Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Un modèle hiérarchique bayésien bivarié pour estimer les taux de location au comptant de terres cultivées au niveau du comté

par Andreea Erciulescu, Emily Berg, Will Cecere et Malay Ghosh

Date de diffusion : le 27 juin 2019

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, juin 2019
Vol. 45, nº 2, pp. 211-229
Statistique Canada, nº 12-001-X au catalogue

Un modèle hiérarchique bayésien bivarié pour estimer les taux de location au comptant de terres cultivées au niveau du comté

Andreea Erciulescu, Emily Berg, Will Cecere et Malay Ghosh[1]

Résumé

Le *National Agricultural Statistics Service* (NASS) du *United States Department of Agriculture* (USDA) est chargé d'estimer les taux moyens de location au comptant au niveau du comté. Par taux de location au comptant, on entend la valeur marchande des terres louées à l'acre contre argent comptant seulement. Les estimations des taux de location au comptant sont utilisées par les agriculteurs, les économistes et les responsables des politiques. Le NASS recueille des données sur les taux de location au comptant au moyen de la Cash Rent Survey. Comme les tailles d'échantillon réalisées au niveau du comté sont souvent trop petites pour permettre des estimateurs directs fiables, des prédicteurs fondés sur des modèles mixtes sont étudiés. Nous spécifions un modèle bivarié pour obtenir des prédicteurs des taux de location au comptant en 2010 pour les terres cultivées non irriguées à l'aide de données provenant de la Cash Rent Survey de 2009 et de variables auxiliaires provenant de sources externes, dont le Recensement de l'agriculture de 2007. Nous utilisons des méthodes bayésiennes pour l'inférence et présentons les résultats pour l'Iowa, le Kansas et le Texas. L'intégration des données de l'enquête de 2009 grâce à un modèle bivarié mène à des prédicteurs dont les erreurs quadratiques moyennes sont plus petites que celles des prédicteurs fondés sur un modèle univarié.

Mots-clés : Bayésien hiérarchique; modèle mixte bivarié; réconciliation.

1 Introduction

Le *National Agricultural Statistics Service* (NASS) du *United States Department of Agriculture* (USDA) réalise des centaines d'enquêtes chaque année pour obtenir des estimations concernant divers aspects de l'agriculture aux États-Unis. La production totale, la superficie récoltée et le rendement des cultures sont des exemples de paramètres qu'estime le NASS. L'estimation pour des domaines à un niveau infra-État, tels que les comtés, est difficile en raison des petites tailles d'échantillon. Nous nous intéressons à l'estimation du taux de location au comptant au niveau du comté, c'est-à-dire la valeur marchande des terres louées à l'acre contre paiement comptant seulement.

Les estimations des taux de location au comptant au niveau du comté ont de multiples utilisations. Les agriculteurs s'appuient sur ces estimations pour établir les ententes de location (Dhuyvetter et Kastens, 2009). Les agronomes les utilisent pour étudier des questions de recherche portant sur les interactions entre les taux de location au comptant et d'autres caractéristiques économiques, dont les prix des produits et les coûts de carburant (Woodard, Paulson, Baylis et Woddard, 2010). Les estimations des taux moyens de location au comptant au niveau du comté publiées par le NASS ont des incidences sur le *Conservation Reserve Program*, une politique qui vise à encourager les propriétaires de terres agricoles à les conserver. Les *Farm Bills* de 2008 et de 2014 exigent que le NASS recueille des données sur les taux de location au comptant pour trois catégories d'utilisation des terres, à savoir les terres cultivées non irriguées, les terres

1. Andreea Erciulescu, National Institute of Statistical Sciences, Washington D.C., États-Unis; Emily Berg, Department of Statistics, Iowa State University, Ames, IA, États-Unis. Courriel : emilyb@iastate.edu; Will Cecere, Westat, Rockville, MD, États-Unis; Malay Ghosh, Department of Statistics, University of Florida, Gainesville, FL, États-Unis.

cultivées irriguées et les pâturages permanents, pour les comtés possédant au moins 20 000 acres de terres cultivées ou de pâturages.

Afin de répondre aux exigences des *Farm Bills* de 2008 et de 2014, le NASS réalise la Cash Rent Survey. Une préoccupation tient au fait que les estimateurs directs des moyennes de comté d'après les données de la Cash Rent Survey pourraient être instables en raison des petites tailles d'échantillon réalisées. Nous étudions l'utilisation de modèles mixtes (Rao et Molina, 2015) pour stabiliser les estimateurs des taux moyens de location au comptant au niveau du comté. Le NASS publie les estimations des taux moyens de location au comptant au niveau de l'État avant que le calcul des estimations au niveau du comté d'après la Cash Rent Survey soit achevé. Pour maintenir la cohérence interne, les prédicteurs au niveau du comté doivent satisfaire une contrainte de réconciliation.

Dans un cadre fréquentiste, Berg, Cecere et Ghosh (2014) utilisent des modèles au niveau du domaine pour prédire les taux de location au comptant au niveau du comté pour tous les États et pour les trois catégories d'utilisation, à savoir les terres cultivées non irriguées, les terres cultivées irriguées et les pâturages permanents. Pour chaque combinaison de catégorie d'utilisation des terres et d'État, la méthode de Berg et coll. (2014) utilise des données provenant de deux années. L'hypothèse que les variances pour les deux années sont les mêmes motive la transformation de Pitman-Morgan, qui convertit le vecteur d'observations pour les deux points dans le temps en une moyenne et une différence. Après l'application de modèles univariés distincts à la moyenne et à la différence, le prédicteur pour chaque point dans le temps s'obtient en ajoutant le prédicteur de la moyenne à la moitié du prédicteur de la différence. On a démontré que la méthode de Berg et coll. (2014) est une approche pratique pour obtenir des prédictions raisonnables dans toute une gamme de conditions. Néanmoins, les effets des hypothèses simplificatrices justifient une étude supplémentaire. Si les variances pour les deux points dans le temps diffèrent, alors, comme le discutent Berg et coll. (2014), l'estimateur de l'erreur quadratique moyenne (EQM) basé sur la transformation de Pitman-Morgan peut présenter un biais négatif. De surcroît, la méthode de Berg et coll. (2014) ne tient pas compte de l'effet de la réconciliation dans l'estimation de l'EQM.

La présente étude aborde les questions des variances non constantes au cours du temps et de l'effet de la réconciliation sur l'efficacité dans le contexte de la Cash Rent Survey du NASS au moyen d'un modèle hiérarchique bayésien (HB) bivarié pour les données au niveau de l'unité. Le modèle est suffisamment flexible pour permettre que les variances diffèrent entre les deux points dans le temps. L'utilisation de méthodes bayésiennes pour l'inférence facilite l'estimation de l'augmentation de l'EQM a posteriori due à la réconciliation. Une autre innovation de l'approche HB bivariée est qu'elle intègre les poids de sondage dans le modèle de variance. Nous cherchons aussi à améliorer l'efficacité des prédicteurs pour des situations particulières, par rapport à Berg et coll. (2014), en permettant que les covariables diffèrent entre les États. Datta, Day et Maiti (1998) examinent des modèles HB bivariés pour les données sur la superficie des cultures au niveau du comté de Battese, Harter et Fuller (1988). Notre modèle étend celui de Datta et coll.

(1998) afin de tenir compte d'une relation entre la pondération et la variance, ainsi que d'une structure de données non équilibrée.

Nous nous concentrons sur la prédiction des taux de location au comptant au niveau du comté pour les terres cultivées non irriguées en utilisant les réponses aux éditions de 2009 et 2010 de la Cash Rent Survey, ainsi que des sources externes d'information auxiliaire. À la section 2, nous discutons des données d'enquête et de l'information auxiliaire en détail. Nous décrivons le modèle HB bivarié à la section 3. À la section 4, nous résumons les résultats pour les terres cultivées non irriguées en Iowa, au Kansas et au Texas. À la section 5, nous donnons un résumé et discutons d'éventuels futurs travaux de recherche applicables à l'estimation des taux de location au comptant de terres cultivées, ainsi qu'à l'estimation sur petits domaines de manière plus générale.

2 Données pour la modélisation des taux de location au comptant de terres cultivées non irriguées

2.1 Cash Rent Survey du NASS

Le NASS a mis en œuvre la Cash Rent Survey (enquête sur la location au comptant) en réponse au *Farm Bill* de 2008. L'objectif particulier de cette enquête est d'obtenir des estimations au niveau du comté des taux moyens de location au comptant pour trois catégories d'utilisation des terres, à savoir les terres cultivées non irriguées, les terres cultivées irriguées et les pâturages permanents. Les données sur lesquelles porte notre étude proviennent des éditions de 2009 et 2010 de la Cash Rent Survey.

2.1.1 Plan de sondage de la Cash Rent Survey du NASS

Les éditions de 2009 et de 2010 de la Cash Rent Survey ont été réalisées selon un plan de sondage stratifié. Pour définir la stratification, neuf groupes ont été formés en se basant sur le montant total de location, en dollars, qu'une exploitation avait déclaré lors d'enquêtes et de recensements précédents. Les strates sont les intersections des neuf groupes et des districts statistiques agricoles. Un district statistique agricole est un groupe de comtés contigus à l'intérieur d'un État, qui sont considérés comme ayant des caractéristiques agricoles similaires. Les fractions d'échantillonnage dans les strates sont définies de manière que les exploitations ayant déclaré des montants monétaires de location élevés lors d'enquêtes et de recensements précédents aient de plus grandes probabilités de sélection. Le même échantillon, dont la taille au niveau national était d'environ 224 000 exploitations, a été utilisé pour les éditions de 2009 et de 2010 de la Cash Rent Survey. Une unité peut n'avoir répondu qu'une seule année en raison d'une non-réponse ou parce que l'exploitation n'avait été partie à une entente de location que l'une des deux années.

2.1.2 Relations entre les loyers au comptant des terres cultivées non irriguées en 2009 et en 2010

Pour une catégorie d'utilisation des terres particulière, un estimateur direct prend la forme d'un ratio d'une somme pondérée des montants en dollars de location à une somme pondérée des nombres d'acres

loués. Le poids associé à un répondant est égal à la taille de la population de la strate contenant le répondant divisée par le nombre d'unités répondantes dans cette strate. Berg et coll. (2014) explorent les liens entre les estimations directes pour deux années. Pour les États considérés dans Berg et coll. (2014), la corrélation entre les estimations directes pour les deux années varie de 0,20 à 0,99, où la corrélation est celle entre les comtés pour un État particulier. Comme nous mettons l'accent sur les modèles au niveau de l'unité, nous nous concentrons sur les relations au cours du temps au niveau de l'unité.

Pour mesurer la corrélation entre les taux de location au comptant déclarés en 2009 et en 2010 au niveau de l'unité (exploitant agricole), nous calculons les différences entre les taux de location au comptant au niveau de l'unité pour les terres cultivées non irriguées et la moyenne d'échantillon pour un comté. Seuls les individus qui déclarent un taux de location au comptant pour les terres cultivées non irriguées les deux années sont utilisés pour calculer les différences. La différence pour l'année t est $y_{ijt} - \overline{y}_{i.t}$, où y_{ijt} est le loyer au comptant par acre pour les terres cultivées non irriguées déclaré par l'exploitant j dans le comté i durant l'année t, et $\overline{y}_{i.t}$ est la moyenne d'échantillon des y_{ijt} dans le comté i pour les exploitants ayant déclaré un taux de location au comptant de terres cultivées non irriguées en 2009, ainsi qu'en 2010. Les écarts entre les taux de location au comptant individuels et les moyennes de comté pour le Kansas sont représentés graphiquement à la figure 2.1. Une relation linéaire existe entre les écarts pour 2009 et 2010 pour le Kansas, et la corrélation entre les écarts pour 2009 et les écarts pour 2010 est de 0,7. Dans la figure 2.1, les valeurs extrêmes reflètent la forte variabilité des taux de location au comptant des terres cultivées non irriguées à l'intérieur d'un comté au Kansas.

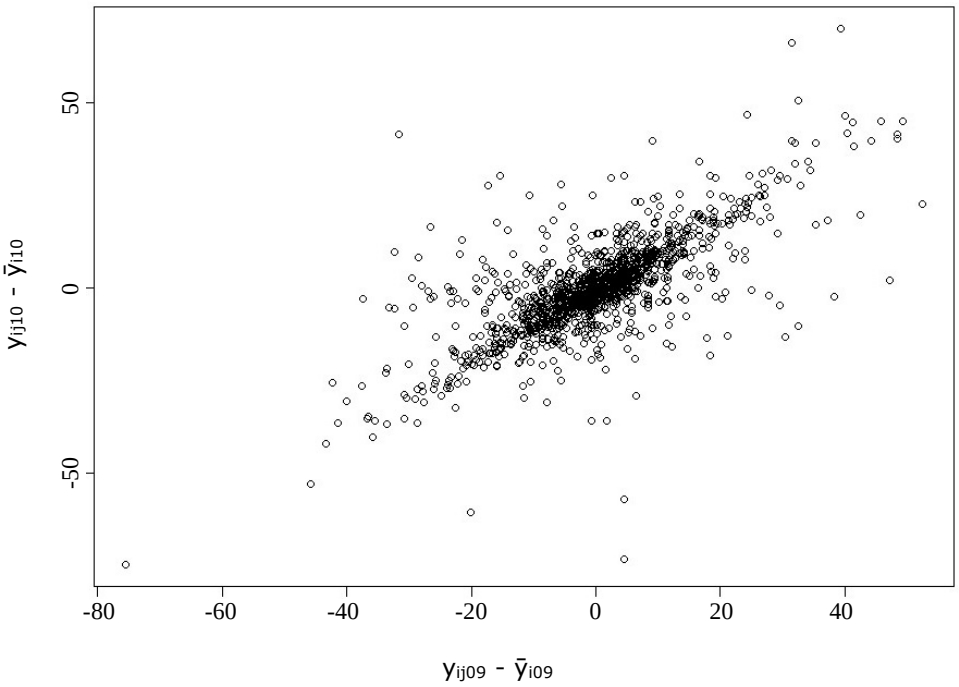

Figure 2.1 Écarts des taux de location au comptant au niveau de l'unité par rapport aux moyennes de comté pour 2009 (axe des x) et pour 2010 (axe des y) pour les unités ayant déclaré des taux de location au comptant de terres cultivées non irriguées les deux années.

2.2 Information auxiliaire

En vue d'améliorer la précision des estimateurs des taux moyens de location au comptant au niveau du comté, nous avons cherché des variables auxiliaires qui expliqueraient la variabilité entre les moyennes de comté, ainsi que la variabilité entre les unités dans un comté. L'information auxiliaire pour la modélisation des taux de location au comptant peut être obtenue auprès de plusieurs sources externes à la Cash Rent Survey. Les covariables possibles se répartissent en trois grandes catégories, selon que la covariable se rapporte principalement à la qualité des terres, à la valeur des produits vendus ou à d'autres caractéristiques agricoles. La liste qui suit résume les trois catégories de covariables, indique si chaque covariable est enregistrée au niveau du comté ou au niveau de l'unité, et précise si la covariable est disponible uniquement pour un État particulier. Les covariables au niveau de l'unité ne sont disponibles que pour les unités présentes dans l'échantillon de la Cash Rent Survey, tandis que les covariables au niveau du domaine sont traitées comme des moyennes de population.

1. Qualité des terres
    - Quatre indices nationaux de productivité des cultures de base (NCCPIs pour *National Commodity Crop Productivity Indexes*) sont des covariables au niveau du comté disponibles pour tous les États. Trois indices propres au climat appelés NCCPI-maïs, NCCPI-blé et NCCPI-coton reflètent la qualité du sol pour les terres cultivées non irriguées dans trois conditions climatiques différentes (Dobos, Sinclair et Robotham, 2012). Le quatrième indice, Max-NCCPI, est le maximum des trois indices propres au climat. Au départ, les indices sont construits au niveau d'une « unité cartographique », c'est-à-dire une région dont les propriétés du sol sont relativement homogènes. Les covariables au niveau du comté sont des moyennes des indices sur l'ensemble des unités cartographiques dans un comté.
    - Un rendement en maïs moyen sur les années 2005 à 2009 est disponible au niveau du comté pour l'Iowa seulement. Tous les comtés de l'Iowa possèdent une estimation du rendement en maïs pour au moins une des années entre 2005 et 2009, et les années pour lesquelles une estimation du rendement manque pour un comté sont exclues de la moyenne pour ce comté.
    - Comme le Kansas présente une plus grande diversité agricole que l'Iowa, aucun rendement pour une culture unique n'est publié pour au moins une année entre 2005 et 2009 pour tous les comtés d'intérêt. Pour obtenir une covariable mesurée pour l'ensemble des comtés, nous avons construit un indice de rendement des terres cultivées non irriguées pour le Kansas. Nous avons d'abord calculé la moyenne des rendements publiés par le NASS pour le maïs, le blé et le sorgho en appliquant la méthode décrite pour les rendements en maïs de l'Iowa. Puis, nous avons standardisé les rendements moyens pour avoir une moyenne nulle et une variance de 1. L'indice de rendement des terres cultivées non irriguées pour un comté est défini comme le plus grand des trois rendements standardisés. (Pour le Texas, l'information disponible sur le rendement des cultures était trop rare pour pouvoir être utilisée pour définir une covariable.)

2. Valeur des produits vendus
    - La valeur totale de la production pour un comté, fondée sur le Recensement de l'agriculture de 2007, est disponible pour tous les États.
    - Les ventes prévues pour une exploitation (unité) consignées dans la base de sondage du NASS sont disponibles pour tous les États au niveau de l'unité.

3. Autres caractéristiques agricoles
    - Le type d'exploitation agricole est une covariable catégorique au niveau de l'unité, disponible pour tous les États. Les exploitations agricoles sont réparties en 17 types dans la base de sondage du NASS. Pour définir une covariable, les types d'exploitation sont agrégés en deux groupes, à savoir 1) céréales et oléagineux et 2) autre.
    - Le nombre d'acres loués pour les terres cultivées non irriguées consigné sur le questionnaire de la Cash Rent Survey du NASS est disponible au niveau de l'unité pour tous les États.

3 Modèle hiérarchique bayésien bivarié

La corrélation entre les taux de location au comptant en 2009 et en 2010 observés à la section 2.1.1 donne à penser que l'utilisation de l'information contenue dans les données pour 2009 pourrait améliorer les prédictions pour 2010. Un modèle hiérarchique bivarié pour un État est spécifié comme moyen d'intégrer les données pour les deux années. Soit $a_{ij,t}$ et $y_{ij,t}$ les nombres d'acres loués et de dollars de location par acre, respectivement, pour l'exploitant j dans le comté i durant l'année $t\,(t = 09, 10)$, et soit $\mathbf{x}_{ij,t}$ le vecteur-colonne de dimension p_t associé aux variables auxiliaires. Pour les covariables qui sont constantes d'une année à l'autre et d'un individu à l'autre, $\mathbf{x}_{ij,t} = \mathbf{x}_{i109}$. Soit $w_{ij,t} = a_{ij,t} N_{g(ijt)} n_{g(ijt)}^{-1}$, où $N_{g(ijt)}$ et $n_{g(ijt)}$ sont la taille de la population et le nombre de répondants, respectivement, dans l'année t pour la strate g qui contient l'unité (ij).

Pour spécifier le modèle, nous répartissons les répondants en trois ensembles :

- l'ensemble 1 contient les unités (ij) qui déclarent un taux de location au comptant de terres cultivées non irriguées en 2009 ainsi qu'en 2010;
- l'ensemble 2 contient les unités (ij) qui déclarent un taux de location au comptant de terres cultivées non irriguées en 2009 seulement;
- l'ensemble 3 contient les unités (ij) qui déclarent un taux de location au comptant de terres cultivées non irriguées en 2010 seulement.

Nous supposons que les observations dans l'ensemble 1 satisfont le modèle bivarié

$$\begin{pmatrix} y_{ij,09} \\ y_{ij,10} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_{ij,09}\boldsymbol{\beta}_{09} + \nu_{i,09} + e_{ij,09} \\ \mathbf{x}'_{ij,10}\boldsymbol{\beta}_{10} + \nu_{i,10} + e_{ij,10} \end{pmatrix}, \tag{3.1}$$

où

$$\begin{pmatrix} e_{ij,09} \\ e_{ij,10} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0},\ \mathbf{D}_{wij}^{-0,5}\boldsymbol{\Sigma}_{ee}\mathbf{D}_{wij}^{-0,5}\right), \tag{3.2}$$

$\mathbf{D}_{wij} = \mathrm{diag}\left(w_{ij,09},\ w_{ij,10}\right)$ et

$$\begin{pmatrix} \nu_{i,09} \\ \nu_{i10} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\nu\nu}\right). \tag{3.3}$$

Nous notons les éléments diagonaux de $\boldsymbol{\Sigma}_{ee}$ correspondant à 2009 et à 2010 par σ_{ee09} et σ_{ee10}, respectivement. Pour les unités (ij) dans l'ensemble 2 ou 3, nous supposons que

$$y_{ij,t} = \mathbf{x}'_{ij,t}\boldsymbol{\beta}_t + \nu_{i,t} + e^*_{ij,t}, \tag{3.4}$$

où $e^*_{ij,t} \sim \mathrm{N}\left(0,\ w_{ij,t}^{-1}\,\tau^2_{e,t}\right)$, $t = 09$ pour l'ensemble 2, et $t = 10$ pour l'ensemble 3. Le modèle permet non seulement que les variances des erreurs au niveau de l'unité diffèrent entre les points dans le temps, mais aussi que les variances des erreurs au niveau de l'unité pour les unités qui répondent aux deux points dans le temps diffèrent de celles pour les unités qui répondent seulement à un point dans le temps. La quantité à prédire pour 2010 est

$$\theta_{i,10} = \bar{\mathbf{x}}'_{N_i,10}\boldsymbol{\beta}_{10} + \nu_{i,10}, \tag{3.5}$$

où $\bar{\mathbf{x}}_{N_i,10}$ est la moyenne de population des covariables pour le comté i.

Les variances des erreurs au niveau de l'unité, $e_{ij,t}$ et $e^*_{ij,t}$, sont supposées être inversement proportionnelles au poids, $w_{ij,t}$, pour deux raisons. Premièrement, l'introduction des poids dans le modèle vise à réduire le biais qui pourrait se produire si le plan est informatif pour le modèle. Comme nous l'expliquons à la section 2, les poids dépendent de la valeur monétaire des terres louées l'année précédente. Par conséquent, il est peu plausible que le plan de sondage puisse être informatif pour un modèle sans les poids. Si $\boldsymbol{\Sigma}_{ee}$ et $\boldsymbol{\Sigma}_{\nu\nu}$ sont diagonales, et si $\tau^2_{e,t} = \sigma_{eet}$, alors dans un cadre fréquentiste, le meilleur prédicteur linéaire sans biais empirique (EBLUP pour *empirical best linear unbiased predictor*) pour la moyenne du comté i durant l'année t est le pseudo-EBLUP convergent sous le plan de You et Rao (2002). La deuxième raison d'introduire les poids dans le modèle est que, d'après les analyses préliminaires, les variances des résidus diminuent à mesure que le nombre d'acres augmente.

Des lois a priori propres, diffuses, sont spécifiées pour les coefficients de régression et les variances inconnus. En particulier, $\boldsymbol{\beta}_t \sim \mathrm{N}\left(\mathbf{0}, 10^6\mathbf{I}\right)$, et $\tau^2_{e,t} \sim \text{Gamma inverse}\left(0{,}001;\ 0{,}001\right)$. Les matrices de covariance, $\boldsymbol{\Sigma}_{ee}$ et $\boldsymbol{\Sigma}_{\nu\nu}$, suivent une loi a priori de Wishart inverse de paramètre de forme 0,01 et de matrice d'échelle diagonale dont les éléments diagonaux sont de 0,001. Les paramétrisations des lois Gamma inverse et Wishart inverse sont tirées de Gelman, Carlin, Stern et Rubin (2009). Nous avons choisi des lois a priori possédant des formes conjuguées pour simplifier les calculs. Les hyperparamètres sont choisis de façon qu'ils soient non informatifs en ce qui concerne les données pour l'application de la Cash Rent Survey.

3.1 Échantillonnage de Gibbs et lois a posteriori

Nous utilisons l'échantillonnage de Gibbs pour obtenir une approximation Monte Carlo de la loi a posteriori. Une analyse de la statistique de Brooks-Gelman-Rubin (BGR) (Gelman et coll., 2009) fondée sur trois chaînes MCMC, chacune comprenant 20 000 itérations, a indiqué que 1 000 itérations suffisaient pour l'apprentissage. Les analyses présentées à la section 4 sont fondées sur une chaîne de longueur 20 000 pour chacun des trois États, Iowa, Kansas et Texas, où les 1 000 premières itérations sont écartées pour l'apprentissage. Étant donné les choix des vraisemblances et des lois a priori, les lois conditionnelles complètes sont des lois connues. Voir l'annexe A.

3.2 Prédiction et estimation de l'EQM

Si $\overline{\mathbf{x}}_{N_i,10}$ est connue, le prédicteur bayésien de $\theta_{i,10}$ pour la perte quadratique est

$$\tilde{\theta}_{i,10}^{B} = E\left[\theta_{i,10} \left| (\mathbf{y}, \mathbf{x}), \overline{\mathbf{x}}_{N_i,10}\right.\right] = \overline{\mathbf{x}}'_{N_i,10}\hat{\boldsymbol{\beta}}_{10} + E\left[v_{i,10} \left| (\mathbf{y}, \mathbf{x})\right.\right], \tag{3.6}$$

où $\hat{\boldsymbol{\beta}}_{10} = E\left[\boldsymbol{\beta}_{10} \left| (\mathbf{y}, \mathbf{x})\right.\right], (\mathbf{y}, \mathbf{x})$ désigne les taux de location au comptant et les covariables pour les deux années, et la deuxième égalité dans (3.6) découle de (3.5) et de la linéarité de l'espérance. L'erreur quadratique moyenne a posteriori de $\tilde{\theta}_{i,10}^{B}$ est

$$E\left[\left(\tilde{\theta}_{i,10}^{B} - \theta_{i,10}\right)^2 \middle| (\mathbf{y}, \mathbf{x}), \overline{\mathbf{x}}_{N_i,10}\right] = V\left\{\theta_{i,10} \left| (\mathbf{y}, \mathbf{x}), \overline{\mathbf{x}}_{N_i,10}\right.\right\}. \tag{3.7}$$

Comme il est discuté à la section 2, la moyenne de population des covariables, $\overline{\mathbf{x}}_{N_i,10}$, n'est pas disponible pour les covariables au niveau de l'unité dans l'application de la Cash Rent Survey. Pour définir un prédicteur, nous ajoutons un modèle pour la moyenne des covariables. Voir Lohr et Prasad (2003) pour une approche commençant par la spécification d'un modèle pour les covariables au niveau de l'unité. Partitionnons $\mathbf{x}_{ij,10}$ en deux sous-vecteurs, $\mathbf{x}_{ij,10}^{(1)}$ et $\mathbf{x}_{ij,10}^{(2)}$, où $\mathbf{x}_{ij,10}^{(1)}$ contient les covariables au niveau du comté, et $\mathbf{x}_{ij,10}^{(2)}$ contient les covariables au niveau de l'unité. Supposons que $\overline{\mathbf{x}}_{wi10} \left| \overline{\mathbf{x}}_{N_i,10} \sim \mathrm{N}\left(\overline{\mathbf{x}}_{N_i,10}, \mathbf{V}_{xxi,10}\right)\right.$, où $\overline{\mathbf{x}}_{wi10} = \left(\sum_{j=1}^{n_{i10}} w_{ij,10}\right)^{-1}\left(\sum_{j=1}^{n_{i10}} w_{ij,10}\mathbf{x}_{ij,10}\right)$, n_{i10} est la somme des nombres d'unités dans l'ensemble 1 et dans l'ensemble 3, et $\mathbf{V}_{xxi,10}$ est connue. Les éléments de $\mathbf{V}_{xxi,10}$ correspondant à $\mathbf{x}_{ij,10}^{(1)}$ sont nuls, et nous expliquons comment nous obtenons les éléments de $\mathbf{V}_{xxi,10}$ correspondant aux covariables au niveau de l'unité à l'annexe B. Le théorème central limite soutient l'hypothèse de normalité pour $\overline{\mathbf{x}}_{wi10}$, même si la distribution des valeurs des covariables au niveau de l'unité n'est pas normale (Kim, Park et Lee, 2017). En supposant que $\overline{\mathbf{x}}_{N_i,10}$ suit une loi a priori uniforme, $\overline{\mathbf{x}}_{N_i,10} \left| \overline{\mathbf{x}}_{wi10} \sim \mathrm{N}\left(\overline{\mathbf{x}}_{wi10}, \mathbf{V}_{xxi,10}\right)\right.$. Le prédicteur bayésien de $\theta_{i,10}$ pour la perte quadratique sous le modèle étendu dans lequel la moyenne de population des covariables est inconnue est

$$\hat{\theta}_{i,10}^{B} = \overline{\mathbf{x}}'_{wi10}\hat{\boldsymbol{\beta}}_{10} + E\left[v_{i,10} \left| (\mathbf{y}, \mathbf{x})\right.\right]. \tag{3.8}$$

L'erreur quadratique moyenne a posteriori de $\hat{\theta}_{i,10}^{B}$ est

$$
\begin{aligned}
E\left[\left(\hat{\theta}_{i,10}^{B}-\theta_{i,10}\right)^{2}\Big|(\mathbf{y},\mathbf{x})\right] &= E\left\{\left(\hat{\theta}_{i,10}^{B}-\tilde{\theta}_{i,10}^{B}+\tilde{\theta}_{i,10}^{B}-\theta_{i,10}\right)^{2}\Big|(\mathbf{y},\mathbf{x})\right\} \\
&= E\left\{\left(\hat{\theta}_{i,10}^{B}-\tilde{\theta}_{i,10}^{B}\right)^{2}\Big|(\mathbf{y},\mathbf{x})\right\} \\
&\quad + 2E\left\{E\left[\left(\hat{\theta}_{i,10}^{B}-\tilde{\theta}_{i,10}^{B}\right)\left(\tilde{\theta}_{i,10}^{B}-\theta_{i,10}\right)\Big|(\mathbf{y},\mathbf{x}),\overline{\mathbf{x}}_{N_i,10}\right]\Big|(\mathbf{y},\mathbf{x})\right\} \\
&\quad + V\left\{\theta_{i,10}\,\middle|\,(\mathbf{y},\mathbf{x})\right\} \\
&= \hat{\boldsymbol{\beta}}_{10}' V\left\{\overline{\mathbf{x}}_{N_i,10}\,\middle|\,\overline{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10} + V\left\{\theta_{i,10}\,\middle|\,(\mathbf{y},\mathbf{x})\right\} \\
&= \hat{\boldsymbol{\beta}}_{10}' V\left\{\overline{\mathbf{x}}_{N_i,10}\,\middle|\,\overline{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10} \\
&\quad + V\left\{\overline{\mathbf{x}}_{wi10}'\boldsymbol{\beta}_{10}+\nu_{i,10}+\left(\overline{\mathbf{x}}_{N_i,10}-\overline{\mathbf{x}}_{wi10}\right)'\boldsymbol{\beta}_{10}\,|(\mathbf{y},\mathbf{x})\right\} \\
&\approx \hat{\boldsymbol{\beta}}_{10}' V\left\{\overline{\mathbf{x}}_{N_i,10}\,\middle|\,\overline{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10} + V\left\{\overline{\mathbf{x}}_{wi10}'\boldsymbol{\beta}_{10}+\nu_{i,10}\,\middle|\,(\mathbf{y},\mathbf{x})\right\}, \qquad (3.9)
\end{aligned}
$$

où l'approximation finale suppose que $\mathbf{Cov}\left\{\overline{\mathbf{x}}_{wi10}'\boldsymbol{\beta}_{10}+\nu_{i,10},\left(\overline{\mathbf{x}}_{N_i,10}-\overline{\mathbf{x}}_{wi10}\right)'\boldsymbol{\beta}_{10}\,|(\mathbf{y},\mathbf{x})\right\}$ est négligeable. Une comparaison de (3.7) et (3.9) montre que le terme $\hat{\boldsymbol{\beta}}_{10}' V\left\{\overline{\mathbf{x}}_{N_i,10}\,\middle|\,\overline{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10}$ tient compte de l'augmentation de l'EQM a posteriori due au remplacement de $\overline{\mathbf{x}}_{N_i,10}$ dans (3.6) par $\overline{\mathbf{x}}_{wi10}$ dans (3.8). Pour quantifier l'EQM a posteriori de $\hat{\theta}_{i,10}^{B}$, nous utilisons

$$
\mathrm{EQM}\left(\hat{\theta}_{i,10}^{B}\right) = \widehat{\mathrm{EQM}}_{1i} + \widehat{\mathrm{EQM}}_{2i}, \qquad (3.10)
$$

où $\widehat{\mathrm{EQM}}_{1i} = V\left\{\overline{\mathbf{x}}_{wi10}'\boldsymbol{\beta}_{10}+\nu_{i,10}\,\middle|\,(\mathbf{y},\mathbf{x})\right\}$ et $\widehat{\mathrm{EQM}}_{2i} = \hat{\boldsymbol{\beta}}_{10}'\mathbf{V}_{xxi,10}\hat{\boldsymbol{\beta}}_{10}$. Dans l'application de la section 4, nous évaluons l'effet de l'inclusion du terme $\widehat{\mathrm{EQM}}_{2i}$, qui tient compte de l'augmentation de l'EQM a posteriori due à l'utilisation de la moyenne d'échantillon au lieu de la moyenne de population de la covariable, sur l'EQM a posteriori du prédicteur.

3.3 Réconciliation en deux étapes

Le NASS obtient des estimations des taux de location au comptant au niveau de l'État en utilisant les données provenant d'une enquête nationale réalisée en juin (la *June Area Survey*) en plus de la Cash Rent Survey. Les estimations au niveau de l'État sont publiées avant que les données au niveau du comté provenant de la Cash Rent Survey soient traitées entièrement. Le NASS établit aussi des estimations des taux de location au comptant pour les districts statistiques agricoles. Afin de maintenir la cohérence interne, les sommes pondérées de manière appropriée des estimations au niveau du comté doivent être égales aux estimations au niveau du district, et les sommes pondérées de manière appropriée des estimations au niveau du district doivent être égales aux estimations au niveau de l'État publiées antérieurement. En notant $\hat{\theta}_{i10}$ le prédicteur réconcilié pour 2010, les contraintes de réconciliation pour un seul point dans le temps sont définies par

$$
\sum_{i\in d_k} w_{i10}\hat{\theta}_{i10} = \hat{\lambda}_{k10}, \qquad (3.11)
$$

et

$$\sum_{k=1}^{K} \eta_{k10} \hat{\lambda}_{k10} = \theta_{\text{pub10}}, \tag{3.12}$$

où $k = 1, \ldots, K$ indice les districts, $w_{i10} = \left(\sum_{i \in d_k} z_{i10}\right)^{-1} z_{i10}$,

$$\eta_{k10} = \left(\sum_{k=1}^{K} \sum_{i \in d_k} z_{i10}\right)^{-1} \sum_{i \in d_k} z_{i10},$$

z_{i10} est l'estimateur direct du nombre d'acres loués dans le comté i durant l'année 2010, d_k est l'ensemble d'indices pour les comtés dans le district k, $\hat{\lambda}_{k10}$ est l'estimation finale du taux moyen de location au comptant pour le district k, et θ_{pub10} est l'estimation publiée du loyer au comptant par acre au niveau de l'État. Nous considérons les estimations pour l'année 2010 dans (3.11) et (3.12), parce que nous nous concentrons sur l'estimation pour 2010 dans l'analyse présentée à la section 4.

Nous utilisons la procédure de réconciliation en deux étapes proposée par Ghosh et Steorts (2013) pour définir les estimations réconciliées. Les estimations réconciliées minimisent la forme quadratique

$$g(\mathbf{c}, \mathbf{b}) = \sum_{k=1}^{K} \sum_{i \in d_k} \xi_i \left(\hat{\theta}^B_{i10} - c_i\right)^2 + \sum_{k=1}^{K} \rho_k \left(\hat{\theta}^B_{k10,w} - b_k\right)^2 \tag{3.13}$$

sous les contraintes dans (3.11) et (3.12), où $\mathbf{c} = (c_1, \ldots, c_D)$, D désigne le nombre total de comtés, $\mathbf{b} = (b_1, \ldots, b_K)$, $\hat{\theta}^B_{k10,w} = \sum_{i \in d_k} w_{i10} \hat{\theta}^B_{i10}$, et (ρ_k, ξ_i) sont des constantes sélectionnées par l'analyste. Nous prenons $\xi_i = w_{i10}$ et $\rho_k = \eta_{k10}$, ce qui donne les estimations réconciliées

$$\hat{\theta}_{i10} = \hat{\theta}^B_{i10} + \hat{\lambda}_{k(i)10} - \hat{\theta}^B_{k(i)10,w}, \tag{3.14}$$

avec

$$\hat{\lambda}_{k(i)10} = \hat{\theta}^B_{k(i)10,w} + \frac{\left(\theta_{\text{pub10}} - \hat{\theta}^B_{w10}\right) \eta_{k(i)10} \left(1 + \eta_{k(i)10}\right)^{-1}}{\sum_{i \in d_{k(i)10}} \eta^2_{k(i)10} \left(1 + \eta_{k(i)10}\right)^{-1}}, \tag{3.15}$$

pour le comté i et le district $k(i)$, respectivement, où $k(i)$ est le district contenant le comté i. Dans (3.15), $\hat{\theta}^B_{w10} = \sum_{k=1}^{K} \eta_{k10} \hat{\theta}^B_{k10,w}$. Chacune des estimations réconciliées dans (3.14) et (3.15) est une somme du prédicteur hiérarchique bayésien et d'un terme d'ajustement. Si le prédicteur hiérarchique bayésien pour l'État est plus grand (plus petit) que le total au niveau de l'État publié antérieurement, alors l'ajustement est négatif (positif), et les estimations réconciliées au niveau du comté et du district sont plus petites (plus grandes) que les prédicteurs hiérarchiques bayésiens. L'erreur quadratique moyenne a posteriori du prédicteur réconcilié pour l'année t est

$$\text{EQM}^{B\text{réc.}}_{i10} = \text{EQM}\left(\hat{\theta}^B_{i10}\right) + \left(\hat{\theta}^B_{i10} - \hat{\theta}_{i10}\right)^2, \tag{3.16}$$

où le terme $\text{EQM}\left(\hat{\theta}_{i10}^{B}\right)$ est défini en (3.10). Voir (You, Rao et Dick, 2004) pour le calcul de l'EQM a posteriori d'un prédicteur réconcilié.

4 Résultats pour les terres cultivées non irriguées en Iowa, au Kansas et au Texas

Le modèle de la section 3 a été ajusté sur les taux de location au comptant de terres cultivées non irriguées déclarés lors des éditions de 2009 et de 2010 de la Cash Rent Survey pour l'Iowa, le Kansas et le Texas. Ces trois États ont été choisis afin de refléter une gamme de situations. Pour l'Iowa, des estimations des rendements en maïs sont disponibles pour tous les comtés, et la location au comptant est un mode relativement fréquent de location de terres cultivées non irriguées. Le Kansas présente une plus grande diversité agricole que l'Iowa. Selon les spécialistes de l'agriculture du NASS, dans de nombreuses régions du Texas, la location en métayage est plus fréquente que la location au comptant, ce qui pourrait expliquer pourquoi les tailles d'échantillon réalisées pour certains comtés texans sont aussi petites que zéro ou un.

4.1 Choix des covariables

Les covariables possibles pour l'Iowa, le Kansas et le Texas sont énumérées à la section 2.2. Pour chaque État, les covariables comprennent quatre variables liées au NCCPI, la valeur totale de la production pour un comté basée sur le Recensement de l'agriculture de 2007, les ventes prévues pour une exploitation enregistrée dans la base de sondage du NASS, le type d'exploitation agricole enregistré dans la base de sondage du NASS, et le nombre d'acres loués pour des terres cultivées non irriguées déclaré lors de la Cash Rent Survey du NASS. Pour l'Iowa, le rendement en maïs au niveau du comté est une covariable supplémentaire. Pour le Kansas, l'indice de rendement des terres cultivées non irriguées est une covariable supplémentaire.

Pour chaque État, les covariables ont été sélectionnées selon la procédure qui suit. D'abord, des modèles univariés ont été ajustés séparément aux données de 2009 et de 2010 en utilisant des estimations du maximum de vraisemblance. Le modèle univarié utilisé pour le choix des covariables est de la forme

$$y_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\alpha}_t + \nu_{it} + \epsilon_{ijt}, \tag{4.1}$$

où $\epsilon_{ijt} \sim \text{N}\left(0, \sigma^2_{\epsilon,t}\right)$, et $\nu_{it} \sim \text{N}\left(0, \sigma^2_{\nu,t}\right)$. Les données pour chaque exploitant agricole ayant déclaré un taux de location au comptant de terres cultivées non irriguées durant l'année t ont été utilisées pour ajuster le modèle univarié pour l'année t, que l'unité ait déclaré ou non un taux de location au comptant pour l'année $s\,(s \neq t)$. La fonction `lmer` du package `nlme` en `R` est utilisée pour l'estimation du maximum de vraisemblance. Pour chaque année, une sélection pas à pas en utilisant la fonction `stepAIC` en `R` est exécutée en utilisant la mesure BIC. Les covariables sélectionnées sont les variables figurant dans les modèles univariés dont le BIC est minimum pour 2009 ainsi que 2010. Nous reconnaissons que le modèle

à BIC minimum est un minimum local identifié par la procédure `stepAIC` plutôt qu'un minimum global. Les covariables sélectionnées pour l'Iowa, le Kansas et le Texas sont les suivantes :

- Iowa : rendement en maïs, ventes prévues, acres non irrigués loués au comptant.
- Kansas : indice de rendement de terres cultivées non irriguées, ventes prévues, type d'exploitation agricole.
- Texas : max-NCCPI, ventes prévues, type d'exploitation agricole.

4.2 Estimations des paramètres de corrélation

L'analyse exploratoire de la section 2.1 porte à croire qu'il existe une corrélation importante entre les taux de location au comptant de terres cultivées non irriguées observés pour 2009 et pour 2010. Le tableau 4.1 résume les lois a posteriori des corrélations dans le modèle HB bivarié défini à la section 3.1. Les colonnes intitulées « Médiane » donnent les médianes a posteriori des corrélations, et les bornes inférieure et supérieure des intervalles de crédibilité à 95 % correspondent aux 2,5^{e} et 97,5^{e} centiles des lois a posteriori des corrélations. Même si les variances de e_{ij09} et e_{ij10} sont proportionnelles aux inverses des poids, la corrélation est une constante, parce que les poids s'annulent dans la définition de la corrélation.

Tableau 4.1
Lois a posteriori des corrélations entre 2009 et 2010

	$\text{Cor}\{\nu_{i09}, \nu_{i10}\}$		$\text{Cor}\{e_{ij09}, e_{ij10}\}$	
État	**Médiane**	**Intervalle de crédibilité à 95 %**	**Médiane**	**Intervalle de crédibilité à 95 %**
Iowa	0,746	[0,611; 0,839]	0,570	[0,548; 0,592]
Kansas	0,919	[0,870; 0,950]	0,727	[0,701; 0,751]
Texas	0,884	[0,831; 0,921]	0,691	[0,667; 0,714]

Les médianes a posteriori des corrélations au niveau du comté et au niveau de l'unité dépassent 0,74 et 0,57, respectivement. Les bornes inférieures des intervalles de crédibilité à 95 % sont supérieures à 0,61 et 0,54 pour les corrélations au niveau du comté et au niveau de l'unité, respectivement. Pour chaque État, les corrélations au niveau du comté sont plus grandes que les corrélations pour les unités individuelles. Les corrélations importantes semblent indiquer la possibilité d'un gain d'efficacité pour les prédicteurs par rapport au modèle univarié.

4.3 Comparaison des prédicteurs pour 2010 pour les modèles bivarié et univarié

Afin de démontrer le gain d'efficacité dû à l'utilisation du modèle bivarié comparativement à un modèle univarié, nous comparons les erreurs quadratiques moyennes a posteriori des prédicteurs pour le modèle bivarié aux erreurs quadratiques moyennes a posteriori des prédicteurs pour un modèle univarié correspondant. Les hypothèses des modèles univariés sont les mêmes que celles des modèles bivariés, sauf

que l'on suppose que les paramètres de covariance dans $\mathbf{\Sigma}_{ee}$ et $\mathbf{\Sigma}_{vv}$ sont nuls. Pour ajuster les modèles univariés, nous utilisons des lois a priori Gamma inverses pour σ_{eet} et σ_{vvt} $(t = 09,10)$.

Pour comparer les modèles bivarié et univarié, nous définissons l'EQM relative a posteriori (EQMrel) pour le comté i par

$$\text{EQMrel}_{i,10} = \frac{\text{EQM}_{i10}^{B\text{réc.}}}{\text{EQM}_{i10}^{\text{UNIréc.}}}, \tag{4.2}$$

où $\text{EQM}_{i10}^{B\text{réc.}}$ est défini en (3.16) et $\text{EQM}_{i10}^{\text{UNIréc.}}$ est l'EQM a posteriori basée sur le modèle univarié correspondant. Les EQM relatives moyennes pour l'Iowa, le Kansas et le Texas valent 88,71 %, 97,27 % et 88,65 %, respectivement, où la moyenne des erreurs quadratiques moyennes relatives pour un État est $D^{-1}\sum_{i=1}^{D}\text{EQMrel}_{i,10}$. Notons que les effets de l'estimation de la moyenne des covariables, ainsi que de la réconciliation sont intégrés dans les formules de l'EQM a posteriori tant pour les modèles bivariés qu'univariés. En raison des corrélations importantes entre les erreurs de modélisation pour les deux points dans le temps, l'EQM a posteriori pour un modèle bivarié est plus petite que l'EQM a posteriori pour le modèle univarié correspondant, et les efficacités relatives moyennes sont inférieures à un.

Pour évaluer l'effet de l'estimation de la moyenne de population des covariables sur l'EQM du prédicteur, nous calculons la moyenne des ratios $\widehat{\text{EQM}}_{2i}\,\widehat{\text{EQM}}_{1i}^{-1}$ pour $i = 1, \ldots, D$, où $\widehat{\text{EQM}}_{2i}$ et $\widehat{\text{EQM}}_{1i}$ sont définies en suivant (3.10). Les ratios sont de 18,21 %, 28,20 % et 21,07 % pour l'Iowa, le Kansas et le Texas, respectivement. Comparativement à l'Iowa et au Texas, la contribution à l'EQM de prédiction due à l'utilisation de la moyenne des covariables dans l'échantillon plutôt que de la moyenne des covariables dans la population est plus importante au Kansas, ce qui est logique puisque le Kansas présente une plus grande diversité agricole. L'EQM relative moyenne plutôt grande pour le Kansas (97,27 %) traduit l'accroissement relativement important de l'EQM a posteriori due à l'estimation de la moyenne des covariables.

4.4 Évaluation du modèle

Afin d'évaluer l'adéquation du modèle, nous utilisons la valeur p prédictive a posteriori, qui mesure les écarts entre les données observées et le modèle. La valeur p prédictive a posteriori compare la loi prédictive a posteriori de certaines statistiques sommaires aux valeurs correspondantes obtenues en utilisant l'échantillon original. Pour l'analyse qui suit, nous utilisons uniquement les éléments observés en 2009 ainsi qu'en 2010 (ensemble 1).

Nous considérons deux statistiques sommaires, à savoir la moyenne pour chaque année et l'asymétrie multivariée. La moyenne pour l'année t est la moyenne des observations dans l'ensemble 1 pour l'année t, et est définie comme étant

$$\bar{y}_t = \left(\sum_{i=1}^{D}|A_i|\right)^{-1}\sum_{i=1}^{D}\sum_{j\in A_i} y_{ijt},$$

où A_i désigne les éléments dans l'ensemble 1 pour le comté i. L'asymétrie multivariée est définie par

$$\hat{\gamma}_{1,p} = \left(\sum_{i=1}^{D} |A_i|\right)^{-1} \sum_{i=1}^{D} \sum_{k=1}^{D} \sum_{j \in A_i} \sum_{\ell \in A_i} m_{ijk\ell}^3,$$

où $m_{ijk\ell} = (\mathbf{y}_{ij} - \overline{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_{k\ell} - \overline{\mathbf{y}})$, $\mathbf{y}_{ij} = (y_{ij,09}, y_{ij,10})'$, $\overline{\mathbf{y}} = (\overline{y}_{09}, \overline{y}_{10})'$ et $\mathbf{S} = \left(\sum_{i=1}^{D} |A_i| - 1\right)^{-1} \sum_{i=1}^{D} \sum_{j \in A_i} (\mathbf{y}_{ij} - \overline{\mathbf{y}})(\mathbf{y}_{ij} - \overline{\mathbf{y}})'$.

La valeur p prédictive a posteriori est définie comme étant la proportion de la statistique sommaire calculée avec des échantillons générés à partir de la loi prédictive a posteriori qui est en excès de la valeur correspondante fondée sur l'échantillon original. Plus précisément, soit $T(\mathbf{y}^{(r)})$ la statistique sommaire basée sur le r^{e} ensemble de données généré à partir de la loi prédictive a posteriori, où la procédure pour générer les données à partir de la loi prédictive a posteriori est définie à l'annexe C. Soit $T(\mathbf{y})$ la statistique correspondante fondée sur l'échantillon original. La valeur p prédictive a posteriori est $R^{-1} \sum_{r=1}^{R} I\left[T(\mathbf{y}^{(r)}) > T(\mathbf{y})\right]$. Une valeur p proche de 0,5 indique un ajustement raisonnable du modèle aux données de l'échantillon.

Le tableau 4.2 donne les valeurs p prédictives a posteriori pour l'Iowa, le Kansas et le Texas. Pour le Kansas, les valeurs prédictives a posteriori indiquent que l'adéquation entre le modèle et les données est bonne. Pour l'Iowa et le Texas, les valeurs p prédictives a posteriori indiquent un manque d'adéquation. Une analyse plus approfondie des résidus donne à penser que le manque d'adéquation peut résulter de valeurs aberrantes. Les valeurs p prédictives a posteriori éloignées de 0,5 peuvent aussi découler du fait que nous utilisons uniquement les observations échantillonnées en 2009 ainsi qu'en 2010 pour calculer les valeurs p prédictives a posteriori, alors que nous utilisons l'ensemble de données complet pour ajuster le modèle.

Tableau 4.2
Valeurs p prédictives a posteriori

État	Statistique	Valeur p
Iowa	Moyenne t = 09	1,000
	Moyenne t = 10	1,000
	Asymétrie	0,931
Kansas	Moyenne t = 09	0,291
	Moyenne t = 10	0,507
	Asymétrie	0,371
Texas	Moyenne t = 09	0,025
	Moyenne t = 10	0,039
	Asymétrie	0,004

5 Conclusion et travaux à venir

Nous utilisons un modèle HB bivarié pour obtenir des prédicteurs des taux de location au comptant au niveau du comté pour les terres cultivées non irriguées en Iowa, au Kansas et au Texas. Le modèle intègre de l'information auxiliaire concernant la qualité des terres, les valeurs des produits et les caractéristiques de

l'exploitation agricole. Des corrélations significatives existent entre les effets aléatoires du modèle en 2009 et en 2010 au niveau tant de l'unité que du comté. En conséquence, l'utilisation de l'information contenue dans les estimations des loyers au comptant en 2009 réduit l'EQM a posteriori comparativement à celle d'un modèle univarié. L'analyse du modèle HB bivarié étaye l'idée qu'une approche plus fine que celle de Berg et coll. (2014) est possible. Afin d'intégrer des covariables au niveau de l'unité dont les moyennes de population sont inconnues, nous ajoutons au modèle hiérarchique un niveau qui justifie l'ajout d'un terme à l'erreur quadratique moyenne a posteriori afin de tenir compte de l'incertitude dans les moyennes de population inconnues des covariables au niveau de l'unité. Contrairement à l'approche de Berg et coll. (2014), le modèle HB bivarié proposé permet que la variabilité évolue au cours du temps et tient compte des effets de la réconciliation sur l'EQM.

L'analyse des résidus et des valeurs p prédictives a posteriori donne à penser que tenir compte des valeurs aberrantes pourrait être un bon moyen d'améliorer considérablement l'adéquation du modèle. Une option consiste à considérer une loi à queues lourdes, telle que la loi t ou un mélange de lois normales, qui peut représenter les réponses observées de manière plus appropriée que la loi normale présumée. Une extension des travaux de Gershunskaya (2010) au cadre bivarié et à l'estimation bayésienne est un moyen possible d'aborder la question des valeurs aberrantes.

Remerciements

Le *National Agricultural Statistics Service* (NASS) du *United States Department of Agriculture* (USDA) a financé la présente étude. Les auteurs remercient Wendy Barboza, Dan Beckler, Angie Considine, Mark Harris, Sharyn Lavender, Joe Parsons, Scot Rumberg, Scott Shimmin, Curt Stock et Linda Young, du *National Agriculture Statistics Service*. En outre, les auteurs remercient Bob Dobos, du *National Resource Conservation Service*, et Rich Iovanna, du *Farm Service Agency*, de leur aide en vue d'obtenir les données du *National Commodity Crop Productivity Index*. Sans la généreuse assistance de ces personnes, la présente étude aurait été impossible. Les opinions exprimées dans le présent article sont celles des auteurs et ne représentent pas forcément celles du NASS ni du USDA.

Annexe A

Afin de spécifier les lois conditionnelles complètes pour l'échantillonnage de Gibbs, nous présentons la notation. Soit Θ_γ l'ensemble de paramètres, à l'exception du paramètre désigné par γ. Soit $\mathbf{X}_{ij} = \left(\mathbf{z}_{ij,09}, \mathbf{z}_{ij,10}\right)'$, où $\mathbf{z}_{ij,09} = \left(\mathbf{x}'_{ij,09}, \mathbf{0}'_{p_{10}}\right)'$, et $\mathbf{z}_{ij,10} = \left(\mathbf{0}'_{p_{09}}, \mathbf{x}'_{ij,10}\right)'$. Soit $\mathbf{y}'_{ij} = \left(y_{ij,09}, y_{ij,10}\right)$. Soit A_i l'ensemble d'unités (exploitants agricoles) dans le comté i qui se trouvent dans l'ensemble 1, $B_{i,09}$, l'ensemble d'unités dans le comté i qui se trouvent dans l'ensemble 2, et $B_{i,10}$, l'ensemble d'unités dans le comté i qui se trouvent dans l'ensemble 3, où l'ensemble 1, l'ensemble 2 et l'ensemble 3 sont définis à la section 3. Les lois conditionnelles complètes sont les suivantes.

1. $\boldsymbol{\beta} \,|\, \left(\Theta_{\boldsymbol{\beta}}, \mathbf{y}\right) \sim \mathrm{N}\left(\boldsymbol{\Sigma}_{\beta\beta}\mathbf{r}_{\beta}, \boldsymbol{\Sigma}_{\beta\beta}\right)$, où

$$\boldsymbol{\Sigma}_{\beta\beta} = \left[\sum_{i=1}^{D}\sum_{j\in A_i}\mathbf{X}'_{ij}\mathbf{D}^{0,5}_{wij}\,\boldsymbol{\Sigma}^{-1}_{ee}\mathbf{D}^{0,5}_{wij}\mathbf{X}_{ij} + 10^{-6}\mathbf{I}_{p_{09}+p_{10}} + \boldsymbol{\Omega}\right]^{-1} \tag{A.1}$$

$$\boldsymbol{\Omega} = \text{diag par blocs}\left(\tau^{-2}_{e,09}\sum_{i=1}^{D}\sum_{j\in B_{i,09}} w_{ij,09}\mathbf{x}_{ij,09}\mathbf{x}'_{ij,09}, \tau^{-2}_{e,10}\sum_{i=1}^{D}\sum_{j\in B_{i,10}} w_{ij,10}\mathbf{x}_{ij,10}\mathbf{x}'_{ij,10}\right)$$

$$\mathbf{r}_{\beta} = \sum_{i=1}^{D}\sum_{j\in A_i}\mathbf{X}'_{ij}\mathbf{D}^{0,5}_{wij}\boldsymbol{\Sigma}^{-1}_{ee}\mathbf{D}^{0,5}_{wij}\left(\mathbf{y}_{ij} - \mathbf{v}_i + \mathbf{r}_{\beta 2}\right),$$

et

$$\mathbf{r}_{\beta 2} = \begin{pmatrix} \sum_{i=1}^{D}\sum_{j\in B_{i,09}} \tau^{-2}_{e,09} w_{ij,09}\mathbf{x}_{ij,09}\left(y_{ij,09} - \nu_{i,09}\right) \\ \sum_{i=1}^{D}\sum_{j\in B_{i,10}} \tau^{-2}_{e,10} w_{ij,10}\mathbf{x}_{ij,10}\left(y_{ij,10} - \nu_{i,10}\right) \end{pmatrix}. \tag{A.2}$$

2. $\boldsymbol{\Sigma}_{ee} \,|\, \left(\boldsymbol{\Theta}_{\boldsymbol{\Sigma}_{ee}}, \mathbf{y}\right) \sim \text{Wishart inverse}\left(\mathbf{A}_e, d_e\right)$, où $d_e = \sum_{i=1}^{D}|A_i| + 0{,}001$, et

$$\mathbf{A}_e = \sum_{i=1}^{D}\sum_{j\in A_i}\mathbf{D}^{0,5}_{w_{ij}}\left(\mathbf{y}_{ij} - \mathbf{v}_i - \mathbf{X}_{ij}\boldsymbol{\beta}\right)\left(\mathbf{y}_{ij} - \mathbf{v}_i - \mathbf{X}_{ij}\boldsymbol{\beta}\right)'\mathbf{D}^{0,5}_{w_{ij}}. \tag{A.3}$$

3. $\boldsymbol{\Sigma}_{\nu\nu} \,|\, \left(\boldsymbol{\Theta}\mathrm{Q}_{\boldsymbol{\Sigma}_{\nu\nu}}, \mathbf{y}\right) \sim \text{Wishart inverse}\left(\mathbf{A}_{\nu}, d_{\nu}\right)$, où

$$d_{\nu} = D + 0{,}001, \tag{A.4}$$

et

$$\mathbf{A}_{\nu} = \sum_{i=1}^{D}\mathbf{v}_i\mathbf{v}'_i. \tag{A.5}$$

$\tau^2_{e,t} \,|\, \left(\boldsymbol{\Theta}_{\tau^2_{e,t}}, \mathbf{y}\right) \sim \text{Gamma inverse}\left(a_{et}, d_{et}\right)$, où

$$d_{et} = \sum_{i=1}^{D}\left|B_{i,t}\right| + 0{,}001, \tag{A.6}$$

et

$$a_{et} = \sum_{i=1}^{D}\sum_{j\in B_{it}}\mathbf{D}_{wi_j}\left(y_{ij,t} - \nu_{i,t} - \mathbf{x}_{ij,t}\boldsymbol{\beta}_t\right)^2. \tag{A.7}$$

4. $\mathbf{v}_i \,|\, \left(\boldsymbol{\Theta}_{\nu_i}, \mathbf{y}\right) \sim \mathrm{N}\left(\boldsymbol{\mu}_{\nu\nu}, \mathbf{M}_i^{-1}\right)$, où

$$\mathbf{M}_i = \left(\boldsymbol{\Sigma}^{-1}_{\nu\nu} + \boldsymbol{\Sigma}^{-1}_{ee,wi} + \boldsymbol{\Omega}^{-1}_{ee,wi}\right)^{-1}, \tag{A.8}$$

$\boldsymbol{\mu}_{\nu\nu} = \mathbf{M}_i^{-1}\left(\mathbf{r}_{i_1} + \mathbf{r}_{i_2}\right)$, $\boldsymbol{\Sigma}_{ee,wi} = \sum_{j\in A_i}\mathbf{D}^{0,5}_{w_{ij}}\boldsymbol{\Sigma}^{-1}_{ee}\mathbf{D}^{0,5}_{w_{ij}}$,

$$\mathrm{W}_{ee,wi} = \mathrm{diag}\left(\tau_{e,09}^{-2}\sum_{j\in B_{i09}} w_{ij,09},\ \tau_{e,10}^{-2}\sum_{j\in B_{i10}} w_{ij,10}\right), \tag{A.9}$$

$$\mathbf{r}_{i_1} = \sum_{j\in A_i} \mathbf{D}_{w_{ij}}^{0,5}\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{D}_{w_{ij}}^{0,5}\left(\mathbf{y}_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}\right),$$

et

$$\mathbf{r}_{i_2} = \begin{pmatrix} \sum_{j\in B_{i,09}} w_{ij,09}\left(y_{ij,09} - \mathbf{x}_{ij,09}'\boldsymbol{\beta}_{09}\right)\tau_{e,09}^{-2} \\ \sum_{j\in B_{i,10}} w_{ij,10}\left(y_{ij,10} - \mathbf{x}_{ij,10}'\boldsymbol{\beta}_{10}\right)\tau_{e,10}^{-2} \end{pmatrix}. \tag{A.10}$$

Annexe B

Nous définissons un estimateur des éléments diagonaux de $V\left\{\overline{\mathbf{x}}_{wi10} \mid \overline{\mathbf{x}}_{N_i,10}\right\} := \mathbf{V}_{xxi,10}$ correspondant aux covariables au niveau de l'unité, $\mathbf{x}_{ijk10}$ pour $k = 1, \ldots, p_{10}$. L'estimateur de variance est fondé sur l'hypothèse de travail qu'un échantillon tiré avec probabilité proportionnelle à la taille avec remise (PPTR) est une approximation raisonnable pour le plan de sondage de la Cash Rent Survey. Comme il est discuté dans Cochran (1977), l'utilisation d'une approximation PPTR est souvent raisonnable si la fraction d'échantillonnage est inférieure à 10 %. Supposons que la probabilité de tirage de l'élément j dans le domaine i pour le plan PPTR est $p_{ij} = n_{i10}^{-1}w_{ij,10}^{-1}$. Parce que $n_{i10} \leq 1$ pour certains comtés, nous définissons l'estimateur des éléments diagonaux de $\mathbf{V}_{xxi10}$ correspondant aux covariables au niveau de l'unité sous forme d'une combinaison convexe d'un estimateur direct de la variance intra-domaine et d'un estimateur de variance qui regroupe l'information sur l'ensemble des comtés dans un État. Pour le domaine i avec $n_{i10} > 1$, l'estimation de la variance intra-domaine de x_{ijk10} sous le plan PPTR supposé (Särndal, Swensson et Wretman, 1992) est donnée par

$$S_{ik10}^2 = \frac{n_{i10}^2}{\left(\sum_{j=1}^{n_{i10}} w_{ij,10}\right)^2 (n_{i10} - 1)}\sum_{j=1}^{n_{i10}} w_{ij,10}^2\left(x_{ijk10} - \overline{x}_{wik10}\right)^2,$$

où $\overline{x}_{wik10}$ est le k^{e} élément de $\overline{\mathbf{x}}_{wi10}$. L'estimateur groupé de la variance est défini par

$$S_{pk10}^2 = \frac{1}{w_{..10}^2\left(\tilde{n}_{10} - \tilde{D}_{10}\right)}\sum_{i=1}^{D}\left(n_{i10}^2\sum_{j=1}^{n_{i10}} w_{ij,10}^2\left(x_{ijk10} - \overline{x}_{wik10}\right)^2\right) I\left[n_{i10} > 1\right],$$

où $w_{..10} = \sum_{i=1}^{D}\left(\sum_{j=1}^{n_i} w_{ij,10}\right) I\left[n_i > 1\right]$, $\tilde{n}_{10} = \sum_{i=1}^{D} n_{i10}\, I\left[n_{i10} > 1\right]$, et $\tilde{D}_{10} = \sum_{i=1}^{D} I\left[n_{i10} > 1\right]$. L'élément de la matrice de covariance diagonale $\mathbf{V}_{xxi,10}$ correspondant à la k^{e} covariable au niveau de l'unité est alors donné par

$$\hat{V}\left\{\overline{x}_{wik10}\right\} = n_{i10}^{-1}\,\hat{S}_{ik10}^2\, I\left[n_{i10} \neq 1\right] + n_{i10}^{-1}\, S_{pk10}^2\, I\left[n_{i10} = 1\right], \tag{B.1}$$

où

$$\hat{S}^2_{ik10} = \frac{n_{i10}}{n_{i10}+1} S^2_{ik10} + \frac{1}{n_{i10}+1} S^2_{pk10}. \tag{B.2}$$

Nous donnons, pour la combinaison en (B.2) une justification heuristique qui est liée à Haff (1980). Soit $S^2 = n^{-1}\sum_{i=1}^{n} X_i^2$, où $X_i \sim \mathrm{N}\left(0, \sigma^2\right)$. Supposons que $\sigma^2 \sim$ Gamma inverse(α, β), où $E\left[\sigma^2\right] := v = \beta(\alpha-1)^{-1}$. Alors,

$$E\left[\sigma^2 \,\middle|\, S^2\right] = \frac{2(\alpha-1)v}{n+2(\alpha-1)} + \frac{nS^2}{n+2(\alpha-1)}.$$

Dans l'application à l'estimation des taux de location au comptant au niveau du comté, S^2_{ik10} joue le rôle de S^2 et S^2_{pk10} joue le rôle de v. En prenant $\alpha = 1{,}5$, on obtient le multiplicateur souhaité.

Annexe C

Simulation des données à partir des lois a posteriori

Considérons les échantillons a posteriori pour $\boldsymbol{\beta}_{09}$, $\boldsymbol{\beta}_{10}$, $\boldsymbol{\Sigma}_{\nu\nu}$ et $\boldsymbol{\Sigma}_{ee}$, désignés par $\boldsymbol{\beta}^s_{09}$, $\boldsymbol{\beta}^s_{10}$, $\boldsymbol{\Sigma}^s_{\nu\nu}$ et $\boldsymbol{\Sigma}^s_{ee}$, respectivement, pour $s = 1, \ldots, S$. Définissons

$$\boldsymbol{\Sigma}^s_{eeij} := \mathbf{D}^{-0,5}_{wij} \boldsymbol{\Sigma}^s_{ee} \mathbf{D}^{-0,5}_{wij},$$

pour $s = 1, \ldots, S$. Tirons les répliques ν^r_{i09}, ν^r_{i10}, y^r_{ij09} et y^r_{ij10}, pour $r = 1, \ldots, R$, conformément au modèle (1-3) et aux propriétés de la loi normale conditionnelle multivariée, comme il suit :

$$\begin{aligned}
\nu^r_{i09} &\sim N\left(\mathbf{0}, \boldsymbol{\Sigma}^r_{\nu\nu,(11)}\right) \\
\nu^r_{i10} &\sim N\left(\left(\boldsymbol{\Sigma}^r_{\nu\nu,(11)}\right)^{-1} \boldsymbol{\Sigma}^r_{\nu\nu,(12)} \nu^r_{i09},\ \left(\boldsymbol{\Sigma}^r_{\nu\nu,(11)}\right)^{-1} \boldsymbol{\Sigma}^r_{\nu\nu,(11)} \boldsymbol{\Sigma}^r_{\nu\nu,(22)} - \left(\boldsymbol{\Sigma}^r_{\nu\nu,(12)}\right)^2\right), \\
\mu^r_{i09} &= \mathbf{x}'_{ij09} \boldsymbol{\beta}^r_{09}, \\
y^r_{ij09} &\sim N\left(\mu^r_{i09} + \nu^r_{i09},\ \boldsymbol{\Sigma}^r_{eeij,(11)}\right) \\
\mu^r_{i10} &= \mathbf{x}'_{ij10} \boldsymbol{\beta}^r_{10}, \\
y^r_{ij10} &\sim N\left(\mu^r_{i10} + \nu^r_{i10} + \left(\boldsymbol{\Sigma}^r_{eeij,(11)}\right)^{-1} \boldsymbol{\Sigma}^r_{eeij,(12)} \left(y^r_{ij09} - \mu^r_{i09} - \nu^r_{i09}\right),\right. \\
&\qquad \left. \left(\boldsymbol{\Sigma}^r_{eeij,(11)}\right)^{-1} \left(\boldsymbol{\Sigma}^r_{eeij,(11)} \boldsymbol{\Sigma}^r_{eeij,(22)} - \left(\boldsymbol{\Sigma}^r_{eeij,(12)}\right)^2\right)\right).
\end{aligned}$$

Bien que le nombre d'échantillons a posteriori soit $S = 20\,000$, nous construisons $R = 1\,901$ répliques, où r est sélectionné dans la série 1 000 à T en sautant tous les 10 échantillons.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Berg, E., Cecere, W. et Ghosh, M. (2014). Small area estimation for county-level farmland cash rental rates. *Journal of Survey Statistics and Methodology*, 2, 1-37.

Cochran, W.G. (1977). *Sampling Techniques*. *3rd Edition*, New York: John Wiley & Sons, Inc.

Datta, G.S., Day, B. et Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā: The Indian Journal of Statistics*, 60, 344-362.

Dhuyvetter, D., et Kastens, T. (2009). *Kansas Land Values and Cash Rents at the County Level*.

Dobos, R.R., Sinclair, H.R. et Robotham, M.P. (2012). *User Guide for the National Commodity Crop Productivity Index (NCCPI), Version 2.0*, NRCS/USDA publication.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2009). *Bayesian Data Analysis: Second Edition*, CRC Press.

Gershunskaya, J. (2010). Robust small area estimation using a mixture model. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Ghosh, M., et Steorts, R. (2013). Two-stage Bayesian benchmarking as applied to small area estimation. *TEST*, 22, 670-687.

Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 586-597.

Kim, J.K., Park, S. et Lee, Y. (2017). Statistical inference using generalized linear mixed models under informative cluster sampling. *Canadian Journal of Statistics*, DOI: 10.1002/cjs.11339.

Lohr, S.L., et Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *Canadian Journal of Statistics*, 31, 383-396.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, New York: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Woodard, S., Paulson, N., Baylis, K. et Woddard, J. (2010). Spatial analysis of Illinois agricultural cash rents. *The Selected Works of Kathy Baylis*. http://works.bepress.com/kathy_baylis/29.

You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchicial Bayes small area estimators with applications in census undercoverage estimation. *Statistics in Transition*, 6, 631-640.