Techniques d'enquête, décembre 1976
Vol. 2, nº 2
Statistique Canada, nº 12-001-X au catalogue

Enquêtes pilotes à double base de sondage sur les porcs en Ontario

D. Serrurier et J.E. Phillips[1]

Résumé

Trois enquêtes pilotes sur les porcs ont été menées en Ontario en 1975 pour tester une méthode d'échantillonnage basée sur l'utilisation simultanée de deux bases listes. Le présent article décrit les différents aspects de cette expérience. On insiste plus particulièrement sur la méthodologie des doubles bases de sondage telle qu'elle a été décrite par Hartley [11]. On considère une répartition optimale de l'échantillon entre les bases de sondage, et l'on révise chaque enquête suivante en fonction de tous les résultats accumulés.

Mots-clés : Base liste; échantillonnage; répartition optimale.

1 Introduction

L'application de méthodes d'estimation précises et robustes dans le domaine du bétail a toujours été particulièrement préoccupante pour les spécialistes de la statistique agricole. Les changements rapides dans la gestion et les délais de réalisation souvent courts du cycle de production, principalement pour les porcs et la volaille, exigent des statistiques fréquentes et actuelles. Pour fournir ces statistiques, dans le domaine des porcs, la Division de l'agriculture de Statistique Canada mène une enquête trimestrielle. Les questionnaires sont postés à tous les fermiers qui ont déclaré des porcs dans le dernier Recensement de l'agriculture. Les questionnaires retournés sont jumelés d'une enquête à l'autre et des estimations de la production porcine sont produites. Malheureusement, le taux de réponse est relativement faible et aucune erreur d'échantillonnage ne peut être réellement associée aux estimations. De plus, à mesure que la liste d'envoi du recensement vieillit, il devient de plus en plus difficile de la garder à jour.

Les enquêtes pilotes sur les porcs en Ontario avaient pour but de tester d'autres procédures en fonction de l'échantillonnage aléatoire. Comme il n'y avait aucune raison de s'attendre à un taux de réponse plus élevé en raison d'un changement de plan de sondage, un objectif important était de garder la taille de l'échantillon assez petite pour permettre un suivi exhaustif à un coût raisonnable.

1. D. Serrurier et J.E. Phillips, Division des méthodes d'enquête - établissements et agriculture.

Tout d'abord, quatre enquêtes pilotes (Furrie et Wills [1]) ont été menées en Ontario de novembre 1973 à octobre 1974. Chacune de ces enquêtes était fondée sur un échantillonnage aléatoire simple stratifié répété reposant sur le Recensement de 1971 comme base liste. Les non-répondants ont fait l'objet d'un suivi téléphonique. Bien que les coefficients de variation aient été raisonnables compte tenu de la taille relativement petite de l'échantillon (moins de 200 fermes), on avait l'impression que les estimations étaient trop faibles. Cela était attribuable à deux raisons. Premièrement, la base du recensement devenait obsolète et bon nombre de fermes avaient changé de strate. Une étude de la robustesse de l'échantillon de porcs (Serrurier [2]) portant sur l'évolution de l'élevage de porcs entre les recensements de l'agriculture de 1966 et de 1971 a confirmé les nombreux changements de strate et leurs répercussions négatives sur tout plan d'échantillonnage qui les ignorait. Deuxièmement, la strate nulle composée de fermes qui n'avaient aucun porc en 1971 a été échantillonnée très légèrement, ce qui a entraîné une possible sous-estimation de la production de ces fermes qui avaient commencé la production de porcs depuis 1971. Le problème de la strate nulle est qu'elle représente un ensemble trop vaste (environ les deux tiers de la population en Ontario) de producteurs potentiels pour seulement quelques fermes qui entreprennent la production de porcs. De plus, elle ne comprend pas les toutes nouvelles exploitations porcines qui peuvent avoir été créées à partir de nouvelles activités agricoles depuis le dernier recensement.

La réponse évidente à ces deux problèmes de changements de strate et de détection des nouveaux producteurs est l'utilisation d'une liste plus à jour comme base de sondage. Malheureusement, une telle liste n'existe pas. Toutefois, la Commission ontarienne de commercialisation du porc (COCP) tient à jour une liste des personnes qui commercialisent des porcs et la rend accessible à Statistique Canada. La liste offre le grand avantage d'être mise à jour tous les ans, mais ne fournit pas de couverture complète de la population dans la province. On a pensé que l'union de cette liste avec la liste du recensement (excluant la strate nulle) fournirait une bonne couverture de la population à l'étude sans les difficultés associées à la strate nulle. Malheureusement, il n'a pas été possible de combiner les deux bases de sondage en une seule, car il n'y avait pas assez de renseignements identificatoires communs pour faire de

cela une opération réalisable à moindre coût. Cependant, des méthodes d'échantillonnage à bases multiples (Hartley [11]) permettent une sélection d'échantillons indépendants dans chaque base de sondage, un repérage de quelles unités sélectionnées appartiennent à l'une ou l'autre des bases de sondage ou aux deux et, par conséquent, une estimation efficace. Selon Hartley (1974), trois enquêtes pilotes à double base de sondage sur les porcs ont été menées en Ontario aux dates de référence suivantes : 1er avril, 1er juillet et 1er octobre 1975. Dans ces enquêtes, on a suivi la pratique antérieure en ce sens qu'il s'agissait d'enquêtes envoi/retour par la poste comprenant un suivi téléphonique. Pour chaque période d'enquête, des échantillons aléatoires simples stratifiés ont été tirés de manière indépendante dans chaque base de sondage. Le présent article a pour but de décrire les méthodes utilisées pour ces enquêtes et de présenter les principaux résultats de l'expérience.

2 Plan d'échantillonnage

2.1 Bases de sondage

Les bases de sondage utilisées pour les enquêtes pilotes étaient le Recensement de l'agriculture de 1971 et une liste des producteurs obtenue de la Commission ontarienne de commercialisation du porc (COCP). En octobre, la liste de 1973 de la COCP a été remplacée par la liste de 1974. Nous décrivons maintenant ces bases de sondage.

La base du recensement est une liste des 30 626 fermes de l'Ontario qui avaient au moins un porc dans le Recensement de l'agriculture de 1971. Les 64 096 fermes de l'Ontario sans porcs au moment du recensement ne sont pas prises en compte, car si elles ont entrepris des activités de production de porcs depuis 1971, on s'attendrait à trouver la plupart d'entre elles dans la liste de la COCP. L'exclusion de la strate nulle (les fermes sans porcs en 1971) est un avantage de taille, car la grande variabilité de ce groupe fait qu'il représente une part importante de l'erreur d'échantillonnage dans les enquêtes fondées sur la base du recensement complète. De plus, la strate nulle comporte en général un faible taux de réponse, ce qui augmente ainsi les coûts de suivi. La liste du recensement comprend la quantité de bétail présent dans l'exploitation au 1er juin 1971 et comprend le nombre de porcs selon certaines

catégories principales. Ces renseignements peuvent être utilisés à des fins de stratification.

Nous décrivons ensuite la base liste de la COCP. La COCP nous a fourni une liste de tous leurs producteurs inscrits. Cependant, la liste ne comprend pas les quelques producteurs situés dans les comtés du nord-ouest de la province. Outre ces derniers, la liste comprend tous ceux qui ont commercialisé, par l'entremise de la Commission, au moins un porc au cours de l'année en question ou de l'année précédente. Lorsqu'un producteur ne commercialise aucun porc pendant deux années consécutives, il est retiré de la liste, mais sera réinscrit s'il commence à commercialiser des porcs lors d'années subséquentes.

Comme on l'a indiqué précédemment, la liste de 1973 a été utilisée pour les enquêtes d'avril et de juillet 1975. La liste de 1974 est devenue accessible à temps pour être utilisée pour l'enquête d'octobre 1975.

Avant que la liste de la COCP puisse être utilisée à des fins d'échantillonnage, elle doit être vérifiée pour y détecter les doublons. Pour l'enquête, l'unité d'échantillonnage est une ferme, mais la liste de la COCP comprend toutes les personnes qui ont commercialisé des porcs. Des problèmes surviennent donc lorsque plusieurs personnes commercialisent des porcs élevés sur la même ferme. Pour réduire leur effet, la première étape a été de retirer de la liste toutes les personnes qui ont commercialisé moins de cinq porcs. Environ 45 % des personnes de la liste de 1973 ont ainsi été éliminées, et 43 % des personnes ont été éliminées de la liste de 1974. Bien entendu, toute ferme qui pourrait avoir été éliminée par erreur à cette étape serait fort probablement prise en compte dans la base du recensement. Ce qu'il restait des listes a ensuite été trié par nom de famille et examiné manuellement afin de détecter tous les doublons possibles. Les doublons possibles ont ensuite été comparés à la liste des fermiers du Recensement de 1971. S'ils faisaient tous partie de la liste du recensement, ils étaient alors tous conservés dans la liste de la COCP en tant que fermes distinctes. Cependant, si une seule figurait sur la liste du recensement, cette ferme était alors conservée dans la liste de la COCP et les autres personnes étaient retirées du fichier, mais le total de leurs porcs commercialisés pour l'année en question était ajouté à la ferme qui était conservée dans la liste.

Au terme du processus de retrait des doublons de la liste de 1973, nous avons obtenu une liste de 15 698 fermes alors que nous avions 31 823 noms au début. La liste de 1974 comprenait initialement 28 573 noms, mais a été réduite à 14 991 fermes.

2.2 Stratification

Comme les enquêtes devaient être menées par la poste et que le suivi devait être fait par téléphone, aucun coût de déplacement n'a dû être pris en compte et l'on a décidé d'utiliser l'échantillonnage aléatoire simple stratifié.

La stratification a été réalisée séparément pour chaque base de sondage au moyen des renseignements disponibles. Cependant, le nombre de strates était limité afin d'avoir au moins deux ou trois fermes échantillonnées dans chaque intersection strate x domaine.

2.2.1 Stratification de la base du recensement

Les renseignements disponibles étaient le nombre de porcs par catégorie principale au moment du recensement. D'après les enquêtes pilotes précédentes (Furrie et Wills [1]), il était évident qu'une stratification croisée par le nombre de truies et de porcs était le meilleur compromis pour les estimations de porcs et de truies au moment de l'établissement du plan de sondage. On a aussi constaté que cette stratification croisée était plus robuste (Serrurier [2]) que les autres fondées sur les porcs à eux seuls ou les truies à elles seules. Il était alors naturel de suivre la stratification utilisée pour les enquêtes précédentes, mais le nombre de strates a été réduit, pour passer de 10 à 8.

Tableau 1
Stratification de la base du recensement

Strate	Limite		Taille de la population	Total de porcs (1971)
	Nombre de porcs	Nombre de truies		
1	Moins de 725	Moins de 25	139	165 523
2	300à 725	Moins de 25	782	345 122
3	26 à 299	Moins de 25	1 592	273 292
4	Moions de 100	1 à 25	1 852	294 780
5	1 à 100	1 à 25	13 226	468 463
6	Moins de 600	0	125	122 476
7	150 à 600	0	1 277	347 118
8	1 à 149	0	11 633	344 893
Total			30 626	2 361 667

2.2.2 Stratification de la base de la Commission ontarienne de commercialisation du porc

Les renseignements disponibles dans les listes de 1973 et de 1974 de la COCP étaient le nombre total de porcs et de truies commercialisés par l'entremise de la Commission pendant l'année pour les fermes qui avaient commercialisé cinq porcs ou plus. Ces données ont été utilisées pour la stratification dans la liste de 1973. Sept strates ont été prises en fonction uniquement du nombre de porcs commercialisés, puisque la sous-stratification en fonction du nombre de truies n'avait donné que peu d'avantages.

Les mêmes limites de strate ont été utilisées pour la liste de 1974 de la COCP.

Tableau 2
Stratification de la base de la Commission ontarienne de commercialisation du porc

Strate	Limite (nombre de ventes de porcs)	Taille de la population	
		Liste de 1973	Liste de 1974
1	1 500 et plus	139	164
2	475 à 1 499	1 186	1 254
3	225 à 474	1 705	1 681
4	90 à 224	3 232	2 931
5	50 à 89	2 425	2 123
6	20 à 49	3 284	3 039
7	5 à 19	3 727	3 799
Total		15 698	14 991

2.3 Estimation des paramètres du plan de sondage

Selon la notation de l'annexe 1, l'union de la base de sondage A (recensement) et de la base de sondage B (COCP) définit trois domaines :

a = Fermes dans la liste du recensement seulement
ab = ba = Fermes dans les deux listes
b = Fermes dans la liste de la COCP seulement.

Comme chaque base de sondage est stratifiée de manière indépendante, les strates traversent les domaines. Si Ah (Ah = 1 ... K) est la h^e strate dans la base de sondage A, elle est formée de deux parties :

aAh = Intersection de la strate Ah avec le domaine a
abAh = Intersection de la strate Ah avec le domaine ab.

Par conséquent, dans la base de sondage B, la j^e strate, Bj (Bj = 1 ... L), est formée de deux parties, bBj et baBj.

Le plan de sondage optimal pour l'estimation d'un élément de contenu, Y, exige une connaissance antérieure des variances de population de Y par strate et intersection strate x domaine.

Pour l'enquête d'avril, les problèmes étaient les suivants : Premièrement, un élément de contenu commun aux deux bases de sondage devait être choisi. Deuxièmement, les variances de population des deux bases de sondage devaient être estimées sans l'avantage d'avoir des données des enquêtes précédentes. Pour juillet, les paramètres nécessaires de la population observée pouvaient être estimés en utilisant les données d'avril. En octobre, les estimations pouvaient être améliorées en utilisant toutes les données accumulées jusqu'à ce jour. Nous décrivons maintenant l'estimation des paramètres séparément pour chaque enquête.

Premièrement, pour le plan de sondage d'avril, aucun renseignement n'était disponible à propos de la portion de chevauchement commune aux deux bases de sondage. Le recensement comprenait le nombre de porcs dans l'exploitation au 1^{er} juin 1971, et la liste de la COCP comprenait les ventes de porcs et de truies pour l'année 1973 au complet. Le nombre total de porcs dans l'exploitation en 1971 a été choisi comme élément de contenu. Pour estimer les variances de population correspondantes pour chaque intersection strate x domaine, nous avons décidé de sélectionner des échantillons préalables stratifiés de taille 100 de manière indépendante dans chaque base de sondage. Chaque échantillon préalable a été jumelé avec l'autre base de sondage, ce qui permet de déterminer toutes leurs intersections strate x domaine. Dans les domaines a, ab et ba, il n'était pas difficile d'estimer les variances de population à partir de ces échantillons préalables, mais les nombres de porcs de 1971 n'étaient pas disponibles dans le domaine b. Ainsi, pour le plan de sondage d'avril, on a présumé que les variances de population dans le domaine b et pour l'ensemble de la base de sondage B étaient égales à celles du domaine ba. En fait, le rôle des échantillons préalables était d'entreprendre un processus permettant, pour chaque plan de sondage suivant, une estimation des paramètres requis à partir des enquêtes précédentes.

Puis, pour le plan de sondage de juillet, l'élément de contenu a été choisi naturellement comme le nombre de porcs au 1^{er} avril 1975 et les

variances de population correspondantes estimées à partir de l'enquête d'avril.

Finalement, pour le plan de sondage d'octobre, les échantillons d'avril et de juillet ont été combinés en considérant le nombre de porcs au 1er avril et au 1er juillet 1975 comme élément de contenu commun. On croyait que les variances de population ne changeraient pas beaucoup en trois mois et, par conséquent, cela valait la peine d'utiliser plus d'unités pour les estimer. Il convient également d'indiquer que compte tenu de la base de sondage de la COCP, ces estimations provenaient de la liste de 1973 alors que la liste de 1974 a effectivement été utilisée en octobre.

2.4 Tailles et répartition de l'échantillon

Une fois les variances de population connues, la méthode, telle qu'on l'a décrite à l'annexe 1, donne lieu à une répartition optimale entre les bases de sondage et entre les strates au sein de chaque base de sondage. Il est également possible d'estimer les variances et les coefficients de variation en fonction de la taille totale de l'échantillon.

En avril, 226 fermes ont été sélectionnées. Pour réduire les erreurs d'échantillonnage, la taille de l'échantillon a été augmentée légèrement en juillet (s'établissant à 250) et substantiellement en octobre (s'établissant à 350). Le graphique montre le coefficient de variation attendu pour le total de porcs (dans le plan de sondage d'octobre) en fonction de la taille de l'échantillon. On peut voir qu'au-delà de 400 à 500 fermes, on peut s'attendre à peu de gains sur le plan de la précision découlant d'une augmentation de l'échantillon. En revanche, la taille de l'échantillon doit permettre un suivi à un coût raisonnable. En tenant compte de ces contraintes, il semble que 400 fermes est le maximum qui peut être envisagé pour une telle enquête.

Le tableau 3 donne les tailles d'échantillon par base de sondage et par strate pour les trois enquêtes. Dans chaque enquête, les tailles d'échantillon « optimales », comme elles sont déterminées par les formules, ont été légèrement modifiées pour garantir un nombre minimum d'unités dans chaque strate. La dernière ligne du tableau donne la valeur de p utilisée dans le plan de sondage. Comme cela est expliqué à l'annexe 1, p est le poids à appliquer à l'estimation du domaine de chevauchement de la base de sondage A. Il y a une relation directe entre p et la proportion, γ,

de l'échantillon total à répartir dans la base de sondage A. Comme on peut aussi le voir dans les résultats, p est très sensible aux variances de population alors que γ est plus robuste. Par conséquent, les échantillons préalables dans le plan de sondage d'avril ont donné p = 0,37 pour le total de porcs, ce qui a entraîné la sélection de 46 % de l'échantillon dans la liste du recensement. Bien entendu, comme on l'a mentionné précédemment, ces estimations d'échantillons préalables des variances de la population étaient de piètre qualité. L'enquête d'avril a fourni des résultats quelque peu différents, ce qui a donné lieu à p = 0,07 pour le plan de sondage de juillet. Finalement, les estimations des variances de population des enquêtes d'avril et de juillet ont donné un poids intermédiaire de p = 0,29 dans le plan de sondage d'octobre associé à 38 % de l'échantillon réparti dans la base de sondage A.

Graphique 1 Coefficient de variation attendu pour le total de porcs en fonction de la taille totale de l'échantillon (pour les deux bases de sondage).

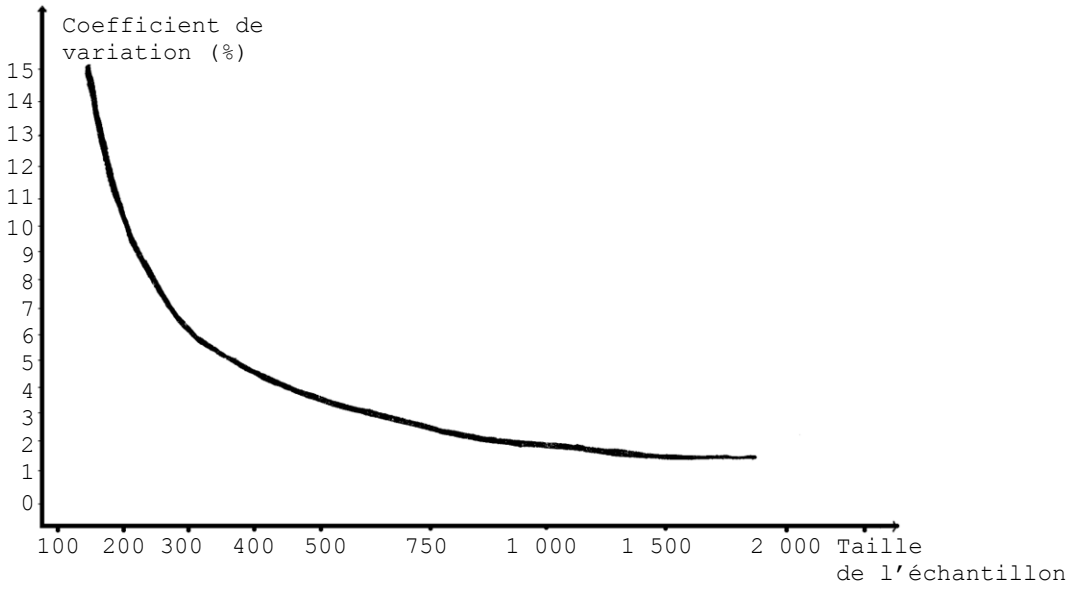


Tableau 3
Tailles de l'échantillon par strate dans les trois enquêtes

Strate	1er avril 1975		1er juillet 1975		1er octobre 1975	
	Recensement	Liste de la COCP de 1973	Recensement	Liste de la COCP de 1973	Recensement	Liste de la COCP de 1974
1	7	6	4	7	10	12
2	14	23	11	25	13	30
3	5	18	14	29	19	40
4	12	51	4	52	10	54
5	26	9	6	7	29	13
6	3	7	4	21	10	17
7	10	8	19	43	26	52
8	27	-	4	-	15	-
Total par base de sondage……	104	122	66	184	132	218
Proportion par base de sondage……	46 %	54 %	26 %	74 %	38 %	62 %
Total pour les deux bases de sondage……	226		250		350	
p optimal utilisé dans le plan de sondage…	0,37		0,07		0,29	

3 Organisation de l'enquête

3.1 Préparation de l'échantillon

Après avoir sélectionné les échantillons du recensement et de la COCP, plusieurs listages ont été produits. Le premier de ces listages a été utilisé pour enregistrer les numéros de téléphone pour les fermes sélectionnées. Le numéro de téléphone n'avait pas été saisi sur la bande magnétique de la COCP; un listage de l'échantillon de la COCP a donc été envoyé au bureau de Toronto de la Commission. Les employés du bureau ont ajouté les numéros de téléphone qu'ils avaient dans leurs dossiers. Pour les fermes de recensement sélectionnées, il a été nécessaire de remonter aux questionnaires du Recensement de 1971 (formule 6) pour obtenir le numéro de téléphone, s'il avait été indiqué.

Il doit être indiqué que les fermes sélectionnées sont ou non dans les deux bases de sondage. L'échantillon de la COCP doit donc être comparé à la liste du Recensement de 1971 et l'échantillon du recensement doit être comparé à la liste sans double compte de la COCP. Pour l'enquête de juillet 1975 reposant sur la liste de 1973 de la COCP, le chevauchement de l'échantillon de la COCP avec la liste du recensement était de 66 % et le chevauchement de l'échantillon du recensement avec la liste de la COCP était de 48 %. Dans l'enquête d'octobre 1975, on a utilisé la liste

de 1974 de la COCP et le chevauchement de l'échantillon de la COCP avec le recensement était de 72 % alors que celui de l'échantillon du recensement qui chevauchait la liste de la COCP était de 43 %.

Les échantillons de la COCP et du recensement ont été comparés l'un à l'autre pour s'assurer qu'aucune ferme qui avait été sélectionnée à partir des deux bases de sondage ne recevrait deux questionnaires. Il y a eu une ferme dans cette situation dans l'enquête d'avril et deux dans les enquêtes de juillet et d'octobre.

3.2 Suivi téléphonique

Les questionnaires ont été postés à temps pour joindre le fermier avant le premier du mois. Le suivi téléphonique auprès des non-répondants commence au milieu du mois, ce qui laisse deux semaines pour retourner les questionnaires. On a essayé une procédure différente pour l'enquête de juillet 1975 où l'on a indiqué au répondant que l'on communiquerait avec lui par téléphone et le suivi téléphonique a débuté plus tôt au début du mois.

La Commission a pu nous fournir 54 % des numéros de téléphone pour l'échantillon d'octobre 1975. Environ 87 % des numéros de téléphone étaient enregistrés dans les formulaires du recensement.

Quand nous n'avions pas de numéro de téléphone pour un fermier, nous devions d'abord obtenir son code régional et passer ensuite par la téléphoniste à l'interurbain pour tenter d'obtenir son numéro. Notre personnel avait à sa disposition un ensemble de bottins téléphoniques et il a semblé avoir plus de succès de cette façon que par l'entremise de la téléphoniste à l'interurbain.

Les appels téléphoniques aux fermiers se faisaient pendant l'heure du dîner (de 11 h 30 à 13 h) et le soir (de 17 h à 20 h). Pour l'enquête d'octobre 1975, nous téléphonions tout au long de la journée. Comme nous faisions affaire avec des fermiers, la saison de l'année semblait avoir une incidence sur le meilleur moment pour appeler. Pendant les mois d'été, le soir était le meilleur moment, mais au printemps et à l'automne, l'heure du dîner était préférable. Pendant l'été, il était difficile de joindre les fermiers à la maison, quel que soit le moment de la journée. On nous a parfois proposé de rappeler tôt le matin (entre 7 h et 7 h 30)

ou après 22 h. Les appels tôt le matin ont été faits, mais pas ceux tard le soir.

À partir de l'enquête d'avril 1975, nous avons calculé que 3,7 appels ont été effectués par heure-personne. Un appel effectué est lorsque l'on arrive à joindre le fermier et qu'un questionnaire est rempli ou qu'un refus est obtenu. Ce taux semble faible, mais il comprend le temps passé à obtenir le code régional ou le numéro de téléphone ainsi que plusieurs rappels.

Au moins trois tentatives étaient effectuées pour communiquer avec un fermier. Comme les estimations devaient être produites avant la fin du mois, nous n'avions qu'environ deux semaines pour terminer les appels.

Pour l'enquête de juillet 1975, nous avons tenté une procédure différente. Nous avons indiqué aux fermiers que nous leur téléphonerions et donc le suivi téléphonique a commencé à la fin de la première semaine. On a décidé de tester cette méthode, car nous téléphonions à la plupart des fermiers de toute manière et il serait utile de profiter de la semaine supplémentaire pour leur téléphoner. Comme prévu, le pourcentage de retours par la poste était plus faible (31 % en juillet par rapport à 44 % en avril et en octobre), de sorte qu'un plus grand nombre de fermiers devaient être joints par téléphone, et un bon nombre de fermiers que nous avons appelés ont dit qu'ils avaient déjà envoyé le questionnaire ou qu'ils l'enverraient, alors nous n'avons pas gagné de temps en utilisant cette procédure.

3.3 Taux de réponse

Le taux de réponse par retour par la poste a été le même (44 %) pour les enquêtes d'avril et d'octobre. Le taux était plus faible (31 %) pour juillet, car nous avions dit aux fermiers que nous leur téléphonerions pour obtenir les renseignements et il n'était pas nécessaire de retourner les questionnaires par la poste. Le pourcentage de questionnaires remplis était le même pour avril et octobre (88 %) tandis que celui de l'enquête de juillet était beaucoup plus bas (79 %). Le taux plus faible en juillet pourrait être attribuable à la procédure différente qui a été suivie ainsi qu'au fait que l'été est le moment le plus occupé pour le fermier, et donc le taux de non-réponse serait plus élevé.

Tableau 4
Retours par catégorie pour les trois enquêtes

	Avril		Juillet		Octobre	
	Recensement	COCP	Recensement	COCP	Recensement	COCP
Questionnaires envoyés par la poste[1]	104	121	66	182	132	215[2]
Questionnaires remplis retournés par la poste[3]…	39	50	17	46	47	93
Questionnaires retournés par le bureau de poste……	} 5	} 6	1	3	5	1
Retiré des affaires……			3	7	3	3
TOTAL DES RETOURS PAR LA POSTE……	44	56	21	56	55	97
Questionnaires remplis par téléphone……	48	49	33	87	51	101
Refus……	3	4	1	7	} 6	} 4
Questionnaire promis, mais pas retourné……	1	6	6	6		
Aucun numéro de téléphone……	} 8	} 6	4	22	8	6
Impossible de prendre contact……			1	4	12	7
TOTAL DES SUIVIS TÉLÉPHONIQUES	60	65	45	126	77	118

1. Peut être différent de la taille de l'échantillon en raison des doublons.
2. Une ferme a été éliminée parce que l'emplacement était à l'extérieur de la province de l'Ontario (au Québec).
3. Juillet n'est pas comparable à avril et à octobre en raison de la procédure spéciale.

Le tableau qui suit fournit les taux de réponse pour les trois enquêtes.

Tableau 5
Taux de réponse

	Avril		Juillet		Octobre	
	Recensement	COCP	Recensement	COCP	Recensement	COCP
	%	%	%	%	%	%
Questionnaire retourné par la poste	42	46	32	31	42	45
TOTAL……	44		31		44	
Total des questionnaires remplis…… (retours par la poste et appels)	88	86	81	78	80	92
TOTAL……	88		79		88	

4 Traitement des données

4.1 Imputation des données manquantes

Premièrement, des imputations intégrales des données ont été effectuées pour les fermes qui n'avaient aucun numéro de téléphone ou avec lesquelles on ne pouvait pas communiquer par téléphone. On a imputé ces enregistrements en ajoutant des zéros pour tous les champs pour l'enquête d'avril. Toutefois, pour les enquêtes de juillet et d'octobre, ces fermes ont été comptabilisées en ne les incluant pas dans le fichier de données, mais en calculant les facteurs d'extension pour les strates selon le nombre réel de questionnaires remplis.

Deuxièmement, les fermes qui étaient des cas de refus ou qui avaient promis d'envoyer un questionnaire, mais ne l'ont pas fait étaient aussi prises en compte en les ignorant et en corrigeant le facteur d'extension en conséquence.

On a imputé toutes les fermes qui ont indiqué s'être retirées des affaires ou dont le questionnaire a été retourné par le bureau de poste en ajoutant des zéros pour tous les champs.

Des imputations partielles ont dû être effectuées principalement pour les dernières questions qui portaient sur le nombre de truies qui devaient mettre bas ou le nombre de porcs de moins de trois mois. Pour ces imputations partielles, nous avons tenté de trouver un enregistrement dans lequel on déclarait à peu près le même nombre pour le « total de porcs » pour imputer les « porcs de moins de trois mois » ou à peu près le même nombre de « truies pour l'élevage » pour imputer les « truies qui devaient mettre bas ». De plus, le bon enregistrement devait être dans la même base de sondage que l'enregistrement nécessitant une imputation. Si l'on trouvait un grand nombre de bons enregistrements, l'enregistrement qui était dans le même domaine (a, b ou ab) avait préséance et si un critère supplémentaire était requis, l'enregistrement dans la même strate était choisi pour faire une imputation à partir de celui-ci. Des imputations partielles ont été effectuées sur moins de 0,1 % des enregistrements.

4.2 Estimation

En raison de la méthode d'imputation qui néglige certaines non-réponses, le nombre de questionnaires traités à des fins d'estimation ne correspond

pas exactement au nombre de fermes sélectionnées. Bien sûr, la procédure d'estimation s'occupe de ces ajustements et à cette étape, la « taille de l'échantillon » signifie en fait le « nombre de questionnaires traités ». Le tableau 6 ci-dessous donne ces nombres par domaine pour chaque enquête.

Tableau 6
Nombre de questionnaires traités

	RECENSEMENT			COCP			Total
	Domaine a	Domaine ab	Total	Domaine b	Domaine ba	Total	
Avril	54	46	100	40	72	112	212
Juillet	23	31	54	45	98	143	197
Octobre	63	43	106	52	147	199	305

Une fois les données saisies, la première étape du processus d'estimation est de calculer les estimations des variances de la population par intersection strate x domaine dans chaque base de sondage.

Les résultats seront analysés à la section suivante, mais nous voilà préoccupés par un aspect particulier de la méthode : Dans quelle mesure les résultats sont-ils sensibles aux variations de l'estimation des paramètres de population ?

La méthode à double base de sondage repose sur le fait que le domaine de chevauchement est estimé par une combinaison pondérée de deux estimations indépendantes. Si ces estimations sont très différentes, ce qui peut être le cas lorsque de petites tailles d'échantillon sont utilisées, un mauvais choix de poids peut grandement changer les résultats. En théorie, le poids est calculé pour minimiser la variance totale, mais le calcul en soi dépend de l'exactitude des intrants, c'est-à-dire des estimations des variances de population. À partir de certains résultats des enquêtes, l'effet des variances de population sur le poids p sera d'abord examiné, puis les conséquences de la variation de p seront abordées.

4.2.1 Effet des estimations des variances de population sur le poids de chevauchement p

Pour l'élément estimé, $p(0 \leq p \leq 1)$ est le poids à appliquer à l'estimation du domaine de chevauchement par l'échantillon de la base de sondage A

(recensement). L'estimation du domaine de chevauchement par l'échantillon de la base de sondage B (COCP) est pondérée par $q = 1 - p$.

Étant donné la variable « total de porcs », une valeur de p = 0,37 a été déduite des échantillons préalables pour le plan de sondage de l'enquête d'avril. Les résultats de cette enquête ont donné une valeur quelque peu différente avec p = 0,07. La différence peut s'expliquer par le mouvement ou le changement de la variable entre 1971 et 1975 et par les difficultés à estimer les variances de population à partir des échantillons préalables pour un élément de contenu commun aux deux bases de sondage (voir la section 2.3). En comparant les estimations des variances de population des deux enquêtes, on a pu observer des écarts importants. On a eu l'impression que de tels changements importants de la variabilité pour le total de porcs n'auraient pu se produire sur une période de trois mois. Nous avons donc décidé de considérer les porcs en avril et en juillet comme une variable commune et de grouper les deux échantillons uniquement pour estimer les variances de population. Nous pensions que ces variances de population seraient relativement constantes avec le temps et que le gain obtenu en augmentant les tailles d'échantillon pour leur estimation compenserait largement l'erreur attribuable au mélange des porcs en avril et en juillet. Cette fois-ci, une valeur de p = 0,29 a été obtenue.

Le tableau 7 donne des exemples de différences d'estimations des variances de population pour chaque base de sondage complète. Au niveau du domaine, les différences sont encore plus importantes.

Tableau 7
Exemples d'estimations de variances de population pour le total de porcs

Strate	Base de sondage A (recensement)				Base de sondage B (COCP)			
	Estimation d'avril	Estimation de juillet	Juillet-avril (%)	Avril et juillet combinés	Estimation d'avril	Estimation de juillet	Juillet-avril (%)	Avril et juillet combinés
1	94 240	142 100	+ 51	98 920	268 500	549 300	+ 105	384 600
2	60 800	78 950	+ 30	65 960	58 390	53 410	- 9	56 630
3	21 190	54 080	+ 155	42 180	34 460	38 550	+ 12	36 220
4	10 450	2 760	- 74	8 845	31 050	19 540	- 37	25 890
5	3 202	1 075	- 66	2 780	1 017	6 000	+ 490	3 336
6	10 900	84 270	+ 673	62 030	4 921	662	- 87	2 055
7	39 580	66 080	+ 67	52 840	17 650	24 180	+ 37	23 250
8	1 031	320	- 69	945	-	-	-	-

Lorsque ce type d'enquête a été entrepris, nous croyions qu'une valeur commune de p pourrait être utilisée pour tous les éléments. En fait, en raison de la sensibilité de p, il a été nécessaire de calculer différents p pour l'estimation de différents éléments. Ainsi, dans l'enquête d'octobre, p variait de 0,12 pour « autres porcs » à 0,69 pour « truies qui devraient mettre bas d'octobre à décembre ». Une conséquence immédiate est que l'estimation directe du total de porcs n'équivaut pas à la somme de ses trois composantes. Toutefois, ce problème n'est pas propre à la méthode et pour les trois enquêtes, les deux estimations du total de porcs étaient très proches.

Après l'enquête de juillet, les échantillons d'avril et de juillet ont été combinés pour estimer les variances de population pour les huit éléments. Les estimations de juillet reposaient sur ces variances et les estimations d'avril ont été révisées pour les incorporer. De plus, le plan de sondage d'octobre a été préparé au moyen de ces mêmes variances. Cependant, le traitement de l'enquête d'octobre était fondé sur les variances de population estimées à partir de l'échantillon d'octobre uniquement. D'un côté, on s'attendait à ce que l'augmentation de l'échantillon fournisse de meilleures estimations des variances de population et, de l'autre, le changement de liste de la COCP rendait les estimations précédentes inexactes dans la base de sondage B.

Un autre aspect à prendre en compte pour déterminer p est que l'expression de la variance à minimiser n'est pas la même à l'étape de traitement qu'à l'étape du plan de sondage (voir l'annexe 1). À l'étape du plan de sondage, on se place dans la situation de trouver un plan de sondage optimal pour l'estimation d'un certain élément de contenu. Cependant, des contraintes pratiques (ajustements, non-réponses, etc.) peuvent altérer ce plan de sondage optimal et l'élément estimé peut être très différent de l'élément de contenu sur lequel le plan de sondage était fondé (par exemple l'estimation des verrats à partir d'un plan de sondage fondé sur le total de porcs).

4.2.2 Effet du poids de chevauchement p sur les estimations

Comme p est très sensible aux variations des paramètres de population et aux écarts par rapport à l'« optimalité » dans le plan de sondage, il est important de mesurer son effet sur les estimations.

Selon les notations de l'annexe 1, l'estimation de la double base de sondage est donnée par

$$\dot{Y} = \hat{Y}_a + \tilde{Y}_b + p\hat{Y}_{ab} + q\tilde{Y}_{ba} \qquad (p + q = 1).$$

L'importance de p est fonction de la différence entre les deux estimations du domaine de chevauchement $\hat{Y}_{ab}$ et $\tilde{Y}_{ba}$. Plus la différence est grande, plus le rôle joué par p est important.

Les deux extrêmes correspondent à p = 1 (chevauchement estimé seulement à partir de la base de sondage A) et à p = 0 (chevauchement estimé seulement à partir de la base de sondage B).

$$p = 1\text{: } \dot{Y} = \hat{Y}_a + \tilde{Y}_b + \hat{Y}_{ab}$$

$$p = 0\text{: } \dot{Y} = \hat{Y}_a + \tilde{Y}_b + \tilde{Y}_{ba}.$$

Le tableau 8 donne, comme exemples, les deux estimations du chevauchement pour quatre éléments dans l'enquête d'octobre ainsi que la valeur de p adoptée et l'estimation définitive correspondante du chevauchement.

Tableau 8
Estimations du chevauchement pour certains éléments dans l'enquête d'octobre

	$\hat{Y}_{ab}$	$\tilde{Y}_{ba}$	p	$p\hat{Y}_{ab} + q\tilde{Y}_{ba}$
Verrats	5 214	8 211	0.44	6 892
Truies	76 662	116 322	0.63	91 336
Autres porcs	883 620	916 109	0.12	912 210
Total de porcs(estimation directe)	965 497	1 040 642	0.16	1 028 619

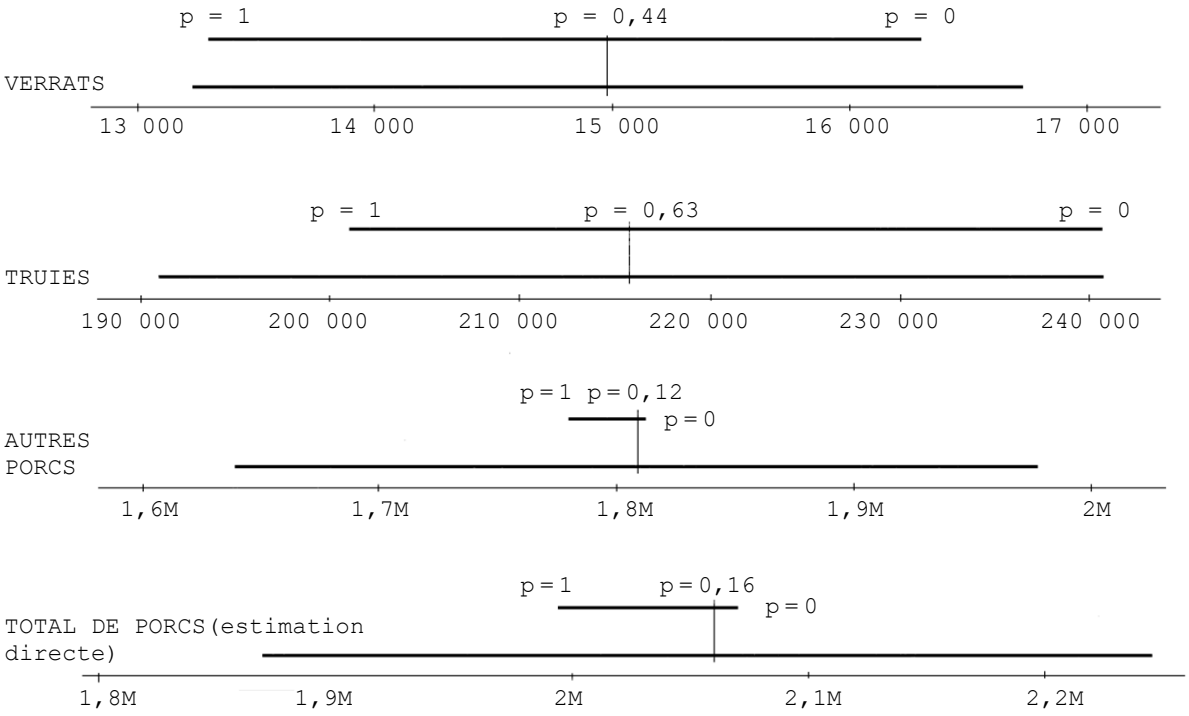
Dans les exemples du tableau 8, le poids p joue un rôle beaucoup plus important pour les verrats et les truies que pour les autres porcs. Cela est attribuable au fait que les deux estimations de base du chevauchement sont relativement proches pour les autres porcs.

Ce fait est illustré au graphique 2 en comparant l'intervalle associé aux valeurs extrêmes de p à l'intervalle de confiance à un écart-type. Dans la figure, les estimations globales à double base de sondage sont prises en compte, c'est-à-dire que les domaines de non-chevauchement a et b sont compris. Pour chaque élément, la première ligne donne l'intervalle des estimations associées aux valeurs de p de 1 à 0, tandis que la seconde ligne donne l'intervalle de confiance à un écart-type associé à

l'estimation à double base de sondage au moyen de p donné dans le tableau 8.

Graphique 2 Estimation dans l'enquête d'octobre en fonction du poids de chevauchement p

(Pour chaque élément, la seconde ligne donne ±1 écart-type de l'estimation associée à la valeur intermédiaire de p.)



Pour les verrats et les truies, l'intervalle possible des estimations ponctuelles, en fonction de p, est dans la mesure de près de deux écarts-types. Pour les autres porcs, le bon choix de p est beaucoup moins critique, car l'intervalle possible est d'environ le quart d'un écart-type. Pour le total de porcs, puisque cette catégorie est composée de près de 90 % des autres porcs, la situation est équivalente à cette dernière.

Il est difficile de tirer une conclusion de ces exemples contradictoires, à part que dans certains cas, un mauvais choix de p peut avoir d'importantes répercussions. Il semble important dans de telles applications d'accorder beaucoup d'attention à ce problème de pondération.

5 Résultats et comparaison avec d'autres estimations

Les résultats sont montrés au tableau 9, tout comme les estimations tirées de l'enquête trimestrielle sur les porcs de la Division de l'agriculture à des fins de comparaison.

L'estimation du total de porcs pour l'enquête pilote de juillet était très proche de celle produite dans le cadre de l'enquête trimestrielle. Cependant, les estimations d'avril et d'octobre étaient plutôt supérieures au résultat trimestriel. Comme on utilise uniquement le Recensement de 1971 comme base de sondage dans l'enquête trimestrielle, on croit que cela pourrait avoir pour effet de sous-estimer les résultats; il n'était donc pas étonnant que les estimations de l'enquête pilote soient plus élevées. De plus, en analysant les estimations individuelles, la plus grande différence entre les enquêtes se produit en général dans la catégorie « Porcs de moins de trois mois », suivie de la catégorie « Truies aux fins d'élevage ».

Les coefficients de variation sont élevés pour les estimations de l'enquête pilote, mais comme on l'a indiqué précédemment, une augmentation importante de la taille de l'échantillon serait nécessaire pour réduire la variance.

La variabilité entre les estimations pour les trois enquêtes pilotes pourrait s'expliquer par le fait que trois échantillons entièrement nouveaux ont été sélectionnés pour chaque enquête. Si une rotation était introduite et qu'une partie de l'échantillon était la même pour les trois enquêtes, les estimations du total de porcs pourraient alors être plus proches.

Tableau 9
Résultats des enquêtes pilotes à double base de sondage sur les porcs de 1975
(Tous les chiffres sont en milliers)

	1er avril 1975				1er juillet 1975				1er octobre 1975			
	Enquête pilote		Est. Div. agr.	Diff. en %[1]	Enquête pilote		Est. Div. agr.	Diff. en %[1]	Enquête pilote		Est. Div. agr.	Diff. en %[1]
	Est.	c.v. (%)			Est.	c.v. (%)			Est.	c.v. (%)		
Verrats pour l'élevage, 6 mois et plus……	15,6	10,8	15,0	+ 4	15,2	14,7	14,0	+ 8	15,0	11,7	14,0	+ 7
Truies et jeunes truies pour l'élevage, 6 mois et plus……	236,8	11,5	205,5	+ 15	228,9	14,1	198,0	+ 16	215,8	11,5	208,0	+ 4
Tous les autres porcs……	1 840	11,2	1 630	+ 13	1 608	11,4	1 620	- 1	1 809	9,3	1 650	+ 10
Porcs de moins de 3 mois……	899,7	13,0	740,0	+ 22	940,2	12,8	870,0	+ 8	911	11,4	780	+ 17
Truies ayant mis bas au cours des 3 derniers mois……	98,4	13,0	91,0	+ 8	114,5	14,8	100,0	+ 14	92,6	11,5	91,0	+ 2
Truies qui devraient mettre bas au cours des 3 prochains mois……	107,7	12,9	96,0	+ 12	95,2	15,9	90,0	+ 6	99,0	12,0	90,0	+ 10
Truies qui devraient mettre bas au cours des 4 à 6 prochains mois……	71,7[2]	16,9[2]	93,0	- 23	102,8	15,4	94,0	+ 9	100,3	12,5	94,0	+ 7
TOTAL DE PORCS[3]……	2 092	10,8	1 850	+ 13	1 852	11,2	1 832	+ 1	2 040	9,2	1 872	+ 9

1. Il s'agit de la différence entre l'estimation de l'enquête pilote et l'estimation de la Division de l'agriculture en pourcentage de l'estimation de la Division de l'agriculture.
2. Il ne s'agit pas d'une estimation fiable, car la question était manquante dans certains questionnaires.
3. Les estimations directes pour le « total de porcs », qui sont les sommes des trois premiers éléments, étaient :
   2 107 en avril, 1 866 en juillet et 2 058 en octobre.

6 Conclusion

Si les enquêtes à bases multiples ne sont pas nouvelles, elles s'appliquent très souvent à une base aréolaire associée à une base liste moins coûteuse (mais incomplète). De plus, dans la plupart des applications, une seule estimation du domaine de chevauchement est prise en compte (c'est-à-dire p = 0 ou 1) ou un poids préalablement attribué est utilisé (par exemple p = 0,5 si les deux bases de sondage sont à coût égal). À ce sujet, nous croyons que l'Enquête pilote sur les porcs de l'Ontario est un type d'application à double base de sondage relativement nouveau. La principale caractéristique de cette enquête est la réduction intentionnelle de la couverture complète de la base du recensement (en éliminant la strate nulle) et, au fond, son remplacement par la liste beaucoup plus petite et mise à jour de la COCP. L'avantage d'une telle procédure est non seulement d'éliminer de la base de sondage environ 50 000 fermes qui ne sont pas susceptibles d'avoir des porcs (les fermes de la strate nulle qui n'ont toujours pas de porcs quelques années plus tard), mais aussi de profiter des renseignements à jour disponibles dans la liste de la COCP. L'autre nouvel aspect dans cette application est une tentative d'optimiser le plan de sondage en répartissant l'échantillon entre les bases de sondage selon la théorie de double base de sondage et en révisant le poids p chaque fois que de nouveaux renseignements deviennent disponibles. Dans cette application, on a découvert que le poids p à appliquer au domaine de chevauchement est très sensible aux changements de paramètres de population et qu'il faut être très attentif à l'estimation de ces paramètres si nous voulons utiliser au mieux la méthode à bases multiples.

La procédure d'enquête s'est révélée viable et nous croyons qu'elle constitue une bonne réponse à l'ancien problème de la strate nulle. Malheureusement, en ce moment, certaines questions demeurent sans réponse et certains résultats ne sont pas satisfaisants. Parmi les questions toujours sans réponse, nous pouvons mentionner l'incapacité dans l'enquête d'avril de mesurer un biais éventuel attribuable à un piètre appariement. On s'attend à ce que les estimations à bases multiples affichent un biais vers le haut. Est-ce ce qui s'est produit dans la présente étude en raison du dédoublement résiduel ou les estimations actuelles sont-elles trop faibles ? Il pourrait être possible d'obtenir une réponse partielle à cette question en demandant à certains fermiers

participant au recensement s’ils avaient commercialisé des porcs par l’intermédiaire de la Commission. Cependant, une justification possible des estimations plus élevées dans l’enquête à double base de sondage est que certaines nouvelles exploitations qui se trouvent dans la liste de la COCP n’existaient pas au moment du recensement. Une autre question d’intérêt est la couverture de la population par l’union des deux listes. Nous avons présumé que les fermes de la strate nulle qui se sont tournées vers l’exploitation des porcs depuis 1971 pourraient se trouver dans la liste de la COCP. Cette hypothèse est probablement vraie pour les grands producteurs, mais dans quelle mesure y a-t-il un sous-dénombrement pour les petites fermes ? Il est beaucoup plus difficile de répondre à cette question étant donné que toute vérification exigerait d’échantillonner la strate nulle en ayant les mêmes difficultés qu’au cours des expériences précédentes.

En effet, la principale préoccupation quant à ces enquêtes pilotes découle des erreurs d’échantillonnage relativement élevées et, par conséquent, de l’instabilité des estimations d’une enquête à l’autre. Le problème ne vient pas de l’approche à double base de sondage, qui serait plutôt bonne pour cela, mais il est attribuable à la grande variabilité du produit à l’étude. Quoi qu’il en soit, le problème est que les utilisateurs potentiels jugeront probablement de telles erreurs d’échantillonnage excessives. Nous avons vu qu’au-delà de 300 à 400 fermes, une légère réduction des coefficients de variation nécessitera une hausse importante de la taille de l’échantillon. Une telle augmentation ne vaut pas la peine d’être envisagée puisque, comme on l’a déjà mentionné, l’avantage d’une telle enquête est d’utiliser des tailles d’échantillon raisonnables qui permettent un vaste suivi et un fardeau de réponse réduit. La réponse à ce problème se trouvera probablement dans les échantillons semi-permanents. En fait, quand une base de référence solide est disponible (par exemple le recensement), il faut mesurer les mouvements ou les changements et c’est ce qui est visé dans l’enquête trimestrielle régulière. Il faut souligner que pour ces trois enquêtes pilotes à double base de sondage sur les porcs, des échantillons complètement nouveaux ont été sélectionnés à chaque occasion, ce qui explique en grande partie l’instabilité élevée des estimations. Nous croyons que des échantillons semi-permanents associés à l’approche à double base de sondage devrait

fournir des estimations plus précises de l'évolution de la production porcine.

En ce qui a trait aux évolutions, un échantillon complètement permanent serait éventuellement préférable, mais cela n'est pas envisageable car, d'une part, il ne serait pas possible de profiter de la mise à jour régulière de la liste de la COCP et, d'autre part, le fardeau imposé au même groupe de fermiers aurait rapidement un effet négatif sur le taux de réponse. Un remplacement partiel des fermes serait un compromis entre un échantillon entièrement nouveau et un échantillon entièrement permanent. Bien entendu, le plan de renouvellement optimal doit encore être élaboré, mais il devrait permettre de marier les avantages actuels de la méthode à double base de sondage avec la production de statistiques actuelles et exactes.

Les auteurs tiennent à remercier M.P. Singh et l'examinateur pour les commentaires utiles fournis.

Bibliographie

[1] Furrie, A.D. et Wills, B.L. (1974), "Ontario Pilot Hog Survey", Rapport de la Census & Institutional Survey Methods Division, Statistique Canada, février 1974.

[2] Serrurier, D. (1975), "Hog Sample Robustness in Ontario", Rapport de la Census & Institutional Survey Methods Division, Statistique Canada, janvier 1975.

[3] Harley, H.O. (1962), "Multiple Frame Surveys", Proceedings of the Social Statistics Section, American Statistical Association meeting, Minneapolis, Minnesota.

[4] Cochran, Robert S. (1964), "Multiple Frame Sample Surveys", Proceedings of the Social Statistics Section, American Statistical Association meeting.

[5] Steinberg, Joseph (1965), "A Multiple Frame Survey for Rare Population Elements", Proceedings of the Social Statistics Section, American Statistical Association meeting.

[6] Cochran, Robert S. (1967), "The Estimation of Domain Sizes when Sampling Frames are Interlocking", American Statistical Association meeting, Washington, D.C.

[7] Lund, Richard E. (1968), "Estimators in Multiple Frame Surveys", Proceedings of the Social Statistics Section, American Statistical Association meeting, Pittsburg, Pennsylvanie.

[8] Fuller, W.A. et Burmeister, L.F. (1972), "Estimators for Samples Selected from Two Overlapping Frames", Proceedings of the Social Statistics Section, American Statistical Association meeting, Montréal, Québec.

[9] Rao, J.N.K. (1972), "Maximum Likelihood Estimation in Multiple Frame Surveys", (note non publiée).

[10] Vogel, Frederic A. (1973), "An Application of a Two-Stage Multiple Frame Sample Design", Proceedings of the Business and Economic Statistics Section, American Statistical Association meeting.

[11] Hartley, H.O. (1974), "Multiple Frame Methodology and Selected Applications", *Sankhyā*, Volume 36, séries C, partie 3, pp. 99-118.

[12] Burmeister, Leon F. (1974), "Sampling from Two Overlapping Frames when the Domain Sizes are Known", Proceedings of the Social Statistics Section, American Statistical Association meeting, St. Louis, Missouri.

[13] Vogel, F.A. (1975), "Surveys with Overlapping Frames - Problems in Application", présenté à la 135th Annual Meeting of the American Statistical Association, Atlanta, Géorgie.

Annexe 1

Méthode à double base de sondage

La méthode décrite dans la présente étude est une application directe du modèle général analysé par Hartley (1974) dans [11]. Des renseignements et des démonstrations supplémentaires sont présentés dans l'article de Hartley.

Nous considérons le cas particulier de fractions en chevauchement de la même population et lorsque des échantillons aléatoires simples stratifiés de taille n_A et n_B sont tirés, respectivement, des bases de sondage A et B. En raison de la procédure d'enquête, on présume que les deux bases de sondage sont de coût égal par unité d'échantillonnage et l'objectif est d'optimiser le plan de sondage pour une taille d'échantillon donné totale $n = n_A + n_B$.

Selon les notations de Hartley, les deux bases de sondage A (recensement) et B (COCP) définissent trois domaines :

a	=	Fermes dans A, mais pas dans B
ab = ba	=	Fermes dans les deux bases de sondage
b	=	Fermes dans B, mais pas dans A

Les deux bases de sondage n'ont pas été appariées et nous sommes dans la situation où le nombre d'unités de population dans le domaine de chevauchement n'est pas connu.

Au niveau de l'échantillon, ab et ba renvoient à l'estimation du domaine de chevauchement à partir de l'échantillon de la base de sondage A et de l'échantillon de la base de sondage B, respectivement.

Supposons que Y est la quantité à estimer (par exemple des porcs ou des truies)

$Y = Y_a + Y_b + Y_{ab}$ et est estimé par :

$$\dot{Y} = \hat{Y}(y_a,\alpha) + \tilde{Y}(y_b,\beta) + p\hat{Y}(y_{ab},\alpha) + q\tilde{Y}(y_{ba},\beta) \qquad (1)$$

$(p+q=1)$

$\hat{Y}$ et $\tilde{Y}$ étant les estimations à base de sondage unique, respectivement, pour A et B.

BASE DE SONDAGE A				
Strate	Taille pop.	Échantillon		
		Taille	Moyenne	Total
1	N_1	n_1	$\bar{y}_1$	y_1
2	N_2	n_2	$\bar{y}_2$	y_2
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
A_h	N_{Ah}	n_{Ah}	$\bar{y}_{Ah}$	y_{Ah}
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
K	N_K	n_K	$\bar{y}_K$	y_K
	N_A	n_A		

BASE DE SONDAGE B				
Strate	Taille pop.	Échantillon		
		Taille	Moyenne	Total
1	N_1	n_1	$\bar{y}_1$	y_1
2	N_2	n_2	$\bar{y}_2$	y_2
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
B_j	N_{Bj}	n_{Bj}	$\bar{y}_{Bj}$	y_{Bj}
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
L	N_L	n_L	$\bar{y}_L$	y_L
	N_B	n_B		

En raison de l'échantillonnage aléatoire simple stratifié :

$$\hat{Y}(y_A,\alpha) = \sum_{Ah=1}^{K} N_{Ah}\,\bar{y}_{Ah},\quad \tilde{Y}(y_B,\beta) = \sum_{Bj=1}^{L} N_{Bj}\,\bar{y}_{Bj}$$

Ensuite, en considérant que chaque domaine a, ab et b traverse les strates :

$$\hat{Y}(y_a,\alpha) = \sum_{Ah=1}^{K} \frac{N_{Ah}}{n_{Ah}}\,y_{aAh},\quad \hat{Y}(y_{ab},\alpha) = \sum_{Ah=1}^{K} \frac{N_{Ah}}{n_{Ah}}\,y_{abAh}$$

$$\tilde{Y}(y_b,\beta) = \sum_{Bj=1}^{L} \frac{N_{Bj}}{n_{Bj}}\,y_{bBj},\quad \tilde{Y}(y_{ba},\beta) = \sum_{Bj=1}^{L} \frac{N_{Bj}}{n_{Bj}}\,y_{baBj}$$

En inscrivant ces valeurs dans (1), nous obtenons l'estimateur $\dot{Y}$[2]. En utilisant les deux relations $y_{A\cdot} = y_a + y_{ab}$ et $y_B = y_b + y_{ba}$ pour indiquer les covariances, la variance de Y est donnée par :

$$\begin{aligned}\mathrm{Var}\,(\dot{Y}) = {} & p\,V_A(y_A,\alpha) + q\,V_A(y_a,\alpha) - pq\,V_A(y_{ab},\alpha) \\ & + q\,V_B(y_B,\beta) + p\,V_B(y_b,\beta) - pq\,V_B(y_{ba},\beta)\end{aligned} \qquad (2)$$

2. En tenant compte des éléments dédoublés dans les échantillons de chevauchement, cet estimateur a été amélioré par Lund [7], Fuller et Burmeister [8] et Rao [9]. Cependant, il y avait trop peu d'éléments dédoublés pour tenir compte de ces améliorations dans les enquêtes sur les porcs de l'Ontario.

$V_A(y_A, \alpha)$ et $V_B(y_B, \beta)$ indiquant, respectivement, les formules de variance pour une base de sondage unique pour A et B.

La formule (2) peut s'écrire, en introduisant une notation simplifiée, comme suit :

$$\text{Var}\ (\dot{Y}) = V_A(\alpha, p) + V_B(\beta, p) \tag{3}$$

où $V_A(\alpha, p)$ et $V_B(\beta, p)$ indiquent, respectivement, les trois expressions de termes de la première et seconde ligne de (2).

α et β étant les vecteurs des tailles d'échantillon des strates, respectivement, dans les bases de sondage A et B, nous devons minimiser $\text{Var}(\dot{Y})$ en fonction d'une taille d'échantillon totale donnée n.

Le problème de minimisation sera réglé en trois étapes :

1. $\min\limits_{\alpha, \beta}[V_A(\alpha, p) + V_B(\beta, p)]$ pour un p, un n_A et un n_B donnés pour produire, en fait, deux variances conditionnellement minimales :

$$V_A(n_A, p),\ V_B(n_B, p).$$

2. $\min\limits_{n_A, n_B}[V_A(n_A, p) + V_B(n_B, p)]$ pour un p et un $n = n_A + n_B$ donnés pour produire une variance minimale conditionnelle à p: V(p).
3. $\min\limits_{p} V(p)$ pour un n donné.

Étape 1

L'étape 1 peut se diviser en deux problèmes de minimisation disjoints :

$$\min_{\alpha} V_A(\alpha, p) \text{ pour un p et un } n_A \text{ donnés}$$

et

$$\min_{\beta} V_B(\beta, p) \text{ pour un p et un } n_B \text{ donnés}$$

Si nous désignons $S^2_{Ah}, S^2_{aAh}, S^2_{abAh}$, les variances de population dans la strate Ah, respectivement, pour la strate complète, le domaine a et le domaine ab, supposons que Σ^2_{Ah} est la quantité :

$$\Sigma^2_{Ah} = p\, S^2_{Ah} + q\, S^2_{aAh} - pq\, S^2_{abAh} \tag{4}$$

alors

$$V_A(\alpha, p) = \Sigma_{Ah=1}^{K} N_{Ah} \frac{N_{Ah} - n_{Ah}}{n_{Ah}} \Sigma_{Ah}^2 \tag{5}$$

S_{aAh}^2 et S_{abAh}^2 peuvent être calculés selon les procédures d'estimation du domaine en utilisant une variable auxiliaire prenant la valeur donnée Y dans le domaine et zéro à l'extérieur du domaine.

La minimisation de $V_A(\alpha, p)$ donne lieu à la solution bien connu de Neyman :

$$\boxed{n_{Ah} = n_A \frac{N_{Ah}\, \Sigma_{Ah}}{\Sigma_{Ah=1}^{K} N_{Ah}\, \Sigma_{Ah}}} \tag{6}$$

Au moyen d'une notation analogue pour la base de sondage B, nous obtenons :

$$\Sigma_{Bj}^2 = q\, S_{Bj}^2 + p\, S_{bBj}^2 - pq\, S_{baBj}^2 \tag{7}$$

$$V_B(\beta, p) = \sum_{Bj=1}^{L} N_{Bj} \frac{N_{Bj} - n_{Bj}}{n_{Bj}} \Sigma_{Bj}^2 \tag{8}$$

et

$$\boxed{n_{Bj} = n_B \frac{N_{Bj}\, \Sigma_{Bj}}{\Sigma_{Bj=1}^{L} N_{Bj}\, \Sigma_{Bj}}} \tag{9}$$

Les variances minimales conditionnelles à p, à n_A et à n_B sont données par :

$$V_A(n_A, p) = \frac{\sum_{Ah=1}^{K} N_{Ah}\, \Sigma_{Ah}^2}{n_A} - \sum_{Ah=1}^{K} N_{Ah}\, \Sigma_{Ah}^2 = \frac{A(p)}{n_A} - a(p)$$

$$V_B(n_B, p) = \frac{\sum_{Bj=1}^{L} N_{Bj} \Sigma_{Bj}}{n_B} - \sum_{Bj=1}^{L} N_{Bj}\, \Sigma_{Bj}^2 = \frac{B(p)}{n_B} - b(p) \tag{10}$$

Étape 2

La deuxième étape consiste à minimiser :

$V_A(n_A, p) + V_B(n_B, p)$ pour n_A et n_B compte tenu de p et $n = n_A + n_B$

Supposons que γ est le ratio : $= \frac{n_A}{n_A + n_B}$, c'est-à-dire la proportion de l'échantillon tiré dans la base de sondage A.

$$\min_{n_A, n_B}[V_A(n_A, p) + V_B(n_B, p)] = \min_{\gamma}[\frac{A(p)}{\gamma n} - a(p) + \frac{B(p)}{(1-\gamma)n} - b(p)] \text{ étant donné p et n}$$

La solution est donnée par :

$$\boxed{\gamma^* = \frac{\sqrt{A(p)}}{\sqrt{A(p)} + \sqrt{B(p)}}} \quad (11)$$

et la variance minimale conditionnelle à p et à n est :

$$V(p) = \frac{1}{n}[\sqrt{A(p)} + \sqrt{B(p)}]^2 - a(p) - b(p) \quad (12)$$

A(p), B(p), a(p), b(p) étant définis dans les relations (10).

Étape 3

La troisième étape consiste à minimiser V(p) comme il apparaît dans (12) pour p compte tenu de n.

Une minimisation analytique de v(p) est faisable uniquement dans certains cas particuliers. Une méthode numérique est donnée par Hartley dans [11].