Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 53–70
Statistique Canada

Propriétés statistiques des estimateurs de la production végétale

CAROL A. FRANCISCO, WAYNE A. FULLER, et RON FECSO[1]

RÉSUMÉ

Le National Agricultural Statistics Service du Département de l'agriculture des États-Unis fait des enquêtes de rendement pour diverses grandes cultures aux États-Unis. Bien que les méthodes d'échantillonnage des champs varient selon les cultures, le plan de sondage demeure le même pour toutes. Cet article présente une analyse de ce plan de sondage et des estimateurs actuellement utilisés. Les auteurs définissent également d'autres estimateurs du rendement et de la production ainsi que des estimateurs de la variance des estimateurs, puis les comparent aux estimateurs courants par une approche théorique et une simulation de Monte Carlo.

MOTS CLÉS: Enquêtes sur les cultures; estimation du rendement; échantillon à deux phases; estimation de variance.

1. INTRODUCTION

Le National Agricultural Statistics Service (anciennement le Statistical Reporting Service) du Département de l'agriculture des États-Unis procède à des enquêtes objectives sur le rendement du maïs, du coton, du soja, du riz, du sorgho à grains, du tournesol et du blé dans les principaux États producteurs. Certains autres pays réalisent le même genre d'enquêtes.

Bien que les méthodes d'échantillonnage des champs varient selon les cultures (taille des parcelles, méthodes de délimitation des parcelles et techniques de mesure appliquées aux légumes et aux fruits), le plan de sondage est le même pour toutes les enquêtes. En effet, les enquêtes objectives sur le rendement reposent sur une méthode d'échantillonnage à quatre degrés. On trouvera à la section 2 une description détaillée de ce plan de sondage. Dans la section 3, nous décrivons et évaluons les estimatuers du rendement moyen des cultures et les estimateurs de la variance. Nous y examinons également d'autres estimateurs. La section 4 renferme les conclusions de l'étude et des recommandations.

2. PLAN DE SONDAGE DE L'ENQUÊTE OBJECTIVE SUR LE RENDEMENT

Les deux premiers degrés de l'échantillonnage du plan de sondage produisent l'échantillon de segments aréolaires qui sert à l'enquête énumérative de juin du National Agricultural Statistics Service (NASS). Dans chaque État, la base aréolaire (base de sondage) est stratifiée suivant le mode d'exploitation du sol. L'État de Californie, par exemple, est divisé en 12 strates d'exploitation. Chacune de ces strates est subdivisée en territoires appelés unités de base. La superficie des unités de base varie; la taille réelle d'une unité de base dépend des renseignements qui existent sur la désignation des frontières, des renseignements complémentaires recueillis, des frontières politiques, etc.. Une fois que les unités de base sont définies, on détermine le nombre de segments aréolaires dans chacune, en divisant la superficie totale d'une unité par la taille de segment désirée. Celle-ci varie selon la strate d'exploitation dans

[1] Carol A. Francisco, Syntex Laboratories Inc., 3401 Hillview Avenue, Palo Alto, Californie 94304; Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa 50011; et Ron Fecso, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C.20250.

laquelle se trouve l'unité de base. En Californie, par exemple, la taille de segment recherchée est ½ mille carré pour les strates de vergers et 1 mille carré pour toutes les autres strates de terres labourables. Une unité de base contient normalement de 1 à 30 segments aréolaires.

Chaque strate d'exploitation est sous-stratifiée en fonction du lieu géographique. Pour former les sous-strates géographiques, on classe les unités de base de chaque strate d'exploitation par comté de manière que des comtés voisins qui ont des caractéristiques agricoles comparables sont fondus en un seul (Fecso 1978), puis on forme de nouveaux groupes de segments aréolaires en suivant l'ordre de classement des unités de base. Ainsi, chaque sous-strate renferme des segments aréolaires qui ont les mêmes caractéristiques agricoles et qui sont voisins. Dans une strate d'exploitation donnée, les sous-strates comptent toutes le même nombre de segments et ont toutes la même superficie, après arrondissement. Pour obtenir des renseignements détaillés sur la conception de la base aréolaire, le lecteur est invité à consulter Fecso et Johnson (1981) et Houseman (1975).

Pour l'estimation de la variance, les sous-strates servent de strates d'échantillonnage. Aussi, les sous-strates d'exploitation seront désignées les strates.

La première étape de l'échantillonnage consiste à prélever des unités de base dans chaque strate. Le nombre d'unités de base prélevées dépend de la nature de la strate. En règle générale, on prélève de 8 à 15 unités dans les strates de terres labourables tandis qu'on en prélève 4 ou 5 dans les strates agro-urbaines, urbaines et non agricoles. Les unités sont prélevées aléatoirement selon une probabilité proportionnelle au nombre de segments aréolaires qu'elles contiennent. La seconde étape consiste à prélever au hasard un segment aréolaire dans chaque unité de base échantillonnée. Ainsi, la probabilité de sélection est la même pour tous les segments aréolaires d'une même strate.

Dans le plan de sondage qui nous intéresse, l'unité de base est l'unité primaire d'échantillonnage. Comme les unités de base sont prélevées selon une probabilité proportionnelle au nombre de segments qu'elles contiennent et qu'un segment est prélevé dans chaque unité de base échantillonnée, on peut considérer le segment comme l'unité primaire d'échantillonnage. Dans notre étude, les deux premiers degrés de l'échantillonnage ne font qu'un et l'échantillon de segments est considéré comme un échantillon aléatoire simple stratifié à un seul degré. Comme la fraction de sondage moyenne est d'environ 1%, nous n'utiliserons pas le terme correctif de la population finie dans notre analyse.

Les troisième et quatrième degrés de l'échantillonnage consistent respectivement à échantillonner des champs et à prélever des parcelles à l'intérieur de ces champs. Dans l'enquête énumérative de juin, on examine chaque segment aréolaire échantillonné pour y recenser les champs qui ont été ensemencés de la plante d'intérêt ou qui doivent l'être. Les champs ainsi recensés sont classés par numéro de segment et par ordre d'énumération à l'intérieur de chaque segment. On prélève ensuite un échantillon systématique de champs selon une probabilité proportionnelle au produit de la superficie du champ par l'inverse de la probabilité de sélection du segment aréolaire qui contient le champ. Ainsi, le nombre de champs échantillonnés par segment varie et il se peut que les grands champs soient échantillonnés plus d'une fois.

Au quatrième et dernier degré de l'échantillonnage, on définit deux parcelles de superficies comparables dans chaque champ échantillonné à l'aide d'une méthode de délimitation aléatoire fondée sur les rangées et les pas. Lorsque les rangées sont difficiles à distinguer l'une de l'autre ou lorsqu'il s'agit de blé, on délimite les parcelles à l'aide d'un nombre aléatoire de pas en bordure du champ et d'un nombre aléatoire de pas vers l'intérieur du champ. Une méthode différente s'applique aussi dans les enquêtes objectives sur le rendement du blé. Pour ce genre d'enquête, on délimite aléatoirement la première parcelle puis on situe la seconde par rapport à la première. Si un grand champ est échantillonné plus d'une fois au troisième degré de l'échantillonnage, on prélève séparément des paires de parcelles additionnelles. Comme les parcelles sont toujours échantillonnées par paire, une paire de parcelles est désignée ici l'unité secondaire. On ne peut avoir plus de huit parcelles (c'est-à-dire quatre unités secondaires) par champ.

3. MÉTHODES D'ESTIMATION

En principe, l'échantillon de l'enquête objective sur le rendement est le résultat d'un sondage à deux phases avec sous-échantillonnage dans la seconde phase. Le tableau 1 donne une description schématique de cet échantillon. L'échantillon produit à la première phase est un échantillon aléatoire simple stratifié de segments. L'échantillon produit à la seconde phase comprend tous les segments qui ne sont pas ensemencés de la culture étudiée de même qu'un échantillon de segments ensemencés de cette culture, ces segments étant prélevés selon une probabilité proportionnelle à la superficie cultivée. L'échantillon de segments est tiré des champs ensemencés de la culture étudiée qui ont été prélevés dans la première phase utilisant l'échantillonnage systématique avec probabilités de sélection proportionnelles à la superficie. Un échantillon d'unités secondaires - où chaque unité secondaire correspond à une paire de parcelles - est ensuite prélevé dans les segments ensemencés qui ont été échantillonnés à la seconde phase.

Puisque l'unité secondaire est toujours une paire de parcelles, nous ne parlerons plus désormais de parcelles mais bien d'unités secondaires. De même, nous oublierons que les champs sont les unités opérationnelles qui servent à délimiter les parcelles et nous parlerons uniquement de segments échantillonnés. Notons que l'on observe deux genres de segments dans la seconde phase; les segments qui sont ensemencés de la culture étudiée et ceux qui ne le sont pas. Le nombre total de segments dans la seconde phase est K. La superficie et la production totale d'un segment qui n'est pas ensemencé de la culture considérée sont connues (toutes deux nulles). En ce qui a trait aux segments ensemencés de la culture considérée, on utilise un sous-échantillon d'unités secondaires pour estimer la production.

Soit M_{hk} le nombre d'unités secondaires incluses dans le segment k de la h-ième strate. On peut supposer, sans perte de généralité, que M_{hk} est égal à A_{hk} (A_{hk} étant la superficie cultivée du segment hk). La validité de cette identité repose essentiellement sur le choix d'une échelle d'équivalence appropriée pour la superficie.

Tableau 1

Méthode d'échantillonnage pour l'enquête objective sur le rendement

Phase/unité d'échantillonnage	Méthode de sélection	Nombre[1] échantillonnées	Données recueillies
Première phase			
Unité primaire d'échantillonnage: segment	probabilité égale à l'intérieur des strates	n_h	superficie cultivée
Seconde phase			
Unité primaire d'échantillonnage: segment	probabilité inégale	K_h	superficie cultivée, production estimée[2]
Unité secondaire d'échantillonnage: paires de parcelles	probabilité égale	m_{hk}	production des parcelles estimée

[1] Par strate pour les unités primaires d'échantillonnage et par segment pour les unités secondaires d'échantillonnage.
[2] La production du segment est nulle si la superficie ensemencée est nulle, autrement elle est estimée à l'aide des données relatives aux parcelles.

Dans la section 3.1, nous étudions l'estimateur de rendement actuellement en usage et voyons à quelles conditions cet estimateur est sans biais pour le rendement moyen d'un État. Dans la section 3.2, nous analysons un estimateur simple de la variance du rendement estimé tandis que, dans la section suivante, nous définissons des estimateurs des variances non conditionnelles des estimateurs du rendement et de la production. Enfin, la section 3.4 décrit une simulation de Monte Carlo appliquée aux estimateurs.

3.1 Estimateur de rendement et estimateur de production actuellement en usage

À l'heure actuelle, on estime le rendement moyen d'un État comme si l'échantillon était un échantillon aléatoire simple d'unités secondaires avec probabilités égales. L'estimateur est le rendement moyen simple des unités secondaires qui sont ensemencés de la culture considérée. En d'autres termes, l'estimation du rendement moyen à l'acre est définie comme étant

$$\bar{y} = D^{-1} \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell}\, \delta_{hk\ell}, \tag{3.1}$$

où

$$\delta_{hk\ell} = 1 \quad \text{si } A_{hk} > 0,$$

$$\delta_{hk\ell} = 0 \quad \text{si } A_{hk} = 0,$$

$$D = \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} \delta_{hk\ell}, \tag{3.2}$$

m_{hk} est le nombre d'unités secondaires prélevées dans le segment hk, L est le nombre de strates et $Y_{hk\ell}$ est l'estimation du rendement à l'acre pour l'unité secondaire ℓ du segment hk. Si la superficie cultivée dans un segment (A_{hk}) est nulle, alors $m_{hk} = 1$ et $Y_{hk\ell} = 0$, par définition. Le nombre total d'unités secondaires observées dans les segments ensemencés de la culture en question est D.

L'équation (3.1) peut être réécrite sous la forme opérationnelle suivante:

$$\bar{y} = D^{-1} \sum_{t=1}^{D} Y_t, \tag{3.3}$$

où l'indice inférieur t remplace l'indice triple $hk\ell$ et où la sommation porte sur les unités secondaires comprises dans les segments ensemencés de la culture en question.

L'estimateur du rendement moyen à l'acre (3.1) est un genre d'estimateur par quotient combiné. On peut vérifier cela en se servant des probabilités de sélection conditionnelles pour réécrire $\bar{y}$. Dans le plan de sondage du NASS, les segments sont prélevés systématiquement avec probabilités proportionnelles à la superficie accrue et ceux dont la superficie accrue est suffisamment élevée sont inclus à coup sûr dans l'échantillon. Le nombre d'unités secondaires attribué à cette catégorie de segments est proportionnel à la taille du segment, mise à part l'erreur d'arrondissement. L'arrondissement est effectué suivant le plan d'échantillonnage systématique. Soit $\pi_{hk\ell}$ la probabilité conditionnelle que l'unité secondaire ℓ du segment k de la strate h soit échantillonnée, étant donné l'échantillon de segments prélevés à la première phase d'échantillonnage. Nous avons

$$\pi_{hk\ell} = D \left(\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk} \right)^{-1} N_h n_h^{-1} \tag{3.4}$$

pour les unités secondaires incluses dans les segments dont $A_{hk} > 0$ où N_h est le nombre global de segments inclus dans la strate h, M_{hk} est le nombre d'unités secondaires incluses dans le segment k de la strate h, et n_h est le nombre de segments inclus dans la strate h qui ont été prélevés à la première phase. La probabilité conditionnelle que l'on observe un segment non ensemencé de la culture en question dans la seconde phase est un.

On peut alors exprimer l'estimateur de la moyenne défini en (3.1) par l'équation suivante:

$$\bar{y} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} Y_{hk\ell}}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} \delta_{hk\ell}}, \tag{3.5}$$

où $N_h n_h^{-1}$ est l'inverse de la probabilité de sélection à la première phase, K_h est le nombre de segments prélevés dans la strate h à la seconde phase, et $K = \Sigma K_h$. Étant donné une échelle appropriée, le numérateur de (3.5) est un estimateur de la production totale et le dénominateur est un estimateur de la superficie totale. Il est possible de montrer que le numérateur est un estimateur sans biais en calculant les espérances mathématiques; pour cela, on définit tout d'abord une relation conditionnelle basée sur les unités échantillonnées de la première phase, puis on fait la moyenne pour l'ensemble de ces échantillons. Le dénominateur est un estimateur stratifié du nombre total d'unités secondaires. Par la nature du sondage, le nombre d'unités d'échantillonnage est proportionnel à la superficie, et on peut choisir l'échelle de telle sorte que le nombre d'unités secondaires soit égal à la superficie. Ainsi, on peut considérer $\bar{y}$ comme le rapport d'un estimateur sans biais de la production totale d'une culture à un estimateur sans biais de la superficie totale consacrée à cette culture.

Pour estimer la production totale d'un État, le NASS multiplie $\bar{y}$ par $\hat{A}$, où $\hat{A}$ est l'estimateur de la superficie cultivée totale, qui prend la forme

$$\hat{A} = \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk} . \tag{3.6}$$

L'estimation de la production totale est donc

$$\hat{Y} = \hat{A}\,\bar{y} . \tag{3.7}$$

3.2 Estimateurs de variance simples

Suivant l'hypothèse d'un échantillonnage aléatoire simple d'unités secondaires parmi toutes les unités secondaires disponibles dans la seconde phase, l'estimation de la variance conditionnelle de y, étant donné les segments de la seconde phase, est

$$\hat{V}_2(\bar{y}) = D^{-1} (D-1)^{-1} \sum_{t=1}^{D} (Y_t - \bar{y})^2 , \tag{3.8}$$

où l'indice inférieur 2 (par rapport à $\hat{V}$) sert à identifier la variance conditionnelle et l'indice inférieur t (par rapport à Y) remplace l'indice triple $hk\ell$. La somme pour t variant de 1 à D est la somme effectuée pour les D unités secondaires incluses dans les segments ensemencés de la culture en question.

À cause de la simplicité de l'expression (3.8), on a proposé de l'utiliser comme estimateur de la variance non conditionnelle. On a également proposé d'estimer la variance de l'estimation de la production totale d'un État par l'équation suivante:

$$\hat{V}_*(\hat{Y}) = \hat{A}^2 \hat{V}_2(\bar{y}) + \bar{y}^2 \hat{V}(\hat{A}) + \hat{V}(\hat{A}) \hat{V}_2(\bar{y}), \tag{3.9}$$

où $\hat{A}$ est défini en (3.6) et $V(\hat{A})$ est l'estimateur habituel de la variance d'un total estimé stratifié,

$$\hat{V}(\hat{A}) = \sum_{h=1}^{L} N_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2, \tag{3.10}$$

et

$$\bar{A}_h = n_h^{-1} \sum_{k=1}^{n_h} A_{hk}.$$

L'estimateur (3.9) est un estimateur de la variance d'un produit, qui est fondé sur l'hypothèse implicite que $\bar{y}$ et $\hat{A}$ sont non corrélés.

Il est difficile d'évaluer dans quelle mesure l'estimateur (3.9) tend à sous-estimer la variance de $\hat{Y}$. La variance non conditionnelle de $\bar{y}$ est

$$\begin{aligned} V(\bar{y}) &= V_1 \{E_2(\bar{y})\} + E_1 \{V_2(\bar{y})\} \\ &= V_1 \{\hat{A}^{-1} \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\} + E_1 \{V_2(\bar{y})\}, \end{aligned} \tag{3.11}$$

où $Y_{hk.} = M_{hk} \bar{Y}_{hk.}$ est le total pour le segment k dans la strate h, et E_1 et V_1 désignent respectivement l'espérance mathématique et la variance suivant le plan d'échantillonnage de la première phase.

Dans le cas d'un échantillonnage aléatoire simple des unités secondaires, l'estimateur $\hat{V}_2(\bar{y})$ est non biaisé pour le second terme du membre de droite de l'équation (3.11). Comme le plan de sondage du NASS prévoit un échantillonnage systématique dans la seconde phase, $\hat{V}_2(\bar{y})$ est un estimateur biaisé de $V_2(\bar{y})$. La nature et l'importance du biais dépendent de la structure de corrélation des unités de la liste servant à l'échantillonnage de la seconde phase. L'estimateur de la variance $\hat{V}_2(\bar{y})$ est également biaisé parce qu'on a établi la formule de la variance en supposant un échantillonnage avec remise dans la seconde phase. Dans la mesure où l'échantillonnage à la seconde phase se fait sans remise (comme les échantillons sont prélevés systématiquement dans la liste des segments élargis, seuls les segments de grande superficie sont échantillonnés plus d'une fois), $\hat{V}_2(\bar{y})$ surestimera $V_2(\bar{y})$.

L'estimateur $\hat{V}_*(\hat{Y})$ ne renferme pas d'estimateur de $A^2 V_1\{E_2(\bar{y})\}$; cela a pour effet de créer un biais négatif. Ce terme n'est toutefois pas facile à estimer, même suivant l'hypothèse simplificatrice d'un échantillonnage avec probabilité proportionnelle à la taille dans la seconde phase. Aussi, nous analyserons l'efficacité de $\hat{V}_*(\hat{Y})$ par la méthode de Monte Carlo (section 3.4).

3.3 Autres estimateurs de variance

Une autre façon d'estimer $V(\bar{y})$ est de considérer l'échantillon comme le résultat d'un sondage à deux phases (voir tableau 1) et de supposer que la probabilité non conditionnelle qu'un segment soit prélevé à l'intérieur d'une strate pour recevoir une unité secondaire est

proportionnelle à sa probabilité conditionnelle de sélection à la seconde phase, étant donné les segments échantillonnés à la première phase.

Soit π_{hk} la probabilité conditionnelle que le segment k de la strate h soit prélevé dans la seconde phase, étant donné l'échantillon de segments de la première phase. Nous avons

$$\pi_{hk} = min\,(1,\ M_{hk}\pi_{hk\ell})\,, \tag{3.12}$$

où $N_{hk\ell}$ est une constante dans le segment hk. Si $\pi_{hk} = 1$ et que le segment est échantillonné pour recevoir plus d'une unité secondaire, on suppose que les unités secondaires sont prélevées séparément.

Soit π^*_{hk} la probabilité non conditionnelle que le segment k de la strate h fasse l'objet d'une observation dans la seconde phase. Si $A_{hk} = 0$ alors π^*_{hk} est la probabilité non conditionnelle que le segment hk soit échantillonné pour recevoir au moins une unité secondaire. Si $A_{hk} = 0$ alors π^*_{hk} est égale à la probabilité que le segment hk soit échantillonné dans la première phase d'échantillonnage. Posons

$$\begin{aligned} \pi^*_{hk} &= \frac{n_h}{N_h} && \text{si } A_{hk} = 0\,, \\ \pi^*_{hk} &= \pi_{hk}\frac{n_h}{N_h} && \text{si } 0 < \pi_{hk} < 1\,, \end{aligned} \tag{3.13}$$

où π_{hk}, définie en (3.12), est la probabilité conditionnelle que le hk-ième segment soit échantillonné à la seconde phase, étant donné l'échantillon de segments de la première phase.

Dans notre analyse, nous supposons que π^*_{hk} est fixe. Cette hypothèse se vérifiera et π^*_{hk} sera la vraie probabilité non conditionnelle si π_{hk} est un multiple déterminé de M_{hk}, qui aura été déterminé avant l'échantillonnage. L'expression (3.13) sera une approximation si π_{hk} est une fonction des segments prélevés au premier degré d'échantillonnage.

L'expression (3.13) est proportionnelle à M_{hk} pour $M_{hk}\pi_{hk\ell} \leq 1$. Si $M_{hk}\pi_{hk\ell} > 1$, le nombre d'unités secondaires échantillonnées est égal ou supérieur à 1. $M_{hk}\pi_{hk\ell}$ est le nombre exact d'unités secondaires qu'il faut attribuer aux segments pour conserver un échantillon auto-pondéré d'unités secondaires. Il n'y a jamais plus qu'une unité d'écart entre $M_{hk}\pi_{hk\ell}$ et le nombre d'unités secondaires effectivement observées par suite d'un échantillonnage systématique avec probabilité proportionnelle à la taille.

Pour simplifier le reste des calculs, nous supposons que l'échantillonnage systématique ne comporte aucune erreur d'arrondissement. En d'autres termes, nous supposons que le nombre d'unités secondaires observées par segment est exactement le nombre nécessaire pour obtenir un échantillon auto-pondéré. Par conséquent, nous supposons que le nombre d'unités secondaires observées dans un segment prélevé à la seconde phase d'échantillonnage est

$$\begin{aligned} m_{hk} &= 1 && \text{si } 0 < \pi_{hk} < 1\,, \\ m_{hk} &= M_{hk}\pi_{hk\ell} && \text{si } \pi_{hk} = 1\,. \end{aligned} \tag{3.14}$$

Suivant cette hypothèse, un estimateur par quotient combiné du rendement moyen avec probabilités inégales équivaut à l'estimateur défini en (3.1). L'estimateur par quotient combiné est

$$\bar{y}_r = \hat{M}_r^{-1} \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi^{*-1}_{hk} M_{hk}\bar{y}_{hk.}\,, \tag{3.15}$$

où

$$\bar{y}_{hk.} = m_{hk}^{-1} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell} \qquad \text{si } A_{hk} > 0,$$

$$\bar{y}_{hk.} = 0 \qquad \text{si } A_{hk} = 0,$$

$$\hat{M}_r = \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk}.$$

Dans l'équation (3.15) et les autres équations données dans le reste de cette section, $\hat{M}_r$ (nombre total d'unités secondaires) équivaut à $\hat{A}_r$ (superficie totale); le lecteur peut à son gré substituer l'un à l'autre.

Dans l'analyse qui suit, nous supposons que l'échantillonnage de segments avec remise selon une probabilité proportionnelle à la superficie ensemencée de la culture en question est une approximation de l'échantillonnage systématique avec probabilité proportionnelle à la taille effectué à la seconde phase. Suivant l'hypothèse de l'échantillonnage avec remise, un estimateur de la variance de $\bar{y}$ est défini comme étant

$$\hat{V}(\bar{y}_r) = \hat{M}_r^{-2} \sum_{h=1}^{L} K_h (K_h - 1)^{-1} \sum_{k=1}^{K_h} (\pi_{hk}^{*-1} u_{hk} - \bar{u}_{h.})^2, \tag{3.16}$$

où

$$u_{hk} = M_{hk} (\bar{y}_{hk.} - \bar{y}_r),$$

$$\bar{u}_{h.} = K_h^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} u_{hk}.$$

Un estimateur de la production totale est défini par

$$\hat{Y}_r = N \bar{M}_n \bar{y}_r, \tag{3.17}$$

où

$$\bar{M}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk}.$$

N est le nombre total de segments dans la population et $W_h = N^{-1} N_h$. La formule d'approximation de Taylor appliquée à la variance non conditionnelle de la distribution approximative de $\hat{Y}_r$ est

$$V\{\hat{Y}_r\} = N^2 [\bar{M}_N^2 V\{\bar{y}_r\} + 2\bar{M}_N \bar{\bar{y}}_N C\{\bar{y}_r, \bar{M}_n\} + \bar{\bar{y}}_N^2 V\{\bar{M}_n\}], \tag{3.18}$$

où $\bar{y}_r$ est défini en (3.15), $\bar{M}_n$ est défini en (3.17),

$$\bar{M}_N = N^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk},$$

$$\bar{\bar{y}}_N = \left(\sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk} \right)^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} Y_{hk.},$$

$Y_{hk.} = M_{hk}\bar{Y}_{hk.}$ est la production totale du k-ième segment de la strate h, et $C\{\bar{y}_r, \bar{M}_n\}$ est la covariance de $\bar{y}_r$ et de $\bar{M}_n$.

En ce qui a trait aux échantillons de taille fixe avec probabilités inégales, l'estimateur $\bar{y}_r (\doteq \bar{y})$ est à peu près conditionnellement non biaisé pour le rendement moyen des $n = \Sigma n_h$ segments de l'échantillon de la première phase. Le rendement moyen des n segments est

$$\bar{\bar{y}}_n = \bar{M}_n^{-1} \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.} .$$

Par conséquent, la covariance de $\bar{y}_r$ et $\bar{M}_n$ est la covariance de $\bar{M}_n^{-1}\,\bar{Y}_n$ et $\bar{M}_n$, où

$$\bar{Y}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.} .$$

À l'aide de la formule d'approximation couramment utilisée pour les rapports, la covariance de $\bar{y}_r$ et $\bar{M}_n$ est donnée, approximativement, par

$$C\{\bar{M}_n^{-1}\,\bar{Y}_n, \bar{M}_n\} \doteq C\{(\bar{Y}_n - \bar{\bar{y}}_N \bar{M}_n)\bar{M}_N^{-1}, \bar{M}_n\}$$

$$= \bar{M}_N^{-1}[C\{\bar{Y}_n, \bar{M}_n\} - \bar{\bar{y}}_N V\{\bar{M}_n\}] . \tag{3.19}$$

Si la probabilité que l'on observe le couple $(Y_{hk.}, M_{hk})$ est proportionnelle à π^*_{hk}, l'estimateur de la covariance de $\bar{Y}_n$ et de $\bar{M}_n$ est défini par

$$\hat{C}\{\bar{Y}_n, \bar{M}_n\} = \sum_{h=1}^{L} W_h^2\, n_h^{-1}\, \hat{S}_{MYh}, \tag{3.20}$$

où

$$\hat{S}_{MYh} = K_h (K_n^{-1})^{-1} \left(\sum_{j=1}^{K_h} \pi^{*-1}_{hj} \right)^{-1} \sum_{k=1}^{K_h} \pi^{*-1}_{hk} (M_{hk} - \bar{M}^*_h)(M_{hk}\bar{y}_{hk.} - \bar{y}^*_{h..}) ,$$

$$\bar{M}^*_h = \left(\sum_{j=1}^{K_h} \pi^{*-1}_{hj} \right)^{-1} \sum_{k=1}^{K_h} \pi^{*-1}_{hk} M_{hk} ,$$

$$\bar{y}^*_{h..} = \left(\sum_{j=1}^{K_h} \pi^{*-1}_{hj} \right)^{-1} \sum_{k=1}^{K_h} \pi^{*-1}_{hk} M_{hk}\bar{y}_{hk.} .$$

L'estimateur $\hat{S}_{MYh}$ équivaut à un estimateur par quotient d'Horvitz-Thompson de la moyenne des produits $(M_{hk} - \bar{M}_h)(Y_{hk.} - \bar{Y}_{h..})$ rajusté pour tenir compte du nombre de degrés de liberté. Le facteur de rajustement $K_h (K_h - 1)^{-1}$ est rendu nécessaire parce que la construction du produit exige que l'on remplace les moyennes de la population par les moyennes d'échantillon.

En substituant les équations (3.15), (3.16) et (3.20) aux termes correspondants dans l'équation (3.18), on obtient

$$\hat{V}\{\hat{Y}_r\} = N^2[\bar{M}_n^2\, \hat{V}\{\bar{y}_r\} + 2\bar{y}_r \hat{C}\{\bar{Y}_n, \bar{M}_n\} - \bar{y}_r^2\, \hat{V}\{\bar{M}_n\}] , \tag{3.21}$$

où $\hat{V}\{\bar{M}_n\}$ est l'estimateur de la variance d'une moyenne stratifiée. L'équation (3.21) définit un estimateur de la variance de l'estimation de la production totale d'un État pour un échantillonnage double stratifié. Contrairement à l'estimateur $\hat{V}_*(\hat{Y})$ défini en (3.9), l'estimateur (3.21) ne repose pas sur l'hypothèse de l'absence de corrélation entre l'estimateur du rendement et l'estimateur de la superficie. L'équation (3.21) comporte également un estimateur non conditionnel de la variance du rendement.

3.4. Comparaison d'estimateurs par la méthode de Monte Carlo

Nous avons procédé à une étude par la méthode de Monte Carlo pour illustrer les différences entre divers estimateurs. À cette fin, nous nous sommes servis des données sur la superficie ensemencée de coton tirées de l'enquête énumérative de juin 1983 en Californie et des données de l'enquête objective sur le rendement correspondante de 1983. Pour les besoins de notre analyse, nous avons considéré que le coton était cultivé dans 28 strates.

Le tableau 2 donne la répartition de cette culture entre les 28 strates, conformément aux données de l'enquête énumérative de 1983. Fecso et Johnson (1981) ont décrit les six modes d'exploitation du sol, qui sont identifiés par les deux premiers chiffres du numéro de la strate; ces modes d'exploitation sont définis ci-dessous avec le code correspondant:

1300 - au moins 50 pour cent des terres sont cultivées; principalement des cultures générales et au plus 10 pour cent de la superficie consacrée à la culture des fruits ou des légumes;
1700 - au moins 50 pour cent des terres sont cultivées; principalement culture des fruits, des noix ou du raisin combinée à des cultures générales;
1900 - au moins 50 pour cent des terres sont cultivées; principalement culture des légumes combinée à des cultures générales;
2000 - de 15 à 50 pour cent des terres sont cultivées; culture extensive et foin;
3100 - zones résidentielles et terres agricoles; plus de 20 logements au mille carré;
4100 - moins de 15 pour cent des terres sont cultivées; principalement de grands pâturages privés.

Nous avons simulé une population à partir des résultats de l'enquête énumérative de juin 1983. Le tableau 2 permet de comparer les caractéristiques de la population simulée aux résultats de l'enquête. Dans la population simulée, on pouvait dire que du coton était cultivé dans le segment k ($k = 1, \ldots, N_h$) de la strate h ($h = 1, \ldots, 28$) si $X_{hk} = 1$, où X_{hk} est une variable (aléatoire) de Bernoulli (p_h) indépendante, p_h étant la proportion de segments de la strate h où, selon l'enquête énumérative de 1983, du coton était cultivé.

La seconde étape de la création de la population consistait à attribuer une superficie (en acres) ensemencée de coton aux segments pour lesquels $X_{hk} = 1$. On a donc calculé une série de ratios mettant en relation la superficie ensemencée de coton dans un segment et la superficie moyenne des segments pour les strates d'exploitation qui renfermaient plus d'un segment ensemencés de coton selon l'enquête énumérative de juin 1983. Cet ensemble de ratios a servi à déterminer la superficie ensemencée de coton dans les segments où du coton était cultivé. Si X_{hk} était égal à 1, un ratio (r_{hk}) était prélevé dans l'ensemble de ratios observés de telle sorte que la probabilité de sélection était la même pour tous les ratios de l'ensemble. La superficie (en acres) ensemencée de coton dans le segment hk, M_{hk}, était définie par

$$M_{hk} = r_{hk}\bar{M}_{h.}, \tag{4.1}$$

où $\bar{M}_{h.}$ était la superficie moyenne consacrée à la culture du coton pour les segments de la strate h où, selon l'enquête énumérative de juin 1983, du coton était cultivé. (Voir tableau 2.)

Les résultats de l'enquête objective sur le rendement de 1983 pour le coton ont servi à simuler les observations de rendement à l'intérieur des segments. Les estimations de rende-

Tableau 2

Estimations de la superficie ensemencée de coton selon l'enquête énumérative de juin 1983 en Californie et la superficie ensemencée de coton pour la population simulée

Strate	Taille du segment visée (acres)	Nombre de segments dans la strate	Nombre de segments échant. en 1983	Proportion de segments ensemencés de coton		Superficie moyenne ensemencée de coton dans les segments ensemencés de coton	
				1983	Population simulée	1983	Population simulée
1314	640	291	10	60	60	197	200
1315	640	291	10	100	100	354	348
1316	640	291	10	90	89	167	173
1317	640	291	10	90	92	149	148
1318	640	291	10	50	53	481	422
1319	640	291	10	20	19	249[1]	260
1320	640	291	10	90	91	154	155
1321	640	291	10	60	61	270	274
1322	640	291	10	70	71	205	210
1323	640	291	10	80	79	288	279
1713	320	432	10	30	28	125	122
1714	320	432	10	30	31	58	57
1715	320	432	10	20	22	86[2]	84
1716	320	432	10	10	8	86[2]	89
1717	320	432	10	40	38	26	27
1718	320	432	10	30	29	144	144
1719	320	432	10	30	31	65	67
1720	320	432	10	30	30	38	35
1721	320	432	10	30	29	133	138
1722	320	432	10	50	47	130	131
1723	320	432	10	40	40	76	76
1906	640	362	10	70	73	117	127
1907	640	362	10	70	74	192	194
1908	640	362	10	80	83	253	246
2010	640	649	10	30	31	303	306
2011	640	649	10	40	41	175	165
3107	160	1,847	5	20	22	25[3]	25
4110	2,560	1,044	10	10	10	178	165

[1] Moins de trois segments échantillonnés; ce chiffre représente la moyenne pour tous les segments des sous-strates de la strate d'exploitation 13.

[2] Moins de trois segments échantillonnés; ce chiffre représente la moyenne pour tous les segments des sous-strates de la strate d'exploitation 17.

[3] Moins de trois segments échantillonnés; ce chiffre représente la superficie approximative ensemencée de coton pour cette strate agro-urbaine.

ment étant difficiles à obtenir, on s'est servi d'une variable qui est une composante importante de ces estimations, soit le nombre de plants aux 100 pi^2. Dans l'enquête objective sur le rendement de 1983, le nombre moyen estimé de plants aux 100 pi^2 pour l'ensemble de la population était de 79.6. Le tableau 3 donne, pour chaque strate étudiée, le nombre moyen de plants aux 100 pi^2 selon l'enquête de 1983. La moyenne pour chaque strate est fondée sur toutes les unités secondaires de la strate qui ont été prélevées par échantillonnage avec probabilité proportionnelle à la taille estimée.

Une analyse de variance portant sur les données de 1983 (tableau 4) a montré que 28% de la variabilité totale entre les unités secondaires était attribuable à des différences entre segments à l'intérieur des strates ($s_b^2 = 378.0$), tandis que 58% de cette variabilité était attribuable à la variation entre unités secondaires à l'intérieur des segments ($s_w^2 = 776.6$). Si la

Tableau 3

Nombre moyen de plants aux 100 pi^2 selon l'enquête objective sur le rendement de 1983 pour le coton en Californie et dans la population simulée

Strate	Nombre moyen de plants aux 100 pi^2	
	Enquête objective sur le rendement de 1983	Population simulée
1314	78	76
1315	80	80
1316	67	68
1317	72	73
1318	80	80
1319	93	93
1320	92	91
1321	70	69
1322	84	84
1323	72	71
1713	118	117
1714	96[1]	95
1715	96[1]	93
1716	96[1]	86
1717	96[1]	96
1718	139	140
1719	96[1]	97
1720	96[1]	97
1721	89	86
1722	79	79
1723	84	85
1906	98	98
1907	67	67
1908	53	53
2010	118	118
2011	47	47
3107	80[2]	79
4110	60	59

[1] Moins de trois unités secondaires observées; ce chiffre représente la moyenne pour les unités secondaires de la strate d'exploitation 17.

[2] Moins de trois unités secondaires observées; ce chiffre représente la moyenne pour les unités secondaires de toutes les strates.

composante de la variance liée à la strate est tenue pour constante, 67% de la variation à l'intérieur des segments est attribuable à la variabilité entre les unités secondaires.

Lorsqu'il avait été établi qu'un segment comprenait des champs de coton, le nombre moyen de plants aux 100 pi^2 pour le segment hk était simulé par

$$\bar{c}_{hk} = \bar{c}_{h.} + e_{hk}, \tag{4.2}$$

où $\bar{c}_{h.}$ est le nombre moyen de plants aux 100 pi^2 pour la strate h, e_{hk} est distribuée suivant une loi normale $N(0, s_b^2)$, et $s_b^2 = 378.0$. Si la moyenne obtenue par la simulation $(\bar{c}_{hk})$ équivalait à moins de 10% de la moyenne pour la strate, on fixait la valeur de c_{hk} à $(.10)\bar{c}_{h.}$. Le tableau 3 permet de comparer les moyennes obtenues pour la population simulée avec celles obtenues par l'enquête objective sur le rendement de 1983. La moyenne globale pour la population simulée était $\bar{\bar{y}}_N = 79.6$.

Tableau 4

Analyse de variance pour les données de l'enquête objective sur le rendement de 1983

Source	Degrés de liberté	Sommes des carrés	Carré moyen	Composante de la variance	Pourcentage de la variabilité totale
Strate	26	80,193	3,084.3	187.3	14
Segments à l'intérieur de la strate	85	124,086	1,459.8	378.0	28
Résidu	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

Cinq cents échantillons de l'enquête énumérative de juin ont été tirés de la population simulée au moyen d'un échantillonnage aléatoire stratifié. Deux cent soixante-quinze segments ont été échantillonnés dans chaque cas. Le nombre de segments prélevés dans chaque strate était identique au nombre prélevé pour l'enquête énumérative de juin 1983 (voir tableau 2). Pour chacun des échantillons simulés, on a calculé la superficie moyenne (en acres) des segments dans la population et les probabilités conditionnelles, $\{\pi_{hk}\}$, de (3.12), que les segments de l'échantillon recevraient des parcelles par suite d'un tirage. Les probabilités conditionnelles ainsi calculées ont été utilisées au second degré de l'échantillonnage systématique avec un seul départ et probabilité proportionnelle à la taille estimée (voir section 2). Afin de simuler des échantillons de l'enquête objective de rendement on a prélevé 220 unités secondaires en se servant de cette méthode d'échantillonnage systématique. Les unités secondaires ainsi prélevées constituent en soi un échantillon de l'enquête objective sur le rendement. Deux échantillons de ce genre ont été simulés pour chacun des 500 échantillons de la population simulée de l'enquête énumérative de juin.

Lorsqu'un segment était sélectionné pour recevoir une unité secondaire, on simulait le rendement (nombre de plants aux 100 pi^2) observé à l'intérieur d'un champ en supposant l'invariabilité du coefficient de variation à l'intérieur de chaque segment. Le nombre de plants observé était défini comme étant

$$y_{hk\ell} = \bar{c}_{hk} + s_w \bar{\bar{y}}_N^{-1} \bar{c}_{hk} f_{hk\ell}, \tag{4.3}$$

où $y_{hk\ell}$ est le nombre moyen estimé de plants aux 100 pi^2 pour la ℓ-ième unité secondaire du segment k de la strate h et $f_{hk\ell}$ est distribuée suivant une loi normale $N(0, 1)$. L'erreur type à l'intérieur des segments est la racine carrée de $s_w^2 = 776.6$, qui figure dans le tableau 4, et $\bar{\bar{y}}_N$ est le nombre moyen global de plants par parcelle. Si $y_{hk\ell}$ équivalait à moins de 10% de la moyenne pour la strate, on fixait sa valeur à $(.10)\bar{c}_{hk}$. De même, si $y_{hk\ell}$ équivalait à plus de 190% de la moyenne pour la strate, sa valeur était fixée à $(1.9)\bar{c}_{hk}$.

Le tableau 5 donne un résumé des résultats des simulations pour les superficies ensemencées de coton. La superficie moyenne estimée (en acres) par segment est

$$\bar{A}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}, \tag{4.4}$$

et la variance estimée correspondante est

$$\hat{V}(\bar{A}_n) = \sum_{h=1}^{L} W_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2. \tag{4.5}$$

Tableau 5

Estimations de la superficie ensemencée de coton établies à partir de 500 échantillons simulés de l'enquête énumérative de juin

	$\bar{A}_n$	$\hat{V}(\bar{A}_n)$
Moyenne	9.93	0.64
Intervalle	8.13 - 12.21	
Variance	0.66	0.016

La superficie moyenne ensemencée de coton (par segment) pour la population simulée est de 9.94 acres tandis que la moyenne des estimations des 500 échantillons était de 9.93 acres. La variance réelle de l'estimateur stratifié $\bar{A}_n$ est 0.63 tandis que la variance moyenne estimée pour les 500 échantillons simulés était 0.64. À cause de la faible variabilité de l'estimation de la superficie consacrée à la culture du coton, π^*_{hk} produit une estimation stable de la probabilité non conditionnelle que le segment k de la strate h soit échantillonné pour recevoir au moins une unité secondaire.

Outre les estimateurs analysés ci-dessus, nous avons construit des estimateurs de la variance par les groupes aléatoires. Deux ensembles de groupes aléatoires ont été formés pour chaque échantillon de l'enquête objective sur le rendement. Le premier ensemble contenait cinq groupes ($\gamma = 5$) et l'autre, dix ($\gamma = 10$). Les groupes aléatoires ont été créés par la formation de sous-ensembles avec les unités primaires d'échantillonnage (c'est-à-dire les segments) à l'intérieur de chaque strate d'exploitation. On a formé le premier groupe de chaque ensemble en tirant un échantillon aléatoire simple sans remise de taille $K_{h(\gamma)} = n_h/\gamma$ dans chaque strate ($h = 1, \ldots, 28$) de l'échantillon parent de l'enquête énumérative de juin. Le second groupe a été formé de la même manière par tirage de $K_{h(\gamma)}$ segments parmi les $n_h - K_{h(\gamma)}$ segments qui restaient dans chaque strate. Les autres groupes aléatoires ont été formés d'une manière semblable. Comme l'échantillon de la strate d'exploitation n° 3107 ne comptait que cinq segments ($n_h = 5$), on a répété les valeurs de superficie et de rendement des cinq segments pour obtenir les dix observations nécessaires à la formation des dix groupes quand $\alpha = 10$.

Soit D_α le nombre d'unités secondaires ensemencées de coton qui ont été prélevées dans le groupe aléatoire α ($\alpha = 1, \ldots, \gamma$) au cours de l'enquête objective sur le rendement. Soit $\bar{y}_{(\alpha)}$ l'estimateur de rendement obtenu pour l'α-ième groupe aléatoire:

$$\bar{y}_{(\alpha)} = D_\alpha^{-1} \sum_{t=1}^{D_\alpha} Y_{t(\alpha)}, \tag{4.6}$$

où $\bar{y}_{(\alpha)}$ est l'équivalent de l'équation (3.3) pour l'α-ième groupe. L'estimateur de la variance de $\bar{y}$ par les groupes aléatoires prend la forme suivante:

$$\hat{V}_{g\gamma}(\bar{y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\bar{y}_{(\alpha)} - \bar{y})^2. \tag{4.7}$$

Cet estimateur est légèrement biaisé dans le cas de l'estimateur de rendement se rapportant à la série de dix groupes parce que la strate n° 3107 ne comporte que cinq observations et que celles-ci se répètent dans les divers groupes.

De même, posons $\hat{Y}_{(\alpha)}$ comme l'estimateur de la production totale pour l'α-ième groupe aléatoire:

$$\hat{Y}_{(\alpha)} = N \bar{M}_{n(\alpha)} \bar{y}_{(\alpha)}, \tag{4.8}$$

ou

$$\bar{M}_{n(\alpha)} = \sum_{h=1}^{L} W_h K_{h(\alpha)}^{-1} \sum_{k=1}^{K_{h(\alpha)}} M_{hk(\alpha)},$$

$M_{hk(\alpha)}$ est la superficie (en acres) consacrée à la culture du coton dans le segment k de la strate h pour le groupe aléatoire α et $K_{h(\alpha)}$ est le nombre de segments compris dans la strate h pour l'α-ième groupe. L'estimateur de la variance de $\hat{Y}$ par les groupes aléatoires est alors défini comme étant

$$\hat{V}_{g\gamma}(\hat{Y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\hat{Y}_{(\alpha)} - \hat{Y})^2. \tag{4.9}$$

Les tableaux 6 et 7 présentent un sommaire des résultats de la simulation de Monte Carlo pour les estimateurs du rendement et de la production. On y trouve les valeurs moyennes des estimations et leurs estimations de variances pour les 1,000 échantillons simulés de l'enquête objective sur le rendement. En simulant deux échantillons de l'enquête objective sur le rendement pour chaque échantillon simulé de l'enquête énumérative de juin, nous avons pu estimer la variance entre enquêtes énumératives et la variance à l'intérieur des enquêtes énumératives.

L'estimateur actuellement en usage, $\bar{y}$, défini en (3.1), et l'estimateur par quotient combiné $\bar{y}_r$, défini en (3.17), qui est fondé sur les probabilités π_{hk}^* établie à partir des résultats de l'enquête énumérative de juin, produisent tous deux des estimations de précision comparable (voir tableau 6). Cette similitude est en partie attribuable à l'exactitude avec laquelle on estime les probabilités de sélection non conditionnelles dans chaque échantillon.

Comme nous l'avons vu dans la section 3.2, l'estimateur de la variance conditionnelle $\hat{V}_2(\bar{y})$ produit une sous-estimation de $V(\bar{y})$. Dans cette étude de simulation $\hat{V}_2(\bar{y})$ a produit une estimation qui est de 38% inférieure à la variance observée de $\bar{y}$. En effet, la variance observée de $\bar{y}$ était de 11.57, comparativement à une moyenne de 7.21 pour $\hat{V}_2(\bar{y})$. Cette sous-estimation de la variance a été constatée dans l'ensemble des échantillons. La variance estimée de $\hat{V}_2(\bar{y})$ était de 0.99, pour des valeurs de $\hat{V}_2(\bar{y})$ allant de 3.85 à 11.24 pour les 1,000 observations. Ainsi, la valeur maximale observée pour l'estimation de la variance conditionnelle était inférieure à la vraie variance.

Suivant l'hypothèse d'un échantillonnage de segments avec probabilité proportionnelle à la taille et remise en seconde phase, $\hat{V}_2(\bar{y})$ est un estimateur sans biais de la variance conditionnelle de $\bar{y}$, étant donné l'échantillon de segments prélevés au premier degré, comme nous l'avons vu dans la section 3.2. Selon la simulation de Monte Carlo, une estimation de l'espérance mathématique de la variance conditionnelle de $\bar{y}$, $V_2(\bar{y})$, est 3.97. L'écart appréciable entre cette estimation et la moyenne indiquée au tableau 6 (3.97 contre 7.21) peut être attribuable au fait que l'estimateur $\hat{V}_2(\bar{y})$ ne tient pas compte des effets de la stratification dans la population (voir tableaux 2 et 3) et que $\hat{V}_2(\bar{y})$ est calculée suivant l'hypothèse d'un échantillonnage de segments avec remise au second degré du sondage.

L'estimateur (3.9), $\hat{V}_*(\hat{Y})$, produit une sous-estimation de la variance non conditionnelle de $\hat{Y}$. En effet, la variance observée de $\hat{Y}$ dans la simulation de Monte Carlo est 49.69 (millions)^2 tandis que la valeur moyenne de $\hat{V}_*(\hat{Y})$ n'est que 40.85 (millions)^2. Cette sous-estimation (18%) de la vraie variance est attribuable à un certain nombre de facteurs. Nous avons vu plus tôt que $\hat{V}_2(\bar{y})$ comporte un biais négatif lorsqu'il sert d'estimateur de $\hat{V}(\bar{y})$.

Tableau 6

Propriétés des estimations du rendement à l'acre et des estimations de variances selon la simulation de Monte Carlo[1]

	Estimateur					
	$\bar{y}$	$\hat{V}_2(\bar{y})$	$\hat{V}_{g5}(\bar{y})$	$\hat{V}_{g10}(\bar{y})$	$\bar{y}_r$	$\hat{V}(\bar{y}_r)$
Moyenne	79.74	7.21	12.62	12.39	79.76	12.39
Variance totale	11.57	0.99	74.58	36.86	11.56	12.51
Variance entre les EEJ	7.60	0.48	6.10	4.56	7.64	7.61
Variance à l'intérieur des EEJ	3.97	0.51	68.48	32.30	3.92	4.90

[1] Deux échantillons de l'enquête objective sur le rendement ont été simulés pour chacun des 500 échantillons de l'enquête énumérative de juin.

Tableau 7

Propriétés des estimations de la production et des estimations de variances selon la simulation de Monte Carlo[1]

	Estimateur[2]					
	$\hat{Y}$	$\hat{V}_*(\hat{Y})$	$\hat{V}_{g5}(\hat{Y})$	$\hat{V}_{g10}(\hat{Y})$	$\hat{Y}_r$	$\hat{V}(\hat{Y}_r)$
Moyenne	73.04	40.85	48.99	48.53	73.07	48.73
Variance totale	49.69	82.52	1245.10	608.80	49.58	222.96
Variance entre les EEJ	46.35	78.17	50.82	208.48	46.30	199.58
Variance à l'intérieur des EEJ	3.34	4.35	1194.28	400.32	3.28	23.38

[1] Deux échantillons de l'enquête objective sur le rendement ont été simulés pour chacun des 500 échantillons de l'enquête énumérative de juin. Il y avait $N = 92{,}240$ segments dans la population simulée.
[2] L'estimateur $\hat{Y}$ est exprimé en millions d'unités et les variances le sont dans les unités correspondantes.

Par ailleurs, cet écart est dû au fait que $\hat{V}_*(\hat{Y})$ ne tient pas compte de la covariance de $\bar{M}_n$ et de $\bar{y}$. Dans l'exemple qui nous occupe, le biais lié au second facteur contrebalance en partie celui lié au premier.

L'utilisation de l'expression (3.16), $\hat{V}(\bar{y}_r)$, pour estimer la variance de $\bar{y}_r$ et de l'expression (3.21), $\hat{V}(\hat{Y}_r)$, pour estimer la variance de $\hat{Y}_r$ a donné des résultats beaucoup plus satisfaisants que ceux obtenus avec les estimateurs présentement en usage. La moyenne des estimations $\hat{V}(\bar{y}_r)$ a été de 12.51 dans la simulation de Monte Carlo, ce qui représente une surestimation d'environ 7% par rapport à la variance observée de $\bar{y}_r$ (11.57). Environ le tiers de cet écart (l'équivalent de 2 à 4 pour cent) peut s'expliquer par l'utilisation d'un échantillonnage sans remise aux deux premiers degrés de l'échantillonnage. Le reste (l'équivalent de 4 pour cent environ) est faible par rapport à l'erreur type de l'écart estimé. On a estimé la variance de l'écart en estimant la variance de la moyenne de z_{tj}, où

$$z_{tj} = (\bar{y}_{n,r(tj)} - 79.76)^2 - \hat{V}(\bar{y}_{n,r(tj)}), \tag{4.10}$$

pour le j-ième échantillon de rendement ($j = 1, 2$) dans le t-ième échantillon de l'enquête énumérative de juin ($t = 1, \ldots, 500$). L'erreur type estimée de l'écart était de 0.58. Ainsi, la valeur moyenne de $\hat{V}(\bar{y}_r)$ se situe à moins de 1.5 erreurs types de la variance estimée de $\bar{y}_r$. Par ailleurs, la valeur moyenne de l'estimation de la variance de $\hat{Y}_r$ se situe à moins de 2% de la variance observée dans la simulation de Monte Carlo.

Les estimateurs de la variance de $\bar{y}$ par les groupes aléatoires avaient un biais peu significatif. Dans la simulation de Monte Carlo, les moyennes des estimateurs $\hat{V}_{g5}(\bar{y})$ et $\hat{V}_{g10}(\bar{y})$ étaient respectivement de 9 et de 7% supérieures aux variances observées. Ces écarts ne sont pas significativement différents de zéro et sont comparables à ceux obtenus pour l'estimateur $\hat{V}(\bar{y}_r)$. Celui-ci, toutefois, est un estimateur de variance beaucoup plus stable, le coefficient de variation pour cet estimateur étant d'environ 30%, comparativement à 75% pour $\hat{V}_{g5}(\bar{y})$. Comme prévu (Wolter 1985), un accroissement du nombre de groupes aléatoires a amené une diminution du coefficient de variation de l'estimateur de la variance par les groupes aléatoires. En effet, le coefficient de variation pour $\hat{V}_{g10}(\bar{y})$ était de 50%. Les variations entre groupes aléatoires et entre échantillons de rendement de l'enquête énumérative de juin ont expliqué en majeure partie la variance des estimateurs de variances par les groupes aléatoires.

4. CONCLUSIONS

Les analyses montrent que les estimateurs du rendement moyen et de la production totale d'un État, qui sont actuellement utilisés par le National Agricultural Statistical Service sont satisfaisants. En revanche, nous avons vu que les estimateurs de variance simples $\hat{V}_2(\bar{y})$ et $\hat{V}_*(\hat{Y})$ étaient entachés d'un biais négatif dont l'importance dépendait de la variance à l'intérieur des segments et des fractions de sondage à l'intérieur des segments. L'estimateur $\hat{V}_2(\bar{y})$ a produit une estimation qui était près de 40% inférieure à la vraie variance de $\bar{y}$, tandis que l'estimateur $\hat{V}_*(\hat{Y})$ en a produit une qui était de 18% inférieure à la vraie variance de $\hat{Y}$ pour la population simulée.

On a élaboré les estimateurs $\bar{y}_r$ et $\hat{Y}_r$ pour l'estimation de rendement d'un échantillonnage à deux phases, selon lequel les segments qui sont identifiés comme des segments de culture du coton dans la première phase (enquête énumérative de juin) sont sous-échantillonnés dans la seconde phase afin d'estimer le rendement. On estime la probabilité non conditionnelle qu'un segment soit prélevé en vue de recevoir une unité secondaire à l'intérieur d'une strate, π^*_{hk}, en supposant que cette probabilité est proportionnelle à la probabilité conditionnelle de sélection de segments dans la seconde phase de l'échantillonnage. Suivant cette hypothèse, on a défini l'estimateur par quotient combiné du rendement moyen avec probabilités inégales, $\bar{y}_r$, et l'estimateur de sa variance, $\hat{V}(\bar{y}_r)$. L'estimateur de l'agrégat $\hat{Y}_r$ est un estimateur par produit de la production moyenne d'un segment pour un sondage à deux phases, où l'estimateur de la moyenne de la variable auxiliaire (superficie consacrée à la culture étudiée) est tiré de l'enquête énumérative de juin (première phase du sondage). L'estimateur de variance $\hat{V}(\hat{Y}_r)$ est un estimateur de la variance de $\hat{Y}_r$ pour un échantillonnage double stratifié (sondage à deux phases).

La simulation de Monte Carlo a montré que $\bar{y}_r$ et $\hat{Y}_r$ produisaient des estimations comparables à celles découlant des estimateurs courants $\bar{y}$ et $\hat{Y}$. $\hat{V}(\bar{y}_r)$ et $\hat{V}(\hat{Y}_r)$ sont tous deux des estimateurs de variance précis pour des échantillons de la taille de ceux qu'utilise normalement le NASS. Ces résultats sont attribuables en partie à l'exactitude avec laquelle les superficies cultivées moyennes sont estimées au moyen de l'enquête énumérative de juin. Des estimations de superficie justes produisent des estimations de probabilités de sélection qui se rapprochent des probabilités de sélection non conditionnelles. En outre, le fait que l'estimateur soit sous forme de quotient amenuise les effets de la substitution d'estimateurs aux probabilités non conditionnelles réelles.

Par ailleurs, les estimateurs de variance pour les groupes aléatoires sont essentiellement des estimateurs non biaisés de la variance de l'estimation du rendement et de la production. Cependant, ces estimateurs sont beaucoup moins stables que $\hat{V}(\bar{y}_r)$ et $\hat{V}(\hat{Y}_r)$. C'est pourquoi ces derniers sont préférés aux estimateurs pour les groupes aléatoires.

L'enquête énumérative de juin constitue la première phase de l'enquête objective sur le rendement. Les méthodes d'échantillonnage utilisées dans cette enquête sont simples et, comme en fait foi la simulation de Monte Carlo, produisent des estimations de superficie justes. Par conséquent, il ne convient pas d'apporter des modifications à la première phase de l'enquête objective sur le rendement.

On peut toutefois envisager un certain nombre de modifications pour la seconde phase de cette enquête. À l'heure actuelle, on estime le rendement au moyen d'un sondage à deux phases, dans lequel est utilisé un estimateur par quotient combiné. Pour les États où l'échantillon est relativement grand, il conviendrait d'envisager un échantillonnage indépendant à la seconde phase pour chaque strate ou des groupes de strates, ainsi que l'utilisation d'un estimateur par quotient distinct.

Pour obtenir des estimateurs non biaisés de la variance, il conviendrait de remplacer l'échantillonnage systématique effectué dans la seconde phase. À l'heure actuelle, l'échantillonnage des segments à la seconde phase pour l'estimation du rendement se fait par ordinateur et vise tous les États américains. Il serait donc relativement facile d'adopter une méthode d'échantillonnage qui utiliserait des probabilités conjointes connues. Des estimateurs semblables à ceux qui sont recommandés pour le plan de sondage qui nous intéresse pourraient toujours convenir si les probabilités de sélection demeuraient les mêmes. Fuller (1970) décrit une méthode d'échantillonnage qui peut être informatisée et qui comprend le calcul de probabilités conjointes de sélection; en outre, cette méthode utilise des probabilités de sélection déterminées et assure un degré de contrôle comparable à celui qu'offre l'échantillonnage systématique.

REMERCIEMENTS

Cette étude a été réalisée en partie grâce à un contrat de recherche coopérative (n° 58-319T-1-0054X) passé avec le National Agricultural Statistics Service du Département de l'agriculture des États-Unis. Nous tenons à exprimer notre reconnaissance aux arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

FECSO, R. (1978). Cluster analysis as an aid in creating paper strata. Statistical Reporting Service, U.S. Department of Agriculture.

FECSO, R., et JOHNSON, V. (1981). The new California area frame: A statistical study. Statistical Reporting Service, U.S. Department of Agriculture.

FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, Sér. B, 32, 209-226.

HOUSEMAN, E.E. (1975). Area frame sampling in agriculture. Statistical Reporting Service, U.S. Department of Agriculture.

PRATT, W.L. (1984). The use of interpenetrating sampling in area frames. Statistical Reporting Service, U.S. Department of Agriculture.

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