Techniques d'enquête, décembre 1991
Vol. 17, n° 2, pp. 229–244
Statistique Canada

Estimation de la production de blé par comté

ELIZABETH A. STASNY, PREM K. GOEL
et DEBORAH J. RUMSEY[1]

RÉSUMÉ

Bien que les enquêtes agricoles du Département de l'agriculture des É.-U. (DAEU) servent à estimer la production végétale à l'échelle du pays et pour chaque État, les estimations par comté sont plus utiles aux décideurs locaux. Ce genre d'estimations sont aussi recherchées par les entreprises qui se spécialisent dans la vente d'engrais, de pesticides, d'assurance-récolte et de matériel agricole. Les États font souvent leurs propres enquêtes dans le but de produire des estimations de la production agricole par comté. En règle générale, ces enquêtes ne reposent pas sur des méthodes d'échantillonnage probabiliste. En outre, la somme des estimations pour chaque comté d'un État doit concorder avec l'estimation calculée par le DAEU pour l'ensemble de l'État. Les méthodes d'estimation classiques pour petites régions ne sont donc pas directement applicables dans les circonstances. Dans cet article, nous allons recourir à des modèles de régression pour estimer la production de blé par comté au Kansas. Nous allons décrire une étude de simulation par laquelle nous comparons les estimations obtenues par régression à celles calculées à l'aide de deux estimateurs classiques pour petites régions: l'estimateur synthétique et l'estimateur direct. Nous allons aussi comparer plusieurs méthodes par lesquelles nous pouvons pondérer les estimations initiales de manière qu'elles concordent avec l'estimation de la production totale de l'État calculée par le DAEU.

MOTS CLÉS: Échantillon non-probabiliste; régression; simulation; estimation pour petits domaines.

1. INTRODUCTION

Les estimations de la production agricole par comté sont de plus en plus recherchées par les organismes publics, qui s'en servent dans l'élaboration des décisions économiques au niveau local, et par les entreprises qui vendent des engrais, des pesticides, de l'assurance-récolte et du matériel agricole. Le Département de l'agriculture des États-Unis (DAEU) est en train de mettre sur pied un programme visant à uniformiser et à améliorer les méthodes d'estimation de la production agricole par comté (Bass et coll., 1989). Jusqu'à maintenant, chaque État avait sa façon propre d'établir des estimations par comté, c'est ce qui explique le peu d'uniformité des méthodes de collecte de données et d'estimation des divers États. Par son programme, le DAEU cherche à mettre à la disposition des États un ensemble de méthodes d'échantillonnage et d'estimation qui favorise une uniformisation de la qualité des estimations dans tout le pays.

Le nouveau programme du DAEU touche tous les aspects de la production d'estimations régionales (ou par comté), depuis la construction de bases de sondage jusqu'à l'estimation proprement dite. En ce qui concerne la présente étude, nous nous penchons uniquement sur l'estimation de la production de boisseaux de blé. Nous espérons toutefois que les méthodes proposées pourront servir à d'autres volets, comme l'estimation de la superficie ensemencée ou de la production d'autres cultures que le blé.

Malgré la diversité des méthodes d'estimation qu'utilisent les États, on note des points communs. Ainsi, dans chaque État, on calcule des estimations initiales à partir des données recueillies dans chaque comté. Ensuite, un expert examine ces estimations, les corrige en

[1] Elizabeth A. Stasny, Prem K. Goel et Deborah J. Rumsey, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.

fonction de ce qu'il sait déjà des fermes contenues dans l'échantillon, des conditions atmosphériques et d'autres facteurs, puis il évalue l'effet de ces corrections sur l'estimation de la production totale pour l'État. Il peut répéter ce processus un certain nombre de fois, jusqu'à ce qu'il juge raisonnables les estimations pour chaque comté et jusqu'à ce que l'estimation de la production totale de l'État concorde avec celle établie par le DAEU. (L'estimation calculée par le DAEU repose sur un grand échantillon probabiliste et pour cette raison, elle est jugée plus précise que l'estimation globale obtenue en faisant la somme des estimations par comté. C'est pourquoi les États font généralement concorder la somme des estimations par comté avec l'estimation de la production totale établie par le DAEU.)

De façon générale, il n'existe pas de documentation sur les méthodes d'estimation appliquées dans les États. Par conséquent, on ne peut vérifier la pertinence des hypothèses et des méthodes qu'utilise l'expert et il est pratiquement impossible d'analyser ces méthodes ou de refaire les calculs. En outre, on ne peut calculer des estimations de la variance ni appliquer la même méthode dans plus d'un État. L'élaboration de nouvelles méthodes d'estimation pour les comtés doit permettre de résoudre ces difficultés.

Les données qui ont servi à notre étude ont été recueillies au Kansas en 1987, soit avant que le DAEU mette en application ses nouvelles techniques d'échantillonnage pour l'estimation par comté. Nous avons utilisé des données de 1987 car c'est l'année où s'est fait le Recensement de l'agriculture des États-Unis et il se pourrait donc que nous recourions à des données du recensement pour les besoins de l'estimation. Si nous avons choisi d'utiliser les données du Kansas, c'est que cet État a un programme de collecte des données régionales parmi les plus complets aux États-Unis. Néanmoins, les données qui servent à l'estimation par comté au Kansas, comme dans la plupart des autres États, ne proviennent pas d'un échantillon aléatoire d'exploitations agricoles. Par conséquent, il ne faudra pas utiliser un échantillon aléatoire de fermes à blé pour la méthode d'estimation que nous proposons. Du reste, cette méthode pourra être utile dans le cadre du nouveau programme d'estimation par comté puisque celui-ci n'oblige pas les États à utiliser des échantillons probabilistes.

L'estimation pour petites régions a fait l'objet de nombreuses études ces dernières années (voir, par exemple, Platek *et al.*, 1987). Selon les méthodes normales d'estimation pour petites régions, les probabilités d'échantillonnage doivent être connues puisqu'on se sert de l'inverse de la probabilité de sélection pour pondérer les observations dans les formules d'estimateurs courants comme l'estimateur synthétique et l'estimateur direct. (Voir, par exemple, Section 2 de Särndal et Hidiroglou (1989) pour une analyse des estimateurs pour petites régions.)

Les méthodes que nous allons examiner ici doivent différer des méthodes habituelles d'estimation pour petites régions parce que primo, l'échantillon de fermes utilisé n'est pas un échantillon aléatoire et secundo, la somme des estimations par comté doit concorder avec l'estimation de la production totale de l'État calculée par le DAEU. De plus, comme le département de l'agriculture de la plupart des États ne dispose pas actuellement d'un gros ordinateur, les calculs pertinents doivent être suffisamment simples pour pouvoir être exécutés sur un ordinateur personnel. Par conséquent, nous éviterons, au départ, d'utiliser des estimateurs comportant de nombreux calculs comme ceux que décrivent Fay et Herriot (1979). C'est pourquoi nous examinerons un estimateur qui comporte peu de calculs et qui repose sur un modèle de régression.

Dans la section 2, nous décrivons les bases de données du Kansas qui ont servi à cette étude. La section suivante expose la méthode de régression utilisée pour estimer la production de blé tandis que la section 4 décrit plusieurs méthodes permettant de pondérer les estimations par régression de manière qu'elles concordent avec l'estimation de la production totale de l'État calculée par le DAEU. Dans la section 5, nous comparons les estimations par régression aux estimations publiées (par comté) et à des estimations calculées au moyen de l'estimateur synthétique et de l'estimateur direct. La section 6 contient les résultats d'une étude de simulation qui visait à comparer les divers estimateurs précités. Enfin, la section 7 renferme les conclusions et propose des sujets de recherche.

2. DONNÉES DU KANSAS

Aux fins de l'enregistrement des données sur la production agricole, chaque État est divisé en neuf ou dix districts. Le Kansas est découpé en neuf districts et en 105 comtés. Le tableau ci-dessous donne la situation géographique de chaque district ainsi que le nombre de comtés correspondant:

District numéro	Situation géographique	Nombre de comtés
1	Nord-Ouest	8
2	Centre ouest	9
3	Sud-Ouest	14
4	Centre nord	11
5	Centre	11
6	Centre sud	13
7	Nord-Est	11
8	Centre est	14
9	Sud-Est	14

Pour notre étude, nous disposions de deux bases de données qui servent à la production d'estimations pour les comtés du Kansas: la base de données sur les emblavures (Planted Acres Data Base) et la base de données sur les menus grains (Small Grain Data Base). Nous nous sommes servis principalement de données de 1987 mais nous avons validé nos résultats à l'aide des données de 1988. La base de données sur les emblavures pour 1987 contient des renseignements touchant 37,094 fermes réparties sur le territoire du Kansas. (Selon la définition établie par le DAEU, une ferme est une entité qui vend pour $1,000 ou plus de produits agricoles par année.) De ces 37,094 fermes, 22,300 comptaient des terres ensemencées en blé; toutes ces fermes sans exception ont été utilisées dans l'étude de simulation décrite dans la section 5. La base de données sur les menus grains pour 1987 contient des renseignements touchant 5,802 fermes où l'on cultive des menus grains. De ces 5,802 fermes, 1,707 produisent du blé et elles ont toutes servi pour notre étude.

Les enregistrements de la base de données sur les emblavures sont un mélange de données agricoles du Kansas qui ont été recueillies auprès de diverses sources à divers moments. Primo, on dresse une liste d'exploitations agricoles (nom et adresse) à l'aide des données recueillies par les conseillers agricoles. Ces données peuvent être révisées à la lumière des résultats des enquêtes agricoles trimestrielles (Quarterly Agricultural Surveys) et des rapports agricoles mensuels (Monthly Farm Reports). Pour les besoins des enquêtes agricoles trimestrielles, on utilise des échantillons systématiques stratifiés d'environ 2,600 exploitations. Le taux de réponse se situe autour de 80%. Quant au rapport agricole mensuel, il est rempli par environ 3,000 agriculteurs qui ont accepté de prêter leur concours. Le même agriculteur peut produire un rapport mensuel pendant de nombreuses années. Les données les plus récentes pour chaque poste se trouvent dans la base sur les emblavures et l'enregistrement relatif à une exploitation quelconque pour une année quelconque peut contenir des données de diverses sources.

La base de données de 1987 sur les menus grains contient des renseignements sur la superficie ensemencée, la superficie moissonnée et la production de boisseaux en ce qui a trait aux exploitations qui participent aux enquêtes agricoles trimestrielles et à l'enquête du Kansas sur les menus grains (Kansas Small Grain Survey). Pour ce qui est de l'édition 1987 de cette enquête, environ 6,000 questionnaires ont été postés à un échantillon aléatoire d'exploitations agricoles; à peu près la moitié de ces questionnaires ont été remplis puis retournés.

Outre le risque d'un biais dû à la non-réponse pour les données de la base sur les menus grains, il y a la probabilité réelle d'un biais de réponse. En effet, dans leur déclaration, les agriculteurs tendent à sous-estimer la production réelle. À cause de tous ces facteurs - échantillon non conforme, biais dû à la non-réponse, biais de réponse, nous avons jugé pertinent d'élaborer la méthode d'estimation décrite dans les sections qui suivent.

3. MODÈLE DE RÉGRESSION

Nous proposons d'élaborer un modèle de régression qui servira à produire des estimations par comté. Il existe un certain nombre de progiciels statistiques conçus pour les ordinateurs personnels et qui permettent l'ajustement de modèles de régression multiple. En outre, l'estimateur que nous proposons tient compte de ce que nous n'utilisons pas un échantillon aléatoire et produira des estimations régionales dont la somme concorde avec le total estimé pour l'État.

Voici comment nous allons procéder:

1) Décrire par un modèle de régression multiple la relation entre la production agricole et certaines variables explicatives en se servant de l'échantillon non probabiliste de fermes.
2) Supposer que cette relation vaut pour toutes les fermes sur le territoire de l'État et estimer la production agricole totale dans chaque comté.
3) Redresser les estimations par comté de la production agricole de manière que leur somme concorde avec l'estimation calculée par le DAEU pour l'État.

Pour décrire le modèle de régression, nous allons utiliser la notation suivante. Pour $i = 1, 2, \ldots, I$(I = 105 - nombre de comtés au Kansas) et $j = 1, 2, \ldots, n_i$, posons

n_i = nombre de fermes du i-ième comté dans l'échantillon;

$n = \sum_{i=1}^{I} n_i$ = taille de l'échantillon global;

N_i = nombre total de fermes dans la i-ième comté dans la population;

$N = \sum_{i=1}^{I} N_i$ = nombre total de fermes dans la population;

Y_{ij} = production de blé de la j-ième ferme du i-ième comté (en boisseaux);

$X_{ij} = (1\ X_{ij1}\ X_{ij2}\ \ldots\ X_{ijp})$ = vecteur de p variables explicatives pour la j-ième ferme du i-ième comté.

Comme nous le verrons plus loin, il importe de choisir des variables explicatives pour lesquelles il existe des totaux de comté ou des estimations très précises de ces totaux. De plus, les variables explicatives doivent renfermer des éléments d'information qui ont trait à la probabilité d'échantillonnage d'une exploitation agricole, comme une mesure de la taille de l'exploitation. De cette manière, nous pouvons utiliser le modèle de régression en tenant compte de ce qu'il ne s'agit pas d'un échantillon aléatoire.

Nous considérons des modèles de régression de la forme

$$Y_{ij} = f(X_{ij} \mid \beta) + \epsilon_{ij},$$

où $\beta = (\beta_0\, \beta_1\, \beta_2\, \ldots\, \beta_p)$ est un vecteur de paramètres et ϵ_{ij} est un terme d'erreur aléatoire de variance σ^2. Nous allons représenter les valeurs ajustées, obtenues à l'aide de données de la base sur les menus grains, par l'expression suivante:

$$\hat{Y}_{ij} = f(X_{ij} \mid \hat{\beta}).$$

On peut alors estimer la production du i-ième comté par la formule:

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} f(X_{ij} \mid \hat{\beta}),$$

où le "+" en indice signifie que la sommation est étendue à tous les j.

Si nous voulions définir une forme générale pour $f(X_{ij} \mid \beta)$, il nous faudrait connaître la valeur de X_{ij} pour toutes les exploitations du i-ième comté. Or, nous savons pertinemment qu'il n'est pas possible d'obtenir des renseignements aussi détaillés. Cependant, si $f(X_{ij} \mid \beta)$ est une fonction linéaire, il nous suffit de connaître la valeur totale des variables explicatives pour chaque comté. Tel est le cas puisque pour une équation de régression linéaire,

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} [\hat{\beta}_0 + \hat{\beta}_1 X_{ij1} + \hat{\beta}_2 X_{ij2} + \ldots + \hat{\beta}_p X_{ijp}]$$

$$= \hat{\beta}_0 N_i + \hat{\beta}_1 X_{i+1} + \hat{\beta}_2 X_{i+2} + \ldots + \hat{\beta}_p X_{i+p},$$

où X_{i+k} est la valeur totale de la k-ième variable explicative pour le i-ième comté.

Les valeurs $\hat{Y}_{i+}$ seront des estimations raisonnables si le modèle de régression décrit la relation entre les variables explicatives et la production totale des fermes de chaque comté de même que la production des fermes de la base de données. Toutefois, la somme de ces estimations ne correspondra pas nécessairement à l'estimation de la production totale de l'État calculée par le DAEU. Dans la section 4, nous envisageons des méthodes qui permettront de résoudre ce problème.

Le modèle de régression proposé ci-dessus permet de calculer non seulement des estimations de la production agricole par comté, mais aussi des estimations de la variance. On peut le voir très facilement en exprimant les estimations de comté sous forme de matrices. Soit

X = matrice $n \times (p + 1)$ des données réelles, où les lignes sont formées des valeurs X_{ij} définies plus haut;

Z = matrice (inconnue) $N \times (p + 1)$ des variables explicatives pour toutes les exploitations agricoles d'un État;

$\hat{Y}$ = vecteur (inconnu) $N \times 1$ des estimations de la production de blé pour les N exploitations d'un État;

B_i = vecteur colonne $N \times 1$ composé des éléments b_{ij};

$$\text{où } b_{ij} = \begin{cases} 1 \text{ si la } j\text{-ième ferme se trouve dans le } i\text{-ième comté} \\ 0 \text{ dans le cas contraire;} \end{cases}$$

$$A = [B_1\, B_2\, B_3 \ldots B_I]_{N \times I}.$$

La méthode d'estimation décrite plus haut ne donne pas $\hat{Y}$ mais un vecteur d'estimations régionales $\hat{Y}_{i+} = A^T\hat{Y}$, où "$T$" désigne la transposée d'une matrice.

La variance des estimations par comté est donc

$$\text{Var}(\hat{Y}_{i+}) = \text{Var}(A^T\hat{Y}) = A^T\text{Var}(\hat{Y})A = A^T\text{Var}(Z\hat{\beta})A = \sigma^2 A^T Z(X^TX)^{-1}Z^TA.$$

Bien que Z soit inconnue, le produit A^TZ est une matrice connue qui est formée uniquement des N_i, (nombre de fermes dans un comté) et des X_{i+k}, (valeur totale d'une variable explicative pour un comté). Par conséquent, si nous utilisons le carré moyen des erreurs (c.m.e.) comme estimation de σ^2, nous pouvons obtenir des valeurs estimées de la variance des estimations par comté, ce qui ne s'était jamais fait jusqu'à maintenant.

L'estimateur par régression que nous venons de décrire satisfait aux conditions que nous avions fixées au départ, à savoir un estimateur comportant peu de calculs et adapté à un échantillon non probabiliste. Dans la section qui suit, nous examinons des méthodes qui permettront de redresser les estimations régionales de manière à faire concorder leur somme avec l'estimation de la production totale de l'État calculée par le DAEU.

4. PONDÉRATION DES ESTIMATIONS EN VUE D'UNE CONCORDANCE AVEC LE TOTAL ESTIMÉ POUR L'ÉTAT

Soit Y l'estimation de la production de blé au Kansas, telle qu'établie par le DAEU. En règle générale, $\sum_{i=1}^{I} \hat{Y}_{i+} \neq Y$. Par conséquent, nous définissons une nouvelle estimation

$$\tilde{Y}_{i+} = c_i \hat{Y}_{i+},$$

où les c_i sont des constantes telles que $\sum_{i=1}^{I} \tilde{Y}_{i+} = \sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$. La question qu'il faut se poser maintenant est de savoir comment on choisit les valeurs c_i. Selon les méthodes en usage actuellement, on pose $c_i = c$ (au niveau du district) et on pondère par conséquent toutes les estimations par le même facteur. Une autre façon de procéder est de choisir c_i de manière à minimiser la somme des carrés des écarts ou des carrés des écarts relatifs entre $\tilde{Y}_{i+}$ et $\hat{Y}_{i+}$. Nous définissons ci-dessous les valeurs de c_i et de $\tilde{Y}_{i+}$ pour les trois scénarios possibles.

1) **Poser $c_i = c$**

Si on définit c_i comme une constante, il est facile de montrer que

$$c_i = c = Y \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}$$

et

$$\tilde{Y}_{i+} = Y\left(\hat{Y}_{i+} \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}\right).$$

2) **Choisir c_i de manière à minimiser la somme des carrés des écarts entre $\tilde{Y}_{i+}$ et $\hat{Y}_{i+}$**

Si nous voulons choisir les c_i de manière à minimiser la somme des carrés des écarts entre $\tilde{Y}_{i+}$ et $\hat{Y}_{i+}$, pourvu que $\sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$, nous devons minimiser $\sum_{i=1}^{I} (\tilde{Y}_{i+} - \hat{Y}_{i+})^2 = \sum_{i=1}^{I} (c_i \hat{Y}_{i+} - \hat{Y}_{i+})^2$ par rapport aux valeurs des c_i au moyen d'un multiplicateur de Lagrange afin d'imposer la condition voulue. De cette manière, nous observons que

$$c_i = 1 + \left[\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \hat{Y}_{i+}^2 \sum_{i=1}^{I} (1/\hat{Y}_{i+})\right]$$

et

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \hat{Y}_{i+} \sum_{i=1}^{I} (1/\hat{Y}_{i+})\right].$$

Notons que l'on obtient l'estimation pondérée $\tilde{Y}_{i+}$, en ajoutant à l'estimation initiale $\hat{Y}_{i+}$, un rapport qui a pour numérateur l'écart entre l'estimation de la production totale de l'État calculée par le DAEU et la somme des estimations initiales par comté, et pour dénominateur une expression fondée sur la moyenne harmonique des estimations initiales. Bien qu'en théorie, certaines des estimations pondérées peuvent avoir une valeur négative, cela risque peu de se produire dans la réalité puisque les agriculteurs ont tendance à sous-estimer leur production dans les rapports qu'ils fournissent. Si jamais la somme des estimations initiales par comté était supérieure à l'estimation calculée par le DAEU pour l'État, l'estimation pondérée se rapportant à un comté pour lequel l'estimation initiale est une valeur peu élevée pourrait être négative.

3) **Choisir les c_i de manière à minimiser la somme des carrés des écarts relatifs entre $\tilde{Y}_{i+}$ et $\hat{Y}_{i+}$**

Si nous voulons choisir les c_i de manière à minimiser la somme des carrés des écarts relatifs entre $\tilde{Y}_{i+}$ et $\hat{Y}_{i+}$, pourvu que $\sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$, nous devons minimiser $\sum_{i=1}^{I} [(\tilde{Y}_{i+} - \hat{Y}_{i+})/\hat{Y}_{i+}]^2 = \sum_{i=1}^{I} (c_i - 1)^2$ par rapport à c_i au moyen d'un multiplicateur de Lagrange afin d'imposer la condition voulue. De cette manière, nous observons que

$$c_i = 1 + \left[\hat{Y}_{i+} \left(Y - \sum_{i=1}^{I} \hat{Y}_{i+} \right) \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}^2 \right]$$

et

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[\hat{Y}_{i+}^2 \left(Y - \sum_{i=1}^{I} \hat{Y}_{i+} \right) \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}^2 \right].$$

Là encore, on obtient l'estimation pondérée $\tilde{Y}_{i+}$, en ajoutant à l'estimation initiale $\hat{Y}_{i+}$, un rapport qui a pour numérateur une expression axée sur l'écart entre l'estimation de la production totale de l'État calculée par le DAEU et la somme des estimations initiales par comté, et pour dénominateur la somme des carrés des estimations initiales. Comme dans le scénario précédent, l'estimation pondérée peut avoir une valeur négative mais cela risque peu de se produire dans la pratique.

Il convient de souligner que dans l'expression à minimiser, nous avons choisi d'exprimer l'écart $\tilde{Y}_{i+} - \hat{Y}_{i+}$ par rapport à $\hat{Y}_{i+}$ plutôt que par rapport à $\tilde{Y}_{i+}$ puisqu'il n'existe pas d'expression en forme analytique fermée pour l'estimateur $\tilde{Y}_{i+}$. Comme nous visions à construire des estimateurs qui comportent peu de calculs, nous avons choisi de considérer le rapport entre l'écart $\tilde{Y}_{i+} - \hat{Y}_{i+}$ et $\hat{Y}_{i+}$.

Dans la section qui suit, nous examinons l'effet de ces trois méthodes de pondération sur les estimations régionales de la production de blé.

5. COMPARAISON DES ESTIMATIONS DE LA PRODUCTION DE BLÉ

Comme nous l'avons vu dans la section 3, nous nous sommes servis d'un modèle de régression linéaire pour décrire la relation entre la production de blé (mesurée en boisseaux) et certaines variables explicatives en ce qui a trait aux fermes de la base de données de 1987 sur les menus grains. Les variables explicatives que nous pouvions considérer étaient les suivantes: superficie ensemencée en blé, superficie moissonnée, production prévue (suivant les estimations par comté de 1986), superficie des emblavures irriguées, superficie des emblavures non irriguées, indicateurs du district où est située la ferme, indicateurs de la région de l'État (est, centre, ouest) et termes d'interaction.

Tableau 1

Ajustement des modèles de régression aux données réelles

	Modèles ajustés	R^2	$\sqrt{\text{cme}}$
Modèle 1	Boisseaux = $-811 + 32(\text{Pla}) + 3{,}248I_1 + 3{,}088I_2 + 2{,}190I_3 + 2{,}526I_4 + 1{,}241I_5 - 562I_6 + 1{,}047I_7 + 399I_8$	85	5,945
Modèle 2	Boisseaux = $-281 + 28(\text{Pla}) + 138I_1 + 1{,}861I_2 + 2{,}328I_3 + 329I_4 - 359I_5 - 334I_6 - 42I_7 + 500I_8 + 11(\text{Pla})I_1 + 5(\text{Pla})I_2 + 3(\text{Pla})I_3 + 11(\text{Pla})I_4 + 9(\text{Pla})I_5 - 0.2(\text{Pla})I_6 + 15(\text{Pla})I_7 - 7(\text{Pla})I_8$	86	5,818

Nota: Pla désigne la superficie ensemencée (en acres); I_i est la variable indicatrice pour le i-ième district.

Les variables les plus importantes pour notre modèle de régression étaient la superficie ensemencée en blé et un indicateur de l'emplacement de la ferme sur le territoire de l'État. La variable fondée sur les estimations par comté de l'année précédente ne semblait pas un bon prédicteur de la quantité de blé produit sur une ferme dans l'année courante. Nous aurions pu aussi choisir d'autres variables comme la superficie des emblavures irriguées mais il est difficile d'en connaître la valeur exacte au niveau des comtés. C'est pourquoi nous avons choisi de faire de la superficie ensemencée en blé la seule variable explicative continue de notre modèle. Nous n'étions pas obligés d'inclure tous les indicateurs de district dans le modèle; certains districts étaient similaires et auraient donc pu être regroupés. Si nous avons décidé de tenir compte de tous les indicateurs, c'est que les groupes de districts risquaient de changer d'une année à l'autre ou pouvaient être différents dans le cas d'autres cultures que le blé.

Nous avons choisi de concentrer notre attention sur deux modèles de régression: le modèle 1, qui comprend la superficie ensemencée en blé et les indicateurs de district, et le modèle 2, qui comprend les mêmes variables plus les termes d'interaction qui mettent en rapport la superficie ensemencée et les variables indicatrices. Ces modèles et les mesures d'ajustement correspondantes figurent dans le tableau 1. Malgré la faible différence entre les racines du carré moyen des erreurs ($\sqrt{\text{cme}}$) des deux modèles, nous étions d'avis que cette différence pourrait s'amplifier si les modèles devaient servir à estimer la production agricole pour tout l'État. C'est pourquoi nous avons cru utile de comparer des estimations tirées des deux modèles.

Pour nous assurer que ces modèles de régression n'étaient pas simplement la conséquence d'une irrégularité dans la base de données de 1987 sur les menus grains, nous avons repris le même ensemble de variables explicatives et avons tenté de construire des modèles de régression acceptables à l'aide des données de 1988. L'ajustement des modèles 1 et 2 aux données de 1988 est comparable à l'ajustement des mêmes modèles aux données de 1987 et nous n'avons trouvé aucun autre modèle qui pouvait produire un meilleur ajustement pour les données de 1988. Les valeurs estimées du paramètre correspondant à la superficie ensemencée en blé étaient assez semblables d'une année à l'autre, alors qu'il y avait des différences notables entre les paramètres pour les indicateurs de district. Nous croyons que ces indicateurs traduisent l'effet des conditions atmosphériques et des différences de pratiques agricoles dans les diverses parties de l'État. Par exemple, on irrigue beaucoup plus les terres dans l'ouest et le centre du Kansas que dans l'est. Les pratiques agricoles sont peu susceptibles de changer considérablement d'une année à l'autre mais il en va tout autrement des conditions atmosphériques. C'est pourquoi il est raisonnable de penser que la contribution de l'indicateur de district dans la prédiction de la production de blé pourra varier sensiblement d'une année à l'autre.

Tableau 2
Estimations par régression pour neuf comtés du Kansas

District	Comté	Production estimée de blé (en milliers de boisseaux)	
		Modèle 1 (sans terme d'interaction)	Modèle 2 (avec termes d'interaction)
1	Decatur	4,944	4,778
		(180)	(179)
2	Trego	4,378	4,229
		(174)	(188)
3	Hodgeman	4,808	4,908
		(123)	(125)
4	Jewell	5,555	5,550
		(275)	(269)
5	Marion	5,144	4,931
		(313)	(315)
6	Comanche	2,615	2,480
		(59)	(63)
7	Leavenworth	231	262
		(53)	(61)
8	Shawnee	232	226
		(106)	(104)
9	Butler	2,374	2,272
		(331)	(338)

Nota: Les erreurs types figurent entre parenthèses au-dessous de chaque estimation.

Nous nous sommes servis des deux modèles proposés pour établir des estimations pour les 105 comtés du Kansas. Le tableau 2 donne les estimations non pondérées et les erreurs types correspondantes calculées pour neuf comtés, soit un par district. Nous avons tiré au hasard un comté dans chaque district de sorte que toutes les régions de l'État soient représentées. En examinant les données du tableau 2, on s'aperçoit que l'erreur type estimée pour le comté de Shawnee est anormale. La variance d'une estimation régionale (ou de comté) dépend du nombre de fermes dans le comté, de la superficie totale ensemencée en blé dans le comté et du nombre de fermes de l'échantillon qui appartiennent au district dans lequel se trouve le comté. Le district no 8, qui comprend Shawnee, compte relativement peu de fermes dans la base de données sur les menus grains. Par ailleurs, Shawnee a un nombre raisonnable de fermes où l'on cultive du blé mais les superficies ensemencées sont modestes. Ces trois facteurs expliquent que l'erreur type des estimations calculées pour le comté de Shawnee soit assez élevée.

Les estimations du tableau 2 se rapprochent assez des estimations publiées (Kansas Agricultural Statistics 1988). Bien qu'il soit réconfortant de voir que nos estimations ne s'écartent pas véritablement de celles de l'État du Kansas, il n'y a rien théoriquement qui justifie de considérer les estimations du Kansas comme la norme. Nous avons donc exécuté une étude de simulation afin de mieux évaluer nos estimateurs. Cette étude est décrite dans la section qui suit.

6. ÉTUDE DE SIMULATION

6.1 Estimateurs comparés

L'étude de simulation a permis de comparer les estimations obtenues au moyen des deux modèles de régression proposés et celles produites par deux estimateurs courants pour petites régions: l'estimateur synthétique et l'estimateur direct. (Voir, par exemple, la section 2 de

Särndal et Hidiroglou (1989) pour une analyse d'estimateurs courants pour petites régions, y compris les estimateurs synthétique et direct.) On obtient des estimations synthétiques en répartissant la production totale de blé pour l'État entre les comtés au prorata de la superficie ensemencée en blé dans chaque comté. Quant à l'estimation directe, la méthode se limite à utiliser les fermes échantillonnées d'un comté pour estimer la production de blé dans ce comté.

Les estimations synthétiques devraient normalement être fortement biaisées puisque les pratiques agricoles et les conditions atmosphériques ne sont pas nécessairement les mêmes dans chaque région d'un État alors que selon l'estimateur synthétique, chaque comté est représentatif de l'État. En revanche, la variance des estimations synthétiques sera relativement faible car ce genre d'estimations sont établies à l'aide de toutes les données pour l'État.

Tableau 3

Nombre de fermes et niveau de production selon le district et la superficie ensemencée

District		Superficie ensemencée (acres)				
		0-99	100-249	250-499	500-999	≥1,000
1	M_i*	354	638	531	302	85
	m_i*	27	45	51	40	9
	bu/pa*	34.68	37.18	37.76	39.21	38.68
2	M_i	266	550	572	377	161
	m_i	27	49	47	55	33
	bu/pa	35.92	33.62	36.78	39.09	34.85
3	M_i	264	549	610	537	264
	m_i	31	80	76	98	61
	bu/pa	26.93	32.84	35.03	36.79	33.13
4	M_i	956	939	626	271	50
	m_i	62	37	23	21	7
	bu/pa	36.81	36.91	39.70	39.87	39.41
5	M_i	1,236	1,529	912	350	54
	m_i	92	93	51	26	3
	bu/pa	31.79	32.25	31.69	36.85	33.65
6	M_i	1,181	1,427	1,160	793	249
	m_i	96	96	81	55	20
	bu/pa	26.24	26.88	28.78	27.87	26.72
7	M_i	957	242	67	9	3
	m_i	62	5	2	0	0
	bu/pa	33.87	40.81**	40.81**	40.81**	40.81**
8	M_i	1,126	251	52	9	1
	m_i	56	11	2	0	0
	bu/pa	26.02	11.48**	11.48**	11.48**	11.48**
9	M_i	1,122	431	166	59	12
	m_i	47	19	7	3	1
	bu/pa	23.57	23.87	27.63**	27.63**	27.63**

* M_i est le nombre de fermes dans la base de données sur les emblavures, m_i est le nombre de fermes dans la base de données sur la production et bu/pa est le rapport entre le nombre de boisseaux produits et la superficie ensemencée (en acres).

** Les cellules pour ce district ont été groupées en vue de calculer le ratio bu/pa.

Comme l'estimation directe calculée pour un comté repose uniquement sur les données d'échantillon recueillies à l'intérieur de ce comté, la variance de cette estimation sera relativement élevée mais son biais devrait être moindre que celui d'une estimation synthétique. L'échantillon doit comprendre au moins une ferme d'un comté donné si l'on veut calculer une estimation pour ce comté, et il doit comprendre au moins deux fermes d'un même comté si l'on veut estimer la variance. Dans la base de données de 1987 sur les menus grains, trois comtés n'étaient pas du tout représentés dans l'échantillon des fermes à blé et trois autres l'étaient par une seule ferme.

Bien que nous comparions "nos" estimateurs par régression aux estimateurs synthétique et direct, il convient de préciser que ces deux derniers n'admettent que des données provenant d'un échantillon probabiliste, ce qui n'est pas le cas des données du Kansas.

6.2 Population simulée et échantillons

Nous avons tout d'abord simulé une population de fermes à blé en générant des niveaux de production pour les 22,300 fermes de la base de données sur les emblavures qui comptaient des terres ensemencées en blé. Comme les niveaux de production semblent varier selon le district et la taille de l'exploitation agricole (voir tableau 3), nous avons généré des ratios "nombre de boisseaux par unité de superficie ensemencée" (bu/pa) à partir de 37 distributions différentes. Celles-ci étaient fondées sur les ratios bu/pa tirés de la base de données sur les menus grains. (Notons qu'en ce concerne les districts de l'est du Kansas - districts 7, 8 et 9 en l'occurrence, plusieurs catégories de taille étaient peu représentées dans l'échantillon ou ne l'étaient pas du tout. Nous avons donc regroupé ces catégories de la manière indiquée dans le tableau 3 afin de simuler des ratios bu/pa.) Nous avons produit un histogramme des ratios empiriques de la base de données sur les menus grains pour chaque district en fonction de cinq catégories de taille: de 0 à 99 acres ensemencés en blé, de 100 à 249, de 250 à 499, de 500 à 999 et 1,000 acres ou plus. Comme ces histogrammes avaient plus ou moins une forme en cloche, nous avons choisi de décrire la distribution des ratios bu/pa au moyen de la distribution normale. Pour moyenne et pour variance de la distribution normale, nous avons utilisé la moyenne et la variance empiriques des ratios bu/pa pour les fermes à blé de la base de données sur les menus grains réparties dans les 37 classes formées par la combinaison des districts et des catégories de taille.

Après avoir tiré de la distribution normale appropriée le ratio bu/pa pour chaque ferme, nous avons calculé la quantité de blé produite sur chaque ferme en multipliant le ratio simulé par la superficie déclarée des emblavures.

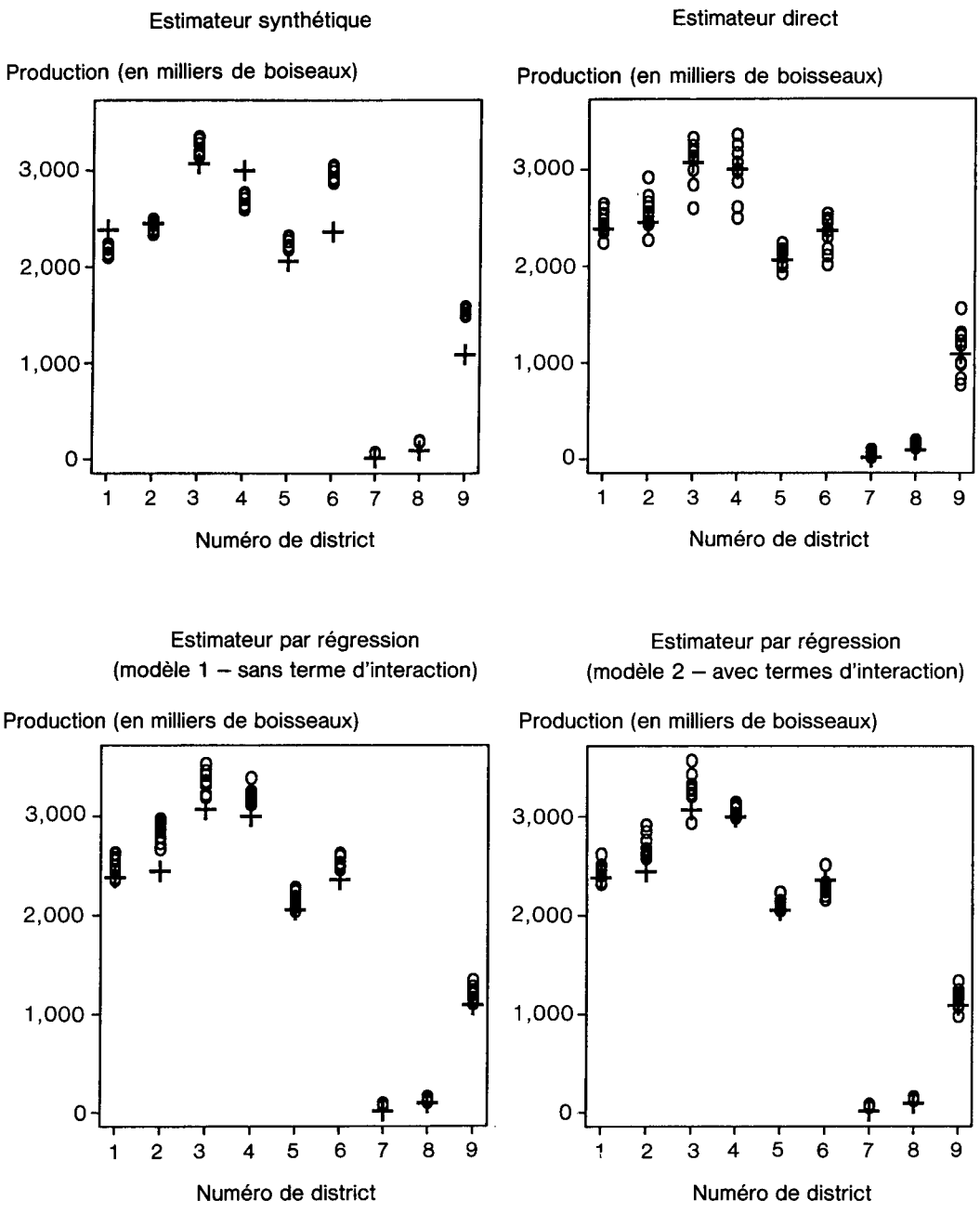
Par la suite, nous avons tiré dix échantillons de cette population simulée. Comme aucun plan d'échantillonnage particulier n'avait été défini, chaque ferme dans les 37 classes a été prélevée avec une probabilité égale au rapport empirique entre le nombre de fermes appartenant à la base de données sur les menus grains et le nombre de fermes de la même classe appartenant à la base de données sur les emblavures. Ainsi, les fermes de la classe C, par exemple, ont été échantillonnées avec une probabilité égale à

$$\frac{\text{Nombre de fermes de la classe } C \text{ dans la base de données sur les menus grains}}{\text{Nombre de fermes de la classe } C \text{ dans la base de données sur les emblavures}}.$$

Par un tel plan, nous cherchions à obtenir des échantillons simulés qui se rapprocheraient le plus possible des échantillons réels, même si nous ne connaissons pas les probabilités de sélection pour les échantillons réels.

6.3 Comparaison des quatre estimateurs pour petites régions

Nous nous sommes servis des quatre estimateurs qui font l'objet d'une comparaison (c'est-à-dire les deux estimateurs par régression, l'estimateur synthétique et l'estimateur direct) et des dix échantillons simulés précédemment pour calculer un niveau de production pour chacun



Nota:Les estimations concernent un comté tiré au hasard dans chacun des districts.
o = estimation calculée pour un des dix échantillons simulés;
+ = valeur réelle tirée de la production simulée.

Figure 1. Comparaison d'estimateurs pour neuf comtés

des 105 comtés. Nous avons ensuite comparé ces estimations aux niveaux de production "réels" établis à l'aide de la population simulée. Cette comparaison nous permet de mesurer le biais et la variabilité des estimations pour chaque comté. Dans la figure 1, nous reproduisons, pour chaque estimateur étudié et les neuf comtés choisis au hasard à raison d'un par district (évoqués dans la section 5), les valeurs calculées pour chacun des dix échantillons ainsi que la production (réelle).

Comme prévu, les estimations synthétiques sont entachées d'un biais considérable. En effet, l'intervalle des estimations issues des dix échantillons simulés ne comprend la valeur réelle de population que dans le cas du district n° 2. L'intervalle des valeurs calculées par l'estimateur direct est plus large que celui des estimations synthétiques pour chacun des neuf districts mais il renferme la valeur réelle de population dans tous les cas. Les intervalles des estimations calculées au moyen de l'un ou l'autre des modèles de régression semblent moins étendus que ceux des estimations directes. Dans le cas du modèle 1, on note un certain biais dans les estimations pour environ la moitié des districts. Les estimations calculées selon le modèle 2 paraissent moins biaisées. Les résultats de la comparaison nous amènent à opter pour le modèle 2, c'est-à-dire le modèle de régression avec termes d'interaction.

6.4 Comparaison des méthodes de pondération

En nous servant des trois méthodes de pondération décrites dans la section 4, nous avons ensuite pondéré les quatre mêmes séries d'estimations régionales tirées des échantillons simulés de manière à les faire concorder avec la production totale de l'État estimée à l'aide de la population simulée. Nous avons comparé les estimations ainsi obtenues aux niveaux de production réels établis pour la population simulée. À cette fin, nous nous sommes servis de la moyenne des valeurs absolues des erreurs relatives, qui est définie:

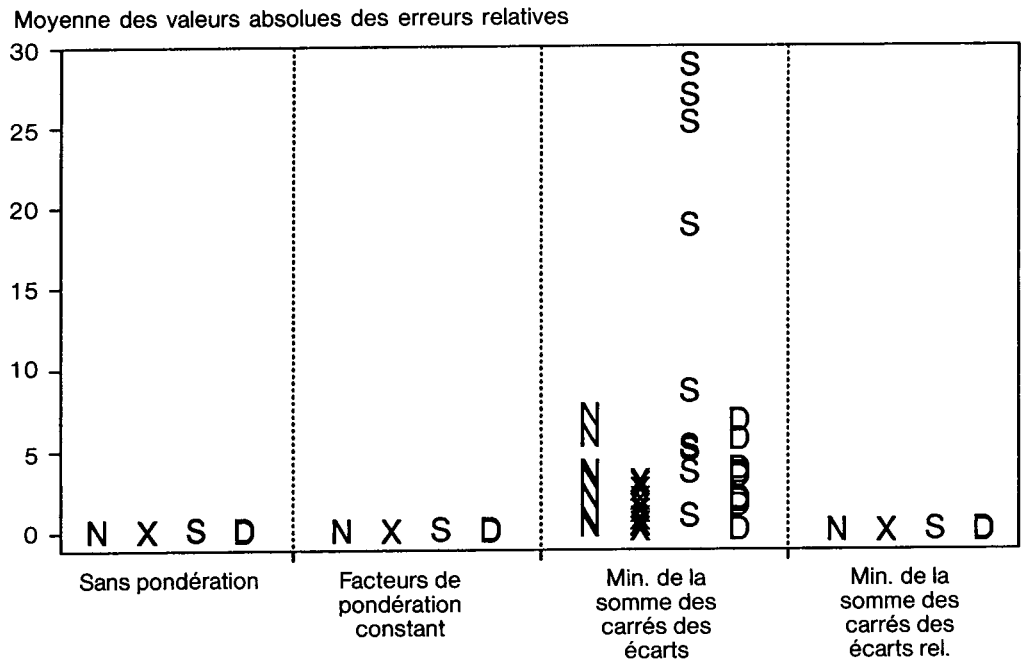
$$(1/I) \sum_{i=1}^{I} \left| (\tilde{Y}_{i+} - Y_{i+})/Y_{i+} \right|.$$

La figure 2 donne les valeurs de cette moyenne tirées des dix échantillons simulés pour les 105 comtés. Nous y comparons les quatre estimateurs dans le cas où il y a absence de pondération et dans trois autres cas qui correspondent à l'application de chacune des trois méthodes de pondération.

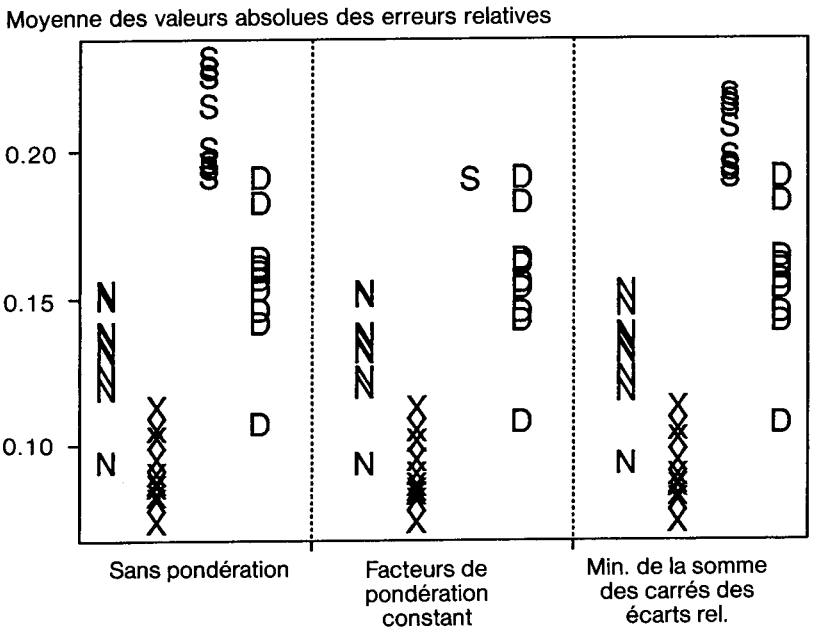
Nous voyons d'après la figure 2A que la méthode qui consiste à minimiser la somme des carrés des écarts donne des estimations finales très médiocres; la moyenne des valeurs absolues des écarts relatifs entre les estimations finales et les niveaux de production des comtés calculés à l'aide de la population simulée est très élevée par rapport aux moyennes observées dans le cas des autres méthodes de pondération. Cela s'explique par la grande différence qui peut exister parfois entre des niveaux de production dans deux comtés. Comme la méthode en question consiste à minimiser la somme des carrés des écarts entre les estimations initiales et les estimations finales, il se pourrait que l'opération modifie considérablement l'estimation initiale pour un comté si cette estimation est très peu élevée. Or, de telles variations semblent injustifiables et c'est pourquoi nous n'insisterons pas davantage sur cette méthode de pondération.

La figure 2B, qui est une version raffinée de la figure 2A, permet une comparaison plus poussée des quatre estimateurs dans le cas de l'absence de pondération et dans le cas de l'application des deux autres méthodes. Nous constatons que le modèle de régression avec termes d'interaction est, parmi les quatre estimateurs étudiés, celui pour lequel l'erreur est généralement la moins élevée. Cela confirme le choix que nous avions fait dans la sous-section précédente. En outre, la figure 2B donne à penser qu'il y a peu de différence entre les estimations initiales non pondérées et les estimations pondérées selon l'une ou l'autre des deux méthodes. De fait,

A.



B.



Nota: Les données proviennent des échantillons simulés.
N = estimateur par régression – modèle 1 (sans terme d'interaction);
X = estimateur par régression – modèle 2 (avec termes d'interaction);
S = estimateur synthétique;
D = estimateur direct.

Figure 2. Comparaison des estimateurs et des méthodes de pondération

la somme des estimations régionales initiales est voisine du total pour la population simulée. Les facteurs de pondération constants, ci, ont donc tous une valeur approximative de un. Enfin, comme les deux méthodes donnent des résultats similaires, il n'y a aucune raison d'utiliser la plus compliquée des deux; nous pouvons donc utiliser celle avec facteur de pondération constant.

7. CONCLUSIONS ET SUJETS DE RECHERCHE POUR L'AVENIR

Nous avons montré que l'on pouvait obtenir des estimations acceptables de la production de blé par comté à l'aide d'un modèle de régression. Les variables explicatives du modèle que nous avons choisi sont la superficie ensemencée (en acres), les indicateurs de district et les termes d'interaction. Le modèle de régression proposé ne nécessite pas un échantillon aléatoire de fermes et, par surcroît, il permet d'estimer la variance d'estimations régionales. Il est possible de pondérer les estimations tirées du modèle de régression de manière à faire concorder leur somme avec la production totale estimée pour l'État; cette pondération peut se faire au moyen d'un facteur constant car nous avons vu que cette méthode ne donne pas des résultats très différents de ceux obtenus avec l'autre méthode, plus complexe.

L'estimation de la production agricole par comté est un sujet qui doit faire encore l'objet de nombreuses recherches. Par exemple, les estimations régionales tirées de l'étude de simulation nous portent à croire que la présence (ou l'absence) des grandes exploitations agricoles (emblavures d'une superficie de 1,000 acres ou plus) dans l'échantillon d'un district pourrait influer largement sur les estimations des comtés de ce district, à plus forte raison pour les districts qui comptent peu de grandes exploitations. Comme les grosses fermes représentent le plus souvent une proportion appréciable de la production agricole, elles mériteraient de faire l'objet d'une opération indépendante dans le calcul d'estimations régionales. Par ailleurs, les États devraient envisager de modifier leur plan d'échantillonnage de sorte que les grandes exploitations fassent partie de l'échantillon à coup sûr.

La recherche doit aussi tenter de déterminer si le modèle de régression que nous avons défini pour le blé peut s'appliquer à d'autres cultures. En particulier, il serait intéressant de savoir si ce genre de modèle peut être utilisé pour les cultures spéciales, au sujet desquelles il existe beaucoup moins de données. De plus, nous découvrirons peut-être que la similitude observée entre le chiffre de production pour l'État et la somme des estimations régionales – pour les données réelles comme pour les données des échantillons simulés – est propre à la production de blé et n'est pas observable pour toutes les cultures. Nous devrions aussi chercher à savoir si, pour les autres cultures que le blé, on doit utiliser une autre méthode de pondération que celle avec facteur de pondération constant.

Nous avons choisi d'amorcer l'étude de l'estimation pour comtés en examinant des méthodes d'estimation de la production. Les recherches à venir devront porter aussi sur l'estimation de la superficie totale ensemencée pour diverses cultures. Pour les besoins de notre étude, nous avons puisé dans la base de données du recensement de l'agriculture de 1987 pour connaître le nombre de fermes et la superficie ensemencée en blé dans chaque comté. Or, le recensement de l'agriculture n'est fait qu'à tous les cinq ans. Dans l'intervalle, il faut estimer la variation du nombre de fermes et de la superficie ensemencée à l'aide de données d'échantillon. Ces variations devraient normalement être modestes en ce qui a trait aux grandes cultures comme le blé au Kansas; toutefois, pour ce qui regarde les cultures moins courantes, il sera vraisemblablement plus difficile d'estimer ces variations.

Enfin, il ne sera probablement plus nécessaire de respecter la condition que nous avions posée au début – à savoir l'utilisation d'un estimateur comportant peu de calculs – et qui nous a amenés à proposer un estimateur par régression, car les divers départements de l'agriculture aux États-Unis seront bientôt raccordés à un grand réseau informatique national. Par conséquent, nos futures études sur l'estimation de la production agricole par comté s'intéresseront à des estimateurs qui comportent plus de calculs.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce au contrat de coopération n° 58-3AEU-9-80040 passé avec le Département de l'agriculture des États-Unis. Les commentaires qui y sont exprimés engagent la responsabilité des auteurs seuls. Ceux-ci tiennent à remercier Gary Keough et Leland Brown, du Département de l'agriculture des É.-U., de même que Ronald Sadler, Melvin Perrott, Eldon Thiessen et M.E. Johnson, du Département de l'agriculture du Kansas, pour les avoir aidés dans ce projet. Ils remercient également l'arbitre et le rédacteur associé qui ont bien voulu exprimer leurs commentaires sur une version antérieure de l'article.

BIBLIOGRAPHIE

BASS, J., GUINN, B., KLUGH, B., RUCKMAN, C., THORSON, J., et WALDROP, J. (1989). Report of the Task Group for Review and Recommendations on County Estimates. USDA National Agricultural Statistics Service, Washington, D.C.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

KANSAS AGRICULTURAL STATISTICS. (1988). Kansas Farm Facts, préparé par le Statistical Division of the Kansas Department of Agriculture avec la coopération du National Agricultural Statistics Service of the U. S. Department of Agriculture, Topeka, Kansas.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., et SINGH M.P. (Éds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.

SÄRNDAL, C.E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.