Techniques d'enquête, décembre 1979
Vol. 5, nº 2
Statistique Canada, nº 12-001-X au catalogue

Test de techniques d'échantillonnage de listes multiples pour des enquêtes sur l'agriculture au Nouveau-Brunswick en 1978

B. Armstrong[1]

Résumé

La question étudiée dans le présent document est l'estimation des diverses variables agricoles selon la méthode des listes multiples. La liste des secteurs de dénombrement comprend intégralement la liste complémentaire. La stratification de la liste complémentaire et de la liste des secteurs de dénombrement repose sur des critères différents. De façon générale, la méthode des listes multiples présente certains avantages sur le plan de la variance par rapport à la liste des secteurs de dénombrement. Cependant, une analyse plus poussée révèle que la liste complémentaire comporte des lacunes, telles que la méthode de stratification et la dégénération des strates dans le temps.

Mots-clés : Liste multiple; estimation; stratification.

1 Introduction

La Division de l'agriculture de Statistique Canada mène l'Enquête descriptive sur l'agriculture (EDA) chaque année en juillet. L'EDA est une enquête polyvalente fondée sur un échantillon aréolaire produisant des estimations pour les cultures, le bétail et les postes de dépenses pour toutes les provinces sauf les Prairies. Un problème récurrent de l'enquête tient au fait que la taille de l'échantillon allouée aux petites provinces est insuffisante pour produire de bonnes estimations provinciales. On a donc décidé de tester l'échantillonnage à bases multiples dans l'une de ces provinces afin de déterminer si cette technique pouvait accroître l'efficacité des estimations et d'étudier les problèmes opérationnels associés à cette technique. La province choisie pour le test était le Nouveau-Brunswick.

C'était notre première expérience avec l'échantillonnage à bases multiples et donc, une expérience d'apprentissage. Le présent article exposera les résultats du test du Nouveau-Brunswick qui, bien qu'ayant réussi à réduire les erreurs d'échantillonnage, a mis en évidence des zones problématiques. Ces problèmes ainsi que les solutions proposées seront abordés.

1. B. Armstrong, Division des méthodes d'enquête-institutions et agriculture, Statistique Canada.

2 Plan d'échantillonnage pour le test de 1978 au Nouveau-Brunswick

Une enquête à bases multiples est une enquête qui s'appuie sur deux bases de sondage ou plus pour produire des estimations pour une population observée particulière. Dans le test de 1978 réalisé au Nouveau-Brunswick, une base aréolaire et une base liste ont été combinées pour produire des estimations pour les dépenses agricoles. L'utilisation de la base aréolaire était essentielle pour assurer une couverture complète de la population. La base liste a été introduite pour améliorer l'efficacité du plan d'échantillonnage.

L'EDA, dans sa forme actuelle, repose sur un type extrême d'échantillonnage à bases multiples. L'échantillon aréolaire de l'EDA est complété par un groupe de très grandes fermes (grandes par rapport à certains éléments clés) extraites de la liste actualisée du Recensement de l'agriculture de 1976 et incluses dans l'échantillon avec une probabilité de 1. Ce groupe de fermes, appelé fermes désignées, est inclus dans l'échantillon en tant que strate de dénombrement complète et distincte pour deux raisons. Puisque l'EDA est une enquête par échantillonnage probabiliste, les valeurs des éléments pour les fermes de l'échantillon sont gonflées pour représenter un plus grand groupe de fermes. Si ces fermes désignées n'ont pas été repérées avant l'enquête et qu'elles ont été choisies par hasard dans l'échantillon aréolaire, les estimations résultantes seraient gonflées hors de proportion puisqu'il ne s'agit pas de fermes typiques. En revanche, puisque ces fermes contribuent de façon importante aux totaux provinciaux, leur exclusion fortuite de l'échantillon entraînerait des estimations plus faibles. Dans l'un ou l'autre cas (inclusion ou exclusion fortuite), les estimations varieraient considérablement, et les fermes désignées seraient donc incluses avec une probabilité de 1. Cette caractéristique du plan d'échantillonnage de l'EDA a également été conservée dans le test à bases multiples, comme il est décrit ci-dessous.

a) Stratification de la base liste

La base liste pour le test à bases multiples du Nouveau-Brunswick était la liste actualisée du Recensement de l'agriculture de 1976. Les très petites fermes ont été exclues de la liste du recensement, et les fermes restantes ont constitué la base de sondage liste. Cette base a été stratifiée comme il est indiqué dans le tableau 1, et un échantillon aléatoire simple a été sélectionné dans chaque strate. En commençant par la strate 1 et en poursuivant de manière

séquentielle, une ferme a été attribuée à la première strate pour laquelle elle répondait au critère de strate. Comme nous le verrons plus loin, cette méthode de stratification de la liste s'est avérée inefficace pour l'estimation d'un des éléments de stratification.

Tableau 1
Stratification de la base liste

Strate	Définition de la strate	Taille de la population	Taille de l'échantillon
1	Fermes désignées telles quelles sont définies dans l'EDA	69	69
2	Ventes totales supérieures à 75 000 $	200	88
3	Superficie totale de pommes de terre supérieure à 50 acres	261	60
4	Nombre total de bovins supérieur à 40	448	60
5	Nombre total de porcs supérieur à 30	83	30
6	Superficie totale de pommes de terre supérieure à 10 acres	204	15
7	Nombre total de bovins supérieur à 25	858	30
8	Nombre total de porcs supérieur à 8	84	15
		2 207	367

La répartition de l'échantillon aux strates s'est effectuée en fonction d'une méthode d'essais et erreurs. La répartition finalement choisie était celle qui donnait la meilleure combinaison de coefficients de variation pour les trois éléments clés (c.-à-d. les pommes de terre, les bovins, les porcs) au Nouveau-Brunswick.

b) Stratification de la base aréolaire

Le plan de sondage de l'échantillon aréolaire était le même que depuis la dernière refonte en 1974. Les secteurs de dénombrement (SD) définis dans le Recensement de l'agriculture de 1971 étaient les unités d'échantillonnage du premier degré. Les SD ont été stratifiés d'une manière semblable à la stratification de la base liste à partir de données de recensement (résumées au niveau du SD) comme le montre le tableau 2. Un échantillon aléatoire stratifié répété de SD a ensuite été sélectionné. Encore une fois, la répartition de l'échantillon s'est effectuée en fonction d'une méthode d'essais et erreurs, la répartition choisie donnant la meilleure combinaison de coefficients de variation pour les éléments clés de la province.

Les SD sélectionnés ont été divisés en terrains ou en segments de taille à peu près égale. Ces segments sont devenus les unités d'échantillonnage du second degré. Un échantillon d'un ou de plusieurs segments a été choisi dans chaque SD sélectionné en fonction de la taille du SD. Tous les exploitants possédant des terres dans les segments sélectionnés ont été dénombrés.

Tableau 2
Stratification de la base aréolaire

Strate	Définition de la strate	Taille de la population	Taille de l'échantillon	Nombre de réitérations	Nombre de SD par réitération	Taille de l'échantillon
	(Niveau SD)	(nombre de SD)	(nombre de SD)			(nombre de fermes)
1	Nombre total de poulets, 25 000 et plus.....	25	12	6	2	29
2	Superficie totale de pommes de terre en acres, 6 000 et plus.....	30	30	15	2	212
3	Nombre total de porcs, 400 et plus	21	14	7	2	32
4	Nombre total de bovins, 500 et plus	62	20	10	2	92
5	X supérieur ou égal à 47.....	56	14	7	2	42
6	X supérieur ou égal à 14.....	136	18	9	2	30
7	SD restants	219	16	16	1	29
8	ANCIENS SD non agricoles.....	221	6	6	1	0
9	NOUVEAUX SD non agricoles.....	57	2	2	1	0
						466

REMARQUE : X était une variable conglomérée construite pour la province à partir d'une combinaison d'éléments relatifs au bétail et aux cultures. Pour les SD qui ne présentaient aucune caractéristique agricole dominante, la variable X a été utilisée pour former des strates contenant des SD dont le niveau d'activité agricole est similaire.

X = 20 000* [(rapport entre le nombre d'animaux d'élevage dans le SD et le nombre total d'animaux d'élevage dans tous les SD agricoles de la province) + (rapport entre la superficie des terres en culture dans le SD et la superficie totale des terres en culture dans tous les SD agricoles de la province)].

Étant donné que 1978 était l'année où a eu lieu notre premier test à bases multiples, un échantillon adéquat devait être alloué à la base aréolaire afin de garantir les estimations de l'échantillon aréolaire habituel de l'EDA en cas de problème avec les estimations à bases multiples. Heureusement, grâce à une subvention du PFFCM (Programme fédéral à forte concentration de main-d'œuvre) du gouvernement fédéral, la répartition de l'échantillon au Nouveau-Brunswick a été augmentée à

un point tel que nous avons pu non seulement ajouter un échantillon tiré d'une liste à l'échantillon aréolaire existant, mais aussi augmenter l'échantillon aréolaire au-dessus de son niveau de 1977.

3 Procédures d'estimation

D'après un article de Hartley (voir référence [1]), un seul estimateur à bases multiples peut être obtenu à partir de deux bases de sondage (dans le cas présent, la base aréolaire et la base liste) en ajoutant une estimation combinée de l'échantillon aréolaire et de l'échantillon tiré d'une liste pour le domaine en chevauchement (c.-à-d. la partie de la population couverte par la base liste) à l'estimation de l'échantillon aréolaire pour le domaine sans chevauchement (c.-à-d. la partie de la population non couverte par la base liste). Cet estimateur de Hartley est :

$$\hat{Y}_H = \hat{Y}_{NOL} + q\ \hat{Y}_{OL} + p\ \hat{Y}_L$$

où $\hat{Y}_{NOL}$ = estimation à base aréolaire pour le domaine sans chevauchement,

$\hat{Y}_{OL}$ = estimation provenant de la base aréolaire pour le domaine en chevauchement,

$\hat{Y}_L$ = estimation provenant de la base liste,

P = facteur de pondération donné aux estimations à base liste,

q = facteur de pondération donné à l'estimation provenant de la base aréolaire

et $p + q = 1$.

On peut déterminer une valeur optimale de p fondée uniquement sur la minimisation de la variance (voir l'annexe 1 pour la dérivation de p_{opt}). Hartley obtient une valeur optimale pour p qui minimise l'estimateur de variance à bases multiples par rapport à une fonction de coût qui dépend des tailles d'échantillon inconnues pour les bases liste et aréolaire. Ces tailles d'échantillons sont ensuite optimisées en même temps que p. Dans le test du Nouveau-Brunswick, cependant, les répartitions d'échantillons provenant de la base liste et de la base aréolaire ont été déterminées en fonction d'autres facteurs pris en considération.

L'échantillon aréolaire devait être suffisamment grand pour produire des estimations adéquates de l'échantillon aréolaire. Ainsi, la taille de l'échantillon aréolaire a été fixée avant l'introduction de la base liste. La taille de l'échantillon affecté à la base liste a ensuite été fixée de manière arbitraire. De même, dans l'EDA, il n'y a pratiquement aucune différence dans les coûts de mise en correspondance et de dénombrement des fermes de la liste et de l'échantillon aréolaire, puisque la collecte de données pour les deux types de fermes a lieu par interview. Pour ces raisons, ni la fonction de coût ni la répartition optimale de l'échantillon entre les deux bases n'ont été utilisées pour déterminer les valeurs optimales de p.

La formule de l'estimation de la variance de l'estimateur de Hartley est donnée à l'annexe 1.

L'estimateur avec sélection à bases multiples est un cas particulier de l'estimateur de Hartley où $p = 1$ et $q = 0$. Dans le cas présent, seul l'échantillon tiré de la liste est utilisé pour estimer le domaine en chevauchement. L'estimation de l'échantillon aréolaire n'est que pour le domaine sans chevauchement. En utilisant la notation définie précédemment, l'estimateur avec sélection prend la forme suivante :

$$\hat{Y}_S = \hat{Y}_L + \hat{Y}_{NOL}.$$

La variance de l'estimation avec sélection est simplement :

$$\mathrm{Var}(\hat{Y}_S) = \mathrm{Var}(\hat{Y}_L) + \mathrm{Var}(\hat{Y}_{NOL}).$$

L'estimateur de l'échantillon aréolaire régulier est, là encore, un cas particulier de l'estimateur de Hartley où $p = 0$ et $q = 1$. Ainsi, l'estimateur de l'échantillon aréolaire est :

$$\hat{Y}_A = \hat{Y}_{OL} + \hat{Y}_{NOL}$$

avec une variance :

$$\mathrm{Var}\ (\hat{Y}_A) = \mathrm{Var}\ (\hat{Y}_{OL}) + \mathrm{Var}\ (\hat{Y}_{NOL}) + 2\ \mathrm{Cov}\ (\hat{Y}_{OL}, \hat{Y}_{NOL}).$$

Pour obtenir de plus amples précisions sur les formules des estimateurs des totaux et des variances de l'échantillon aréolaire, consulter l'annexe 2.

4 Estimations pour le Nouveau-Brunswick en 1978

Les estimations à bases multiples avec sélection pour le test du Nouveau-Brunswick de 1978 sont présentées dans le tableau 3. Si l'on compare les estimations régulières à base aréolaire de l'EDA avec les estimations à bases multiples, il y a deux observations importantes à souligner. Tout d'abord, les coefficients de variation des estimations à bases multiples étaient nettement inférieurs à ceux des estimations à base aréolaire (souvent jusqu'à 50 %). La deuxième observation est que le niveau des estimations à bases multiples semble être généralement plus élevé que celui des estimations à base aréolaire. Dans seulement 4 des 21 estimations affichées dans le tableau 3, les estimations à bases multiples étaient inférieures aux estimations à base aréolaire.

La première observation doit faire l'objet d'un petit commentaire. La base liste est une base de sondage plus efficace et nous nous attendions donc à ce que les coefficients de variation diminuent fortement avec l'introduction d'un échantillon tiré d'une liste.

C'est la deuxième observation qui a soulevé le plus de préoccupations. Cependant, bien qu'il semble y avoir une tendance à ce que les estimations à bases multiples soient plus élevées que les estimations à base aréolaire, il est intéressant de souligner que la même tendance a pu être observée en ce qui a trait aux estimations publiées. Les estimations publiées sont compilées par des experts en la matière en tenant compte des estimations de toutes leurs sources (dont l'EDA fait partie). Des 17 cas où l'estimation à bases multiples était supérieure à l'estimation à base aréolaire, 11 des estimations publiées étaient également supérieures à l'estimation à base aréolaire. Dans les 4 cas où les estimations à bases multiples étaient inférieures aux estimations à base aréolaire, les estimations publiées étaient également inférieures aux estimations à base aréolaire. Ainsi, le problème de « niveau » des estimations à bases multiples ne semble pas aussi extrême si l'on tient compte de cet aspect, même si nous l'examinerons plus en détail dans la section 10.

Tableau 3
Estimations du Nouveau-Brunswick, 1978

Élément	Estimation provenant de la base aréolaire	Coefficient de variation	Estimation à bases multiples avec sélection	Coefficient de variation	Chiffres publiés
Superficie totale (acres)	1 090 235	8,5	1 271 419	6,0	1 090 200
Pommes de terre (acres)	63 355	15,4	70 318	7,2	58 000
Céréales mélangées (acres)	3 945	46,5	5 077	37,8	6 800
Avoine (acres)	37 357	17,2	45 771	9,4	42 000
Orge (acres)	5 210	26,6	5 946	19,5	7 200
Foin cultivé (acres)	158 628	12,0	197 167	7,5	180 000
Blé de printemps et blé d'hiver (acres)	7 903	-	8 405	-	9 600
Maïs-grain (acres)	380	31,9	510	33,8	500
Superficie totale des cultures (acres)	298 238	9,8	356 746	5,4	298 200
Terres améliorées pour le pâturage (acres)	97 987	13,3	119 155	10,0	-
Nombre total de bovins	109 350	12,0	118 844	5,4	113 000
Vaches laitières (moins de 2 ans)	35 277	20,2	34 801	9,6	31 000
Vaches de boucherie (plus de 2 ans)	19 300	13,8	23 815	11,8	23 000
Génisses laitières (l à 2 ans)	11 071	23,4	9 804	11,0	7 600
Taureau	2 600	16,6	2 743	16,1	2 700
Bouvillon	7 200	12,1	8 821	11,1	8 200
Veau	25 500	11,1	29 365	11,1	29 700
Nombre total de porcs	47 610	18,8	60 925	12,1	43 000
Truies et cochettes	10 205	44,1	7 741	14,1	5 000
Verrats	700	38,5	587	17,8	400
Cochons de marché (moins de 3 mois)	21 200	16,9	29 148	14,1	18 000

5 Comparaison des estimations à bases multiples avec sélection et de celles de Hartley

Le tableau 4 présente les estimations provenant de la base aréolaire, avec sélection et de Hartley pour les quatre éléments clés au Nouveau-Brunswick. De même, les valeurs p et q de l'estimation de Hartley sont données.

Tableau 4
Comparaison des différentes estimations pour le Nouveau-Brunswick, 1978
(coefficients de variation entre parenthèses)

	Estimation provenant de la base aréolaire	Erreur-type	Estimation avec sélection	Erreur-type	Estimation de Hartley	Erreur-type	p	q
Total	1 090 235	92 730	1 271 419	75 726	1 217 572	71 487	0,70	0,30
Superficie (acres)	(8,5%)		(6,0%)		(5,9%)			
Pommes de terre	63 355	9 788	70 318	5 050	69 900	5 023	0,94	0,06
Acres	(15,4%)		(7,2%)		(7,2%)			
Total	109 351	13 124	118 844	6 410	117 577	6 159	0,75	0,25
Bovins	(12,0%)		(5,4%)		(5,2%)			
Total	47 610	8 960	60 925	7 365	55 581	5 751	0,59	0,41
Porcs	(18,8%)		(12,1%)		(10,4%)			

Les résultats concernant la superficie totale, la superficie de pommes de terre et le nombre total de bovins sont conformes aux résultats prévus. Les erreurs-types ont été légèrement réduites en utilisant l'estimation

de Hartley, et les niveaux des estimations de Hartley et des estimations avec sélection sont comparables. Ce sont les estimations du nombre total de porcs qui sont intéressantes. L'estimation de Hartley est inférieure de 5 000 à celle avec sélection. L'erreur-type de l'estimation de Hartley est également nettement inférieure, le coefficient de variation étant de 10,4 % (contrairement à 12,1 % pour l'estimation avec sélection). La raison en est l'estimation provenant de la base liste du nombre total de porcs. Étant donné que la variance de l'estimation provenant de la base liste était relativement élevée (bien qu'elle constitue encore environ la moitié de celle de l'estimation à base aréolaire pour le domaine de chevauchement), le résultat a été qu'un facteur de pondération plus faible de p a été affecté à l'estimation provenant de la base liste du domaine de chevauchement et, par conséquent, des gains d'efficacité ont été réalisés en utilisant la combinaison des estimations pour le domaine de chevauchement dans l'estimation de Hartley. La valeur de p pour le nombre total de porcs était de 0,59.

6 Estimations provenant de la base liste — comparaison avec le Recensement de 1976

Pour que l'échantillonnage à bases multiples soit efficace et qu'il produise de bonnes estimations, il est essentiel que l'échantillon tiré d'une liste fournisse une bonne estimation de la partie de la population couverte par la base liste. Comme première étape de l'évaluation des estimations provenant de la base liste, les estimations de la base liste de 1978 ont été comparées par strate aux totaux correspondants de la mise à jour du Recensement de 1976.

Tableau 5
Comparaison des totaux actualisés du Recensement de 1976 et des estimations provenant de la base liste de 1978, Nouveau-Brunswick

Strate	Superficie totale (acres)		Acres de pommes de terre		Nombre total de bovins		Nombre total de porcs	
	1976	1978	1976	1978	1976	1978	1976	1978
1	50 954	68 888	2 123	3 967	13 416	13 359	18 443	19 427
2	99 644	93 498	17 721	19 148	7 145	5 082	4 191	1 809
3	87 284	102 155	26 243	25 461	4 005	2 462	561	3 306
4	203 822	245 601	678	866	39 186	37 438	3 293	12 484
5	19 518	19 973	168	357	1 895	1 494	9 278	9 907
6	43 453	58 303	6 193	8 418	2 790	3 006	268	1 210
7	242 836	229 229	237	86	34 303	26 255	1 745	4 976
8	13 964	12 634	39	34	912	1 058	1 228	885
Total	761 475	830 281	53 402	58 336	103 652	90 154	39 007	54 005
Coefficient de variation	-	5,5%	-	6,3%	-	4,6%	-	12,9%

Le tableau 5 présente les chiffres correspondants de 1976 et de 1978 pour quatre éléments clés au Nouveau-Brunswick. Comme il a été mentionné dans la dernière section, le coefficient de variation de l'estimation provenant de la base liste pour le nombre total des porcs était élevé. Cela n'est pas étonnant puisque, au stade de la conception, on a accordé aux porcs la priorité la plus faible de toutes les variables de stratification. Cette faible priorité peut être à l'origine de la grande différence entre le total du recensement et l'estimation provenant de la base liste. Il convient de mentionner que les strates bovines (strates 4 et 7) ont contribué à hauteur de 17 460 à l'estimation totale des porcs alors que le total du recensement pour ces deux strates n'était que de 5 038. La priorité accordée aux bovins étant plus élevée que celle accordée aux porcs lors de la stratification, si une ferme répondait à la fois aux critères de la grande strate bovine et de la grande strate porcine, elle était placée dans la grande strate bovine. Ainsi, la strate résultante était homogène pour le nombre total de bovins, mais pas pour le nombre total de porcs. Le résultat était que, puisque quelques grandes fermes porcines ont été prises dans l'échantillon, la strate a produit une estimation gonflée pour le nombre total de porcs. Pour les prochaines enquêtes, il serait judicieux d'envisager de donner aux porcs une priorité égale à celle des bovins.

L'estimation provenant de la base liste pour la superficie en acres de pommes de terre était de 58 336. La partie de la liste de recensement utilisée comme base liste au Nouveau-Brunswick représentait 96,18 % de toute la superficie en acres de pommes de terre de la province en 1976. Ainsi, si nous n'avions eu que l'échantillon tiré d'une liste à partir duquel faire les estimations, les résultats auraient été semblables à ceux de l'enquête sur le rendement prévu en pommes de terre au Nouveau-Brunswick en 1978 (une enquête basée sur un échantillon tiré d'une liste seulement). Il n'en demeure pas moins que nous avions également un échantillon aréolaire qui a montré que les fermes de l'échantillon aréolaire qui ne figuraient pas sur la liste ont contribué à l'estimation provinciale à hauteur de 11 982 acres de pommes de terre. Les fermes qui figuraient sur la liste du recensement, mais pas sur la liste réduite utilisée au Nouveau-Brunswick, représentaient 4 680 acres. Cela signifie que l'estimation provenant des fermes qui ne se trouvaient pas dans la liste de recensement était de 7 302 acres. Il n'est donc pas étonnant

que les estimations provenant de bases multiples soient plus élevées que celles de l'enquête sur le rendement prévu en pommes de terre au Nouveau-Brunswick, ces dernières étant fondées uniquement sur une liste.

7 État de la base liste

Le principal problème de l'utilisation de n'importe quelle base liste pour sélectionner un échantillon est que de telles bases de sondage deviennent très rapidement obsolètes. Les données du recensement sur lesquelles repose la stratification au Nouveau-Brunswick dataient de deux ans lorsque l'enquête a été menée. Le tableau 6 montre pour les fermes de l'échantillon tiré d'une liste à quelle strate chaque ferme a été affectée avant l'enquête et à quelle strate elle aurait dû être affectée en fonction des données de l'enquête.

Tableau 6
Changements de strates d'échantillons tirés d'une liste, 1976 et 1978

	Strate de 1978										
Strate de 1976	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9***	**Retiré des affaires**	**Total**
1	51	8	1	9	-	-	-	-	-	-	69
2	2	66	9	2	1	-	-	-	6	2	88
3	1	11	35	-	-	4	-	-	-	9	60
4	2	3	1	52	-	-	-	-	-	2	60
5	1	-	1	4	14	-	5	1	2	2	30
6	-	-	4	-	-	6	-	1	2	2	15
7	-	-	-	8	1	-	11	-	5	5	30
8	-	-	-	-	2	-	-	4	8	1	15
Total	57	88	51	75	18	10	16	6	23	23	367

*La strate 9 concerne les fermes qui ne répondent plus aux critères d'aucune strate de la liste.

Comme l'indique le tableau 6, les strates de grandes fermes (strates 1 à 5) avaient tendance à être plus stables que les strates de fermes plus petites, ce qui a réduit ou éliminé les avantages de la stratification pour ces strates inférieures. Ainsi, il serait indiqué de fixer des limites plus strictes aux strates à inclure dans la base liste. Les fermes de plus petite taille sont celles qui peuvent plus facilement « sauter » des strates, de sorte que l'estimation de la base aréolaire serait presque aussi bonne pour ces fermes. Cela permettrait également de laisser l'échantillon tiré d'une liste intact pour estimer plus efficacement les fermes de plus grande taille.

8 Retrait des strates moyennes de la liste

L'élimination d'une strate de la couverture de la liste modifie l'échantillon à bases multiples de deux façons. Premièrement, les fermes de l'échantillon tiré d'une liste dans cette strate ne contribuent pas à l'estimation de la liste. Deuxièmement, les fermes de l'échantillon aréolaire qui chevauchent cette strate dans la base de population (et qui faisaient donc partie du domaine de chevauchement aréolaire) ne font pas partie du domaine sans chevauchement. Pour l'estimation avec sélection décrite à la section 3, le nombre de fermes contribuant à la partie liste de l'estimation diminue, tandis que le nombre de fermes contribuant à la partie aréolaire de l'estimation augmente.

Le tableau 7 indique l'effet sur les estimations à bases multiples (sélection) de la suppression des strates de taille moyenne (strates 6, 7, 8) de la couverture de la base liste. Les coefficients de variation pour tous les éléments n'ont augmenté que légèrement (par rapport aux estimations à bases multiples complètes), mais cette augmentation était importante si l'on considère que la taille de l'échantillon a également augmenté. Les coefficients étaient toujours inférieurs à ceux des estimations à base aréolaire. Il convient également de mentionner que permettre l'utilisation de la base aréolaire pour estimer ces strates a réduit le niveau des estimations résultantes pour tous les éléments. Par conséquent, les estimations à base aréolaire étaient inférieures à celles de la base liste pour ces strates de liste. Cela peut être attribuable au problème de stratification périmée dans les strates de listes inférieures, comme il a été mentionné dans la section précédente. Étant donné que la base aréolaire semblait produire de meilleures estimations que la base liste pour les strates inférieures, la conclusion de la section 7 est renforcée ici, c'est-à-dire que les strates de taille moyenne devraient être retirées de la couverture de la base liste pour l'enquête de 1979.

Tableau 7
Estimations pour la couverture de la base liste réduite
(coefficient de variation entre parenthèses)

Élément	Estimation provenant de la base aréolaire	Estimation à bases multiples complètes	Retrait des strates 6, 7, 8 de la base liste
Superficie totale (acres)	1 090 235	1 271 419	1 220 732
	(8,5%)	(6,0%)	(6,4%)
Acres de pommes de terre	63 355	70 318	66 600
	(15,4%)	(7,2%)	(7,2%)
Nombre total de bovins	109 350	118 844	114 483
	(12,0%)	(5,4%)	(6,5%)
Nombre total de porcs	47 610	60 925	57 082
	(18,8%)	(12,1%)	(12,1%)
Taille de l'échantillon	535	569	591

9 Élimination des strates par produits

Le tableau 8 indique ce qui s'est passé lorsque les strates de pommes de terre (3 et 6), les strates de bovins (4 et 7) et les strates de porcs (5 et 8) ont été, à leur tour, exclues de la couverture de la base liste. Au fur et à mesure que les strates basées sur un certain élément sont éliminées de la base liste, on pourrait s'attendre à ce que l'efficacité de l'estimation pour cet élément diminue, puisque l'échantillon aréolaire produit maintenant des estimations presque entièrement pour cet élément. Comme l'indique le tableau 8, c'est en fait ce qui s'est passé. Lorsque les strates de pommes de terre ont été retirées, le coefficient de variation pour les pommes de terre est passé de 7,2 à 11,1; lorsque les strates de bovins ont été retirées, le coefficient pour les bovins est passé de 5,4 à 11,7; et lorsque les strates de porcs ont été retirées, le coefficient pour le total des porcs est passé de 12,1 à 16,6.

Ce qui est intéressant de souligner dans le tableau 8, c'est ce qui est arrivé à l'estimation du nombre total de porcs lorsque les strates de bovins (4 et 7) ont été retirées de la couverture de la liste. En permettant à l'échantillon aréolaire d'estimer les strates de bovins, l'estimation du nombre total de porcs a été ramenée au niveau de l'estimation à base aréolaire de 1978. De même, le coefficient de variation de cette estimation était inférieur à celui de l'estimation provenant de la base aréolaire et de l'estimation à bases multiples complètes. Cela a tendance à confirmer le doute soulevé dans la section 6 selon lequel l'échantillon tiré d'une liste pour les strates de bovins a surestimé le nombre total de porcs. Cela a également renforcé les recommandations selon lesquelles, aux fins de stratification, la priorité

du nombre total de porcs devrait être égale ou supérieure à celle du nombre total de bovins pour la conception future des listes.

Tableau 8
Élimination des strates de la liste pour certains produits
(coefficient de variation entre parenthèses)

Élément	Estimation provenant de la base aréolaire	Estimation à bases multiples complètes	Élimination des strates 3 et 6 (pommes de terre)	Élimination des strates 4 et 7 (bovins)	Élimination des strates 5 et 8 (porcs)
Superficie totale (acres)	1 090 235	1 271 419	1 216 582	1 145 174	1 263 446
	(8,5%)	(6,0%)	(6,0%)	(8,1%)	(6,0%)
Acres de pommes de terre	63 355	70 318	62 191	70 151	69 944
	(15,4%)	(7,2%)	(11,1%)	(7,1%)	(7,2%)
Nombre total de bovins	109 350	118 844	118 585	106 582	119 271
	(12,0%)	(5,4%)	(5,2%)	(11,7%)	(5,6%)
Nombre total de porcs	47 610	60 925	56 868	48 100	60,945
	(18,8%)	(12,1%)	(12,0%)	(9,7%)	(16,6%)
Taille de l'échantillon	535	569	591	585	537

10 Détermination des chevauchements

Le niveau élevé d'estimations à bases multiples indique qu'il pourrait y avoir un problème dans la détermination du chevauchement entre l'échantillon aréolaire et la base liste. Rappelons que seules les fermes de l'échantillon aréolaire qui ne figurent pas dans la base liste contribuent à la partie régionale de l'estimation à bases multiples. Si les fermes de l'échantillon aréolaire qui figurent dans la base liste ne sont pas désignées comme telles, les estimations à bases multiples qui en résultent seront gonflées.

Pour vérifier la détermination du chevauchement, une liste de fermes de l'échantillon aréolaire qui n'ont pas été appariées à la base liste au bureau central a été envoyée au bureau régional pour vérifier qu'il s'agissait de véritables « non-appariements » à la base liste. La liste a été renvoyée avec des commentaires indiquant que plusieurs fermes figuraient dans la base liste, mais étaient désormais exploitées sous des noms différents. Alors que le bureau régional supposait que nous appariions des fermes (c.-à-d. des terrains), nous appariions en fait des exploitants agricoles. Si une ferme de l'échantillon aréolaire figurait dans la base liste sous le nom d'un exploitant différent, il était pratiquement impossible pour le bureau central de les désigner

comme étant la même ferme. Nous avons donc dû nous contenter d'apparier des fermes plutôt que des terrains.

Cependant, après cette confusion avec le bureau régional au sujet des « fermes » par rapport aux « exploitants agricoles », il nous est apparu que la même confusion s'était probablement produite avec les agents recenseurs des fermes de l'échantillon tiré d'une liste. Encore une fois, puisque c'est un exploitant agricole qui est sélectionné dans la base liste, la ferme de l'échantillon tiré d'une liste devrait être classée comme étant retirée des affaires si l'exploitant n'est plus le même. Comme cela n'avait jamais été dit explicitement aux agents recenseurs, ils avaient tendance à ne pas faire cette distinction. Les fermes ont été dénombrées comme d'habitude, même si l'exploitant agricole avait changé. Il a donc été nécessaire de passer en revue les questionnaires de la liste pour sélectionner toutes ces fermes (au nombre de 12). Le tableau 9 donne les estimations révisées produites en retirant la contribution de ces fermes des estimations à bases multiples.

Tableau 9
Estimations à bases multiples révisées

Élément	Estimation à bases multiples initiale	Estimation à bases multiples révisée	Estimation pondérée	Estimation publiée
Superficie totale	1 271 419	1 241 888	1 090 235	1 090 200
	(6,0)	(6,2)	(8,5)	
Acres de pommes de terre	70 318	66 479	63 335	58 000
	(7,2)	(7,5)	(15,4)	
Nombre total de bovins	118 844	115 334	109 350	113 000
	(5,4)	(5,6)	(12,0)	
Nombre total de porcs	60 925	58 388	47 610	43 000
	(12,1)	(12,3)	(18,8)	

Les estimations à bases multiples révisées montrent une amélioration par rapport aux estimations initiales relativement au niveau d'estimations. Ainsi, une grande partie de notre problème de « chevauchement » était principalement d'ordre définitionnel. L'année prochaine, les agents recenseurs recevront des instructions plus explicites et des définitions plus claires des fermes de la liste des fermes retirées des affaires. Il faudra cependant effectuer des travaux supplémentaires pour vérifier la qualité de la détermination des chevauchements pour l'enquête de 1979.

11 Conclusions et recommandations

Le test de 1978 au Nouveau-Brunswick a permis d'obtenir beaucoup de renseignements. De manière générale, le test s'est bien déroulé et les résultats sont encourageants. Il semble que l'échantillonnage à bases multiples est une technique viable pour la collecte de données et la production d'estimations dans une province comme le Nouveau-Brunswick. Il a déjà été décidé que l'utilisation de la base multiple sera étendue aux trois provinces maritimes pour l'enquête de 1979. Ce sera une année pour tester davantage les bases multiples puisqu'une fois encore, des estimations provenant de la base aréolaire seront produites en plus des estimations à bases multiples.

À la lumière de l'analyse présentée dans le présent article, les recommandations suivantes ont été formulées et adoptées pour l'enquête de 1979.

1. Puisque les données doivent être recueillies pour les fermes de la zone de chevauchement en 1979, des estimations de Hartley ainsi que des estimations avec sélection seront produites. Bien que, pour la plupart des éléments, le gain d'efficacité soit faible lorsque les estimations de Hartley sont utilisées, si les données sont disponibles, les calculs supplémentaires nécessaires doivent être effectués. Cela s'avérera particulièrement utile pour les éléments pour lesquels l'estimation à base liste est mauvaise, car c'est avec ceux-ci que l'estimation de Hartley présente la plus grande amélioration.

2. Les strates de taille moyenne seront supprimées de la base liste. Les estimations à base liste pour ces strates ne sont pas aussi bonnes que celles pour les strates plus importantes. Il a été démontré qu'un grand nombre de fermes sélectionnées dans ces strates ont changé de strate au moment de l'enquête, ce qui réduit l'efficacité de la stratification. De même, les fermes de petite ou moyenne taille figurant dans la base liste et disposant de peu de données peuvent être plus difficiles à apparier avec les fermes de l'échantillon aréolaire. Ainsi, l'étape cruciale de la détermination des chevauchements pourrait être mise en danger par l'inclusion de ces fermes dans la base liste.

3. À l'étape de la stratification, la priorité accordée aux porcs doit être la même que celle des bovins lors de la définition des strates de liste. Les données de 1978 indiquent que le fait de donner la priorité aux bovins a entraîné des problèmes d'estimation provenant de la base liste des porcs. Une priorité égale sera accordée aux bovins et aux porcs par l'utilisation de techniques de stratification multiple (ou profonde).

4. Les intervieweurs tenteront de déterminer le chevauchement entre l'échantillon aréolaire et la base liste sur le terrain en 1979. Ils emporteront avec eux une copie de la base liste et, après chaque interview avec une ferme de l'échantillon provenant de la base aréolaire, ils noteront si cette ferme figure ou non dans la base liste. On espère que les intervieweurs seront en mesure de faire cette détermination avec plus de précision que le personnel du bureau central ne l'a fait en 1978. Pour 1979, cependant, l'ensemble de l'appariement entre l'échantillon aréolaire et la base liste sera vérifié au bureau central afin d'évaluer dans quelle mesure les intervieweurs ont pu réaliser cette étape.

   En outre, on a ajouté une question au questionnaire de 1979, ce qui aidera l'intervieweur à déterminer si la ferme de l'échantillon aréolaire est susceptible de se trouver dans la base liste. Voici la question [Traduction] :

   « Cette ferme était-elle exploitée sous le nom actuel le 1er juin 1976? »

   Si la réponse est « non », la ferme est automatiquement une ferme sans chevauchement; si la réponse est « oui », la ferme aurait dû être incluse dans la liste de recensement (mais pas nécessairement dans la base liste) et la base liste serait donc soigneusement vérifiée pour trouver un appariement.

En conclusion, on estime que le test à bases multiples a certainement été utile. Nous poursuivons la vérification en 1979 et si les modifications apportées à l'enquête de 1979 (énumérées ci-dessus) sont fructueuses, nous serons en mesure d'améliorer considérablement nos estimations dans les enquêtes à venir.

12 Remerciements

L'auteur tient à remercier l'examinateur pour ses commentaires et ses suggestions utiles.

Bibliographie

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

Annexe 1

Formules s'appliquant aux bases multiples

a) Notation

$\hat{Y}_L$ = estimation de la population de la base liste à partir de l'échantillon tiré d'une liste.

$\hat{Y}_A$ = estimation de l'échantillon aréolaire pour l'ensemble de la population.

$\hat{Y}_{OL}$ = estimation de l'échantillon aréolaire de la population de la base liste (domaine de chevauchement).

$\hat{Y}_{NOL}$ = estimation de l'échantillon aréolaire de la population non couverte par la base liste (domaine sans chevauchement).

p = facteur de pondération donné à l'estimation provenant de la base liste (pour l'estimation de Hartley).

q = facteur de pondération donné à l'estimation à base aréolaire de la population de la base liste.

$p + q = 1$.

b) Estimation provenant de la base aréolaire

$$\hat{Y}_A = \hat{Y}_{OL} + \hat{Y}_{NOL} .$$

$$\mathrm{Var}\,(\hat{Y}_A) = \mathrm{Var}\,(\hat{Y}_{OL}) + \mathrm{Var}\,(\hat{Y}_{NOL}) + 2\,\mathrm{Cov}\,(\hat{Y}_{OL}, \hat{Y}_{NOL}) .$$

c) Estimation à bases multiples avec sélection

$$\hat{Y}_S = \hat{Y}_L + \hat{Y}_{NOL} .$$

$$\mathrm{Var}\,(\hat{Y}_S) = \mathrm{Var}\,(\hat{Y}_L) + \mathrm{Var}\,(\hat{Y}_{NOL}) .$$

d) Estimation à bases multiples de Hartley

$$\hat{Y}_H = \hat{Y}_{NOL} + q\,\hat{Y}_{OL} + p\,\hat{Y}_L \quad .$$

L'estimation provenant de la base aréolaire est un cas particulier de l'estimation de Hartley lorsque p = 0 et q = 1.

L'estimation avec sélection est un cas particulier de l'estimation de Hartley lorsque p = 1 et q = 0.

L'estimation de Hartley peut s'écrire comme suit :

$$\begin{aligned}\hat{Y}_H &= \hat{Y}_{NOL} + (1-p)\,\hat{Y}_{OL} + p\,\hat{Y}_L \\ &= \hat{Y}_{NOL} + \hat{Y}_{OL} - p\,\hat{Y}_{OL} + p\,\hat{Y}_L \\ &= \hat{Y}_A + p(\hat{Y}_L - \hat{Y}_{OL}).\end{aligned}$$

La variance de cette estimation est :

$$\begin{aligned}Var\,(\hat{Y}_H) &= Var(\hat{Y}_{NOL}) + q^2 Var(\hat{Y}_{OL}) + p^2 Var(\hat{Y}_L) + 2\,q\,Cov(\hat{Y}_{NOL},\hat{Y}_{OL}) \\ &= Var(\hat{Y}_{NOL}) + Var(\hat{Y}_{OL}) - 2\,p\,Var(\hat{Y}_{OL}) + p^2 Var(\hat{Y}_{OL}) + p^2 Var(\hat{Y}_L) \\ &\qquad + 2\,Cov(\hat{Y}_{NOL},\hat{Y}_{OL}) - 2\,p\,Cov(\hat{Y}_{NOL},\hat{Y}_{OL}) \\ &= [Var(\hat{Y}_{NOL}) + Var(\hat{Y}_{OL}) + 2\,Cov(\hat{Y}_{NOL},\hat{Y}_{OL})] \\ &\quad - 2\,p[Var(\hat{Y}_{OL}) + Cov(\hat{Y}_{OL},\hat{Y}_{NOL})] + p^2[Var(\hat{Y}_{OL}) + Var(\hat{Y}_L)] \\ &= Var(\hat{Y}_A) - 2\,p\,[Cov(\hat{Y}_A,\hat{Y}_{OL})] + p^2[Var(\hat{Y}_{OL}) + Var(\hat{Y}_L)].\end{aligned}$$

La valeur optimale de p est maintenant déterminée :

$$\frac{\delta\,Var(\hat{Y}_H)}{\delta\,p} = -\,2\,Cov(\hat{Y}_A,\hat{Y}_{OL}) = 2\,p\,[Var(\hat{Y}_{OL}) + Var(\hat{Y}_L)] = 0 \quad .$$

$$\therefore\ p_{opt} = \frac{Cov(\hat{Y}_A,\hat{Y}_{OL})}{[Var(\hat{Y}_{OL}) + Var(\hat{Y}_L)]} \quad .$$

En utilisant cette valeur de p, la variance de l'estimation de Hartley devient :

$$\begin{aligned}\mathrm{Var}\,(\hat{Y}_H) &= \mathrm{Var}(\hat{Y}_A) - 2\,p_{opt}\,[\mathrm{Cov}(\hat{Y}_A,\hat{Y}_{0L})] + p_{opt}^2\,[\mathrm{Var}(\hat{Y}_L) + \mathrm{Var}(\hat{Y}_{0L})] \\ &= \mathrm{Var}(\hat{Y}_A) - 2\,p_{opt}^2\,[\mathrm{Var}(\hat{Y}_{0L}) + \mathrm{Var}(\hat{Y}_L)] + p_{opt}^2\,[\mathrm{Var}(\hat{Y}_L + \mathrm{Var}(\hat{Y}_{0L})] \\ &= \mathrm{Var}(\hat{Y}_A) - p_{opt}^2\,[\mathrm{Var}(\hat{Y}_{0L}) + \mathrm{Var}(\hat{Y}_L)].\end{aligned}$$

Annexe 2

Estimateur d'échantillon aréolaire

Comme l'indique le tableau 2 (section 2), la sélection des unités d'échantillonnage de premier degré dans l'EDA est répétée. Les données de tous les SD d'une répétition sont gonflées au niveau de la strate. L'estimateur pour la strate h à partir de la répétition k est le suivant :

$$\hat{y}_{hk} = \frac{M_h}{m_h} \sum_{i=1}^{m_h} \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{n_{hi}} y_{hij},$$

où M_h = le nombre de SD dans la strate h,

m_h = le nombre de SD sélectionnés par répétition pour la strate h,

N_{hi} = le nombre de segments dans le i^e SD sélectionnés de la strate h,

n_{hi} = le nombre de segments sélectionnés dans le i^e SD sélectionné de la strate h,

y_{hij} = valeur des données pour le j^e segment du i^e SD dans la strate h.

Habituellement, une partie des terres d'une ferme dénombrée dans un segment est située à l'intérieur du segment et une partie à l'extérieur. Pour ces fermes, les valeurs des données sont ramenées au niveau du segment en appliquant un facteur de pondération égal au rapport entre les terres de la ferme à l'intérieur du segment (à l'exclusion des terres boisées) et la totalité des terres exploitées par la ferme (à l'exclusion des terres boisées). C'est cette valeur « pondérée » des données pour

chaque ferme d'un segment qui est additionnée pour donner le total du segment y_{hij}.

Après avoir obtenu des estimations pour toutes les répétitions, l'estimation de la strate peut être calculée; elle correspond à :

$$\hat{y}_h = \frac{\hat{y}_{h1} + \hat{y}_{h2} + \ldots \hat{y}_{hr_h}}{r_h}$$

où r_h = le nombre de répétitions dans la strate h.

Pour obtenir des estimations provinciales, on additionne les estimations des strates :

$$\hat{Y}_A = \sum_{h=1}^{H} \hat{Y}_h$$

où H = le nombre de strates dans la province.

Les estimations de la variance prennent la forme suivante :

$$\hat{V}(\hat{Y}_A) = \sum_{h=1}^{H} \frac{1}{r_h(r_h-1)} \{(\hat{Y}_{h1} - \hat{Y}_h)^2 + \ldots \quad (\hat{Y}_{hr} - \hat{Y}_h)^2\}.$$