Nº 11-522-X au catalogue
ISSN 1709-822X

**Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société**

# Méthode de désagrégation fondée sur un modèle pour l'estimation des compétences des adultes


par Andreea L. Erciulescu, Weijia Ren, Jianzhu Li, Tom Krenzke, Leyla Mohadjer et Robert Fay

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

# Méthode de désagrégation fondée sur un modèle pour l'estimation des compétences des adultes

Andreea L. Erciulescu, Weijia Ren, Jianzhu Li, Tom Krenzke, Leyla Mohadjer et Robert Fay[1]

## Résumé

L'estimation à de fins niveaux d'agrégation est nécessaire pour mieux décrire une société. Les approches fondées sur un modèle d'estimation sur petits domaines qui combinent des données d'enquête parcimonieuses à des données riches provenant de sources auxiliaires se sont révélées utiles pour améliorer la fiabilité des estimations sur petits domaines. Nous examinons ici un scénario où des estimations basées sur un modèle pour petits domaines, produit à un niveau d'agrégation donné, devaient être désagrégées pour mieux décrire la structure sociale à des niveaux plus fins. Pour ce scénario, nous avons élaboré une méthode de répartition afin de mettre en œuvre la désagrégation, surmontant les problèmes associés à la disponibilité des données et à l'élaboration de modèles à des niveaux de cette finesse. La méthode est appliquée à l'estimation de la littératie et de la numératie des adultes au niveau du comté par groupe, au moyen des données du Programme pour l'évaluation internationale des compétences des adultes (PEICA) des États-Unis. Dans cette application, les groupes sont définis en fonction de l'âge ou de la scolarité, mais la méthode pourrait être appliquée à l'estimation d'autres groupes en quête d'équité.

Mots clés : répartition; compétences des adultes; statistiques officielles; Programme pour l'évaluation internationale des compétences des adultes; estimation sur petits domaines.

## 1. Introduction

L'élaboration de politiques et de lois efficaces fondées sur des données probantes en matière d'éducation des adultes exige des recherches solides s'appuyant sur les données fiables les plus pertinentes pour des administrations comme les comtés, les États et les groupes démographiques au sein des comtés et des États. Dans le cadre d'une étude internationale multicycle à laquelle participent plus de 30 pays sous la direction de l'Organisation de coopération et de développement économiques, le premier cycle du Programme pour l'évaluation internationale des compétences des adultes (PEICA) a été conçu pour fournir des estimations nationales des compétences des adultes en littératie, en numératie et en résolution de problèmes. L'enquête du PEICA fournit des estimations nationales de grande qualité au moyen d'un plan de sondage probabiliste à plusieurs degrés avec des collectes de données en personne comprenant un questionnaire de sélection, un questionnaire de base et une évaluation des compétences des adultes. Aux États-Unis, le PEICA est parrainé par le National Center for Education Statistics de l'Institute of Education Sciences. De 2012 à 2017, environ 12 330 adultes américains, âgés de 16 à 74 ans et vivant dans des ménages, ont été interrogés dans le cadre du PEICA des États-Unis. Dans la suite de l'article, nous désignerons le PEICA des États-Unis simplement par PEICA.

Parce que la taille de l'échantillon du PEICA était trop petite pour appuyer la production d'estimations à des niveaux désagrégés, Krenzke et coll. (2020) et Li et coll. (2022) ont élaboré une méthodologie d'estimation sur petits domaines (EPD) fondée sur un modèle bayésien hiérarchique pour produire des estimations par comté, par État et par État selon l'âge et le groupe de scolarité pour les scores moyens en littératie et en numératie, et divers niveaux de compétence. En utilisant les données de l'enquête du PEICA conjointement avec les données de l'American Community Survey

[1]Andreea L. Erciulescu, Westat, 1600 Research Blvd, Rockville, MD, États-Unis, 20850 (AndreeaErciulescu@westat.com); Weijia Ren; Meta, États-Unis; Jianzhu Li, FINRA, 1735 K Street, NW, Washington DC, États-Unis, 20006; Tom Krenzke, Westat, 1600 Research Blvd, Rockville, MD, États-Unis, 20850; Leyla Mohadjer, Westat, 1600 Research Blvd, Rockville, MD, États-Unis, 20850; Robert Fay, Westat, 1600 Research Blvd, Rockville, MD, États-Unis, 20850. Les travaux du Weijia Ren et Jianzhu Li ont été menés pendant qu'ils travaillaient chez Westat.

(ACS, Enquête sur les collectivités américaines), les estimations EPD fournissent des statistiques officielles fiables des États-Unis sur la littératie et la numératie des adultes dans les 50 États, les 3 141 comtés et le District de Columbia, et dans les 50 États et dans le District de Columbia selon six groupes d'âge (16 à 24 ans, 25 à 34 ans, 35 à 44 ans, 45 à 54 ans, 55 à 64 ans et 65 à 74 ans) et selon quatre groupes d'éducation, à savoir : n'ayant pas terminé ses études secondaires, n'ayant pas obtenu de diplôme d'études secondaires ou de formation générale (GED), n'ayant pas terminé ses études collégiales (n'ayant pas obtenu de diplôme collégial ni de diplôme connexe), et titulaire d'un baccalauréat ou d'un diplôme supérieur. Ces estimations sont disponibles dans la carte des compétences à l'adresse https://nces.ed.gov/surveys/piaac/skillsmap/ pour tous les résultats d'intérêt, les proportions égales ou inférieures au niveau 1 (P1), au niveau 2 (P2) et au niveau 3 (P3) ou plus, et les moyennes.

Le présent article porte sur l'estimation par groupe des compétences des adultes, où les quantités d'intérêt sont les mêmes que celles étudiées dans Krenzke et coll. (2020) et Li et coll. (2022), et les groupes d'intérêt sont les mêmes que ceux étudiés dans Li et coll. (2022). Une méthode déterministe sert à répartir les estimations d'État par groupe au comté par niveau de groupe, pour tous les comtés imbriqués dans l'État correspondant, au moyen des ratios des estimations de comté à État. La répartition est appliquée au niveau de l'échantillon a posteriori, ce qui donne des pseudo-distributions a posteriori pour les quantités d'intérêt.

Le plan de l'article est le suivant. À la section 2, nous décrivons les données disponibles provenant du PEICA, de la carte des compétences et de l'ACS, qui servent de données d'entrée dans le processus d'estimation au niveau du comté par niveau de groupe. La méthode de répartition est présentée à la section 3 et les mesures de validation sont fournies à la section 4. Un résumé est donné dans la section 5. À moins d'indication contraire, les résultats choisis sont présentés dans le manuscrit seulement pour une quantité d'intérêt, soit la proportion de littératie de niveau 1 ou moins, et pour un groupe démographique, la population qui n'a pas terminé ses études secondaires. Les résultats pour toutes les quantités d'intérêt et pour tous les domaines d'intérêt sont comparables et sont disponibles dans la carte des compétences et dans le rapport d'Erciulescu et coll. (2022).

## 2. Données disponibles à différents niveaux d'agrégation

Les données d'enquête au niveau de la personne sont tirées du PEICA et servent à produire des estimations d'enquête pour le comté par domaines de groupe avec des données d'échantillonnage. Comme ces données sont parcimonieuses à des niveaux désagrégés, comme comté par groupe, une méthode d'estimation indirecte est élaborée. La méthode de l'estimation indirecte utilise les données produites comme données de sortie des modèles d'EPD pour des niveaux d'agrégation plus élevés, c.-à-d. comté, État et État par groupe. Les estimations d'enquête et certaines données auxiliaires sont utilisées dans le processus de validation des estimations indirectes du comté par groupe. Des détails sur l'enquête, le modèle et les données auxiliaires du PEICA sont donnés dans la suite de la section.

### 2.1 Estimations d'enquête

D'après les conclusions de Li et coll. (2022), la méthode d'imputation multiple est mise en œuvre pour construire les estimations d'enquête et les estimations de la variance correspondantes. Les données du PEICA à un microniveau comprennent 10 valeurs plausibles générées à partir d'une distribution a posteriori par la combinaison d'une mise à l'échelle selon la théorie de la réponse d'item pour les items cognitifs avec un modèle de régression latente au moyen de l'information du questionnaire de base dans un modèle de population. Pour chaque valeur plausible, les estimations de type Háyek et les estimations de la variance associées approximées au moyen des séries de Taylor sont établies pour le comté selon les quantités d'intérêt au niveau du groupe, les poids d'enquête étant calés sur les totaux de contrôle de l'ACS de 2013-2017 pour les groupes d'âge, les niveaux de scolarité, le sexe, et la race ou l'origine ethnique au sein de l'État. Ensuite, on calcule la moyenne des estimations pour les 10 valeurs plausibles afin de produire les estimations d'enquête au niveau du comté par niveau de groupe. On estime les variances des estimations du comté par niveau de groupe au moyen de la méthode de l'imputation multiple en combinant les variances intra-imputation et entre imputations.

Les données du PEICA comprennent 185 comtés comptant au moins un répondant dans chaque comté, mais certains comtés n'ont pas de répondant pour un ou plusieurs groupes d'âge ou de scolarité. Le tableau 2.1-1 présente le nombre de domaines de comté par groupe avec données d'échantillonnage et les distributions des tailles d'échantillon des domaines de comté par groupe. En raison du petit nombre d'estimations d'enquête disponibles au niveau du comté

pour les groupes, il faut une méthode d'estimation indirecte afin de produire des estimations pour tous les domaines du comté selon le groupe d'âge/de scolarité. Les médianes des tailles d'échantillon de domaine de comté par groupe varient de 6,5 à 16, ce qui indique que les estimations de l'enquête sont sujettes à une grande incertitude dans la plupart des domaines de comté par groupe. En plus de produire des estimations pour tous les domaines d'intérêt, la méthode d'estimation indirecte décrite dans la section suivante aide à réduire l'incertitude des estimations de l'enquête.

## 2.2 Estimations fondées sur un modèle

Des distributions a posteriori pour les quantités d'intérêt des compétences des adultes au niveau du comté, de l'État et de l'État par groupe ont été produites par Krenzke et coll. (2020) et Li et coll. (2022). Bien que seules les estimations ponctuelles, avec les estimations de la variance correspondantes et les intervalles de confiance, soient téléchargeables directement à partir de la carte des compétences, 4 500 échantillons a posteriori pour les quantités de compétence d'intérêt des adultes sont également disponibles sur demande. Nous travaillons à partir de ces échantillons a posteriori pour construire les estimations de comté par niveau de groupe.

**Tableau 2.1-1**
**Distributions de la taille d'échantillon au niveau du comté pour les groupes d'âge et de scolarité : PEICA 2012/2014/2017**

| Groupes d'âge et de scolarité | Nombre de comtés avec échantillon | Nombre de répondants | | | | | |
|---|---|---|---|---|---|---|---|
| | | Minimum | 10e centile | Médiane | 90e centile | Maximum | Moyenne |
| 16-24 | 172 | 1 | 2 | 10 | 31 | 82 | 15 |
| 25-34 | 177 | 1 | 3 | 11 | 33 | 96 | 16 |
| 35-44 | 180 | 1 | 2 | 9,5 | 21,5 | 56 | 11 |
| 45-54 | 178 | 1 | 2 | 10 | 21 | 72 | 11 |
| 55-64 | 180 | 1 | 2 | 10 | 18 | 45 | 10 |
| 65-74 | 176 | 1 | 1 | 6,5 | 14 | 31 | 7 |
| Niveau inférieur aux études secondaires | 175 | 1 | 2 | 8 | 25 | 67 | 12 |
| Diplôme d'études secondaires ou de formation générale (GED) | 180 | 1 | 4 | 16 | 42,5 | 86 | 20 |
| Études collégiales non terminées | 182 | 1 | 3 | 16 | 38 | 114 | 19 |
| Baccalauréat ou diplôme de niveau supérieur | 179 | 1 | 3 | 13 | 36 | 115 | 18 |

Soit $Y_j^{(b)}, Y_{jk}^{(b)}$ et $Y_{g,j}^{(b)}$ qui désignent les échantillons a posteriori respectivement au niveau de l'État, du comté et de l'État par niveau de groupe. La variable $Y$ désigne l'une des quantités d'intérêt, P1, P2, P3, ou la moyenne, $j$ est un indice pour les États, $k$ est un indice pour les comtés, $g$ est un indice pour les groupes et $b$ est un indice pour les échantillons a posteriori. Ces échantillons sont disponibles pour tous les comtés et États, et pour tous les groupes d'âge et de scolarité d'intérêt. La qualité des estimations de comté par niveau de groupe est dictée par la qualité des estimations au niveau du comté, de l'État et de l'État par niveau de groupe, puisque les dernières sont les seules à entrer dans le modèle utilisé pour produire les précédentes, comme cela est décrit dans la section suivante.

## 2.3 Données auxiliaires

Deux variables de l'ACS liées aux compétences des adultes sont maintenant sélectionnées dans la validation des estimations de comté par niveau de groupe. Elles sont disponibles à partir du bassin initial de variables examiné par Krenzke et coll. (2020) pour les modèles d'estimation sur petits domaines. La première variable est la proportion de

la population sous le seuil de pauvreté et la deuxième est la proportion de la population ayant un emploi. Ces variables sont disponibles pour tous les domaines d'intérêt de comtés par groupe.

## 3. Méthode de répartition

Les estimations au niveau du comté pour les groupes sont établies au moyen d'une méthode de répartition. Cette méthode consiste à utiliser les estimations basées sur un modèle au niveau du comté et de l'État par niveau de groupe et à les attribuer aux domaines d'intérêt de comté par groupe au moyen d'un modèle déterministe. À l'instar du processus de modélisation de l'EPD élaboré aux fins d'estimation du comté, de l'État et de l'État selon le groupe d'âge/de scolarité, on construit d'abord les estimations fondées sur un modèle pour P1 et P3, puis on soustrait leur somme de 1 pour obtenir les estimations fondées sur un modèle pour P2. De plus, les mesures de la littératie et de la numératie sont estimées indépendamment. Les données de sortie du modèle des modèles d'EPD élaborées aux fins d'estimation des mesures des compétences par comté, par État et par État selon le groupe d'âge/de scolarité servent de données d'entrée dans le processus d'estimation de comté par groupe.

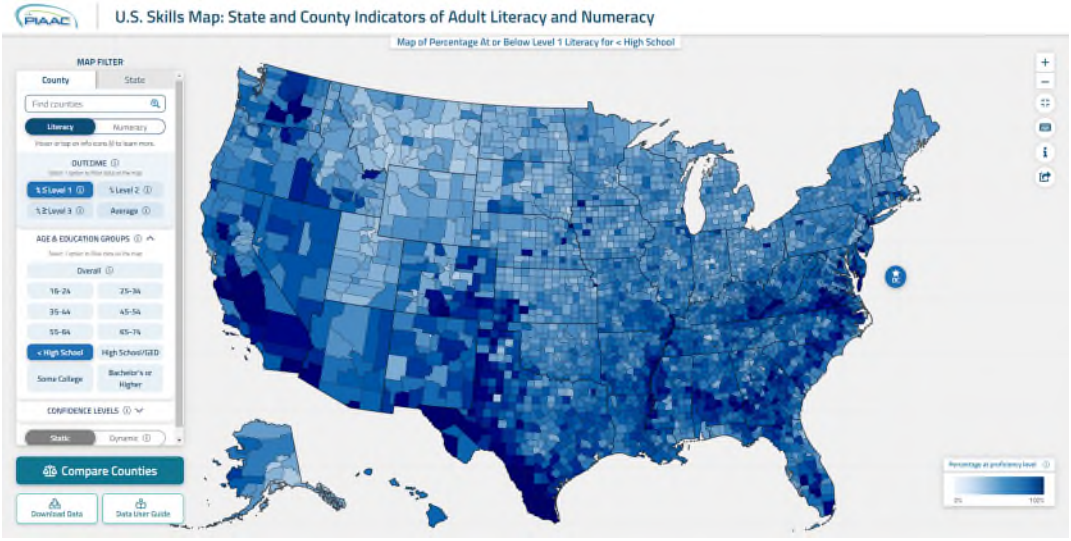
En suivant la notation de la section précédente, supposons que les pseudo-échantillons a posteriori pour les quantités d'intérêt de comté par niveau de groupe soient définis comme suit :

$$Y_{g,jk}^{(b)} := Y_{g,j}^{(b)} \frac{Y_{jk}^{(b)}}{Y_{j}^{(b)}}, \tag{1}$$

avec une note indiquant que $Y$ désigne maintenant P1, P3 ou la moyenne. L'hypothèse de travail implicite est que le ratio comté-État des échantillons a posteriori est constant pour tous les groupes d'âge et de scolarité. Les pseudo-échantillons a posteriori pour le comté par niveau de groupe P2 sont ensuite définis comme étant $1 - P1_{g,jk}^{(b)} - P3_{g,jk}^{(b)}$.Des résumés a posteriori, comme les moyennes, les variances et les intervalles de confiance, sont construits au moyen des pseudo-distributions postérieures décrites ci-dessus. Les moyennes a posteriori pour P1, P2 et P3 inférieures à 0 ou supérieures à 1 sont respectivement réglées à 0 ou 1.

Parmi les 175 comtés ayant des données d'échantillonnage pour le groupe des personnes sans diplôme d'études secondaires, il y avait 3 estimations de la littératie fondées sur le modèle P1 supérieures à 1. Parmi les 2 967 comtés sans données d'échantillonnage pour le groupe des personnes sans diplôme d'études secondaires, il y avait 42 estimations de la littératie fondées sur le modèle P1 supérieures à 1. Aucune des estimations fondées sur le modèle P1 au niveau du comté pour le groupe sans diplôme d'études secondaires n'était inférieure à 0. Aucun ajustement n'a été nécessaire pour les estimations fondées sur un modèle de score moyen, car elles étaient toutes positives. La carte de la figure 3-1 illustre les estimations de comté par niveau de groupe basées sur le modèle de littératie P1 pour la population sans diplôme d'études secondaires.

**Figure 3-1**
**Carte des estimations basées sur un modèle au niveau du comté pour la proportion de personnes se situant au niveau 1 ou sous le niveau 1 en littératie pour le groupe des personnes sans diplôme d'études secondaires (en pourcentage)**




## 4. Validation des estimations finales

Les estimations d'enquête construites pour le comté par groupes d'âge/de scolarité servent de données d'entrée dans le processus de validation, de même que les estimations sélectionnées de l'ACS liées à la compétence qui sont décrites dans la section 2. Trois mesures de validation sont considérées : (1) des affichages visuels de l'ampleur et de la direction des estimations d'enquête aux estimations basées sur un modèle, (2) des affichages visuels de la relation entre les estimations des compétences et les variables auxiliaires de l'ACS, et (3) des mesures de l'incertitude pour les estimations basées sur un modèle.

Les graphiques de rétrécissement de la figure 4-1 montrent l'ampleur et la direction des estimations d'enquête par rapport aux estimations basées sur un modèle des proportions de littératie pour le groupe des personnes sans diplôme d'études secondaires, par taille d'échantillon. Les flèches courtes correspondent à de petites différences entre les estimations d'enquête et les estimations basées sur un modèle, tandis que les flèches longues correspondent à de grandes différences entre les estimations d'enquête et les estimations basées sur un modèle. Le rétrécissement est plus important dans les domaines ayant de plus petites tailles d'échantillon que dans ceux à grandes tailles d'échantillon. Les flèches pointant vers le haut correspondent aux différences négatives entre les estimations d'enquête et les estimations basées sur un modèle, tandis que les flèches pointant vers le bas correspondent aux différences positives entre les estimations d'enquête et les estimations basées sur un modèle.

Les diagrammes de dispersion des estimations de la littératie P1 pour le groupe des personnes sans diplôme d'études secondaires sont représentés dans la figure 4-2, par rapport aux variables sélectionnées de l'ACS (proportion de la population sous le seuil de pauvreté, proportion de la population ayant un emploi). En général, on observe une fourchette semblable des estimations d'enquête et des estimations fondées sur un modèle ainsi qu'une relation semblable entre les variables de l'ACS et les estimations, cette relation étant plus claire entre les estimations basées sur un modèle et les variables de l'ACS qu'entre les estimations d'enquête et les variables de l'ACS. Ce résultat est attendu parce que les estimations basées sur un modèle de comté par groupe sont des fonctions des estimations fondées sur un modèle de comté, d'État et d'État par groupe, qui sont elles-mêmes des fonctions des variables de l'ACS examinées ici ou qui y sont liées.

**Figure 4-1**
**Proportion de littératie (sans diplôme d'études secondaires) – Graphiques de rétrécissement des estimations ponctuelles selon la taille de l'échantillon**

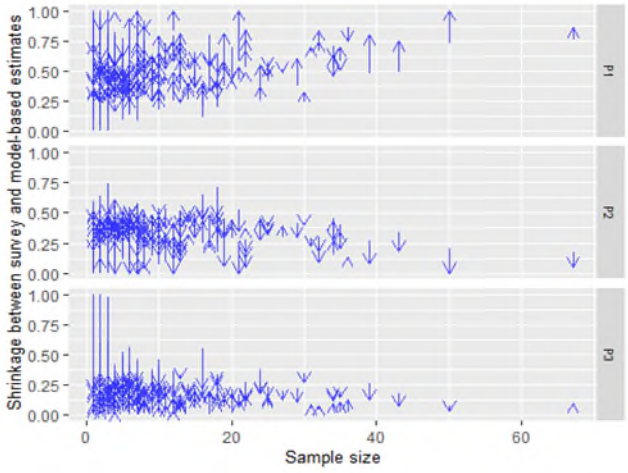
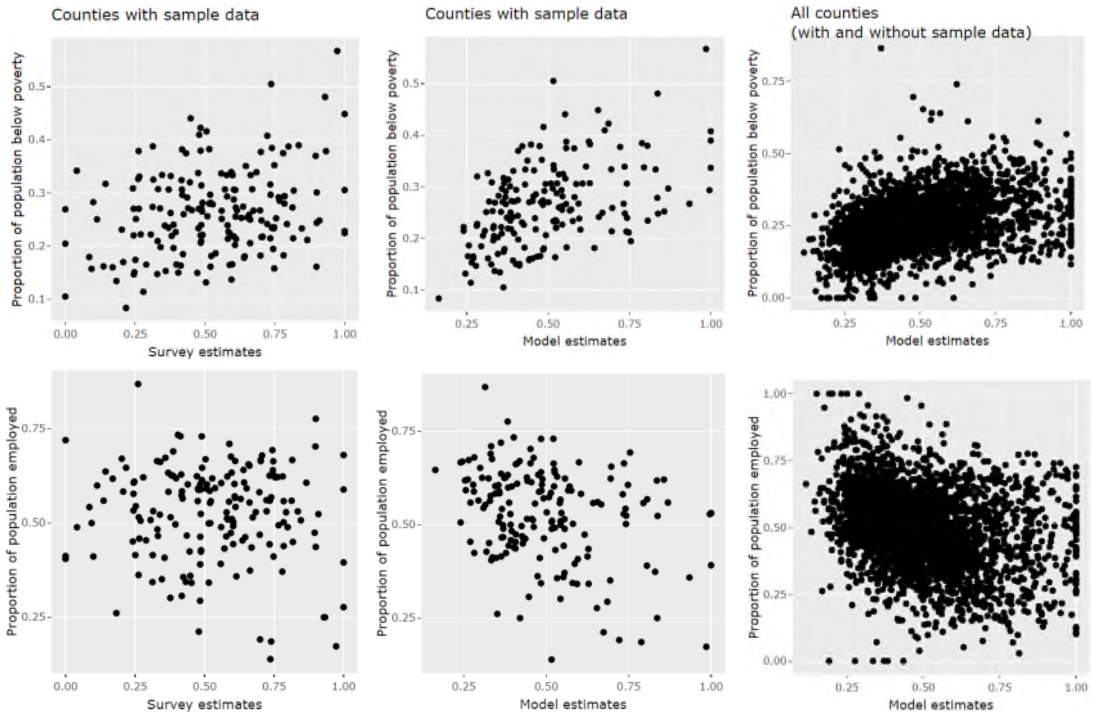



**Figure 4-2**
**Proportion de littératie égale ou inférieure au niveau 1 pour le groupe de personnes sans diplôme d'études secondaires par rapport aux variables sélectionnées de l'ACS – estimations du comté par niveau de groupe : PEICA 2012/2014/2017**




Comme l'indique le tableau 4-1, la largeur médiane des intervalles de confiance est de 24,5 % pour les estimations au niveau du comté de la littératie P1 pour le groupe sans diplôme d'études secondaires, où la médiane est prise sur l'ensemble du comté par domaine de groupe aux États-Unis. Lorsque ces domaines de comté par groupe sont classés par disponibilité de l'échantillon, les largeurs médianes des intervalles de confiance sont de 22,7 % et de 24,7 % respectivement pour les domaines dans l'échantillon et les domaines hors échantillon.

Les coefficients de variation (c.v.) pour les estimations basées sur le modèle de littératie P1 au niveau du comté pour le groupe sans diplôme d'études secondaires sont également résumés dans le tableau 4-1 par disponibilité d'échantillon. Pour la plupart de ces domaines de comté par groupe, les c.v. sont inférieurs à 20 %. Tous les c.v. indiqués dans le tableau 4-1 pour les domaines dans l'échantillon sont inférieurs à la moitié des c.v. correspondants pour les estimations d'enquête indiquées dans le tableau 2-2 d'Erciulescu et coll. (2022). Ce résultat est attendu parce que l'incertitude dans les estimations basées sur un modèle au niveau du comté, de l'État et de l'État par groupe utilisées comme données d'entrée dans la méthode de répartition était inférieure à l'incertitude dans les estimations d'enquête correspondantes (voir Krenzke et coll. (2020) et Li et coll. (2022)). De plus, les c.v. indiqués dans le tableau 4-1 sont plus importants que les c.v. pour les estimations de la littératie P1 basées sur un modèle au niveau de l'État pour le groupe de scolarité sans diplôme d'études secondaires, indiquées dans le tableau 3-5 de Li et coll. (2022) : les estimations au niveau de l'État par groupe ont un c.v. médian de 11,8 %, comparativement à 12,9 % pour les estimations au niveau du comté par groupe. Ce résultat est attendu parce que le niveau d'agrégation dans le présent article est plus fin que le niveau d'agrégation dans le rapport cité (comté par groupe au lieu d'État par groupe).

**Tableau 4-1**
**Distribution des largeurs d'intervalle de confiance et des coefficients de variation pour les estimations basées sur un modèle au niveau du comté pour la population sans diplôme d'études secondaires pour une proportion de littératie égale ou inférieure au niveau 1 : PEICA 2012/2014/2017**

| | Centile | | | | |
|---|---|---|---|---|---|
| Statistiques pour la population sans diplôme d'études secondaires | 20 | 40 | 50 (Médiane) | 60 | 80 |
| Estimations de comté pour tous les domaines | | | | | |
| Largeur d'intervalle de confiance de 95 % (pourcentage) | 21,1 | 23,4 | 24,5 | 25,9 | 29,4 |
| Coefficient de variation (pourcentage) | 10,7 | 12,1 | 12,9 | 14,0 | 17,7 |
| Estimations de comté pour les domaines dans l'échantillon | | | | | |
| Largeur d'intervalle de confiance de 95 % (pourcentage) | 19,6 | 21,5 | 22,7 | 24,0 | 27,9 |
| Coefficient de variation (pourcentage) | 10,3 | 11,6 | 12,3 | 12,8 | 16,3 |
| Estimations de comté pour les domaines hors échantillon | | | | | |
| Largeur d'intervalle de confiance de 95 % (pourcentage) | 21,2 | 23,5 | 24,7 | 25,9 | 29,5 |
| Coefficient de variation (pourcentage) | 10,7 | 12,1 | 13,0 | 14,1 | 17,7 |

## 5. Résumé

Nous avons élaboré une méthode de répartition pour produire des estimations basées sur un modèle au niveau du comté pour six groupes d'âge et quatre groupes de niveau de scolarité. Les estimations sur petits domaines basées sur un modèle antérieures, établies au niveau du comté, de l'État et de l'État par groupe, ont servi de données d'entrée dans la méthode de répartition. Les estimations calculées sont disponibles dans la carte des compétences.

## Bibliographie

Krenzke, T., Mohadjer, L., Li, J., Erciulescu, A., Fay, R., Ren, W., Van de Kerckhove, W., Li, L., et J.N.K. Rao (2020), « Program for the International Assessment of Adult Competencies (PIAAC): State and County Estimation Methodology Report (NCES 2020-225). », U.S. Department of Education, National Center for Education Statistics, Washington, DC: U.S. Government Printing Office. Extrait le 26 octobre 2020 à partir du site : https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020225.

Li, J., Krenzke, T., Ren, W., Mohadjer, L., Fay, R. et A. Erciulescu (2022), « Program for the International Assessment of Adult Competencies (PIAAC): State-level Estimation for Age and Education Groups Methodology Report (NCES 2022-050). » U.S. Department of Education, National Center for Education Statistics, Washington, DC: U.S. Government Printing Office. Extrait le 2 septembre 2022 à partir du site : https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2022050.

Erciulescu, A., Ren, W., Li, J., Krenzke, T., Mohadjer, L. et R. Fay (2022), « Program for the International Assessment of Adult Competencies (PIAAC): County-Level Estimation for Age and Education Groups Methodology Report (NCES 2023-004). » U.S. Department of Education, National Center for Education Statistics, Washington, DC: U.S. Government Publishing Office. Extrait le 31 octobre 2022 à partir du site : https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2023004.