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

Techniques d’enquête

Confidentialité différentielle bayésienne pour les petits comtés et les produits individuels

par Balgobin Nandram, Habtamu Benecha et Linda J. Young

Date de diffusion : le 29 juin 2026

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 infostats@statcan.gc.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

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 2026
Vol. 52, nº 1, pp. 67-103
Statistique Canada, nº 12-001-X au catalogue

Confidentialité différentielle bayésienne pour les petits comtés et les produits individuels

Balgobin Nandram, Habtamu Benecha et Linda J. Young[1]

Résumé

Nous construisons une méthode bayésienne hybride, comprenant un mécanisme de confidentialité différentielle, pour masquer les totaux de comtés du recensement pour un État des États-Unis concernant la superficie d'un produit agricole. Nous utilisons des substituts aux données recueillies au niveau des exploitations agricoles dans le cadre d'un précédent recensement de l'agriculture aux États-Unis pour illustrer notre procédure. Nous utilisons deux modèles bayésiens pour petits domaines (modèle paramétrique et modèle de mélange) afin de tenir compte des comtés de petite taille comportant peu d'exploitations ainsi que de certains comtés de grande superficie. Dans ces modèles, la distribution de Laplace fournit un mécanisme de confidentialité différentielle. Au stade du prétraitement, nous incorporons également les poids du recensement pour former la superficie totale observée, un facteur d'échelle du mécanisme de Laplace pour chaque comté, une transformation par la racine carrée de la superficie totale observée pour éviter des estimations masquées négatives, en particulier pour les petits comtés, ainsi que la règle du pourcentage p et la règle du maximum de trois pour répartir les comtés en comtés supprimés, en comtés non sensibles et en comtés sensibles. En raison des difficultés liées à la spécification et à l'ajustement du budget de confidentialité (un paramètre inconnu), afin d'assurer un équilibre entre la sécurité et l'utilité, nous précisons une distribution *a priori* pour le budget de confidentialité, où les valeurs ne sont pas précisées, et utilisons l'échantillonneur de Gibbs pour ajuster les modèles hiérarchiques bayésiens. Au stade du post-traitement, nous recourons à l'inférence prédictive bayésienne pour obtenir la superficie masquée des comtés, ce qui comprend un étalonnage afin que le total masqué de l'État corresponde au total observé de l'État. Comme mesure de fiabilité de la procédure bayésienne, nous utilisons les coefficients de variation *a posteriori* pour les moyennes *a posteriori* masquées des comtés. Comme mesure d'utilité, nous utilisons les erreurs relatives absolues pour chaque comté, ainsi que d'autres mesures globales. Pour les comtés sensibles, on observe certaines différences entre les deux modèles pour petits domaines, mais les deux sont nettement supérieurs à un modèle pour domaine individuel; le modèle de mélange constitue le meilleur compromis entre la sécurité et l'utilité.

Mots-clés : Distribution de Laplace; étalonnage; inférence prédictive bayésienne; modèle de mélange avec poids « stick-breaking »; post-traitement; règle du pourcentage p; utilité.

1. Introduction

Il est probable que la confidentialité différentielle devienne importante pour les organismes gouvernementaux. Le concept de confidentialité différentielle a été utilisé dans le recensement de 2020 par le Bureau du recensement des États-Unis, mais plusieurs défis subsistent quant à son incidence sur la qualité et l'utilisabilité des données. Toutefois, on reconnaît que la confidentialité différentielle ϵ est le modèle *officiel* de pointe pour la diffusion de renseignements de nature délicate tout en protégeant la confidentialité (par exemple Williams et Bowen, 2023). Drechsler (2023) a fait remarquer que malgré l'enthousiasme du milieu universitaire et de l'industrie, la plupart des organismes, à l'exception notable du Bureau du recensement des États-Unis, ont été réticents à même envisager le concept pour leur stratégie de diffusion de données. Voir également la discussion de Reiter (2019) sur la confidentialité différentielle et les diffusions de données fédérales. Cela laisse entendre qu'il reste beaucoup de travail à faire pour rendre la confidentialité différentielle plus pratique et accessible. Pourtant, il s'agit d'un problème mathématique difficile à résoudre et

1. Balgobin Nandram, Habtamu Benecha et Linda J. Young, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. Courriels : balnan@wpi.edu; habtamu.benecha@usda.gov et linda@youngstatconsulting.com.

son adaptation au domaine statistique constitue un champ de recherche actif; voir la formulation originale et les développements en cryptographie dans Dwork (2006), Dwork, McSherry, Nissim et Smith (2006) et Dwork, Naor, Reingold, Rothblum et Vadhan (2009). Nous fournissons une méthode bayésienne hybride, comprenant un mécanisme de confidentialité différentielle, pour masquer les totaux de comtés du recensement pour un État des États-Unis concernant la superficie d'un produit agricole.

Le fonctionnement des mécanismes de confidentialité différentielle se résume à l'introduction de bruit (ou une composante aléatoire) soigneusement calé dans les données confidentielles observées, afin de protéger la vie privée des répondants individuels (par exemple les agriculteurs). L'utilité désigne la valeur informative des données bruitées (ou des résultats statistiques) après l'application de tout mécanisme de préservation de la vie privée. L'utilité et la confidentialité sont liées par un compromis fondamental et doivent être soigneusement équilibrées. Nous souhaitons que la différence relative entre les données confidentielles observées et les données masquées soit faible, tout en veillant strictement à ne pas révéler l'identité des agriculteurs.

De nombreux organismes gouvernementaux, y compris le National Agricultural Statistics Service (NASS) du département de l'Agriculture des États-Unis (USDA), ont recours à la suppression de cellules (primaires et secondaires) pour protéger la confidentialité des réponses (voir Cox, 1980, 1995, et Cox, Kelly et Patil, 2004). Cette approche peut être utilisée pour protéger à la fois les données de fréquence et les données continues. Par comparaison, la méthode de contrôle de la divulgation que nous proposons consiste à appliquer du bruit aléatoire aux cellules sensibles, provenant possiblement d'un mécanisme de confidentialité différentielle, afin que les tableaux publiés puissent comporter des valeurs pour toutes les cellules. Nous analysons des superficies récoltées simulées (une variable économique); d'autres variables peuvent être étudiées de la même façon ou être liées aux superficies récoltées.

Duncan et Lambert (1986, 1989), même si leurs travaux ne sont pas entièrement liés au nôtre, ont élaboré un cadre décisionnel théorique d'évaluation du risque qui tient compte des objectifs et de la stratégie d'un intrus visant à compromettre une base de données, et abordé l'information que l'intrus obtient. Ils ont décrit deux types de divulgation de microdonnées, soit la divulgation distincte de l'identité d'un répondant et la divulgation d'attributs d'un répondant à la suite d'une identification non autorisée. Ils ont également présenté une formule pour estimer le risque de divulgation de l'identité et ont évalué une approximation simple.

Rubin (1993) a proposé de diffuser uniquement des microdonnées synthétiques, construites au moyen de l'imputation multiple. L'objectif est de produire des microdonnées pouvant être analysées à l'aide d'un logiciel statistique standard. Cette méthode est suivie de près par Raghunathan, Reiter et Rubin (2003). La procédure de base est de simuler de multiples copies de la population à partir de laquelle les répondants ont été sélectionnés, puis de diffuser un échantillon aléatoire de chacune de ces populations synthétiques. Les utilisateurs peuvent analyser les ensembles de données des échantillons synthétiques à l'aide de logiciels standard pour données complètes conçus pour des échantillons aléatoires simples, puis obtenir des inférences valides en combinant les estimations ponctuelles et celles de la variance à l'aide des méthodes

proposées; voir Little (1993). Plusieurs copies de l'ensemble de données original peuvent également être générées à l'aide de la confidentialité différentielle. Nous ne pouvons pas diffuser de microdonnées (par exemple des données au niveau des exploitations agricoles), mais nous pouvons diffuser des données au niveau du comté après avoir effectué un contrôle statistique de la divulgation.

Bowen et Liu (2020) ont examiné les techniques de synthèse de données incorporant la confidentialité différentielle qui sont actuellement utilisées pour diffuser des données de substitution au niveau individuel plutôt que les données originales. Claire Bowen et Fang Liu ont comparé ces techniques sur le plan conceptuel et évalué l'utilité statistique ainsi que les propriétés inférentielles des données synthétiques issues de chaque technique au moyen d'études par simulation approfondies. Toutefois, aucune discussion n'est présentée dans le cadre bayésien; leur examen porte uniquement sur les techniques non bayésiennes actuelles. En revanche, Hu et Bowen (2024) décrivent des méthodes bayésiennes à la section 3 de leur article et comparent des modèles bayésiens et non bayésiens ainsi que des modèles paramétriques et non paramétriques (y compris le processus de Dirichlet); elles n'abordent cependant pas les méthodes bayésiennes en confidentialité différentielle. Essentiellement, elles ont utilisé des modèles bayésiens classiques (par exemple des modèles de régression) qui reflètent les principales caractéristiques des données confidentielles et ont eu recours à l'inférence prédictive *a posteriori* à des fins de masquage. La distribution de Laplace, un mécanisme de randomisation, est une composante importante de nos modèles bayésiens (c'est-à-dire que la confidentialité différentielle est utilisée pour former une méthode hybride).

Les méthodes bayésiennes en confidentialité différentielle évoluent rapidement. Par exemple, Zhang, Rubinstein et Dimitrakakis (2016) ont étudié la façon de communiquer les résultats d'une inférence bayésienne à des tiers, tout en préservant la solide garantie de confidentialité différentielle. Leurs principales contributions portent sur quatre algorithmes distincts d'inférence bayésienne confidentielle au moyen de modèles graphiques probabilistes. La distribution de Laplace est l'un des mécanismes qu'ils ont utilisés pour étudier la confidentialité différentielle. Ils ont également examiné des données binaires, mais ont utilisé une distribution de Laplace non négative, tronquée à l'extrémité supérieure à la taille de l'échantillon. L'inférence différentiellement confidentielle pour des données binomiales est utile en soi; voir Awan et Slavkovic (2019). Hu, Williams et Savitsky (2025) ont élaboré un synthétiseur pseudobayésien et abordé le compromis entre l'utilité et le risque. Un synthétiseur entièrement bayésien est difficile à obtenir dans leur cadre. Au Bureau du recensement des États-Unis (données de fréquence), Janicki, Holan, Irimata, Livsey et Raim (2024) décrivent une méthode bayésienne permettant de produire un ensemble de mesures imprécises pour des données tabulaires de fréquence, puis effectuent le post-traitement (leur principale contribution) de ces mesures imprécises à l'aide d'un modèle par inférence *a posteriori*. Cette méthode intègre des contraintes connues pour les dénombrements et les ratios, et les auteurs démontrent une amélioration de l'inférence des données partiellement débruitées. Ils disposent de données confidentielles, Y, et souhaitent diffuser publiquement des données masquées. Ils ont construit $Z = Y + E$, où E est le bruit (analyse standard non bayésienne), puis ont utilisé le modèle bayésien, $Z \mid Y, \boldsymbol{\theta}$, pour ajouter les contraintes à Y et mettre à jour Y.

Rinott, O'Keefe, Shlomo et Skinner (2018) ont examiné la confidentialité différentielle lors de la diffusion de tableaux de fréquences. Ils ont étudié la protection de la confidentialité pour des tableaux de fréquences perturbés, y compris le compromis avec l'utilité analytique. Toutefois, notre étude ne porte pas sur les tableaux de fréquences; notre contribution consiste à élaborer une méthodologie de confidentialité différentielle pour le masquage d'un produit agricole, dans l'objectif de proposer une méthodologie plus étendue applicable à des tableaux couplés et hiérarchiques pour le recensement de l'agriculture des États-Unis.

Nous soulignons qu'il existe d'autres façons de masquer les superficies récoltées, déjà obtenues dans le cadre du recensement de l'agriculture, aux fins de diffusion publique. Par exemple, un organisme gouvernemental peut avoir recours à la convolution plutôt qu'à la confidentialité différentielle. On peut multiplier les superficies récoltées par $U \sim$ Uniforme$(1-a, 1+a)$, où a est proche de 0, disons $a = 0{,}10$, en choisissant a de manière à obtenir un niveau raisonnable de sécurité et d'utilité; voir Evans, Zayata et Slanta (1996). Cette idée est liée à la réponse aléatoire (un mécanisme de confidentialité différentielle) dans les enquêtes-échantillons, lorsqu'une variable devant être protégée est convoluée (par exemple multipliée) avec une autre variable au moyen d'un mécanisme aléatoire. Voir les travaux pionniers sur la réponse aléatoire appliquée aux variables continues en échantillonnage d'enquête de Poole (1974); voir également l'examen de Nandram et Yu (2019) portant sur les données discrètes. Encore une fois, ce n'est pas l'orientation de notre étude. L'approche standard consiste à utiliser la suppression de cellules, mais il est souhaitable d'examiner les caractéristiques utiles de la confidentialité différentielle, même si elle est difficile à mettre en œuvre dans des modèles bayésiens. Notre méthodologie fournit un cadre général applicable à de nombreux problèmes comportant des structures de données plus complexes (par exemple des tableaux multidirectionnels, des tableaux couplés et des tableaux hiérarchiques); bien sûr, les modèles varieront selon les cas.

Nous soulignons que les superficies de chaque exploitation agricole sont pondérées à l'aide des poids du recensement. Ces poids sont estimés pour chaque exploitation agricole répondante du recensement au moyen d'une méthode d'estimation de système dual pour tenir compte de la sous-couverture dans la base liste du recensement ainsi que de la non-réponse et de la classification erronée des exploitations agricoles. Après le calage, les poids sont additionnés pour les comtés ou les États afin d'obtenir le nombre d'exploitations agricoles à ce niveau géographique. La somme pondérée des valeurs déclarées pour un produit agricole d'une zone géographique d'intérêt donne l'estimation totale de ce produit. (Il convient de mentionner que l'USDA définit une exploitation agricole comme tout lieu ayant produit ou vendu, ou qui aurait normalement vendu, pour au moins 1 000 $ de produits agricoles au cours d'une année.)

Des données simulées sont utilisées pour toutes les applications figurant dans la présente étude. Les superficies récoltées pour un produit agricole sont simulées pour un État hypothétique comprenant 75 comtés; chaque comté compte un certain nombre d'exploitations agricoles, qui sont donc des microdonnées. Nous désignerons le produit agricole par « PRO » et le tableau confidentiel original (des superficies par comté et État) par « TAB ». Dans ce cas-ci, le TAB comporte en tout 76 cellules, y compris le total de l'État; le total de l'État n'est pas confidentiel. L'objectif est de protéger les cellules du tableau en ajoutant du bruit approprié aux valeurs des cellules.

Dans le présent article, nous examinons le PRO pour l'ensemble des comtés de l'État. Soit n_i, le nombre d'exploitations agricoles dans le i^{e} comté. Soit (w_{ij}, y_{ij}^*), $j = 1,\ldots,n_i, i = 1,\ldots,\ell$, les poids du recensement et les superficies récoltées observées associés à la j^{e} ferme dans le i^{e} comté. Puis, $y_{ij} = w_{ij} y_{ij}^*$ représente les superficies récoltées du recensement associées aux w_{ij} exploitations représentées par la j^{e} exploitation observée dans le i^{e} comté. Même si nous n'utilisons pas de covariables, nous décrivons la façon de les intégrer dans la section de conclusion. Dans une étude ultérieure, nous examinerons des tableaux à deux entrées (comté par produit) comprenant plusieurs comtés et produits agricoles (et non un seul) en superficies.

Nous utilisons le mécanisme de Laplace afin d'assurer la confidentialité différentielle du TAB, en combinaison avec des hypothèses supplémentaires dans les modèles hiérarchiques bayésiens afin de refléter les principales caractéristiques des données. Grâce à l'utilisation de la confidentialité différentielle et avec un petit budget, ϵ, il peut être possible de diffuser les données en toute sécurité; cela dépendra de l'organisme. Le principal problème est que l'utilité des données pourrait ne pas être acceptable. Par conséquent, notre méthode est une procédure bayésienne hybride parce qu'elle ne repose pas uniquement sur la confidentialité différentielle. Nous utilisons des modèles pour petits domaines, car les exploitations agricoles d'un comté, et les comtés d'un État, peuvent avoir des pratiques semblables. De plus, certains comtés peuvent comporter peu de données, ce qui peut devenir problématique. À titre d'objectif secondaire, nous montrerons l'utilité des modèles pour petits domaines en les comparant à un modèle pour domaine unique (c'est-à-dire pour comté individuel).

Nous supposons un modèle de population (sans bruit), et un modèle d'échantillon (avec bruit de Laplace) en découle. Le modèle de population est ajusté par l'ajout de bruit de Laplace, et la prédiction est effectuée à l'aide du modèle de population une fois le modèle d'échantillon ajusté. Nous utilisons un modèle de base pour motiver nos deux modèles hiérarchiques bayésiens, qui sont entièrement décrits à la section 3. Le modèle de base est

$$y_{ij} \mid \mu_i, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}(\mu_i, \sigma^2), j = 1,\ldots,n_i, i = 1,\ldots,\ell.$$

Ce modèle est utilisé pour prédire les y_{ij}, afin d'obtenir les totaux de comtés après l'ajout de bruit aux μ_i et aux σ^2 à l'aide d'un mécanisme de confidentialité différentielle. Malgré le manque de robustesse du modèle à la non-normalité, l'effet de mélange de la distribution de Laplace le rendra, dans une certaine mesure, robuste à la non-normalité dans le modèle d'échantillon.

Le modèle d'échantillon est constitué en ajustant notre modèle de population pour tenir compte des données confidentielles observées comme suit :

$$y_{ij} \mid \mu_i, r_i, t_i, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left\{\mu_i + (2r_i - 1)\, t_i, \sigma^2\right\}, j = 1,\ldots,n_i, i = 1,\ldots,\ell,$$

où $(2r_i - 1)\, t_i$ est une représentation pratique de la variable aléatoire de Laplace, conçue de façon probabiliste comme suit :

$$r_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), \quad t_i \mid \epsilon \overset{\text{ind}}{\sim} \text{Gamma}(1, \epsilon)$$

(ce qui est équivalent à la densité exponentielle), et ϵ est le budget de confidentialité pour lequel $(2r_i - 1)\, t_i$ a une moyenne nulle; voir l'annexe A pour obtenir la définition de la confidentialité différentielle. Il convient de prendre note que $f(t_i \mid \epsilon) = \epsilon e^{-\epsilon t_i}$, $t_i > 0$ avec $E(t_i) = 1/\epsilon =$ écart-type(t_i). Ainsi, plus ϵ est petit, plus la variabilité est grande, ce qui entraîne un risque plus faible (une sécurité accrue) et une utilité moindre. Lorsque ϵ est choisi pour être petit (par exemple $\epsilon = 1, 2$), la procédure de confidentialité différentielle vise à assurer un faible risque, mais pas nécessairement une grande utilité. Le compromis entre le risque et l'utilité ne peut être évité.

Une robustesse supplémentaire à la non-normalité est obtenue lorsque des distributions *a priori* sont spécifiées pour les μ_i, les σ^2 et le ϵ; voir la distribution *a priori* pour ϵ à la section 2 et les deux modèles hiérarchiques bayésiens à la section 3. Les modèles hiérarchiques bayésiens, en particulier le modèle de mélange, introduisent une robustesse par effet de mélange.

Il convient de prendre note que le modèle est ajusté aux microdonnées (données au niveau des exploitations agricoles) pour assurer une protection robuste des renseignements au niveau individuel. La confidentialité différentielle est utilisée pour ajouter simultanément du bruit à tous les comtés sensibles, et non un à la fois. Autrement dit, l'ensemble de données comprend toutes les exploitations agricoles de tous les comtés. La définition de la confidentialité différentielle s'applique à l'ensemble des comtés pertinents en même temps. Nous élaborons ainsi une méthode hybride novatrice qui intègre un mécanisme de confidentialité différentielle dans un modèle bayésien.

Il convient de souligner que la fonction de densité de probabilité conjointe de toutes les données observées, $\mathbf{y}$, est

$$f(\mathbf{y} \mid \boldsymbol{\mu}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell} n_i \left\{\tilde{s}_i^2 + \left(\overline{y}_i - \mu_i - (2r_i - 1)\, t_i\right)^2\right\}\right\},$$

où $\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $\tilde{s}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$, $i = 1, \ldots, \ell$. Il est bon que $f(\mathbf{y} \mid \boldsymbol{\mu}, \sigma^2)$ ne soit pas une fonction des données individuelles au niveau des exploitations agricoles; cela est clairement vrai pour toute densité *a posteriori* fondée sur cette vraisemblance. Nous soulignons également qu'il ne s'agit pas directement d'une fonction des totaux de comtés (c'est-à-dire qu'il s'agit d'une fonction de $\tilde{s}_i^2$ et de $(\overline{y}_i - \mu_i - (2r_i - 1)\, t_i)^2$). Dans la densité *a posteriori* conjointe, qui contient tous les renseignements, $\tilde{s}_i^2$ et $\overline{y}_i$ demeurent fixes.

Nous utilisons un modèle pour petits domaines et la transformation par la racine carrée pour la superficie des exploitations agricoles. Ce modèle permet aux données masquées d'être positives après la rétrotransformation, ainsi la contrainte de positivité est satisfaite. Il est possible d'utiliser des contraintes d'inégalité (par exemple voir Nandram, Cruze et Erciulescu, 2023), mais il est plus simple d'utiliser la transformation par la racine carrée, qui contribue également à la robustesse à la non-normalité. Notre procédure de masquage comprend les étapes suivantes :

a. Le modèle d'échantillon est ajusté à la racine carrée des données observées pour maintenir la contrainte de positivité. Du bruit est ajouté à ce modèle au moyen du mécanisme de confidentialité différentielle de Laplace.

b. Nous utilisons l'inférence prédictive bayésienne pour obtenir les données masquées. Une fois le modèle d'échantillon ajusté, malgré plusieurs options possibles, l'inférence prédictive bayésienne est effectuée à l'aide du modèle de population lors du post-traitement.

c. Dans le cadre d'une analyse des données de sortie, les données masquées sont étalonnées en fonction des données observées au moyen de la méthode itérative du quotient. Autrement dit, les données masquées des comtés sont étalonnées de sorte que le total masqué au niveau de l'État soit identique au total observé au niveau de l'État. Cela se fait à l'étape du post-traitement. Il convient de mentionner que le total observé au niveau de l'État ne fait pas partie de la procédure de masquage (ajustement du modèle).

Cette procédure produira des données masquées ayant une précision relativement accrue par rapport à l'inférence prédictive bayésienne à l'étape b parce que l'étalonnage à l'étape c augmente la précision. Cette procédure d'étalonnage n'aura pas d'incidence sur la confidentialité, car les superficies agricoles observées ne sont pas utilisées dans l'analyse (Dwork et Roth, 2014, proposition 2.1). Williams et Bowen (2023) ont indiqué que le post-traitement offrait des possibilités d'amélioration de l'utilité, car les connaissances publiques ou spécialisées disponibles peuvent être utilisées pour réduire la quantité de bruit sans entraîner de perte supplémentaire de confidentialité (théorème du post-traitement). Ils ont également souligné qu'il conviendrait d'intégrer davantage de renseignements dans une analyse de confidentialité différentielle, mais, à notre connaissance comme à la leur, de nombreux chercheurs ne le font pas.

Il convient de rappeler et il est important de noter que les résultats et les démonstrations présentés dans cet article sont fondés sur des données simulées pour reproduire la structure et les caractéristiques du Recensement de l'agriculture. Cet ensemble de données simulées a été généré exclusivement pour illustration méthodologique et ne contient aucunes microdonnées confidentielles ou réelles. L'approche et les découvertes présentées ici ont pour objet de mettre en avant la faisabilité et la performance des méthodes proposées dans un environnement réaliste mais entièrement simulé.

La présente étude comprend cinq sections, dont la présente section d'introduction, ainsi que trois annexes contenant certains renseignements techniques pertinents. À la section 2, nous présentons des renseignements généraux. Plus précisément, nous avons inclus un facteur d'échelle dans la distribution de Laplace, une caractéristique qui fait en sorte que le budget global du modèle est ϵ, conformément à ce qui est normalement précisé dans la définition de la confidentialité différentielle (Williams et Bowen, 2023), ainsi qu'une évaluation des comtés sensibles accompagnée de discussions sur la sensibilité locale et globale. À la section 3, nous décrivons deux modèles : un modèle paramétrique et un modèle de mélange robuste (c'est-à-dire un modèle de mélange avec des poids « stick-breaking ») pour tenir compte des grandes superficies (valeurs aberrantes) et de la non-normalité. À la section 4, nous présentons une comparaison approfondie des deux modèles; les deux modèles sont également comparés à un modèle pour domaine individuel. De plus, nous présentons partiellement des données masquées pour le TAB de l'État. À la section 5, nous formulons des remarques finales. L'annexe A fournit une brève description de la notion de confidentialité différentielle ϵ, l'annexe B décrit la règle de sensibilité du pourcentage p et l'annexe C présente un modèle applicable aux pourcentages afin d'illustrer une plus grande généralité.

2. Renseignements généraux

Nous examinons le budget total; le nombre d'exploitations agricoles par comté et un facteur d'échelle nécessaire à la distribution de Laplace ont une incidence sur le budget total. Nous traitons également de l'approche prédictive visant à masquer les données pondérées du recensement et de la transformation par la racine carrée visant à fournir des données masquées positives au niveau des comtés.

En utilisant le mécanisme de confidentialité différentielle de Laplace, nous appliquons un système de pondération élémentaire pour obtenir le budget total (voir l'annexe A) parce que la composition séquentielle est inappropriée et que la composition parallèle n'est pas disponible pour l'estimation sur petits domaines. Nous estimons qu'une seule requête est utilisée, donc la composition séquentielle est inappropriée. Nous estimons qu'il y a des similarités au sein des comtés pour l'estimation sur petits domaines, de sorte que les comtés ne sont pas réellement sans chevauchement. Par conséquent, la composition parallèle est aussi inappropriée. Williams et Bowen (2023) ont présenté une bonne discussion sur les deux théorèmes de composition utilisés en confidentialité différentielle. Soit n_i le nombre d'exploitations agricoles dans le i^{e} comté, $i = 1,\ldots,\ell$. Afin d'aller de l'avant, nous appliquons le système de pondération simple pour obtenir le budget global, ce qui permet aux comtés de plus grande taille de bénéficier d'une part plus grande du budget. Parce que nous voulons que le budget global soit ϵ, $a_o < \epsilon < b_o$, pour $i = 1,\ldots,\ell$, nous définissons $a_{oi} = \frac{n_i}{\sum_{i'=1}^{\ell} n_{i'}}$, et donc nous assumons que le budget global est $\sum_{i=1}^{\ell} a_{oi}\epsilon = \epsilon$.

Nous examinons maintenant la difficulté liée à la mise à l'échelle, qui découle de la méthode bayésienne de confidentialité différentielle. Il convient d'observer que nous ajoutons simplement $(2r_i - 1)\,t_i$ au total moyen des comtés et qu'il y a un problème d'échelle avec t_i; par exemple, voir Janicki et coll. (2024). L'échelle a une valeur de un pour tous les comtés et cela n'est pas suffisamment souple, mais les superficies totales pour les comtés sont variables; l'échelle et ϵ influencent la sécurité et l'utilité. L'une des façons de surmonter ce problème est d'introduire un facteur d'échelle, t_{oi},

$$t_i \mid \epsilon \overset{\text{ind}}{\sim} \text{Gamma}\left\{1, \frac{\epsilon}{t_{oi}}\right\}, i = 1,\ldots,\ell,$$

où $\text{E}(t_i) = \frac{t_{oi}}{\epsilon} =$ écart-type (t_i) et t_{oi} sont spécifiés. De toute évidence, t_{oi} ne sont pas déterminables et, par conséquent, ϵ ne l'est pas non plus. Il faudrait une autre source de données pour préciser ces paramètres d'échelle et peut-être que des données antérieures, s'il y en a, pourraient être utilisées pour respecter le paradigme bayésien.

Pour des raisons de commodité, nous utilisons les données pondérées du recensement, $y_{ij}, j = 1,\ldots,n_i$, $i = 1,\ldots,\ell$, pour obtenir t_{oi} et nous supposons que nous les connaissons. Nous utilisons l'écart absolu médian, robuste aux valeurs aberrantes et à la normalité. En posant $\tilde{y}_i =$ méd $\{y_{ij}, j = 1,\ldots,n_i\}$, nous utilisons

$$t_{oi} = 1{,}4826 \text{ méd } \{|\, y_{ij} - \tilde{y}_i \,|, j = 1,\ldots,n_i\}, i = 1,\ldots,\ell,$$

un estimateur convergent de l'écart-type. Pour les t_{oi} nuls, il est simple d'apporter un ajustement. D'autres estimateurs (par exemple étendue, écart-type échantillonnal) sont clairement possibles, mais ils sont fortement influencés par les valeurs aberrantes; l'écart interquartile constitue un concurrent. Ainsi, pour tenir compte du budget global de ϵ, nous devons intégrer à la fois les tailles d'échantillon et les facteurs d'échelle. Ces deux éléments influencent le risque et l'utilité et doivent donc être intégrés simultanément au modèle. Une possibilité consiste alors à prendre

$$a_{oi} = \frac{n_i / t_{oi}}{\sum_{i'=1}^{\ell} n_{i'} / t_{oi'}}, \ i = 1, \ldots, \ell,$$

où $\sum_{i=1}^{\ell} a_{oi} = 1$. Cela signifie que nous devons mettre à jour notre hypothèse comme suit :

$$t_i \mid \epsilon \overset{\text{ind}}{\sim} \text{Gamma}\{1, a_{oi}\epsilon\}, i = 1, \ldots, \ell.$$

Toutefois, veuillez prendre note que si les facteurs d'échelle, $t_{oi}, i = 1, \ldots, \ell$, sont les mêmes, ils peuvent être retirés de a_{oi} pour donner $a_{oi} = \frac{n_i}{\sum_{i'=1}^{\ell} n_{i'}}, i = 1, \ldots, \ell$ (c'est-à-dire que les facteurs d'échelle n'ont pas d'importance).

Les données dont nous disposons réellement sont $y_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, \ell$, où n_i est le nombre d'exploitations agricoles pour le PRO dans chaque comté et $y_{ij}, i = 1, \ldots, \ell$ désignent les superficies récoltées (pondérées) dans ces exploitations. Il s'agit de microdonnées. Nous pouvons transformer y_{ij} (les superficies sont positives) de sorte que, lors de la rétrotransformation, nous nous situons sur une échelle positive (par exemple $s = \sqrt{y}$ puis $y = s^2$ est positif). Nous n'utilisons pas la transformation logarithmique parce que lorsque le logarithme est transformé de nouveau, la distribution prédictive bayésienne n'est pas bien définie; voir par exemple Manandhar et Nandram (2021, 2025), qui ont utilisé des distributions gamma généralisées plutôt que la transformation logarithmique. Bien sûr, d'autres transformations menant à la positivité sur l'échelle originale sont possibles.

La bonne nouvelle est que, pour de petites valeurs de ϵ, le modèle de confidentialité différentielle est fiable et le risque est acceptable. Malheureusement, cela dépend du choix de ϵ. En général, l'utilité est faible pour tout modèle de confidentialité différentielle, et un post-traitement est nécessaire; voir la discussion de Williams et Bowen (2023) sur le théorème de post-traitement. Comme il est indiqué, il n'est pas nécessaire de préciser une valeur particulière de ϵ, et une simple distribution *a priori* suffit. Il s'agit en effet d'une nouveauté du paradigme bayésien, et une étude de sensibilité quant au choix de ϵ n'est pas requise. Un modèle de mélange est souple et peut être configuré pour traiter les valeurs aberrantes. Nous tenons à souligner que Lee et Clifton (2011) ont utilisé les données observées pour calculer une borne supérieure pour ϵ pour exprimer l'incertitude concernant ϵ, mais nous devons préciser la probabilité que les renseignements d'un agriculteur soient divulgués. À la suite de Lee et Clifton (2011), Hsu, Gaboardi, Haeberlen, Khanna, Narayan, Pierce et Roth (2014) ont obtenu un intervalle pour ϵ (bornes inférieure et supérieure) fondé sur

un modèle de menace simple dans le cadre duquel des variables économiques sont utilisées plutôt que la variable à l'étude réelle; toutefois, ces variables économiques ne sont pas appropriées dans notre application.

Nous avons calculé la sensibilité locale des données observées (totaux de comtés), qui repose uniquement sur les totaux de comtés observés. Elle correspond simplement à l'étendue des données, soit 3 498. Il convient de mentionner que les trois quartiles des données sont 60,5, 241,4 et 671,4, et que l'écart interquartile est de 610,9 (c'est-à-dire que la sensibilité locale est de 5,73 = 3 498 / 610,9 unités standard). Si les valeurs minimales et maximales de tous les ensembles de données étaient identiques, il s'agirait de la sensibilité globale, mais ce n'est pas le cas; même si l'on peut fixer le minimum à zéro, la borne supérieure est inconnue. Toutefois, la sensibilité globale ne peut pas être infinie simplement parce que les superficies doivent être bornées. Par conséquent, la sensibilité globale doit être d'au moins (ou d'environ) 3 498 (manifestement pas infinie), et de nombreux comtés existent et sont individuellement sensibles. Nous remarquons que les totaux de comtés sont beaucoup plus variables que les moyennes de comtés; si l'on utilisait les moyennes de comtés, la sensibilité locale/globale serait nettement plus faible. Il est possible d'obtenir la sensibilité globale à l'aide d'une analyse fondée sur un modèle, mais cela n'est pas nécessaire.

De toute évidence, tout comté comptant une ou deux exploitations agricoles est sensible, et il est possible qu'un comté comptant trois exploitations agricoles le soit également; c'est la raison pour laquelle certains organismes gouvernementaux appliquent la règle du maximum de trois, mais il existe aussi d'autres règles de sensibilité; voir l'examen de Bring et Wang (2014). Nous devons déterminer les comtés comptant plus de trois exploitations agricoles qui sont sensibles, en utilisant la règle de sensibilité du pourcentage p. Dans ce cas-ci, nous effectuons une analyse des données fondée sur la règle de sensibilité du pourcentage p, et $p = 10$ pour illustrer notre méthodologie. Nous définissons la règle de sensibilité du pourcentage p à l'annexe B.

Enfin, sur les conseils du NASS, nous divisons les comtés en trois ensembles. Le premier ensemble comprend tous les comtés qui satisfont à la règle du maximum de trois, selon laquelle tous les comtés comptant au plus trois exploitations agricoles sont supprimés. Cet ensemble de comtés supprimé est exclu de toute modélisation. Le deuxième ensemble comprend tous les comtés non sensibles; aucun bruit n'est ajouté à ces comtés. Le troisième ensemble comprend les comtés sensibles, auxquels du bruit est ajouté au moyen de la confidentialité différentielle. Les deuxième et troisième ensembles comprennent les comtés non supprimés, et la partition de ces comtés est déterminée par la règle de sensibilité du pourcentage p; les modèles pour petits domaines sont appliqués à ces comtés non supprimés, mais, comme il est mentionné ci-dessus, aucun bruit n'est ajouté aux observations provenant des comtés non sensibles (non supprimés).

Sur les 75 comtés, on compte 17 comtés supprimés (premier ensemble) satisfaisant à la règle du maximum de trois; 38 comtés non sensibles (deuxième ensemble) et 20 comtés sensibles (troisième ensemble) satisfont à la règle du pourcentage p. Il est important de prendre note que seuls les comtés sensibles sont étalonnés en fonction du total des superficies observées pour l'ensemble de ces comtés. Même si ce total agrégé devrait être sensible (non divulguable), la règle du pourcentage p montre qu'il ne l'est pas. Pour ce

faire, on imagine que toutes les exploitations agricoles de ces comtés sensibles proviennent d'un seul grand comté (c'est-à-dire une seule sous-population). L'étalonnage de tous les comtés non supprimés en fonction du total observé des comtés non supprimés est problématique, parce que les comtés non sensibles ont tendance à être plus grands que les comtés sensibles, ce qui crée un biais.

3. Modèles bayésiens

Nous construisons deux modèles bayésiens et avons utilisé l'inférence prédictive pour produire les données masquées finales. À la section 3.1, nous utilisons la version bayésienne du modèle à effets aléatoires à un facteur et un mécanisme de confidentialité différentielle de Laplace. Il s'agit du modèle paramétrique présenté dans Nandram, Toto et Choi (2011) pour les analyses statistiques, et non pour la confidentialité différentielle. À la section 3.2, pour tenir compte des valeurs aberrantes, nous utilisons un modèle de mélange comportant un nombre aléatoire de composantes et des poids aléatoires « stick-breaking » (c'est-à-dire une distribution de Dirichlet généralisée); voir l'examen complet sur les distributions *a priori* « stick-breaking » d'Ishwaran et James (2001). Il s'agit du modèle de mélange, qui devrait être robuste aux valeurs aberrantes relatives à la superficie des exploitations agricoles, et à la non-normalité. À la section 3.3, nous abordons une caractéristique supplémentaire simple de nos modèles visant à atténuer le rétrécissement excessif, ainsi que la façon d'effectuer l'inférence prédictive bayésienne pour fournir les données masquées. Il convient de prendre note que nos modèles ne sont pas simples, puisqu'ils comportent des paramètres déterminés de manière imprécise.

3.1 Modèle paramétrique

Dans notre modèle d'échantillon, nous supposons

$$y_{ij} \overset{\text{ind}}{\sim} \text{Normal}\left\{\mu_i + (2r_i - 1)\, t_i, \sigma^2\right\},\ j = 1, \ldots, n_i,$$

$$r_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right),\ t_i \overset{\text{ind}}{\sim} \text{Gamma}\left(1, a_{oi}\epsilon\right),\ \mu_i \overset{\text{ind}}{\sim} \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), i = 1, \ldots, \ell,$$

où nous préférons utiliser Gamma $\left(1, a_{oi}\epsilon\right)$ plutôt que Exponentielle $\left(a_{oi}\epsilon\right)$, comme cela se fait habituellement (elles sont équivalentes), et les facteurs d'échelle, a_{oi}, sont spécifiés; voir la section 2. Veuillez prendre note que les μ_i représentent les effets au niveau des comtés, que les $(2r_i - 1)\, t_i$ correspondent au bruit de Laplace ajouté aux effets au niveau des comtés et que nous n'utilisons pas d'effets au niveau des exploitations agricoles, parce que ceux-ci ne seraient pas déterminables. *A priori*,

$$\epsilon \sim \text{Gamma}\,(1,1), a_o < \epsilon < b_o;\ \pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2},$$

où (a_o, b_o) sont également précisés pour saisir les valeurs de ϵ ($\epsilon = 1, 2$), les valeurs habituellement utilisées dans la définition de la confidentialité différentielle; des valeurs plus petites sont préférables pour offrir une sécurité accrue, mais au détriment de l'utilité. Nous avons retenu $a_o = 0$, $b_o = 2{,}25$; a_o ne peut pas être inférieur à 0 et b_o constitue une borne supérieure. Ces spécifications assurent un niveau élevé de sécurité et, par conséquent, une faible utilité; un mécanisme de compromis entre l'utilité et le risque sera construit ultérieurement.

Soit $n = \sum_{i=1}^{\ell} n_i$, le nombre total d'exploitations agricoles. Alors, à l'aide du théorème de Bayes, la densité *a posteriori* conjointe est

$$\pi(\boldsymbol{\mu}, \mathbf{r}, \mathbf{t}, \epsilon, \theta, \sigma^2, \rho \mid \mathbf{y}) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} \left(\frac{1-\rho}{\rho\sigma^2}\right)^{\ell/2} e^{-\epsilon} I_{\epsilon}(a_o, b_o) \times$$
$$\prod_{i=1}^{\ell} \left[a_{oi} \epsilon \exp\{-a_{oi}\epsilon t_i\} \prod_{j=1}^{n_i} \exp\left\{ -\frac{1}{2\sigma^2} \left\{ y_{ij} - \mu_i - (2r_i - 1)\, t_i \right\}^2 \right\} \right] \prod_{i=1}^{\ell} \exp\left\{ -\frac{1-\rho}{2\rho\sigma^2} (\mu_i - \theta)^2 \right\}.$$

Nous présentons ensuite les densités *a posteriori* conditionnelles (DPC) nécessaires pour exécuter l'échantillonneur de Gibbs.

En posant $\lambda_i = \frac{\rho n_i}{\rho n_i + 1 - \rho}$, $i = 1, \ldots, \ell$, il est facile de montrer que les DPC des μ_i sont

$$\mu_i \overset{\text{ind}}{\sim} \text{Normal}\left\{ \lambda_i \left(\overline{y}_i - \frac{(2r_i - 1)\, t_i}{n_i} \right) + (1 - \lambda_i)\, \theta,\ (1 - \lambda_i) \frac{\rho}{1-\rho} \sigma^2 \right\},\ i = 1, \ldots, \ell.$$

En posant

$$q_i = \prod_{j=1}^{n_i} \left\{ \frac{\phi\left(\frac{y_{ij} - \mu_i + t_i}{\sigma}\right)}{\phi\left(\frac{y_{ij} - \mu_i + t_i}{\sigma}\right) + \phi\left(\frac{y_{ij} - \mu_i - t_i}{\sigma}\right)} \right\},$$

où $\phi(\cdot)$ est la densité normale standard, les DPC des r_i sont

$$r_i \overset{\text{ind}}{\sim} \text{Bernoulli}\,(q_i), i = 1, \ldots, \ell.$$

Il est également vrai que la DPC de ϵ est

$$\epsilon \sim \text{Gamma}\left(\ell + 1, \sum_{i=1}^{\ell} a_{oi} t_i + 1 \right) I_{\epsilon}(a_o, b_o).$$

Les t_i sont transformés en $t_i = \frac{\gamma_i}{1 - \gamma_i}$, $i = 1, \ldots, \ell$, et la méthode de la grille est utilisée pour échantillonner leurs DPC (des densités normales tronquées peuvent être utilisées). Il reste à échantillonner les DPC de θ, σ^2, ρ, et leur DPC conjointe est

$$\pi(\theta, \sigma^2, \rho) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} \left(\frac{1-\rho}{\rho\sigma^2}\right)^{\ell/2} \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ y_{ij} - \mu_i - (2r_i - 1) t_i \right\}^2 \right\} \exp\left\{ -\frac{1-\rho}{2\rho\sigma^2} \sum_{i=1}^{\ell} (\mu_i - \theta)^2 \right\}.$$

Les DPC pour σ^2 et θ sont respectivement des densités gamma inverse et normale, mais la DPC de ρ n'est pas simple; on peut utiliser la méthode de la grille pour l'échantillonner.

Nous avons utilisé un échantillonneur de Gibbs plus efficace, obtenu à partir de la décomposition suivante :

$$\pi(\boldsymbol{\mu},\mathbf{r},\mathbf{t},\epsilon,\theta,\sigma^2,\rho\,|\,\mathbf{y}) = \pi_1(\boldsymbol{\mu},\theta,\sigma^2,\rho\,|\,\mathbf{r},\mathbf{t},\epsilon,\mathbf{y})\,\pi_2(\mathbf{r},\mathbf{t},\epsilon\,|\,\mathbf{y}),$$

où l'on peut échantillonner $(\boldsymbol{\mu},\theta,\sigma^2,\rho)$ simultanément à partir de $\pi_1(\boldsymbol{\mu},\theta,\sigma^2,\rho\,|\,\mathbf{r},\mathbf{t},\epsilon,\mathbf{y})$ en appliquant la règle de multiplication des probabilités; voir Nandram, Toto et Choi (2011). Il s'agit d'un échantillonneur de Gibbs bloqué, plus efficace qu'un échantillonneur de Gibbs à balayage systématique (par exemple voir Tan et Hobert, 2009), ce qui permet de réduire le nombre d'exécutions nécessaires de l'échantillonneur de Gibbs.

Remarques

a. Nous pouvons démontrer que la densité *a posteriori* conjointe est valide. Nous pouvons éliminer par intégration $\mu_i,\theta,\sigma^2,\epsilon,$ et il nous restera $\pi(\mathbf{r},\mathbf{t},\rho\,|\,\mathbf{y})$ si nous supposons que $\ell \geq 2.$ Nous pouvons ensuite transformer $t_i = \frac{\gamma_i}{1-\gamma_i}$ (c'est-à-dire que $0 < \gamma_i < 1$) et, en supposant que γ_i est strictement bornée loin de 0 et de 1 et qu'il en est de même pour $\rho,$ nous pouvons démontrer que la densité *a posteriori* conjointe est valide.

b. Le total masqué des comtés sensibles doit être exactement égal au total observé. Cela est réalisé par étalonnage (méthode itérative du quotient) lors du post-traitement; voir les commentaires précédents sur la sensibilité de la cible. Le post-traitement permet de maintenir le même niveau de protection de la confidentialité que les mesures imprécises (Dwork et Roth, 2014, proposition 2.1). Le total observé de l'État dans le TAB n'est pas utilisé dans la procédure de masquage (ajustement du modèle) et peut être considéré comme une donnée externe (Williams et Bowen, 2023), et les valeurs réelles des superficies au niveau des exploitations ne sont pas utilisées dans la méthode itérative du quotient (étalonnage).

c. L'étalonnage se fait à l'échelle d'origine comme suit. Soit $\tilde{T}_1,\ldots,\tilde{T}_\ell,$ les valeurs prédites des totaux de comtés et $B,$ le total des données confidentielles observées. Nous voulons $T_1,\ldots,T_\ell$ de sorte que $\sum_{i=1}^{\ell} T_i = B,$ qui sont

$$T_i = \left(\frac{B}{\sum_{i=1}^{\ell}\tilde{T}_i}\right)\tilde{T}_i, i = 1,\ldots,\ell.$$

Il convient de mentionner que $\sum_{i=1}^{\ell} T_i = B$ et qu'à cette étape de post-traitement, les données confidentielles observées ne sont pas utilisées. Cela se fait dans l'analyse des données de sortie pour chacune des « bonnes itérations » issues de l'échantillonneur de Gibbs, afin d'obtenir les distributions empiriques des données masquées. Cette méthode itérative du quotient (répartition proportionnelle) est courante dans les enquêtes avec échantillon, mais elle peut être améliorée à l'aide d'une analyse fondée sur un modèle.

3.2 Modèle de mélange robuste

Pour la variable à l'étude, nous pouvons utiliser un modèle de mélange avec des poids « stick-breaking ». Ce modèle de mélange robuste offre une meilleure robustesse à la non-normalité et aux superficies importantes (valeurs aberrantes), et est, en ce sens, supérieur au modèle paramétrique. Il s'inscrit dans le même esprit que le processus de Dirichlet.

Soit $n=\sum_{i=1}^{\ell} n_i,$ le modèle de mélange regroupe les $y_{ij}, j=1,\ldots,n_i, i=1,\ldots,\ell$ en grappes. Soit $n=\sum_{i=1}^{\ell} n_i$ et soit le nombre de grappes des y_{ij}, $n_o,$ où $n_o \le n.$ Nous supposons alors l'indépendance des y_{ij} avec

$$p(y_{ij}, d_{ij} \mid n_o) = \prod_{s=1}^{n_o} \left[p_s \text{ Normal}\left\{ \mu_i + (2r_i - 1)\, t_i + \eta_s, \sigma^2 \right\} \right]^{I(di_j = s)}, n_o \le n,$$

où les d_{ij} sont des variables latentes et $d_{ij}=s, j=1,\ldots,n_i, i=1,\ldots,\ell,$ signifie que la j^{e} exploitation agricole du i^{e} comté est affectée à la s^{e} grappe. De plus, les p_s sont des poids « stick-breaking » (par exemple Ishwaran et James, 2001),

$$p_1 = \nu_1,\ p_2 = \nu_2(1-\nu_1), \ldots,\ p_s = \prod_{k=1}^{s-1}(1-\nu_k),\ \sum_{k=1}^{s} p_k = 1,$$

$$\nu_s \overset{\text{ind}}{\sim} \text{Beta}\left\{ \delta_1^s \frac{\delta_2}{1-\delta_2}, (1-\delta_1^s)\frac{\delta_2}{1-\delta_2} \right\}, 0 < \delta_1 < \frac{1}{2} < \delta_2 < 1.$$

Autrement dit, les $(p_1,\ldots,p_s)$ suivent une distribution de Dirichlet généralisée. Les distributions des ν_s, spécifiées dans ce cas-ci, sont semblables à celles du processus à deux paramètres de Pitman-Yor. Le processus « stick-breaking » s'explique comme suit : à partir d'un bâton (« stick ») d'une longueur de 1, on retranche $p_1=\nu_1,$ donc il reste $1-\nu_1,$ puis on retranche ν_2 du reste pour obtenir $p_2=\nu_2(1-\nu_1),$ et ainsi de suite. Il convient de prendre note que $\sum_{s=1}^{n_o} p_s = 1.$ Le processus de Dirichlet est un autre processus « stick-breaking » (Sethuraman, 1994; Ishwaran et James, 2001), mais il est moins général que celui décrit dans la présente étude.

La densité conjointe de $(\mathbf{y},\mathbf{d})$ est alors

$$p(\mathbf{y},\mathbf{d} \mid n_o) = \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \prod_{s=1}^{n_o} \left[p_s \text{ Normal}\left\{ \mu_i + (2r_i - 1)\, t_i + \eta_s, \sigma^2 \right\} \right]^{I(d_{ij}=s)},$$

et, *a priori* pour la mise en grappes, nous supposons

$$\eta_s \overset{\text{ind}}{\sim} \text{Normal}\left\{ 0, \frac{\rho_1}{1-\rho_1}\sigma^2 \right\}, s=1,\ldots,n_o.$$

Contrairement au modèle paramétrique, ce modèle de mélange permet de prendre en compte les valeurs aberrantes (les grandes superficies) par la mise en grappes probabiliste des données; il est également robuste à la non-normalité. Il convient de mentionner que nous supposons que les $y_{ij}-\mu_i,$ et non les $y_{ij},$ ont la

confidentialité différentielle standard, puisque $\mathrm{E}(y_{ij} - \mu_i \mid n_o) = (2r_i - 1)\, t_i, j = 1,\ldots,n_i, i = 1,\ldots,\ell,$ le mécanisme de confidentialité différentielle de Laplace.

Le reste du modèle (c'est-à-dire les hypothèses ci-dessous) est semblable au modèle paramétrique. Autrement dit :

$$\mu_i \overset{\text{ind}}{\sim} \text{Normal}\left\{\theta, \frac{\rho_2}{1-\rho_2}\sigma^2\right\},\ r_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left\{\frac{1}{2}\right\},\ t_i \mid \epsilon \overset{\text{ind}}{\sim} \text{Gamma}\left\{1, a_{oi}\epsilon\right\}, i = 1,\ldots,\ell,$$

$$\pi(\theta, \sigma^2, \epsilon, \rho_1, \rho_2) \propto \frac{1}{\sigma^2} e^{-\epsilon}, a_o < \epsilon < b_o$$

(par exemple $a_o = 0, b_o =$ 2,25, comme dans le modèle paramétrique). Nous pouvons ajuster le modèle complet à l'aide de l'échantillonneur de Gibbs bloqué; par exemple, il y a un bloc pour $\mathbf{d}$ et un autre pour $(\mathbf{v}, \delta_1, \delta_2)$. En fixant initialement $n_o =$ 15, par exemple, nous pouvons établir un ensemble préliminaire de grappes en partitionnant les données et obtenir une matrice de partition, $E,$ une matrice d'incidence $n \times n_o$. Cela permet d'obtenir $(\boldsymbol{\mu}, \mathbf{r}, \mathbf{t}, \boldsymbol{\eta}, \sigma^2),$ et les poids « stick-breaking » peuvent être calculés. Soit $\mathbf{e}'_{ij}, j = 1,\ldots,$ $n_i, i = 1,\ldots,\ell,$ la $(ij)^{\text{e}}$ ligne de E. Alors,

$$y_{ij} \mid d_{ij} \overset{\text{ind}}{\sim} \text{Normal}\left\{\mu_i + (2r_i - 1)\, t_i + \mathbf{e}'_{ij}\boldsymbol{\eta}, \sigma^2\right\}, j = 1,\ldots,n_i, i = 1,\ldots,\ell.$$

Il est simple d'écrire le reste des DPC nécessaires pour exécuter l'échantillonneur de Gibbs.

Toutefois, nous remarquons que la DPC de tous les paramètres (étant donné $\mathbf{d}$) est

$$\begin{aligned}\pi(\boldsymbol{\mu}, \mathbf{r}, \mathbf{t}, \boldsymbol{\eta}, \epsilon, \theta, \rho_1, \rho_2, \sigma^2 \mid \mathbf{d}, \mathbf{y}) \propto{}& \left(\frac{1}{\sigma^2}\right)^{(n+n_o+\ell)/2+1} \left(\frac{1-\rho_1}{\rho_1}\right)^{n_o/2} \left(\frac{1-\rho_2}{\rho_2}\right)^{\ell/2} e^{-\epsilon} I_{[a_o, b_o]}(\epsilon) \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}(y_{ij} - \mu_i - (2r_i - 1)\, t_i - \mathbf{e}'_{ij}\boldsymbol{\eta})^2\right\} \times \exp\left\{-\frac{1-\rho_1}{2\rho_1\sigma^2}\sum_{s=1}^{n_o}\eta_s^2\right\} \\ &\times \exp\left\{-\frac{1-\rho_2}{2\rho_2\sigma^2}\sum_{i=1}^{\ell}(\mu_i - \theta)^2\right\} \times \prod_{i=1}^{\ell}(a_{oi}\epsilon) \times \exp\{-a_{oi}\epsilon t_i\}.\end{aligned}$$

3.3 Rétrécissement excessif et prédiction

Nous décrivons une procédure simple permettant de remédier au rétrécissement excessif (un problème récurrent en estimation sur petits domaines) dans les deux modèles. Il existe d'autres approches (par exemple le regroupement global-local), mais cette complication n'est pas nécessaire. Il s'agit d'une extension des deux modèles. Nous montrons également la façon d'effectuer la prédiction dans les deux modèles avec un ajustement visant à accroître l'utilité à l'aide de l'erreur relative absolue (ERA).

Pour atténuer le rétrécissement excessif et éviter les problèmes d'indéterminabilité, nous ajoutons maintenant une caractéristique commune aux deux modèles. Autrement dit, nous avons partitionné les comtés en déciles à partir de la moyenne pondérée des superficies agricoles. Plus précisément, nous examinons

$$\overline{y}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},\ i = 1,\ldots,\ell,$$

où w_{ij} sont les poids du recensement, et nous partitionnons les $\overline{y}_i$ en déciles (10 intervalles). Soient q_0 et q_{10}, les valeurs minimale et maximale des $\overline{y}_i$, et $q_1,\ldots,q_9$, les $10^{e},\ldots,90^{e}$ centiles des $\overline{y}_i$. Les déciles sont $[q_0,q_1),[q_1,q_2),\ldots,[q_9,q_{10}]$. Soit $C_s, s = 1,\ldots,10,$ l'ensemble de comtés dans ces déciles. Pour $s = 1,\ldots,10,$ g_s désigne la cardinalité de ces C_s, qui est d'environ 8.

Nous supposons alors que

$$\mu_i \overset{\text{ind}}{\sim} \text{Normal}\left(\theta_s, \frac{\rho}{1-\rho}\sigma^2\right),\ i \in C_s, s = 1,\ldots,10.$$

Dans le modèle paramétrique original, il n'y a qu'un seul θ. Cela s'ajoute à nos hypothèses standard du modèle paramétrique et du modèle de mélange robuste. Par exemple, dans le modèle paramétrique,

$$y_{ij} \overset{\text{ind}}{\sim} \text{Normal}\{\mu_i + (2r_i - 1)\, t_i, \sigma^2\},\ j = 1,\ldots,n_i, i = 1,\ldots,\ell,$$

avec les mêmes hypothèses pour r_i, t_i et ϵ que dans le modèle paramétrique original. Nous avons également

$$\pi(\boldsymbol{\theta}) = 1,$$

où les θ_s sont ordonnés comme suit : $-\infty < \theta_1 < \ldots < \theta_{10} < \infty.$ Un ajustement semblable est apporté au modèle de mélange robuste.

L'inférence prédictive bayésienne est utilisée pour produire des estimations masquées au niveau des comtés. Ces estimations sont ensuite étalonnées de la façon décrite précédemment. Nous utilisons le modèle de population pour effectuer la prédiction; le modèle d'échantillon, auquel est ajouté du bruit de Laplace, fournit déjà l'ajustement des paramètres du modèle de population.

Pour le modèle paramétrique, la prédiction est effectuée à l'aide du modèle de population,

$$Y_{ij} \mid \mu_i, \sigma^2 \sim \text{Normal}(\mu_i, \sigma^2),\ j = 1,\ldots,n_i, i = 1,\ldots,\ell,$$

et cela est soumis à toute transformation utilisée. Les estimations au niveau des comtés sont alors

$$Y_i = \sum_{j=1}^{n_i} Y_{ij},\ i = 1,\ldots,\ell,$$

et les distributions prédictives *a posteriori* des superficies agricoles masquées au niveau des comtés sont

$$\pi(Y_i \mid \mathbf{y}) = \int f(Y_i \mid \mu_i, \sigma^2, \mathbf{y})\, \pi(\mu_i, \sigma^2 \mid \mathbf{y})\, d\mu_i\, d\sigma^2 = \int f(Y_i \mid \mu_i, \sigma^2)\, \pi(\mu_i, \sigma^2 \mid \mathbf{y})\, d\mu\, d\sigma^2,$$

où la densité *a posteriori*, $\pi(\mu_i, \sigma^2 \mid \mathbf{y}),$ provient du modèle d'échantillon, et

$$f(Y_i \mid \mu_i) \overset{\text{ind}}{\sim} \text{Normal}\left(n_i\mu_i, n_i\sigma^2\right), i = 1,\ldots,\ell.$$

Bien entendu, cela se fait lors du post-traitement de l'échantillonneur de Gibbs conjointement avec l'étalonnage comme nous l'avons décrit précédemment. En présence d'une transformation, il faut prédire individuellement les Y_{ij}. Autrement dit, $Y_{ij} \mid \mu_i \sim \text{Normal}(\mu_i, \sigma^2)$, $j = 1,\ldots,n_i, i = 1,\ldots,\ell,$ puis la rétrotransformation est effectuée.

Tout comme le modèle paramétrique, la prédiction est effectuée dans le modèle de population (modèle de mélange robuste),

$$y_{ij} \mid \mu_i, \sigma^2, \boldsymbol{\eta}, \mathbf{p}, n_o \overset{\text{ind}}{\sim} \sum_{s=1}^{n_o} p_s \,\text{Normal}\left(\mu_i + \eta_s, \sigma^2\right), j = 1,\ldots,n_i, i = 1,\ldots,\ell.$$

Ce modèle de population est alimenté par le modèle d'échantillon, et le masquage des données se fait exactement de la même façon que dans le modèle paramétrique.

Nous pouvons accroître l'utilité en imposant une borne aux ERA. Sur l'échelle transformée (c'est-à-dire la racine carrée) de chaque exploitation agricole, nous pouvons supposer que

$$\left|\frac{P^2 - O^2}{O^2}\right| < a, 0 < a < 1,$$

où P est la superficie prédite et O est la superficie observée sur l'échelle transformée. Ainsi, $O\sqrt{1-a} < P < O\sqrt{1+a}$. Par exemple, si nous choisissons $a = 1$, $a = 0{,}75$ ou $a = 0{,}5$, nous obtenons $0 < P < O\sqrt{2}$, $O\sqrt{0{,}25} < P < O\sqrt{1{,}75}$ ou $O\sqrt{0{,}5} < P < O\sqrt{1{,}5}$. Dans chaque cas, il s'agit d'une légère entorse au théorème de post-traitement que nous pouvons tolérer afin d'augmenter légèrement l'utilité. Par prudence, nous avons retenu $a = 0{,}75$; $a = 1$ entraîne des ERA supérieures à l'unité, tandis que $a = 0{,}50$ produit des ERA artificiellement faibles sur l'échelle originale non transformée. Cet ajustement est également nécessaire pour éviter que les totaux masqués au niveau des comtés ne soient trop petits (ou trop grands) par rapport aux observations confidentielles d'origine.

4. Comparaison des deux modèles

Nous comparons les deux modèles à l'aide de données simulées sur les superficies récoltées pour le PRO dans le TAB de l'État; l'État comprend le PRO cultivé dans les 75 comtés de l'État, et le nombre pondéré (N) d'exploitations dans un comté varie de 2 à 97. (Dans les deux modèles, nous avons maintenu ϵ dans (0,000; 2,250).) Rappelons qu'il y a 17 comtés supprimés et 58 comtés non supprimés répartis en 38 comtés non sensibles et 20 comtés sensibles. Comme il a été mentionné précédemment, les superficies des comtés non sensibles demeurent non masquées par rapport aux superficies observées initialement lors du recensement, et les superficies pondérées sont diffusées pour ces comtés. Les comtés supprimés ne sont jamais diffusés. Les modèles sont ajustés aux 58 comtés non supprimés. Nous comparons également ces deux

modèles pour petits domaines à un modèle pour domaine (comté) individuel. Dans nos tableaux et nos graphiques, nous présentons uniquement les résultats pour les 20 comtés sensibles.

L'échantillonneur de Gibbs a été utilisé pour ajuster les deux modèles au moyen d'une exécution de longue durée. Le suivi des échantillonneurs de Gibbs a été effectué à l'aide des tests de Geweke (Geweke, 1992) ainsi que des tailles d'échantillon efficaces pour les hyperparamètres afin d'évaluer la convergence et un mélange robuste. Encore une fois, le post-traitement, au cours duquel nous avons étalonné les données masquées en fonction du total observé au niveau de l'État, permet de maintenir le même niveau de protection de la confidentialité que les mesures imprécises (Dwork et Roth, 2014, proposition 2.1) et est compris dans le temps de calcul. Tous les calculs ont été effectués à l'aide de la grappe informatique à haut rendement (Turing) du Worcester Polytechnic Institute. (La grappe comprend 3 492 unités centrales de traitement, allant des unités centrales Intel E5-2695 v4 aux unités centrales Intel Scalable Gold 6248 et AMD EPYC 7543, ainsi que 40,5 To de mémoire. La grappe comporte des segments accélérés par 60 unités de traitement graphique, y compris 20 unités A100 de 80 Go et 2 unités H100.)

Pour les statistiques *a posteriori* sommaires, nous avons utilisé la moyenne *a posteriori* (MP), l'écart-type *a posteriori* (ETP), l'erreur-type numérique (ETN), le coefficient de variation *a posteriori* (CVP) et l'intervalle de la plus haute densité *a posteriori* à 95 % (IHDP; l'unimodalité est supposée). Nous avons utilisé le CVP comme mesure de la fiabilité de la procédure bayésienne. Si CVP $\leq$ 0,30, la fiabilité est élevée, si 0,30 $<$ CVP $\leq$ 0,75, la fiabilité est moyenne et si CVP $>$ 0,75, la fiabilité est faible. Sur l'échelle d'origine, pour chaque comté, nous définissons encore l'utilité comme étant l'ERA :

$$\text{ERA} = |\ (\text{MP} - \text{Obs.}) / \text{Obs.}\ |,$$

où Obs. désigne les données observées; de petites valeurs indiquent une bonne utilité. Afin d'illustrer différents degrés d'utilité (d'excellente à faible), nous avons examiné les ERA et défini les catégories suivantes : ERA $\leq$ 0,05 (excellente); 0,05 $\leq$ ERA $<$ 0,10 (très bonne); B0,10 $\leq$ ERA $<$ 0,25 (bonne); ERA $\geq$ 0,25 (mauvaise). (L'utilité pour des comtés individuels peut être qualifiée d'utilité locale.)

De plus, nous avons considéré deux mesures globales de l'utilité. Premièrement, nous avons utilisé l'erreur quadratique moyenne des scores de propension (EQMSP) décrite dans Snoke, Raab, Nowok, Dibben et Slavkovic (2018); des valeurs de l'EQMSP proches de zéro indiquent une utilité globale élevée. Il s'agit d'une bonne mesure de l'utilité globale; d'autres mesures sont présentées par Snoke et coll. (2018), comme le ratio de l'EQMSP et la EQMSP normalisée, mais celles-ci ne sont pas appropriées, car elles reposent sur un échantillonnage aléatoire simple. Les scores de propension sont calculés au moyen d'une régression logistique et la variable de réponse (régression quadratique) est la covariable (non-ignorabilité); aucune autre covariable n'est utilisée dans la présente analyse. Lorsque nous ajoutons le carré de la variable de réponse, nous obtenons une légère amélioration. Deuxièmement, nous avons utilisé le test de Wilcoxon pour observations appariées (TWOA), dans le cadre duquel nous examinons à quel point la moyenne *a posteriori* (ou chaque « bonne » itération issue de l'échantillonneur de Gibbs) est proche des superficies observées (paires appariées). Le TWOA fournit une évaluation globale de l'utilité; une grande valeur de p indique une

utilité élevée, tandis qu'une petite valeur de p indique une utilité plus faible. Nous privilégions l'utilisation des valeurs de p parce qu'elles sont calées, contrairement aux statistiques de test. (Un réviseur a souligné à juste titre que le test de Kolmogorov-Smirnov n'est pas approprié parce que les deux échantillons sont fortement corrélés.)

Pour le modèle paramétrique, nous avons exécuté l'échantillonneur de Gibbs 300 000 fois, utilisé 50 000 itérations pour la période de rodage et prélevé un échantillon systématique toutes les 250 exécutions. Cela a pris environ 2 minutes au moyen de Turing. Les tests de Geweke et les tailles d'échantillon effectives montrent un bon rendement de l'échantillonneur de Gibbs pour $\boldsymbol{\theta}, \sigma^2, \rho, \epsilon;$ la valeur de p du test de Geweke et la taille d'échantillon effective pour ϵ sont de 0,937 et de 1 000.

Pour le modèle de mélange robuste, puisqu'il est beaucoup plus complexe que le modèle paramétrique, on s'attendrait à ce qu'un plus grand nombre d'exécutions soient nécessaires pour assurer une convergence adéquate, mais ce n'est pas le cas. En fait, nous avons exécuté l'échantillonneur de Gibbs 200 000 fois, utilisé 50 000 itérations pour la période de rodage, prélevé un échantillon systématique toutes les 150 exécutions et surveillé la convergence (test de Geweke et tailles d'échantillon effectives). Cela a pris environ 13 minutes au moyen de Turing. Là encore, les tests de Geweke et les tailles d'échantillon effectives indiquent un bon rendement de l'échantillonneur de Gibbs pour $\boldsymbol{\theta}, \sigma^2, \rho_1, \rho_2, \delta_1, \delta_2, n_o, \epsilon;$ la valeur de p du test de Geweke et la taille d'échantillon effective pour ϵ sont de 0,902 et de 1 000.

Le modèle de mélange est très souple et renforce la crédibilité de notre procédure bayésienne, puisqu'il peut tenir compte de superficies agricoles extrêmes. Cela peut être observé en examinant le nombre de composantes (c'est-à-dire de grappes), $n_o,$ un paramètre important du modèle de mélange robuste. La valeur de p du test de Geweke et la taille d'échantillon effective pour n_o sont de 0,587 et de 1 000. Pour $n_o,$ MP = 5,227, ETP = 1,383, ETN = 0,038, CVP = 0,264 et IHDP à 95 % est (3,00; 8,00); l'inférence *a posteriori* pour le nombre de grappes est fiable, mesurée par le CVP. Par conséquent, il est clair que le modèle de mélange robuste peut séparer les superficies des comtés en différents groupes (données ordinaires, valeurs aberrantes faibles, valeurs aberrantes extrêmes, etc.), ce que le modèle paramétrique ne peut manifestement pas faire. Toutefois, les deux modèles nécessitent un grand nombre d'exécutions pour atteindre la convergence; le modèle paramétrique requiert davantage d'exécutions, mais le temps de calcul pour atteindre la convergence est beaucoup plus court que celui du modèle de mélange robuste.

Nous avons calculé la sensibilité locale (approximativement globale) pour les données masquées (dans ce cas-ci, les moyennes *a posteriori* des comtés). Selon le modèle paramétrique, nous avons obtenu 3 553 (5,967 unités standard) pour tous les comtés et 585 (6,797 unités standard) pour les comtés sensibles. Selon le modèle de mélange robuste, nous avons obtenu 3 492 (5,660 unités standard) pour tous les comtés et 583 (6,623 unités standard) pour les comtés sensibles. Ces résultats sont semblables et, comme prévu, ils sont comparables à ceux découlant des données confidentielles initialement observées.

À des fins de comparaison, nous examinons un modèle pour domaine (comté) individuel. Soit $y_j,$ $j = 1,\ldots,n,$ les superficies pondérées au niveau des exploitations agricoles, où n est le nombre d'exploitations

agricoles dans ce comté. Il convient de mentionner que nous ne considérons que les comtés sensibles et que nous utilisons une version du modèle paramétrique pour petits domaines pour un seul comté. Par conséquent, nous supposons

$$y_j \mid \mu, r, t, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\,(\mu + (2r-1)\,t, \sigma^2),\ j = 1, \ldots, n.$$

Soit t_o, l'échelle (la même que dans les deux modèles pour petits domaines) pour ce comté seulement. Pour le mécanisme de confidentialité différentielle, nous supposons

$$r \sim \text{Bernoulli}\left(\frac{1}{2}\right), t \sim \text{Gamma}\left(1, \frac{\epsilon}{t_o}\right), \epsilon \sim \text{Gamma}\,(1,1), a_o < \epsilon < b_o,$$

où, à nouveau, nous prenons $a_o = 0{,}0$ et $b_o = 2{,}25$, et *a priori*, $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$. Selon le théorème de Bayes, la densité *a posteriori* conjointe est

$$\pi(\mu, \sigma^2, r, t, \epsilon \mid \mathbf{y}) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left\{-\frac{1}{2\sigma^2}\{(n-1)\,s^2 + n(\bar{y} - (2r-1)\,t - \mu)^2\}\right\} \frac{\epsilon}{t_o} e^{-\frac{\epsilon}{t_o}t} e^{-\epsilon} I_{(0,b_o)}(\epsilon), t > 0.$$

Cette densité *a posteriori* peut être ajustée à l'aide de l'échantillonneur de Gibbs, mais il faut faire preuve de prudence, car $(r, t, \mu, \sigma^2, \epsilon)$ ne sont pas entièrement déterminables. Pour atténuer cette indéterminabilité, nous utilisons les deux blocs, (μ, σ^2) et (r, t, ϵ), pour exécuter l'échantillonneur de Gibbs.

L'échantillonneur de Gibbs est exécuté 65 000 fois, une période de rodage de 15 000 itérations est utilisée et un échantillon systématique est prélevé toutes les 50 exécutions afin d'obtenir un échantillon de taille 1 000. Les 20 échantillonneurs de Gibbs ont nécessité environ 11 minutes au moyen de Turing. Les valeurs de p de Geweke et les tailles d'échantillon effectives sont généralement acceptables lorsqu'ils sont exécutés pour $(\mu, \sigma^2, t, \epsilon)$; pour ϵ, les valeurs de p sont toutes suffisamment élevées et les tailles d'échantillon effectives sont pour la plupart de 1 000. L'étalonnage pour les 20 comtés est effectué comme dans les deux modèles pour petits domaines. Dans les tableaux 4.1 à 4.3 et les figures 4.1 à 4.4, nous comparons les trois modèles à l'aide des ERA, des CVP, des MP et des ETP. Nous n'incluons que les 20 comtés sensibles dans ces tableaux et ces figures, puisque les comtés supprimés ne sont jamais présentés et que les superficies pondérées des comtés non sensibles sont diffusées sans masquage.

Dans les tableaux 4.1 à 4.3, plusieurs constats ressortent. Dans le tableau 4.1, le modèle de mélange est généralement supérieur au modèle paramétrique lorsque nous examinons les ERA et les CVP; les ETP sont également plus faibles pour le modèle de mélange. Dans le tableau 4.2, le modèle paramétrique est supérieur au modèle pour domaine individuel et, dans le tableau 4.3, le modèle de mélange est supérieur au modèle pour domaine individuel, lorsque nous examinons les ERA et les CVP; les ETP sont également plus faibles pour le modèle de mélange. Par conséquent, le modèle de mélange est nettement plus fiable que les deux autres modèles concurrents, en particulier le modèle pour domaine individuel.

Tableau 4.1
Comparaison du modèle paramétrique et du modèle de mélange à l'aide des statistiques *a posteriori* sommaires concernant les données masquées pour les comtés sensibles

C	*N*	Obs.	MP	ERA	Utilité	ETP	ETN	CVP	Fiab.	IHDP à 95 %
a. Modèle paramétrique										
8	14	95,400	107,505	0,127	B	25,169	0,664	0,234	E	(62,272; 155,343)
12	7	58,000	61,849	0,066	TB	25,518	0,914	0,413	M	(26,282; 112,268)
15	5	60,500	63,825	0,055	TB	24,326	0,831	0,381	M	(27,497; 110,899)
17	5	44,800	50,455	0,126	B	14,785	0,603	0,293	E	(23,225; 76,857)
18	7	89,300	94,899	0,063	TB	33,710	1,066	0,355	M	(41,524; 163,502)
19	10	25,500	29,714	0,165	B	7,679	0,198	0,258	E	(15,437; 45,922)
28	12	84,300	92,165	0,093	TB	27,763	0,726	0,301	M	(45,029; 144,265)
31	24	178,000	165,012	0,073	TB	50,435	1,452	0,306	M	(82,448; 271,373)
43	59	518,000	525,721	0,015	E	79,725	2,613	0,152	E	(383,559; 689,560)
46	5	102,500	98,845	0,036	E	41,108	1,206	0,416	M	(43,179; 186,845)
48	5	60,500	67,643	0,118	B	20,298	0,606	0,300	M	(34,336; 108,325)
49	16	121,400	133,035	0,096	TB	33,157	1,205	0,249	E	(73,781; 198,713)
50	7	25,600	29,343	0,146	B	10,556	0,370	0,360	M	(11,478; 48,428)
57	11	218,400	193,913	0,112	B	56,509	1,646	0,291	E	(116,698; 319,623)
59	5	41,300	46,875	0,135	B	15,068	0,478	0,321	M	(21,675; 75,556)
61	9	50,500	57,227	0,133	B	16,414	0,519	0,287	E	(26,724; 89,722)
62	29	455,300	439,421	0,035	E	84,016	2,731	0,191	E	(276,284; 596,880)
70	6	110,600	113,344	0,025	E	37,461	1,292	0,331	M	(51,437; 188,677)
72	15	58,000	66,782	0,151	B	17,617	0,570	0,264	E	(37,101; 100,482)
76	31	417,800	378,128	0,095	TB	75,786	3,027	0,200	E	(251,244; 539,671)
b. Modèle de mélange										
8	14	95,400	106,886	0,120	B	23,110	0,745	0,216	E	(62,678; 150,491)
12	7	58,000	60,972	0,051	TB	22,601	0,710	0,371	M	(24,259; 102,284)
15	5	60,500	65,248	0,078	TB	22,653	0,671	0,347	M	(27,306; 105,988)
17	5	44,800	49,520	0,105	B	13,503	0,467	0,273	E	(26,225; 76,459)
18	7	89,300	93,059	0,042	E	29,809	0,890	0,320	M	(41,293; 149,528)
19	10	25,500	28,793	0,129	B	6,846	0,200	0,238	E	(15,460; 41,969)
28	12	84,300	91,069	0,080	TB	24,121	0,814	0,265	E	(45,356; 136,532)
31	24	178,000	171,431	0,037	E	49,926	1,230	0,291	E	(95,786; 277,416)
43	59	518,000	519,965	0,004	E	70,480	2,315	0,136	E	(384,103; 654,173)
46	5	102,500	98,059	0,043	E	41,151	1,429	0,420	M	(41,235; 177,710)
48	5	60,500	66,654	0,102	B	17,569	0,596	0,264	E	(36,033; 102,083)
49	16	121,400	131,711	0,085	TB	30,520	1,000	0,232	E	(73,449; 188,965)
50	7	25,600	28,741	0,123	B	9,433	0,241	0,328	M	(11,420; 45,431)
57	11	218,400	201,056	0,079	TB	57,575	2,009	0,286	E	(116,466; 326,679)
59	5	41,300	45,305	0,097	TB	13,797	0,496	0,305	M	(20,169; 70,867)
61	9	50,500	56,933	0,127	B	15,100	0,497	0,265	E	(27,731; 84,088)
62	29	455,300	443,668	0,026	E	73,115	2,319	0,165	E	(309,219; 592,471)
70	6	110,600	111,124	0,005	E	35,230	1,212	0,317	M	(53,610; 177,132)
72	15	58,000	63,627	0,097	TB	15,276	0,428	0,240	E	(37,934; 94,967)
76	31	417,800	381,878	0,086	TB	70,250	1,582	0,184	E	(247,661; 516,522)

Note : ERA = |(MP-Obs.)/Obs.|. Nous utilisons pour l'utilité (E : excellente, TB : très bonne, B : bonne, M : mauvaise); Fiab. (fiabilité) (E : élevée, M : moyenne, F : faible). Erreur relative absolue (ERA); intervalle de la plus haute densité *a posteriori* (IHDP); erreur-type numérique ETN); données observées (Obs.), coefficient de variation *a posteriori*, (CVP); moyenne *a posteriori* (MP); écart-type *a posteriori* (ETP).

Tableau 4.2
Comparaison du modèle paramétrique et du modèle individuel à l'aide des statistiques *a posteriori* sommaires concernant les données masquées pour les comtés sensibles

C	*N*	Obs.	MP	ERA	Utilité	ETP	ETN	CVP	Fiab.	IHDP à 95 %
a. Modèle paramétrique										
8	14	95,400	107,505	0,127	B	25,169	0,664	0,234	E	(62,272; 155,343)
12	7	58,000	61,849	0,066	TB	25,518	0,914	0,413	M	(26,282; 112,268)
15	5	60,500	63,825	0,055	TB	24,326	0,831	0,381	M	(27,497; 110,899)
17	5	44,800	50,455	0,126	B	14,785	0,603	0,293	E	(23,225; 76,857)
18	7	89,300	94,899	0,063	TB	33,710	1,066	0,355	M	(41,524; 163,502)
19	10	25,500	29,714	0,165	B	7,679	0,198	0,258	E	(15,437; 45,922)
28	12	84,300	92,165	0,093	TB	27,763	0,726	0,301	M	(45,029; 144,265)
31	24	178,000	165,012	0,073	TB	50,435	1,452	0,306	M	(82,448; 271,373)
43	59	518,000	525,721	0,015	E	79,725	2,613	0,152	E	(383,559; 689,560)
46	5	102,500	98,845	0,036	E	41,108	1,206	0,416	M	(43,179; 186,845)
48	5	60,500	67,643	0,118	B	20,298	0,606	0,300	M	(34,336; 108,325)
49	16	121,400	133,035	0,096	TB	33,157	1,205	0,249	E	(73,781; 198,713)
50	7	25,600	29,343	0,146	B	10,556	0,370	0,360	M	(11,478; 48,428)
57	11	218,400	193,913	0,112	B	56,509	1,646	0,291	E	(116,698; 319,623)
59	5	41,300	46,875	0,135	B	15,068	0,478	0,321	M	(21,675; 75,556)
61	9	50,500	57,227	0,133	B	16,414	0,519	0,287	E	(26,724; 89,722)
62	29	455,300	439,421	0,035	E	84,016	2,731	0,191	E	(276,284; 596,880)
70	6	110,600	113,344	0,025	E	37,461	1,292	0,331	M	(51,437; 188,677)
72	15	58,000	66,782	0,151	B	17,617	0,570	0,264	E	(37,101; 100,482)
76	31	417,800	378,128	0,095	TB	75,786	3,027	0,200	E	(251,244; 539,671)
b. Modèle individuel										
8	14	95,400	79,194	0,170	B	37,097	1,764	0,468	M	(22,099; 147,552)
12	7	58,000	37,640	0,351	M	16,299	0,697	0,433	M	(16,472; 74,560)
15	5	60,500	43,954	0,273	M	18,193	0,617	0,414	M	(16,337; 80,381)
17	5	44,800	40,429	0,098	TB	22,084	1,045	0,546	M	(10,400; 76,628)
18	7	89,300	62,966	0,295	M	21,326	0,883	0,339	M	(27,800; 103,893)
19	10	25,500	22,946	0,100	B	11,066	0,571	0,482	M	(5,438; 41,253)
28	12	84,300	64,234	0,238	B	26,314	1,085	0,410	M	(21,397; 115,674)
31	24	178,000	130,865	0,265	M	62,570	2,948	0,478	M	(41,944; 255,148)
43	59	518,000	351,244	0,322	M	114,566	6,006	0,326	M	(170,364; 587,641)
46	5	102,500	69,292	0,324	M	29,474	1,136	0,425	M	(28,365; 131,002)
48	5	60,500	49,534	0,181	B	17,686	0,568	0,357	M	(16,756; 81,967)
49	16	121,400	93,514	0,230	B	41,769	1,852	0,447	M	(31,586; 177,478)
50	7	25,600	22,040	0,139	B	12,263	0,435	0,556	M	(5,788; 42,880)
57	11	218,400	142,618	0,347	M	54,434	1,718	0,382	M	(61,851; 267,875)
59	5	41,300	30,302	0,266	M	11,468	0,431	0,378	M	(12,711; 54,605)
61	9	50,500	41,648	0,175	B	18,513	0,648	0,445	M	(10,245; 74,589)
62	29	455,300	319,743	0,298	M	110,496	4,874	0,346	M	(151,884; 550,197)
70	6	110,600	78,447	0,291	M	26,990	1,073	0,344	M	(35,761; 133,664)
72	15	58,000	43,728	0,246	B	16,018	0,659	0,366	M	(16,028; 76,549)
76	31	417,800	280,711	0,328	M	101,139	6,141	0,360	M	(125,131; 500,439)

Note : ERA = |(MP-Obs.)/Obs.|. Nous utilisons pour l'utilité (E : excellente, TB : très bonne, B : bonne, M : mauvaise); Fiab. (fiabilité) (E : élevée, M : moyenne, F : faible). Erreur relative absolue (ERA); intervalle de la plus haute densité *a posteriori* (IHDP); erreur-type numérique ETN); données observées (Obs.), coefficient de variation *a posteriori*, (CVP); moyenne *a posteriori* (MP); écart-type *a posteriori* (ETP).

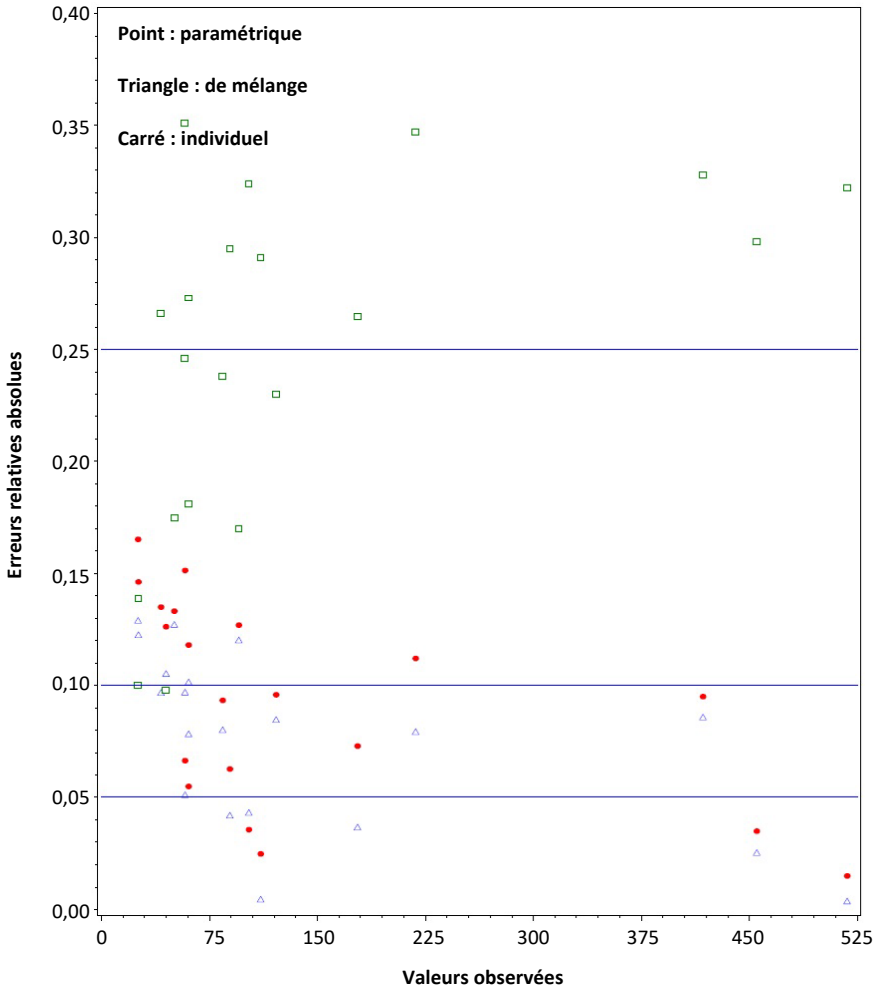
Tableau 4.3
Comparaison du modèle de mélange et du modèle individuel à l'aide des statistiques *a posteriori* sommaires concernant les données masquées pour les comtés sensibles

C	*N*	Obs.	MP	ERA	Utilité	ETP	ETN	CVP	Fiab.	IHDP à 95 %
a. Modèle de mélange										
8	14	95,400	106,886	0,120	B	23,110	0,745	0,216	E	(62,678; 150,491)
12	7	58,000	60,972	0,051	TB	22,601	0,710	0,371	M	(24,259; 102,284)
15	5	60,500	65,248	0,078	TB	22,653	0,671	0,347	M	(27,306; 105,988)
17	5	44,800	49,520	0,105	B	13,503	0,467	0,273	E	(26,225; 76,459)
18	7	89,300	93,059	0,042	E	29,809	0,890	0,320	M	(41,293; 149,528)
19	10	25,500	28,793	0,129	B	6,846	0,200	0,238	E	(15,460; 41,969)
28	12	84,300	91,069	0,080	TB	24,121	0,814	0,265	E	(45,356; 136,532)
31	24	178,000	171,431	0,037	E	49,926	1,230	0,291	E	(95,786; 277,416)
43	59	518,000	519,965	0,004	E	70,480	2,315	0,136	E	(384,103; 654,173)
46	5	102,500	98,059	0,043	E	41,151	1,429	0,420	M	(41,235; 177,710)
48	5	60,500	66,654	0,102	B	17,569	0,596	0,264	E	(36,033; 102,083)
49	16	121,400	131,711	0,085	TB	30,520	1,000	0,232	E	(73,449; 188,965)
50	7	25,600	28,741	0,123	B	9,433	0,241	0,328	M	(11,420; 45,431)
57	11	218,400	201,056	0,079	TB	57,575	2,009	0,286	E	(116,466; 326,679)
59	5	41,300	45,305	0,097	TB	13,797	0,496	0,305	M	(20,169; 70,867)
61	9	50,500	56,933	0,127	B	15,100	0,497	0,265	E	(27,731; 84,088)
62	29	455,300	443,668	0,026	E	73,115	2,319	0,165	E	(309,219; 592,471)
70	6	110,600	111,124	0,005	E	35,230	1,212	0,317	M	(53,610; 177,132)
72	15	58,000	63,627	0,097	TB	15,276	0,428	0,240	E	(37,934; 94,967)
76	31	417,800	381,878	0,086	TB	70,250	1,582	0,184	E	(247,661; 516,522)
b. Modèle individuel										
8	14	95,400	79,194	0,170	B	37,097	1,764	0,468	M	(22,099; 147,552)
12	7	58,000	37,640	0,351	M	16,299	0,697	0,433	M	(16,472; 74,560)
15	5	60,500	43,954	0,273	M	18,193	0,617	0,414	M	(16,337; 80,381)
17	5	44,800	40,429	0,098	TB	22,084	1,045	0,546	M	(10,400; 76,628)
18	7	89,300	62,966	0,295	M	21,326	0,883	0,339	M	(27,800; 103,893)
19	10	25,500	22,946	0,100	B	11,066	0,571	0,482	M	(5,438; 41,253)
28	12	84,300	64,234	0,238	B	26,314	1,085	0,410	M	(21,397; 115,674)
31	24	178,000	130,865	0,265	M	62,570	2,948	0,478	M	(41,944; 255,148)
43	59	518,000	351,244	0,322	M	114,566	6,006	0,326	M	(170,364; 587,641)
46	5	102,500	69,292	0,324	M	29,474	1,136	0,425	M	(28,365; 131,002)
48	5	60,500	49,534	0,181	B	17,686	0,568	0,357	M	(16,756; 81,967)
49	16	121,400	93,514	0,230	B	41,769	1,852	0,447	M	(31,586; 177,478)
50	7	25,600	22,040	0,139	B	12,263	0,435	0,556	M	(5,788; 42,880)
57	11	218,400	142,618	0,347	M	54,434	1,718	0,382	M	(61,851; 267,875)
59	5	41,300	30,302	0,266	M	11,468	0,431	0,378	M	(12,711; 54,605)
61	9	50,500	41,648	0,175	B	18,513	0,648	0,445	M	(10,245; 74,589)
62	29	455,300	319,743	0,298	M	110,496	4,874	0,346	M	(151,884; 550,197)
70	6	110,600	78,447	0,291	M	26,990	1,073	0,344	M	(35,761; 133,664)
72	15	58,000	43,728	0,246	B	16,018	0,659	0,366	M	(16,028; 76,549)
76	31	417,800	280,711	0,328	M	101,139	6,141	0,360	M	(125,131; 500,439)

Note : ERA = |(MP-Obs.)/Obs.|. Nous utilisons pour l'utilité (E : excellente, TB : très bonne, B : bonne, M : mauvaise); Fiab. (fiabilité) (E : élevée, M : moyenne, F : faible). Erreur relative absolue (ERA); intervalle de la plus haute densité *a posteriori* (IHDP); erreur-type numérique ETN); données observées (Obs.), coefficient de variation *a posteriori*, (CVP); moyenne *a posteriori* (MP); écart-type *a posteriori* (ETP).

Dans la figure 4.1, nous comparons les trois modèles à l'aide des ERA des 20 comtés sensibles. Les lignes droites horizontales correspondent à ERA = 0,05, 0,10, 0,25. Le modèle paramétrique et le modèle de mélange ont des ERA inférieures à la ligne de 0,05, contrairement au modèle pour domaine individuel, dont la plupart des points se situent au-dessus de la ligne de 0,25. Pour le modèle de mélange, les points sont généralement plus bas que pour le modèle paramétrique.

Figure 4.1 Graphique des erreurs relatives absolues des comtés par rapport aux valeurs observées, selon le modèle



Dans la figure 4.2, nous comparons les trois modèles à l'aide des CVP des 20 comtés sensibles. Les lignes droites horizontales correspondent à CVP = 0,30 et CVP = 0,75. Il est bon que les CVP du modèle de mélange robuste soient généralement inférieurs à ceux du modèle paramétrique; par conséquent, selon ce modèle, le modèle de mélange robuste s'avère légèrement plus fiable. Tous les points du modèle pour domaine individuel se situent au-dessus de la ligne de 0,30; par conséquent, ce modèle est très peu fiable.

Dans la figure 4.3, nous représentons graphiquement les MP par rapport aux valeurs observées pour les 20 comtés sensibles. Certains points s'écartent de la ligne droite à 45^o; ceux du modèle paramétrique se situent légèrement en dessous de ceux du modèle de mélange. Les points issus du modèle pour domaine individuel se trouvent bien en dessous de la ligne droite à 45^o et ceux associés aux valeurs observées plus faibles sont plus proches de la ligne droite à 45^o. Cela indique encore une fois que le modèle pour domaine individuel produit de piètres ERA. Ce modèle pour domaine individuel donne de piètres résultats pour les valeurs aberrantes.

Dans la figure 4.4, nous comparons les trois modèles à l'aide des ETP des 20 comtés sensibles. La ligne droite la mieux ajustée à travers les points du modèle pour domaine individuel est présentée (ordonnée à l'origine = 7,27, pente = 0,222, $R^2 =$ 0,971). Une fois de plus, le modèle pour domaine individuel affiche les ETP les plus élevés, tandis que le modèle de mélange présente généralement les ETP les plus faibles; les valeurs aberrantes ont des ETP nettement plus petits selon le modèle de mélange que selon le modèle paramétrique, ce qui démontre la force du modèle de mélange.

Figure 4.2 Graphique des coefficients de variation *a posteriori* des comtés par rapport aux valeurs observées, selon le modèle

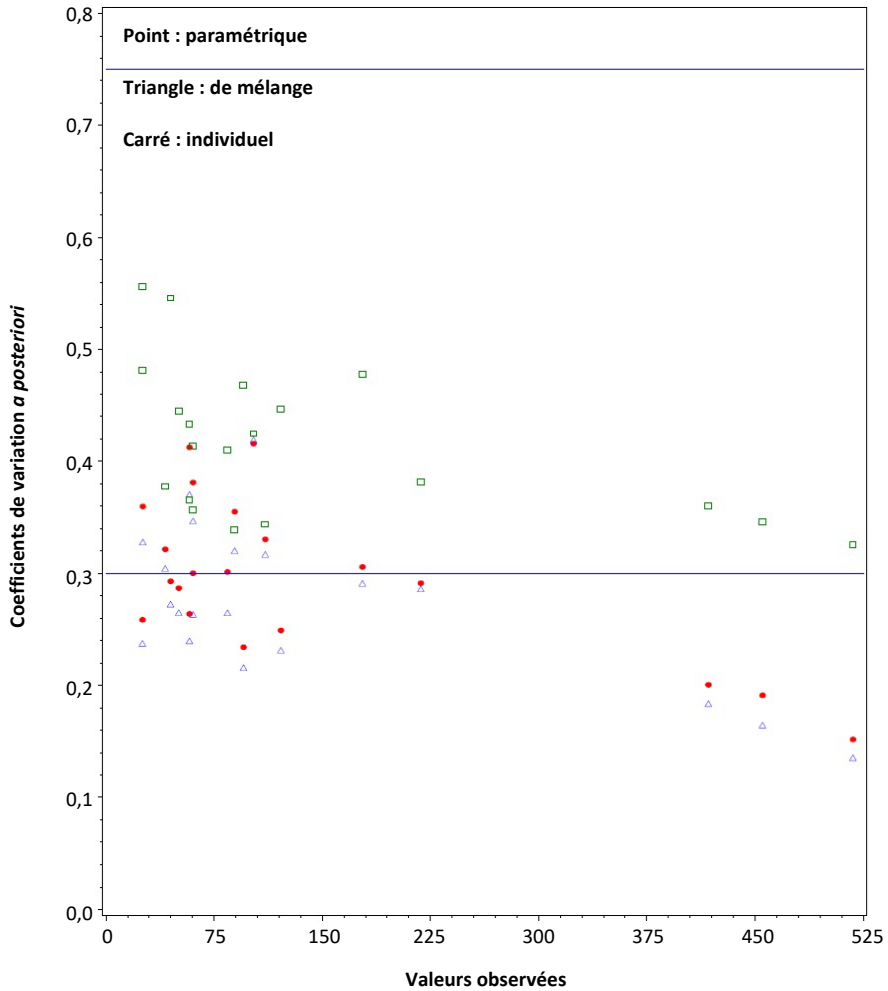
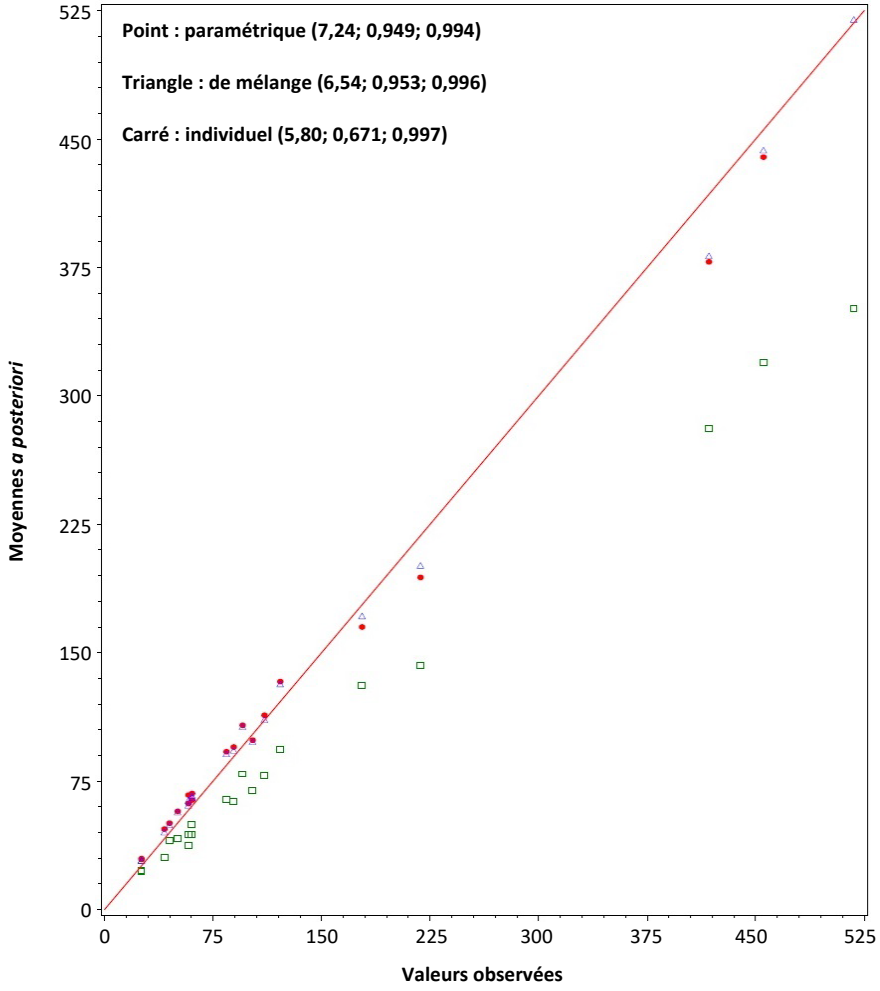
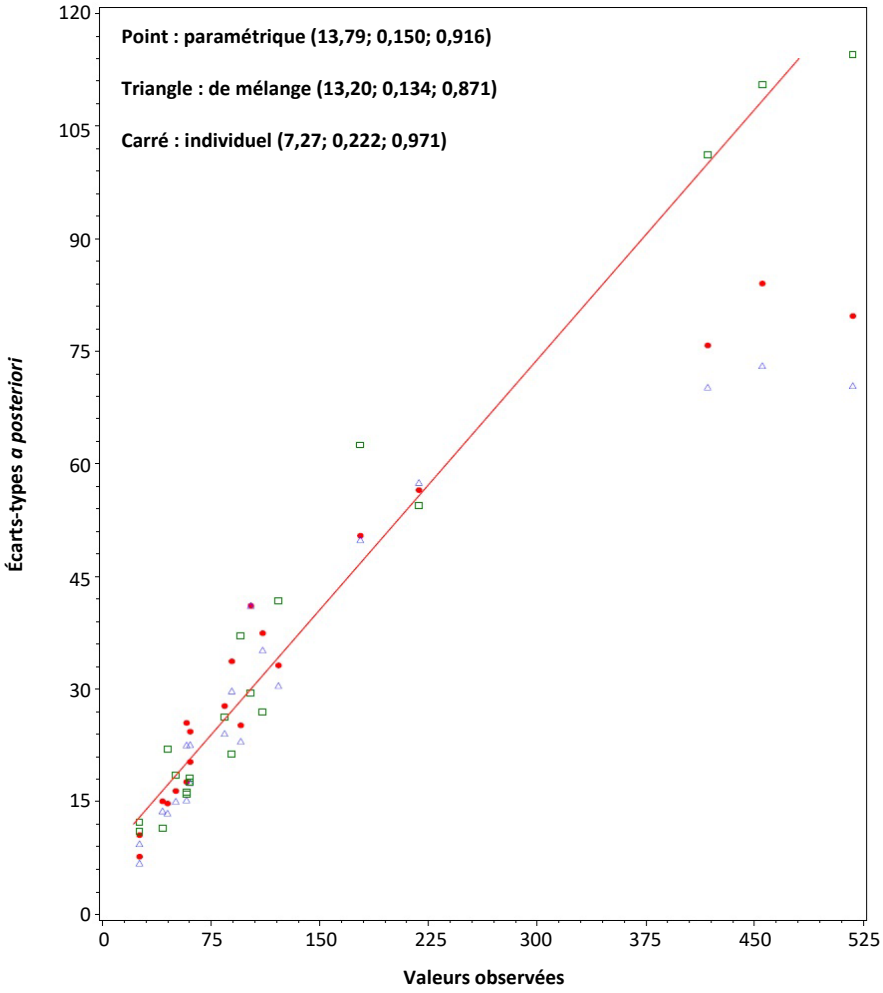


Figure 4.3 Graphique des moyennes *a posteriori* des comtés par rapport aux valeurs observées, selon le modèle



Compte tenu de ces résultats (tabulaires et graphiques), il est donc nécessaire d'ajuster les modèles pour petits domaines, en particulier le modèle de mélange robuste. Il n'est pas judicieux de modéliser chaque comté séparément; il est essentiel d'intégrer la similarité entre les exploitations agricoles au sein d'un même comté et entre les comtés d'un même État. Le modèle de mélange offre le meilleur compromis entre la sécurité et l'utilité; l'une des principales raisons est sa capacité à prendre en compte des comtés dont les superficies sont très différentes (petites et grandes, y compris des valeurs aberrantes) grâce à une mise en grappes automatique des comtés. De plus, de nombreux organismes sont préoccupés par la précision, et le modèle de mélange donne également de bons résultats à cet égard; les deux modèles pour petits domaines surpassent le modèle pour domaine individuel.

Figure 4.4 Graphique des écarts-types *a posteriori* des comtés par rapport aux valeurs observées, selon le modèle



Enfin, nous présentons les tableaux 4.4, 4.5 et 4.6, dans lesquels nous comparons davantage les trois modèles du point de vue de la sécurité et de l'utilité (à la fois globale et locale). Des commentaires figurent dans les notes de chaque tableau; nous formulons ici des observations générales. Dans le tableau 4.4, nous comparons la sécurité en fonction du budget total, ϵ. On peut constater que le modèle pour domaine individuel indique qu'un comté type présente le niveau de sécurité le plus élevé, mais qu'il n'est pas fiable. Pour cette même mesure, le modèle paramétrique et le modèle de mélange sont semblables. Dans le tableau 4.5, nous comparons les ERA (utilité locale), et il est clair que le modèle de mélange est le meilleur, tandis que le modèle pour domaine individuel est le pire. Toutefois, tous les modèles présentent une forte utilité globale selon les valeurs de p et les EQMSP. Dans le tableau 4.6, nous comparons les trois modèles à l'aide des 1 000 échantillons de Gibbs. À partir des cinq statistiques sommaires, il pourrait être possible de diffuser certains de ces ensembles de données synthétiques (et non seulement les moyennes *a posteriori*) si les données originales sont réelles; voir Rubin (1993) et Raghunathan, Reiter et Rubin (2003).

Tableau 4.4
Comparaison de la sécurité des trois modèles pour les comtés sensibles

Modèle	MP	ETP	ETN	CVP	IHDP
P	2,204	0,044	0,001	0,020	(2,112; 2,250)
M	2,206	0,043	0,001	0,019	(2,112; 2,250)
I-T	0,811	0,576	0,016	0,710	(0,034; 1,983)
I-C	1,863	0,284	0,011	0,152	(1,326; 2,250)

Note : Il s'agit de statistiques sommaires du paramètre de confidentialité différentiel, ϵ, le budget total. P : modèle paramétrique, M : modèle de mélange. I-T renvoie à un comté type et I-C renvoie aux comtés combinés. P et M sont très semblables et offrent un niveau de sécurité élevé. Un comté type (I-T) semble offrir un niveau de sécurité, mais n'est pas fiable (CVP : 0,710). Lorsque tous les comtés sont combinés (I-C) selon une composition parallèle, le budget est comparable (légèrement plus petit avec une ETP et un CVP plus grands) aux deux modèles pour petits domaines. Intervalle de la plus haute densité *a posteriori* (IHDP); erreur-type numérique (ETN); coefficient de variation *a posteriori* (CVP); moyenne *a posteriori* (MP); écart-type *a posteriori* (ETP).

Tableau 4.5
Comparaison des utilités locales (erreurs relatives absolues) des trois modèles pour les comtés sensibles

Modèle	E	M	F	TF
P	4	7	9	0
M	6	8	6	0
I	0	1	8	11

Note : Les valeurs correspondent au nombre de comtés sensibles présentant une utilité élevée (E), moyenne (M), faible (F) et très faible (TF). P : modèle paramétrique, M : modèle de mélange, I : modèle individuel. Le modèle de mélange est supérieur au modèle paramétrique, et le modèle individuel est moins bon que ces deux modèles. Pour les modèles paramétrique, de mélange et individuel, les utilités globales (mesurées par les erreurs quadratiques moyennes des scores de propension) sont respectivement de 0,007, de 0,008 et de 0,008, et (mesurées par les valeurs de p du test de Wilcoxon pour observations appariées) de 0,951, de 0,938 et de 0,951; tous les modèles présentent donc une forte utilité globale.

Tableau 4.6
Comparaison des utilités globales des trois modèles pour les comtés sensibles pour les 1 000 échantillons de Gibbs

Modèle	Min	Q_1	Q_2	Q_3	Max
a. Valeurs de p du TWOA					
P	0,867	0,927	0,942	0,957	0,996
M	0,856	0,928	0,944	0,956	0,999
I	0,892	0,931	0,941	0,952	0,998
b. EQMSP de Snoke, Raab, Nowok, Dibben et Slavkovic (2018)					
P	0,000	0,000	0,001	0,003	0,030
M	0,000	0,000	0,001	0,003	0,020
I	0,006	0,007	0,009	0,013	0,058

Note : Il s'agit de cinq statistiques sommaires pour les 1 000 échantillons de Gibbs. Tous les modèles ont une très forte utilité globale. Si les données sont réelles (non simulées), il pourrait être possible de diffuser certains de ces échantillons de Gibbs, puisque les plus petites valeurs de p du TWOA sont élevées et les plus grandes EQMSP ne sont pas trop grandes. Erreur quadratique moyenne des scores de propension (EQMSP); test de Wilcoxon pour observations appariées (TWOA).

5. Mot de la fin

Nous avons élaboré une méthode bayésienne hybride, qui intègre un mécanisme de confidentialité différentielle, pour masquer les totaux de comtés du recensement pour un État des États-Unis concernant la superficie d'un produit agricole, et ce, à partir de données simulées au niveau des exploitations agricoles. Nous avons inclus un facteur d'échelle dans la distribution de Laplace ainsi qu'une caractéristique pour que

le budget global lors de la modélisation soit ϵ, comme la définition de la confidentialité différentielle le préciserait normalement. Enfin, nous avons appliqué deux modèles pour petits domaines (le modèle paramétrique et le nouveau modèle de mélange robuste) au PRO de l'État dans le TAB; ceux-ci sont comparés à un modèle pour domaine individuel. Il est important d'isoler les comtés sensibles. Nous n'avons pas examiné la confidentialité différentielle (ϵ, δ), mais il est possible d'élaborer une méthodologie semblable. L'objectif de la présente étude est de montrer une façon de faire en présence de structures de données plus complexes; notre approche est toutefois pertinente dans de nombreuses applications pratiques. Nous proposons une méthodologie bayésienne générale pouvant être utilisée pour divers problèmes faisant intervenir des modèles plus complexes (par exemple des tableaux couplés et hiérarchiques).

Une autre nouveauté de la présente étude est la spécification de la distribution *a priori* pour ϵ. Il convient de prendre note qu'il n'est pas nécessaire d'attribuer de distribution *a priori* à ϵ, mais cela s'intègre mal au paradigme bayésien. On peut certes fixer ϵ à des valeurs comme $\epsilon =$ 1, 2, comme cela se fait couramment, mais il faudrait alors effectuer une analyse de sensibilité, ce qui représente un fardeau supplémentaire. En revanche, on peut attribuer une distribution *a priori* discrète à ϵ, $P(\epsilon = 1) = P(\epsilon = 2) = \frac{1}{2}$, de la même manière que nous l'avons fait dans le présent article. Cela semble avoir un certain sens dans le paradigme bayésien. Bien entendu, les données de sortie sont protégées par la confidentialité différentielle, mais l'utilité peut être limitée, particulièrement pour les très petits comtés.

Nous mentionnons deux extensions possibles, qui peuvent accroître l'utilité. Premièrement, il est possible d'étendre cette approche à $r \times c$ tableaux comportant r comtés et $c \geq 2$ produits agricoles. Dans le présent travail, nous avons traité le cas où $c = 1$ (c'est-à-dire un seul produit). Nous souhaitons maintenant que les marges des totaux masqués de comtés correspondent aux totaux marginaux observés, lesquels peuvent toutefois être sensibles. Cependant, on s'attend à ce que la rareté des données pose des difficultés; cette rareté pourrait être évitée si l'on se limite aux principaux produits. Deuxièmement, pour améliorer l'utilité, on peut ajouter une structure spatiale pour les comtés. On peut utiliser un modèle autorégressif conditionnel pour les μ_i de notre modèle, mais cela nécessite une matrice d'incidence pour les comtés de l'État ayant un produit particulier, PRO. Ces deux extensions sont utiles en pratique, mais leur mise en œuvre n'est pas simple. La méthode bayésienne pour petits domaines présentée dans la présente étude constitue un modèle pour des structures de données plus complexes.

À la demande d'un réviseur, pour être concrets, nous montrons la façon de généraliser notre méthode à des données comportant des covariables et quand la variable à l'étude est binaire (par exemple le Bureau du recensement des États-Unis et d'autres organismes s'intéressent également aux pourcentages). Nous donnons une brève description de notre modèle paramétrique avec des covariables, comme le sexe de l'exploitant, la taille de l'exploitation, le régime foncier, la diversité des cultures et l'irrigation, et la variable à l'étude demeure la superficie récoltée. Une autre généralisation aux pourcentages (données binaires) est décrite à l'annexe C. Nous supposons que les covariables sont $\mathbf{x}_{ij}, j = 1, \ldots, n_i, i = 1., \ldots, \ell$, et que la variable à l'étude, y_{ij}, pour les superficies, est pondérée comme nous l'avons indiqué pour les trois modèles. Les $(\mathbf{x}_{ij}, y_{ij})$ sont des microdonnées qui ne sont pas diffusées. De la même façon, nous déterminons les comtés

supprimés et non supprimés (non sensibles et sensibles). Ensuite, pour le modèle de population, nous pouvons supposer que

$$y_{ij} \overset{\text{ind}}{\sim} \text{Normal}\,(\mathbf{x}'_{ij}\boldsymbol{\beta}, \sigma^2), j = 1, \ldots, n_i, i = 1, \ldots, \ell,$$

et pour le modèle d'échantillon,

$$y_{ij} \overset{\text{ind}}{\sim} \text{Normal}\,(\mathbf{x}'_{ij}\boldsymbol{\beta} + (2r_i - 1)\, t_i, \sigma^2),\ j = 1, \ldots, n_i,$$

$$r_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), t_i \overset{\text{ind}}{\sim} \text{Gamma}\,(1, a_{oi}\epsilon), i = 1, \ldots, \ell,$$

$$\epsilon \sim \text{Gamma}\,(1,1), a_o < \epsilon < b_o, \pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2},$$

où les a_o, b_o et les a_{oi} sont spécifiés de la façon décrite dans l'étude. La méthode consistant à compromettre le niveau de sécurité afin d'accroître l'utilité s'applique presque exactement comme ce qui est indiqué dans la présente étude : exécuter l'échantillonneur de Gibbs, effectuer la prédiction et procéder à l'étalonnage.

Enfin, comme l'a indiqué le rédacteur en chef, nous présentons sept mises en garde concernant notre étude. Celles-ci portent principalement sur les tentatives d'établir un compromis entre la sécurité et l'utilité.

1. La composition parallèle de la confidentialité différentielle n'est pas disponible dans l'estimation pour petits domaines. Pour aller de l'avant, nous avons donc utilisé un système de pondération simple pour partitionner le budget total entre les comtés.
2. Il est nécessaire de partitionner les comtés non supprimés en comtés non sensibles et sensibles, et ces derniers nécessitent une protection particulière. Il est avantageux que tous les comtés non supprimés soient ajustés aux modèles, mais seuls les comtés sensibles font l'objet d'un étalonnage. Il pourrait être possible d'apporter un ajustement aux modèles pour les comtés non sensibles en utilisant un terme de décalage.
3. Il est nécessaire de borner les ERA pour obtenir une utilité et une précision acceptables. L'utilisation d'une borne supérieure qui est inférieure à l'unité (une valeur supérieure à 0,25 indique une utilité médiocre) semble raisonnable. Comme nous l'avons mentionné dans l'étude, le compromis entre l'utilité et la sécurité est inévitable.
4. La cible formée par le total des observations lorsque les comtés sensibles sont étalonnés peut elle-même être sensible. Dans ce cas, il est possible d'utiliser une autre procédure de masquage afin d'obtenir une cible masquée, pour laquelle l'ERA serait, par exemple, de 5 %.
5. Bien que nous ayons utilisé les totaux de comtés, qui sont les quantités d'intérêt, il est possible d'effectuer l'analyse à partir des moyennes de comtés (données au niveau des exploitations), lesquelles sont moins sensibles. Nous pouvons alors prédire les moyennes de comtés et, par

conséquent, les totaux de comtés; si le niveau de sécurité et d'utilité est bon, nous pourrions les diffuser.

6. Nous pouvons améliorer la spécification *a priori* de ϵ en empruntant l'intervalle, $(0, b_o)$, construit par Lee et Clifton (2011) à l'aide des données confidentielles observées. Une distribution *a priori* exponentielle tronquée (ou une distribution *a priori* uniforme) pourrait alors être utilisée pour ϵ. La difficulté réside dans le fait qu'il est nécessaire de connaître la probabilité qu'un participant soit identifié dans l'ensemble de données, ce qui n'est pas réalisable en pratique.
7. L'étalonnage par la méthode itérative du quotient (répartition proportionnelle) peut être amélioré à l'aide d'une analyse fondée sur un modèle, mais, faute d'espace, nous ne la présentons pas dans la présente étude.
8. Il est possible d'obtenir une bonne valeur de sensibilité globale à l'aide d'une analyse fondée sur un modèle, mais il n'est pas nécessaire de s'y attarder davantage dans la présente étude.

Avertissement et remerciements

Toutes les analyses présentées dans cet article ont été conduites en utilisant des données simulées conçues à des fins illustratives pour reproduire la structure et les caractéristiques du Recensement de l'agriculture. Aucunes microdonnées confidentielles ou réelles du Recensement de l'agriculture n'ont été utilisées pour cet article. L'ensemble de données simulées représente un état hypothétique ayant 75 comtés et il a été généré uniquement à des fins de démonstration de la méthodologie. Les résultats et conclusions dans cette publication préliminaire n'ont pas été formellement diffusés par le U.S. Department of Agriculture et ne doivent pas être interprétés comme représentant une quelconque décision ou politique de l'agence. Cette recherche a été financée en partie par le programme de recherche interne du U.S. Department of Agriculture (USDA), National Agriculture Statistics Service (NASS). Le travail de Balgobin Nandram a été complété en tant que scientifique principal de la recherche sur un IPA (Intergovernmental Personnel Act) à la NASS. Les auteurs remercient plusieurs réviseurs à l'interne à la NASS (Yang Cheng, Denise Abreu, Valbona Bejleri et Luca Sartore) pour leurs commentaires, ce qui a permis d'obtenir un manuscrit nettement amélioré. Nous remercions également les deux évaluateurs pour leurs commentaires, et le rédacteur en chef pour ses encouragements.

Annexe

A. Mécanisme de confidentialité différentielle de Laplace

Soit ϵ, un nombre réel positif représentant le budget de confidentialité et soit A, un mécanisme randomisé (par exemple une distribution de Laplace) associant l'ensemble de données, D, à $A(D)$. Supposons que D_1 et D_2 diffèrent par un seul élément (une exploitation agricole ou un comté). Dwork (2006) qualifie que A est différentiel ϵ si, pour tous les sous-ensembles S de $A(D)$,

$$P(A(D_1) \in S) \leq e^{\epsilon} P(A(D_2) \in S), \epsilon \geq 0,$$

où $P(\cdot)$ désigne la distribution de randomisation; voir également Dwork et Roth (2014). Par symétrie, il est aussi vrai que

$$P(A(D_2) \in S) \leq e^{\epsilon} P(A(D_1) \in S).$$

Ainsi, en posant $R = \frac{P(A(D_1) \in S)}{P(A(D_2) \in S)}$, on obtient $-\epsilon \leq \log(R) \leq \epsilon$.

Le mécanisme de confidentialité différentielle possède la propriété de composition, qui comprend des compositions parallèles, séquentielles et de groupe. Il est robuste au traitement *a posteriori* (McSherry, 2009) si les données confidentielles observées ne sont pas utilisées lors du traitement *a posteriori*; quant à elle, l'opinion d'experts peut être utilisée. Voir l'examen de Williams et Bowen (2023).

Le mécanisme de Laplace ajoute du bruit provenant d'une distribution de Laplace. Soit Y, le bruit, où $g(y) = \lambda e^{-\lambda|y|}, -\infty < y < \infty$; nous soulignons que $\text{Var}(Y) = \frac{2}{\lambda^2}$. Nous ajoutons y à $f(D)$ pour obtenir

$$z = f(D) + y,$$

où $f(D)$ est une requête possible (peut correspondre aux données d'origine) et où le bruit, y, est $z - f(D)$. Alors, pour toute valeur z,

$$R = \frac{g\{z - f(D_1)\}}{g\{z - f(D_2)\}},$$

et selon le mécanisme de Laplace,

$$R = \exp\left\{\lambda\left\{\left|z - f(D_2)\right| - \left|z - f(D_1)\right|\right\}\right\}.$$

Maintenant, à l'aide de l'inégalité triangulaire inversée,

$$R \leq \exp\left\{\lambda\left\{\left|f(D_2) - f(D_1)\right|\right\}\right\} \leq \exp\{\lambda\Delta(f)\},$$

où $\Delta(f) = \text{Sup}_{D_1, D_2}\left|f(D_2) - f(D_1)\right|$ est la sensibilité de la fonction $f(\cdot)$, et l'on peut interpréter $\lambda\Delta(f)$ comme le paramètre de confidentialité différentielle, ϵ. Par conséquent, il convient de prendre note que pour le mécanisme de Laplace, il faut avoir $\lambda = \frac{\epsilon}{\Delta(f)}$ pour que A soit un algorithme de confidentialité différentielle ϵ. Étant donné que ϵ et λ sont tous deux inconnus, le choix d'une valeur appropriée pour ϵ constitue un problème difficile (par exemple voir Lee et Clifton, 2011), mais il est judicieux d'opter pour une petite valeur de ϵ.

B. Règle de sensibilité du pourcentage p

Considérons un seul comté composé de n exploitations agricoles ayant des superficies, $y_1, \ldots, y_n$, que nous ordonnons maintenant de la plus petite à la plus grande, $y_{(1)} \leq \ldots \leq y_{(n)}$, et supposons que $T = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_{(i)}$. La règle du pourcentage p stipule qu'un comté est sensible si

$$T - (y_{(n-1)} + y_{(n)}) \leq \frac{p}{100} y_{(n)},$$

où $0 < p < 100$. De façon équivalente, un comté est non sensible si $T - (y_{(n-1)} + y_{(n)}) > \frac{p}{100} y_{(n)}$. Cette règle énonce simplement qu'un comté est sensible si $\sum_{i=1}^{n-2} y_{(i)} \leq \frac{p}{100} y_{(n)}$ et non sensible si $\sum_{i=1}^{n-2} y_{(i)} > \frac{p}{100} y_{(n)}$ et $n \geq 3$.

Cette partition des exploitations en deux ensembles montre que si l'exploitation de $y_{(n-1)}$ peut permettre de très bien estimer $y_{(n)}$ et que la somme des comtés restants (pas celui au premier rang ou au deuxième rang en ce qui concerne la grandeur) est faible, le comté sera alors sensible. Un attaquant disposant de sa valeur $y_{(n-1)}$ et de T peut estimer $y_{(n)}$ dans l'intervalle,

$$\frac{T - y_{(n-1)}}{1 + \frac{p}{100}} \leq y_{(n)} \leq T - y_{(n-1)}.$$

C. Pourcentages et régression logistique

Comme nous nous intéressons à des pourcentages, il est naturel d'envisager une régression logistique. Nous désignons les microdonnées par $(w_{ij}, \mathbf{x}_{ij}, y_{ij}), j = 1,\ldots,n_i, i = 1,\ldots,\ell,$ où y_{ij} sont les variables binaires à l'étude. Nous appliquons les poids du recensement aux covariables $\tilde{\mathbf{x}}_{ij} = w_{ij}\mathbf{x}_{ij},$ plutôt qu'aux variables binaires. Il s'agit de la technique utilisée pour la régression logistique pondérée (par exemple Yang, Nandram et Choi, 2023).

La règle du maximum de trois est utilisée de la même façon pour supprimer les comtés comptant moins de trois exploitations agricoles. Nous pouvons maintenant partitionner les comtés non supprimés en comtés non sensibles et sensibles à l'aide des scores de propension issus du modèle,

$$y_{ij} \mid \boldsymbol{\beta} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta}}}{1 + e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta}}}\right), j = 1,\ldots,n_i, i = 1,\ldots,\ell.$$

Nous pouvons calculer l'estimateur du maximum de vraisemblance (identique au mode *a posteriori* selon une distribution *a priori* uniforme pour $\boldsymbol{\beta}$), $\hat{\boldsymbol{\beta}}$. Les scores de propension sont alors

$$\pi_{ij} = \frac{e^{\tilde{\mathbf{x}}_{ij}'\hat{\boldsymbol{\beta}}}}{1 + e^{\tilde{\mathbf{x}}_{ij}'\hat{\boldsymbol{\beta}}}}, j = 1,\ldots,n_i, i = 1,\ldots,\ell.$$

Pour le i^{e} comté, nous appliquons la règle du pourcentage p aux π_{ij} pour déterminer s'il est non sensible ou sensible.

Pour le modèle de population, nous supposons

$$y_{ij} \mid \boldsymbol{\beta} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta}}}{1 + e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta}}}\right), j = 1,\ldots,n_i, i = 1,\ldots,\ell.$$

Pour le modèle d'échantillonnage, nous supposons

$$y_{ij} \mid \boldsymbol{\beta} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta} + (2r_i - 1)t_i}}{1 + e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta} + (2r_i - 1)t_i}}\right), j = 1,\ldots,n_i, i = 1,\ldots,\ell,$$

$$r_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right),\ t_i \overset{\text{ind}}{\sim} \text{Gamma}\,(1, a_{oi}\epsilon), i = 1,\ldots,\ell,$$

où a_{oi} sont les paramètres d'échelle. Ces a_{oi} peuvent être choisis pour être

$$a_{oi} = \frac{n_i / t_{oi}}{\sum_{i'=1}^{\ell} n_{i'} / t_{oi'}}, \quad \text{où} \quad t_{oi} = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}} \quad \text{et} \quad \hat{p}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, i = 1,\ldots,\ell.$$

Enfin, comme dans les modèles présentés dans la présente étude, en supposant l'indépendance de $\boldsymbol{\beta}$ et ϵ, nous supposons

$$\pi(\beta) = 1,\ \epsilon \sim \text{Gamma}\,(1,1), a_o < \epsilon < b_o,$$

où a_o et b_o sont spécifiés. Ce modèle peut être ajusté à l'aide d'un échantillonneur de Gibbs à balayage systématique grâce à l'augmentation de données Pólya-Gamma et des blocs; voir Polson, Scott et Windle (2013).

Pour la prédiction, nous échantillonnons

$$z_{ij} \sim \text{Bernoulli}\left(\frac{e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta}}}{1+e^{\tilde{\mathbf{x}}_{ij}'\boldsymbol{\beta}}}\right), j = 1,\ldots,n_i, i = 1,\ldots,\ell,$$

pour assurer une utilité acceptable, nous utilisons la contrainte (comme dans la présente étude),

$$(1-a)\,\overline{y}_i < \overline{z}_i < (1+a)\,\overline{y}_i,\ 0 < a < 1,\ \overline{z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij}, i = 1,\ldots,\ell,$$

où a est spécifié. Les $\overline{z}_i$ sont les proportions à étalonner et les $\overline{y}_i$ sont les proportions observées. Les proportions de comtés étalonnées (multipliées par 100 pour obtenir des pourcentages) sont alors

$$P_i = \overline{z}_i \frac{\sum_{i=1}^{\ell} n_i \overline{y}_i}{\sum_{i=1}^{\ell} n_i \overline{z}_i}.$$

Bien entendu, nous effectuons en réalité un étalonnage des comtés sensibles en fonction de leur total observé (ce total peut lui-même être sensible; voir le texte).

Bibliographie

Awan, J. et Slavkovic, A. (2019). *Differentially Private Inference for Binomial Data*. Accessible à l'adresse : https://arXiv:1904.00459v1[math.ST], 31 mars 2019, 1-39.

Bowen, C.M. et Liu, F. (2020). Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2), 280-307.

Bring, J. et Wang, Q. (2014). Comparison of different sensitivity rules for tabular data and presenting a new rule – The interval rule, (Éd., J. Domingo-Ferrer): PSD 2014, LNCS 8744, 36-47, Springer International Publishing Switzerland.

Cox, L.H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75(370), 377-385.

Cox, L.H. (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association*, 90, 1453-1462.

Cox, L.H., Kelly, J.P. et Patil, R.J. (2004). Computational aspects of controlled tabular adjustment: Algorithm and analysis. *The Next Wave in Computing, Optimization, and Decision Technologies*, 45-59.

Drechsler, J. (2023). Differential privacy for government agencies – Are we there yet? *Journal of the American Statistical Association*, 118(541), 761-773. https://doi.org/10.1080/01621459.2022.2161385.

Duncan, G.T. et Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393), 10-18.

Duncan, G.T. et Lambert, D. (1989). The risk of disclosure of microdata. *Journal of Business and Economic Statistics*, 7(2), 207-217.

Dwork, C. (2006). Differential privacy. *Microsoft Research*, 1-12.

Dwork, C., McSherry, F., Nissim, K. et Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Dans *Theory of Cryptography Conference (TCC),* (Éds., S. Halevi et T. Rabin), Springer, 265-284.

Dwork, C., Naor, M., Reingold, O., Rothblum, G. et Vadhan, S. (2009). On the complexity of differentially private data release: Efficient algorithms and hardness results. *Proceedings of the forty-first annual ACM Symposium on Theory of Computing*, 381-390. https://doi.org/10.1145/1536414.1536467.

Dwork, C. et Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407. https://doi.org/10.1561/0400000042.

Evans, T., Zayata, L. et Slanta, J. (1996). Using noise for disclosure limitation of establishment tabular data. Proceedings: *Annual Research Conference and Technology Interchange*, Bureau of the Census, 65-68.

Geweke, J. (1992). Evaluating the accuracy of the sample-based approach to calculation of posterior moments. Dans *Bayesian Statistics*, (Éds., J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith), 4, 169-193.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B.C. et Roth, A. (2014). Differential privacy: An economic method for choosing epsilon. Accessible à l'adresse : https://arxiv.org/abs/1402.3329v1 et https://do1.org/10.48550/arXiv.1402.332, 13 février 2014, 1-29.

Hu, J. et Bowen, C.M. (2024). Advancing microdata privacy protection: A review of synthetic data methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1), e1636, 1-19.

Hu, J., Williams, M.R. et Savitsky, T.D. (2025). Mechanisms for global differential privacy under Bayesian data synthesis. *Statistica Sinica*, 35(1), 563-584.

Ishwaran, H. et James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 161-172.

Janicki, R., Holan, S.H., Irimata, K.M., Livsey, J. et Raim, A. (2024). *Bayesian Methods to Improve the Accuracy of Differentially Private Measurements of Constrained Parameters*. Accessible à l'adresse : https://arXiv:2406.04448v1[Stat.ME], 6 juin 2025, 1-7.

Lee, J. et Clifton, C. (2011). How much is enough? Choosing ϵ for differential privacy. *Proceedings of the International Conference on Information Security*, Springer, Berlin, Heidelberg, 325-340.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9(2), 407-426.

Manandhar, B. et Nandram, B. (2021). Hierarchical Bayesian models for continuous and positively skewed data from small areas. *Communications in Statistics-Theory and Methods*, 50(4), 944-962.

Manandhar, B. et Nandram, B. (2025). Hierarchical Bayesian models for small area estimation with GB2 distribution. *Journal of Applied Statistics*, 52(1), 1-30.

McSherry, F.D. (2009). Privacy integrated queries: An extensible platform for privacy-preserving data analysis. Dans *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/1559845.1559850.

Nandram, B., Cruze, N.B. et Erciulescu, A.L. (2023). Modèles bayésiens pour petits domaines sous contraintes d'inégalité avec réconciliation et rétrécissement double. *Techniques d'enquête*, 49(2), 559-591. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00004-fra.pdf.

Nandram, B., Toto, M.C.S. et Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, 81(11), 1593-1608.

Nandram, B. et Yu, Y. (2019). Bayesian analysis of a sensitive proportion for a small area. *Revue Internationale de Statistique*, 87, S1, S104-S120. https://doi.org/10.1111/insr.12286.

Polson, N.G., Scott, J.G. et Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association,* 108(504), 1339-1349.

Poole, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69(348), 1002-1005.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1), 1-16.

Reiter, J.P. (2019). Differential privacy and federal data releases. *The Annual Review of Statistics and its Application*, 6, 85-101.

Rinott, Y., O'Keefe, C.M., Shlomo, N. et Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, 33(3), 358-385.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2), 461-468.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Snoke, J., Raab, G.M., Nowok, B., Dibben, C. et Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society*, Series A, 181(3), 663-688.

Tan, A. et Hobert, J.P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *Journal of Computational and Graphical Statistics*, 18(4), 861-878.

Williams, A.R. et Bowen, C.M. (2023). The promise and limitations of formal privacy. *Computational Statistics*, 15(2), e1615, 1-17.

Yang, L., Nandram, B. et Choi, J.W. (2023). Bayesian predictive inference under nine methods for incorporating survey weights. *International Journal of Statistics and Probability*, 12(1), 33-53.

Zhang, Z., Rubinstein, B.I.P. et Dimitrakakis, C. (2016). On the differential privacy of Bayesian inference. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Association for the Advancement of Artificial Intelligence* (Accessible à l'adresse : https://www.aaai.org), 2365-2371.