tirage partiel. Dans la majorité des cas, les strates sont formées à l'intérieur de régions infraprovinciales selon neuf variables clés représentant les trois catégories usuelles: culture, bétail et dépenses. Le nombre de variables dans chaque catégorie est respectivement de un, six et deux.

Le principe sous-jacent à la stratification est le suivant. Chaque ferme est caractérisée par neuf variables et on réunit les fermes qui sont voisines entre elles, le voisinage étant défini en termes de distance euclidienne. Deux algorithmes de classification automatique multidimensionnelle (multivariate clustering) sont employés à cette fin. Ces deux algorithmes seront appelés FASTCLUS et CLUSTER car ils sont disponibles dans les procédures du même nom du progiciel SAS, version 5.

L'algorithme FASTCLUS partitionne un ensemble d'observations en un nombre prédéterminé de grappes mutuellement disjointes. Pour ce faire, l'algorithme choisit d'abord des observations qui servent de noyaux initiaux des grappes. Chaque observation est alors assignée au noyau le plus près et, cela fait, les noyaux sont mis à jour par les moyennes des grappes ainsi formées. Le processus est répété et prend fin lorsque les changements dans les noyaux deviennent petits. Cet algorithme est basé sur les travaux de Hartigan (1975) et MacQueen (1967).

L'algorithme CLUSTER regroupe un ensemble d'observations en grappes mutuellement disjointes de façon hiérarchique. Au départ, chaque observation forme une grappe par elle-même. En utilisant une méthode inspirée de Ward (1963), les deux grappes les plus semblables sont alors agglomérées en une seule et remplacées par cette dernière. Le processus continue et prend fin lorsqu'il ne reste plus qu'une seule grappe. Massart et Kaufman (1983) donnent une introduction à ce genre de classification. Ainsi, on forme autant de partitions de l'ensemble des observations qu'il y a d'observations au départ, et chaque partition correspond à une stratification.

Ces algorithmes sont employés successivement de la façon suivante. On utilise d'abord FASTCLUS pour regrouper les fermes en 250 grappes; ensuite, on fusionne progressivement ces grappes à l'aide de CLUSTER pour finalement former les strates. On effectue une classification préliminaire avec FASTCLUS car l'emploi direct de CLUSTER avec un nombre élevé de dossiers requiert un temps d'ordinateur beaucoup trop important.

Au moment de la stratification, il est impératif que chacune des trois catégories de variables contribue avec le même degré d'importance à la formation des strates. Pour ce faire, on procède à des transformations des variables initiales de stratification. Ces transformations sont exécutées de façon à ce que la somme des variables transformées dans chaque catégorie ait une moyenne de 0 et une variance déterminée, généralement 1. Pour la catégorie culture où il y a une seule variable, il suffit de la standardiser de façon usuelle en lui soustrayant sa moyenne et en divisant par son écart type. Pour les deux autres catégories, on effectue indépendamment dans chacune d'elles deux transformations successives. Soit $X_i$ les variables initiales d'une catégorie donnée C. On effectue d'abord une analyse en composantes principales pour obtenir des variables transformées $Y_i$. Ces nouvelles variables, de moyenne $\mu_i$ et de variance $\sigma_i^2$ sont des combinaisons linéaires des anciennes et mutuellement indépendantes. Ensuite, on standardise les $Y_i$ pour obtenir des variables de stratification finales $Z_i$ de la façon suivante:

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

Ainsi, on constate que $\sum_{i \in C} Z_i$ possède une moyenne de 0 et une variance de 1.

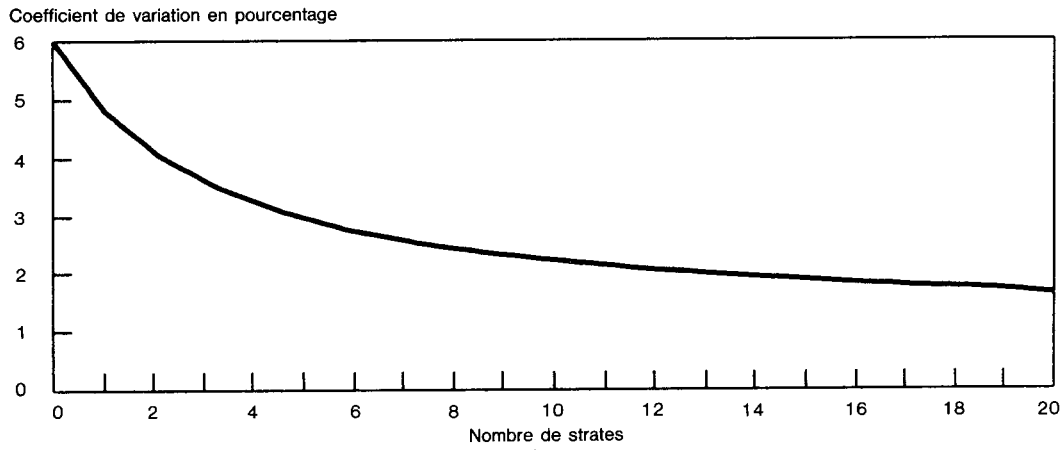


**Figure 1.** Courbe générale du coefficient de variation en fonction du nombre de strates

Une approche empirique est employée pour décider du nombre de strates. On effectue d'abord plusieurs stratifications et allocations en variant le nombre de strates. Cela fait, on construit la courbe du coefficient de variation obtenu en fonction du nombre de strates, et ce, pour toutes les variables clés et bien d'autres. Ces courbes ont généralement l'allure de la figure 1. On considère que les gains dûs à la stratification sont à toutes fins pratiques complètement réalisés au point où la majorité des courbes deviennent quasi horizontales. Le choix du nombre de strates est un compromis entre le point susmentionné et le désir de ne pas former trop de strates pour atténuer les effets des classifications initiales erronées et des changements de strates dans le temps, deux causes importantes d'observations aberrantes.

L'allocation de l'échantillon est multidimensionnelle et utilise généralement les mêmes variables clés que la stratification. L'algorithme d'allocation consiste à minimiser une combinaison linéaire du carré des coefficients de variation des variables clés, sous la contrainte que la taille totale d'échantillon est fixe. En fait, soit $c_i$ le coefficient de variation d'une variable clé, $a_i > 0$ une constante et $n_o$ la taille totale de l'échantillon, il s'agit de minimiser $\sum a_i c_i^2 = f(n)$ sous la contrainte que $n = n_o$. L'algorithme utilisé est décrit dans Bethel (1986). Cela fait, on procède à des ajustements de façon à ce que la taille minimale soit de 4 et le facteur de pondération maximal soit de 50 dans chaque strate.

Enfin, lorsque l'allocation est complétée, l'échantillon est prélevé dans chaque strate de façon circulaire systématique après avoir ordonné les fermes selon leur région infraprovinciale et leurs dépenses totales d'exploitation. Pour la liste L1 de la région de la CCB, on choisit d'abord l'échantillon complet duquel on sélectionne, toujours de façon circulaire systématique, le sous-échantillon noyau.

## 6. MÉTHODES D'ÉCHANTILLONNAGE ARÉOLAIRE

La sélection des échantillons de type aréolaire est basée sur un plan d'échantillonnage stratifié à deux degrés. Les secteurs de dénombrement du recensement et les segments constituent respectivement les unités primaires et secondaires d'échantillonnage.

Étant donné que le plan d'échantillonnage aréolaire n'a pas été modifié au Québec et en Ontario, les paragraphes qui suivent s'appliquent seulement à la région de la CCB.

La première étape consiste à obtenir une mesure de l'activité agricole dans chaque SD de la base en agrégeant au niveau du SD les données des fermes du recensement qui n'appartiennent pas à la liste L1. L'exclusion des fermes de la liste L1 des agrégations produit des distributions de SD qui reflètent fidèlement les caractéristiques des petites fermes. L'emploi subséquent de ces distributions permet de sélectionner un échantillon aréolaire qui complète la liste L1 par rapport aux petites fermes avec une efficacité accrue.

Une fois les agrégations terminées, chaque SD est traité comme une ferme pour les fins d'échantillonnage. La stratégie et les méthodes employées pour la sélection des SD sont très similaires à celles qu'on applique à la liste L1 de la région de la CCB. En effet, on détermine d'abord des SD autoreprésentatifs avec la règle de l'écart sigma. On répartit ensuite, à l'intérieur de régions infraprovinciales, le reste des SD en strates à tirage partiel au moyen de l'algorithme de classification multidimensionnelle CLUSTER. Une classification préliminaire avec FASTCLUS n'est pas nécessaire dans ce cas-ci en raison du nombre relativement peu élevé de SD à traiter, soit jamais plus de 3,000 dans une province. De plus, les transformations appliquées aux variables clés se limitent aux standardisations usuelles. Il n'est pas nécessaire de recourir aux composantes principales car la contribution de la base aréolaire aux estimations provinciales n'est pas suffisante pour justifier une telle approche.

L'allocation aux strates est exécutée avec le même algorithme que celui de la liste et la taille minimale est également fixée à 4. Cela complété, on divise la taille d'échantillon par quatre dans chaque strate et on prélève quatre répliques indépendantes de façon circulaire systématique. Le recours à des répliques facilite le calcul de la variance car il arrive souvent qu'on choisisse une seule unité secondaire par unité primaire.

Une fois les SD sélectionnés, on délimite leur contour sur des cartes topographiques et on parcellise chacun de ceux-ci en segments d'environ 7,5 $km^2$ (3 $mi^2$). Ce faisant, on tente, dans la mesure du possible, d'utiliser des limites naturelles comme des routes ou des rivières afin de faciliter ultérieurement le travail des interviewers sur le terrain. Puis, un échantillon aléatoire simple sans remise de segments est prélevé au taux minimum de un sur trente dans chaque SD sélectionné. On note cependant quelques exceptions à la règle: d'abord, on prélève des segments additionnels de façon à ce que le facteur de pondération global n'excède jamais 180; un minimum de deux segments sont choisis dans chacun des SD appartenant aux strates qui font l'objet d'un tirage complet au premier degré; enfin, lorsqu'un même SD figure dans plus d'une réplique, des dispositions sont prises pour éviter le tirage d'un même segment plus d'une fois. Toutes ces exceptions constituent cependant des cas plutôt rares.

## 7. RÉSULTATS DU PLAN DE SONDAGE

Le tableau 1 contient les résultats du plan de sondage des bases de liste. On y retrouve les quantités suivantes: le nombre de fermes dans la liste ($N$); le nombre de strates ($H$); la taille de l'échantillon de fermes ($n$); et enfin, dans les provinces où cela s'applique, le nombre de fermes dans le sous-échantillon noyau ($n$-noyau).

Le tableau 2 renferme les résultats du plan de sondage aréolaire dans les provinces où un tel plan est utilisé. On y retrouve les quantités suivantes: le nombre de SD dans la base ($N$); le nombre de strates ($H$); le nombre total de SD échantillonnés ($n$); le nombre de SD échantillonnés où chaque SD est compté une seule fois lorsqu'il apparaît dans plus d'une réplique ($n$-uniques); et enfin le nombre de segments prélevés ($m$).

**Tableau 1**

Résultats du plan de sondage des bases de liste

| Province | Liste L1 | | | | Liste L2 | | |
|---|---|---|---|---|---|---|---|
| | *N* | *H* | *n* | *n*-noyau | *N* | *H* | *n* |
| Î.-P.-É. | 2830 | 26 | 451 | | | | |
| N.-É. | 4273 | 35 | 550 | | | | |
| N.-B. | 3544 | 39 | 498 | | | | |
| Québec | 41380 | 80 | 6096 | | | | |
| Ontario | 72598 | 78 | 8401 | | | | |
| Manitoba | 6712 | 48 | 1364 | 490 | 18058 | 29 | 2267 |
| Saskatchewan | 15668 | 48 | 3625 | 1106 | 45798 | 41 | 4573 |
| Alberta | 13928 | 63 | 2981 | 909 | 38504 | 25 | 2973 |
| C.-B.(Paix)[a] | 494 | 25 | 190 | 190 | 1187 | 6 | 170 |
| C.-B.(reste)[b] | 17042 | 41 | 1999 | | | | |
| Total | 178469 | 479 | 26155 | 2695 | 103547 | 101 | 9983 |

[a] District de la rivière de la Paix en Colombie-Britannique.
[b] Colombie-Britannique, sauf le district de la rivière de la Paix.

**Tableau 2**

Résultats du plan de sondage aréolaire

| Province | *N* | *H* | *n* | *n*-uniques | *m* |
|---|---|---|---|---|---|
| Québec | 2065 | 43 | 191 | 182 | 230 |
| Ontario | 2687 | 49 | 195 | 185 | 259 |
| Manitoba | 794 | 21 | 277 | 264 | 305 |
| Saskatchewan | 1496 | 26 | 328 | 308 | 477 |
| Alberta | 1623 | 32 | 328 | 319 | 434 |
| C.-B.(Paix)[a] | 54 | 7 | 36 | 32 | 58 |
| Total | 8719 | 178 | 1355 | 1290 | 1763 |

[a] District de la rivière de la Paix en Colombie-Britannique.

**Tableau 3**

Taux de non-réponse totale par province en pourcentage

| Province | Refus | Non-contact | Total |
|---|---|---|---|
| Î.-P.-É. | 0.00 | 3.55 | 3.55 |
| N.-É. | 0.00 | 2.18 | 2.18 |
| N.-B. | 0.00 | 1.61 | 1.61 |
| Québec | 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 |
| C.-B. | 1.78 | 10.28 | 12.06 |
| Total | 2.32 | 8.11 | 10.43 |

## 8. FACTEURS AFFECTANT LA PRÉCISION DES ESTIMATIONS

Pour mieux apprécier les résultats obtenus à la suite de l'enquête de 1988, il est nécessaire d'apporter des précisions sur trois facteurs qui affectent la fiabilité des estimations. Ces facteurs sont la taille de l'échantillon, le traitement de la non-réponse totale et la méthodologie de l'estimation.

D'abord, la taille de l'échantillon pour la liste L1 de la région de la CCB a été réduite de 10 % comparativement à l'équivalent de cette liste dans l'ancien plan de sondage. Cette réduction est surtout motivée par le désir de réaliser des économies.

En second lieu, la méthodologie pour traiter la non-réponse totale à été modifiée en 1988. Auparavant, on utilisait les données d'une autre ferme dans la même strate pour imputer des données à une ferme qui n'avait rien répondu à l'enquête. Ces données imputées permettaient alors de compléter l'échantillon à sa taille originale. En 1988, au lieu d'imputer les cas de non-réponse totale, on utilise seulement l'échantillon des répondants et on ajuste à la hausse les facteurs de pondération. Ainsi, l'échantillon effectif est réduit comparativement à l'ancienne méthode.

Lors de l'enquête de 1988, on a observé un taux de non-réponse totale variant entre 2 % et 13 % selon la province. Au niveau national, ce taux s'établissait à 10 %. On présente au tableau 3 les détails des taux de non-réponse.

Le dernier facteur est la méthodologie de l'estimation. Dans les bases de liste, on utilise les estimateurs usuels correspondant à un échantillonnage aléatoire simple stratifié. En ce qui concerne la base aréolaire, on emploie un estimateur décrit dans Wolter (1986 pp. 19-26) et correspondant à un plan de sondage avec répliques indépendantes. Les estimations provinciales sont obtenues en additionnant la contribution des bases de liste et aréolaire car, rappelons-le, ces deux bases sont indépendantes et représentent des domaines mutuellement exclusifs. Les détails de l'estimation se retrouvent dans Lynch (1988).

## 9. ÉVALUATION DE LA PERFORMANCE DU NOUVEAU PLAN

Pour évaluer la performance du nouveau plan, la précision des estimations obtenues en 1988 est comparée, dans un premier temps, à celle de l'enquête de 1987 et, dans un deuxième temps, à la précision espérée lors du développement du plan de l'enquête.

### 9.1 Enquête de 1988 contre enquête de 1987

Deux tendances s'opposent lorsque l'on compare la précision des estimations de 1988 à celle des estimations de 1987. D'un côté, les estimations de 1988 devraient être plus précises puisque le plan de sondage de 1987 était vieux de 4 ans déjà. Par contre, les deux facteurs de réduction de la taille de l'échantillon décrits à la section 8 militent en faveur d'une précision moins élevée des estimations de 1988.

La précision est comparée en employant le coefficient de variation (CV) des estimations du niveau provincial provenant de l'union des bases de liste L1 et aréolaire. Ces estimations sont celles de plusieurs variables clés dont le CV en 1987 n'excédait pas 20 %.

La comparaison de la précision de 234 estimations est présentée à l'aide de graphiques de la figure 2. Sur ces graphiques, chaque carré représente le CV d'une estimation tel qu'atteint en 1987, en abcisse, et atteint en 1988, en ordonnée. De plus, on présente la fréquence (en pourcentage) des variables clés situées à l'intérieur de chaque zone délimitée par les droites $Y = X/2$, $Y = X$ et $Y = 2X$.

On observe que près de 60 % des estimations des cultures sont plus précises en 1988 qu'en 1987. Celles qui le sont moins, le sont par peu dans la plupart des cas. Pour le bétail, près de 95 % des estimations sont plus précises en 1988. En particulier, 32 % des estimations sont même deux fois plus précises. Enfin, plus de 60 % des estimations des dépenses sont plus précises

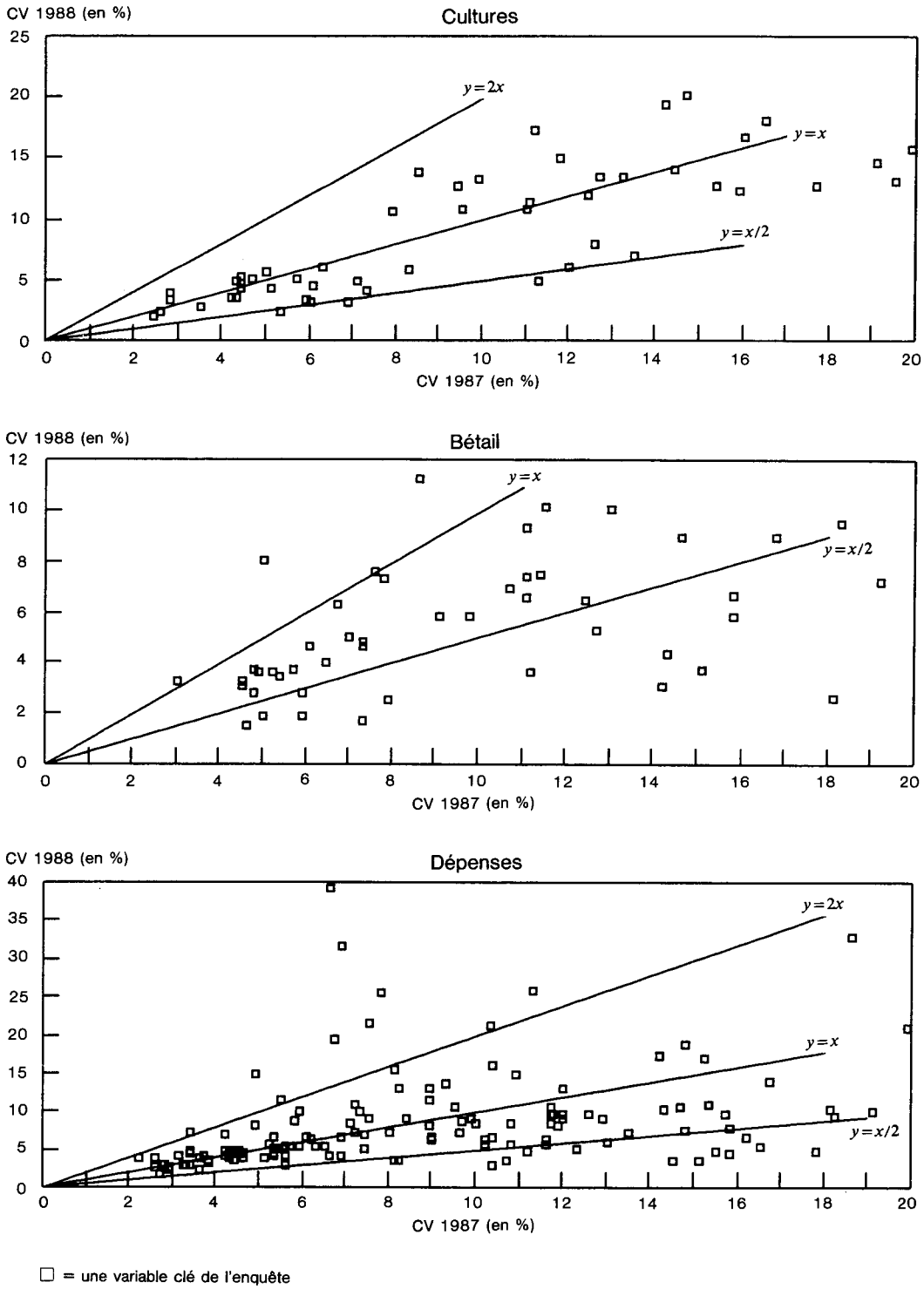


**Figure 2.** Comparaison de la précision des estimations de variables clés de l'enquête de 1988 à celle de 1987 par catégories de questions.

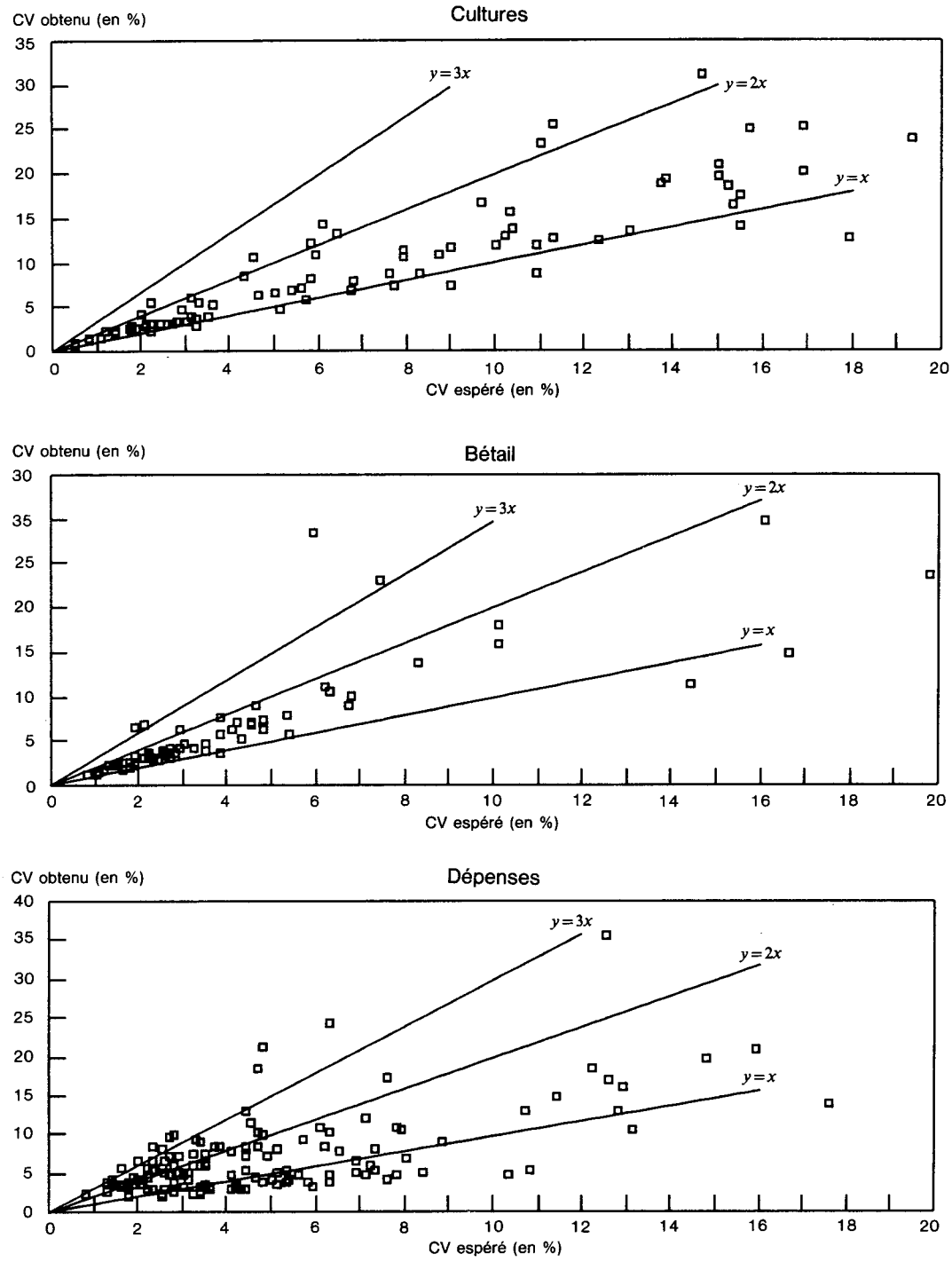


**Figure 3.** Comparaison de la précision des estimations de variables clés de l'enquête de 1988 à celle qui était espérée lors du développement du plan de sondage.

en 1988. Par contre, certaines de ces estimations sont beaucoup moins précises, et 7% de l'ensemble sont même deux fois moins précises. Ces estimations proviennent du Québec et de l'Ontario où on collecte les dépenses d'exploitation seulement chez les fermes incorporées. Or, dans ces provinces, la forme juridique d'une ferme est difficile à identifier, autant lors du recensement que lors de l'enquête.

Malgré la réduction de l'échantillon effectif en raison de la non-réponse totale et des coupures au moment du développement du plan de sondage, on conclut que l'enquête de 1988 a fourni en général des estimations plus précises pour chaque catégorie de variables.

### 9.2 Précision atteinte contre précision espérée

On s'attend à ce que la précision atteinte soit moins bonne que la précision espérée, et ce, pour deux raisons. Premièrement, l'ajustement des facteurs de pondération pour compenser la non-réponse totale entraîne une augmentation de la variance.

Deuxièmement, les données qui ont servi à créer la base de sondage proviennent du recensement agricole de 1986. D'une part, ces données sont sujettes à des erreurs, et, d'autre part, la base de sondage se détériore, à la suite des changements dans l'activité agricole.

La précision est comparée en employant le coefficient de variation des estimations du niveau provincial provenant de la base de liste L1 seulement. Ces estimations sont celles de plusieurs variables clés dont le CV espéré n'excédait pas 20%.

La comparaison de la précision de 288 estimations est présentée à l'aide de graphiques de la figure 3. Sur ces graphiques, chaque carré représente, pour une estimation, le CV espéré, en abcisse, et atteint en 1988, en ordonnée. De plus, on présente la fréquence (en pourcentage) des variables clés situées à l'intérieur de chaque zone délimitée par les droites $Y = X$, $Y = 2X$ et $Y = 3X$.

Pour les cultures et le bétail, environ 90% des estimations ont une précision acceptable, compte tenu du taux de non-réponse, puisque la plupart des variables clés se situent plus près de la droite $Y = X$ que de la droite $Y = 2X$. Pour les dépenses, on distingue deux tendances. D'abord, fait remarquable, le CV atteint est inférieur dans 28% des cas à celui qui était espéré; ces cas proviennent en grande majorité de la région de la CCB. Par contre, 31% de l'ensemble présente une précision plus que deux fois plus faible qu'espérée. Ces cas proviennent du Québec et de l'Ontario pour les raisons mentionnées à la section 9.1.

Enfin, on a effectué une étude complémentaire où on a comparé la précision atteinte à la précision espérée basée sur la taille de l'échantillon effectivement observé. On a constaté que la fréquence des estimations deux fois moins précises que prévu ou pire passait de 12% à 5% pour les cultures, de 9% à 5% pour le bétail et de 31% à 7% pour les dépenses.

On en conclut qu'en général la précision atteinte est acceptable et diffère de la précision espérée principalement à cause du traitement pour la non-réponse totale. Ceci indique que la plan de sondage est robuste et que la base de sondage de liste L1 est adéquate. Par contre, on a obtenu des estimations moins précises pour les dépenses à cause d'un problème dans l'identification des fermes incorporées au Québec et en Ontario lors du recensement et lors de l'enquête. Enfin, on a noté une certaine détérioration de la base de liste, vieille de deux ans déjà lors de l'enquête, et ce, surtout en raison des faillites et des ventes de ferme.

## 10. CONCLUSION

De façon générale, les résultats de l'enquête ont été sensiblement améliorés suite à l'utilisation du nouveau plan de sondage. Également, la réduction des tailles échantillonnales a permis de réaliser des économies et de réduire de façon appréciable le fardeau de réponse des fermiers enquêtés. Cependant, certaines difficultés persistent, surtout au niveau des dépenses des fermes incorporées au Québec et en Ontario. Des travaux additionnels sont envisagés pour résoudre ces difficultés.

## REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur en chef de la revue et les arbitres qui, par leurs précieux commentaires, ont contribué à l'amélioration de cet article.

## BIBLIOGRAPHIE

BETHEL, J. (1986). An optimum allocation algorithm for multivariate surveys. Rapport technique du United States Department of Agriculture, Statistical Reporting Service, Statistical Research Division, numéro SF et SRB-89.

GERMAIN, M.-F., DOLSON, D., et MARANDA, F. (1989). Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988. Document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles, Statistique Canada.

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

INGRAM, S., et 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. Document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles, Statistique Canada.

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

MASSART, D.L., et 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.

STATISTIQUE CANADA (1987). Dictionnaire du Recensement de 1986. Statistique Canada, n° 99-101F au catalogue.

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.