Institut de la statistique Québec

Contributions de Louis-Paul Rivest aux travaux menés à l'Institut de la statistique du Québec

La statistique au service de la collectivité

Hommage à Louis-Paul Rivest

29 août 2014

Éric Gagnon

Institut de la statistique du Québec

Institut de la statistique Québec

Plan de la présentation

- Introduction
- Études de couverture du recensement canadien
- Stratification dans les enquêtes-entreprises
- Estimation pour petits domaines
- Conclusion

Institut de la statistique Québec

Introduction

- Liens de l'Institut avec Louis-Paul Rivest:
  - Direction de la méthodologie et de la qualité de l'Institut compte une vingtaine de statisticiens:
    - une quinzaine de l'Université Laval
  - Majorité de l'équipe a bénéficié de plusieurs cours avec Louis-Paul:
    - notamment en échantillonnage

Institut de la statistique Québec

Introduction (suite)

- Liens (suite):
  - Directeur de maîtrise de deux collègues
  - Préparation d'examen pour concours de recrutement
  - Colloque francophone sur les sondages 2005 à Québec:
    - Organisé conjointement par l'Université Laval et l'Institut
    - Président du comité scientifique: Louis-Paul Rivest

Institut de la statistique Québec

Études de couverture du recensement

- Estimations de population de 1991:
  - Incorpore pour la première fois des études de couverture du recensement
- À partir des données provisoires des études:
  - Ministère des finances a réalisé des études d'impact
  - Perte de plus de 100 millions par année en péréquation pour le Québec
- Bureau de la statistique du Québec:
  - A formé un groupe d'experts en démo. et en statistique
  - Trouver le meilleur estimateur de la population tenant compte des études de couverture

Institut de la statistique Québec

Études de couverture (suite)

- Travaux de ce groupe ont mené à la proposition d'un estimateur composite de la sous-couverture (Rivest,1995) inspiré de celui de Royce (1992):

$$\hat{\theta}_{comp_prov_i} = (\alpha)\hat{\theta}_{dir_prov_i} + (1-\alpha)\hat{\theta}_{dir_can}(w_{prov_i})$$

- Rivest a développé un estimateur de α
- Autres travaux de Rivest et Belmonte (2000) utilisant les études de couverture:
  - Lien avec les estimateurs pour petits domaines.

Institut de la statistique Québec

Stratification dans les enquêtes

- L'Institut réalise plusieurs enquêtes auprès d'entreprises
- Avant de procéder à l'échantillonnage d'entreprises, on doit procéder à la stratification de la population:
  - Avoir des unités homogènes à l'intérieur d'une strate
  - S'assurer d'avoir toutes les entreprises importantes dans l'échantillon (strate à tirage complet)
- But: obtenir une stratification optimale permettant d'obtenir la plus petite taille d'échantillon pour une précision désirée.

Institut de la statistique Québec

Stratification (suite)

- Important d'avoir une variable auxiliaire corrélée à la variable mesurée lors de l'enquête:
  - Variable disponible pour toute la population
    - Ex: nombre d'employés de l'entreprise
- Lavallée et Hidiroglou (1988) ont développé un algorithme permettant d'obtenir une taille d'échantillon optimale dans cette situation.
  - Utilisé pendant plusieurs années à l'Institut

Institut de la statistique Québec

Stratification (suite)

- Baillargeon, Rivest et Ferland (2007) ont suggéré une méthode généralisée de Lavallée-Hidiroglou:
  - Tient compte d'une corrélation imparfaite entre la variable auxiliaire et la variable mesurée
  - Tient compte de la variabilité des taux de réponse entre les strates
  - Tient compte de la mortalité des entreprises
  - Donne la possibilité d'avoir une strate à tirage nul

Institut de la statistique Québec

Stratification (suite)

- Baillargeon et Rivest (2011) ont développé le «package» stratification en R qui incorpore plusieurs méthodes de stratification.

- L'Institut a utilisé ce «package» pour résoudre différents problèmes de stratification dans ces enquêtes notamment:
  - pour l'enquête sur la vente de livres neufs

Institut de la statistique Québec

Estimation pour petits domaines

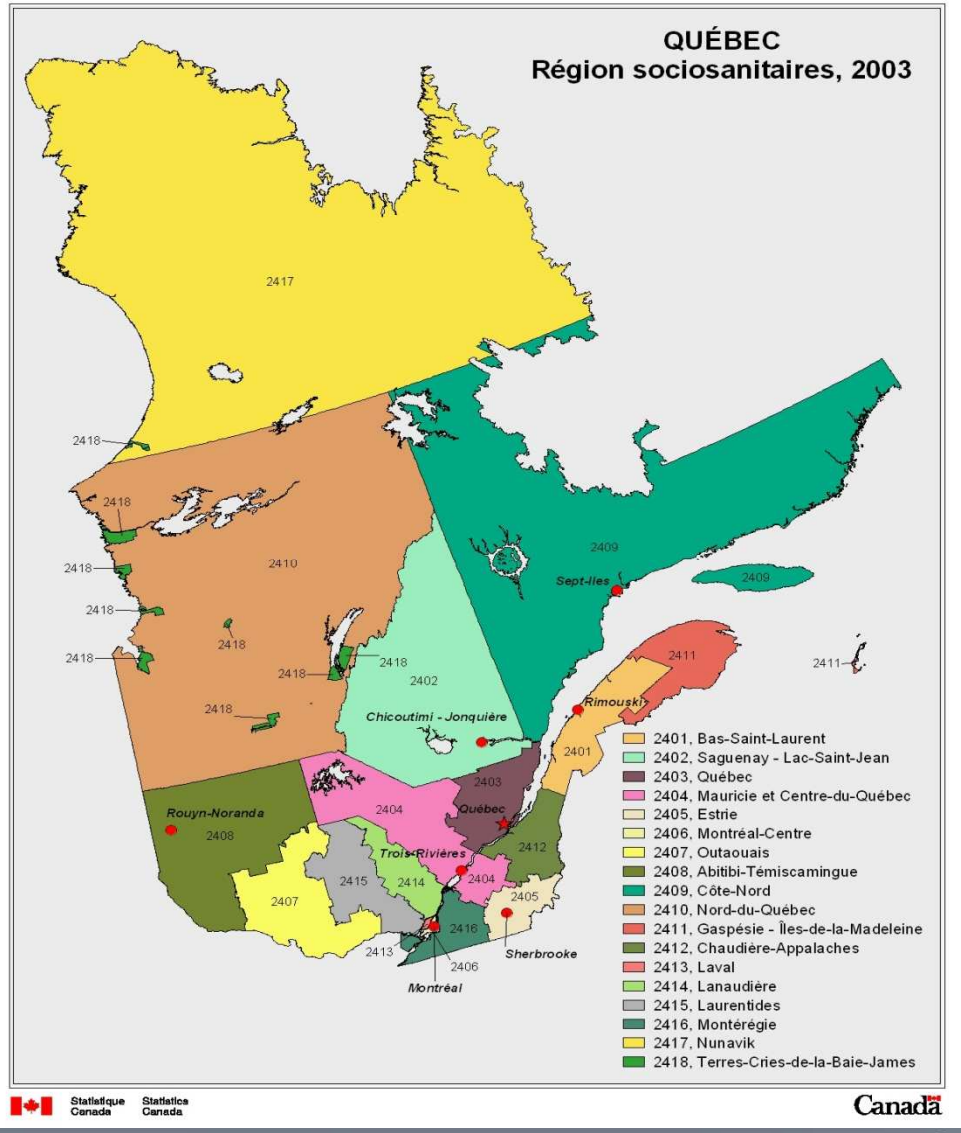
- Projet qui s'est déroulé en 2009 avec la collaboration de Louis-Paul Rivest

- Utilisation des méthodes d'estimation pour petits domaines afin de calculer le taux d'emploi régional québécois chez les personnes ayant une incapacité.

Institut de la statistique Québec

Objectifs de l'étude d'estimation pour petits domaines

- Produire des estimations du taux d'emploi des personnes ayant une incapacité par région sociosanitaire québécoise:
  - à la demande de l'Office des personnes handicapées du Québec (OPHQ)



Institut de la statistique Québec

Objectifs de l'étude

- Utiliser des données de l'Enquête sur la participation et les limitations d'activités (EPLA) réalisée en 2006 par Statistique Canada (SC):
  - estimations peu précises pour certaines régions.
- Solution: utiliser une méthode d'estimation pour petits domaines pour obtenir des estimations plus précises:
  - répondants de l'EPLA sélectionnés à partir du recensement canadien 2006
  - grande quantité de données auxiliaires disponibles
  - s'inspirer des travaux de Bizier et coll. (2009) de SC

Institut de la statistique Québec

Choix de la méthode

- Méthode d'estimation composite:
  - Estimation combinant:
    - Une estimation modélisée (synthétique) avec variables externes à l'EPLA
    - L'estimation directe provenant de l'EPLA
  - Plus stable que l'estimation de l'EPLA
  - Moins biaisé que l'estimation modélisée prise seule
  - Objectif: obtenir une EQM plus petite que la variance de l'estimateur de l'EPLA
- Modélisation de niveau régional:
  - Statistiques régionales provenant du recensement utilisées comme variables explicatives
  - Utilisation de l'ensemble des régions sociosanitaires canadiennes (120)

Institut de la statistique Québec

Choix de la méthode

- Modèle linéaire utilisé pour incorporer l'information auxiliaire régionale (Ghosh et Rao, 1994):

$$\theta_i = x_i'\beta + \nu_i$$

- où θ_i est le paramètre d'intérêt
- où $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ est un vecteur de variables auxiliaires de la région *i*
- où β est un vecteur des coefficients de régression
- où ν_i est l'effet aléatoire propre à la région *i* qui est supposé iid avec $E(\nu_i) = 0$ et $V(\nu_i) = \sigma_\nu^2$
- L'hypothèse de normalité des ν_i est souvent incluse.

- Ce modèle est appelé: « modèle de liaison »

Institut de la statistique Québec

Choix de la méthode

- Pour créer un estimateur composite, un modèle d'échantillonnage est nécessaire pour incorporer l'estimation provenant de l'enquête:

$$\hat{\theta}_i = \theta_i + e_i$$

- où $\hat{\theta}_i$ est l'estimateur direct obtenu à l'EPLA
- où e_i est l'erreur d'échantillonnage de l'estimateur direct. Nous supposons les e_i normales et indépendantes avec $E(e_i \mid \theta_i) = 0$ et $V(e_i \mid \theta_i) = \sigma_i^2$

Institut de la statistique Québec

Choix de la méthode

- L'estimateur composite est obtenu en combinant le modèle de liaison et d'échantillonnage:

$$\widehat{\theta}_i = x_i'\beta + \nu_i + e_i$$

- Modèle connu sous le nom de Fay-Herriot :
  - Les taux à estimer varient de façon générale de 40 à 60%. En présence d'une telle distribution, l'hypothèse de normalité du modèle de Fay-Herriot ne devrait pas être compromise.
- Paramètres inconnus: β et σ_ν^2

Institut de la statistique Québec

Choix de la méthode

- À l'aide de l'approche EBLUP, on obtient l'estimateur de θ_i suivant pour le modèle de Fay-Herriot:

$$\hat{\theta}_i^{FH} = \gamma_i \hat{\theta}_i + (1-\gamma_i) x_i^{'} \hat{\beta}$$

- où

$$\gamma_i = \frac{\hat{\sigma}_v^{\,2}}{\hat{\sigma}_v^{\,2} + \sigma_i^{\,2}}$$

- L'estimation des paramètres du modèle est obtenue de façon itérative à l'aide de la méthode des moments itérative de Fay-Herriot (1979).
  - Méthode recommandée par Datta, Rao et Smith (2005).
  - Les méthodes ML et REML ont également été testées mais celles-ci ont produit une estimation négative de $\sigma_v^{\,2}$

Institut de la statistique Québec

Choix de la méthode

- Calcul de la précision de $\hat{\theta}_i^{FH}$:
  - Se présente sous la forme d'une *EQM* (Rao, 2003) :

$$EQM(\hat{\theta}_i^{FH})_{fhi} = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) - b_{\hat{\sigma}_v^2}(\hat{\sigma}_v^2)(1-\hat{\gamma}_i)^2$$

- Où $g_{1i}(\hat{\sigma}_v^2)$ est le terme principal. Lorsque σ_i^2 est grand par rapport à $\hat{\sigma}_v^2$ alors $g_{1i}(\hat{\sigma}_v^2)$ tend vers $\hat{\sigma}_v^2$
- Où $g_{2i}(\hat{\sigma}_v^2)$ est une composante due à l'estimation de β
- Où $g_{3i}(\hat{\sigma}_v^2)$ est une composante due à l'estimation de σ_v^2
- Où $-b_{\hat{\sigma}_v^2}(\hat{\sigma}_v^2)(1-\hat{\gamma}_i)^2$ est un ajustement à l'*EQM* due à l'utilisation de la méthode des moments de Fay-Herriot pour estimer σ_v^2

- Rivest et Vandal (2003) ont proposé l'ajout d'un terme supplémentaire à l'*EQM* pour tenir compte que σ_i^2 est estimée

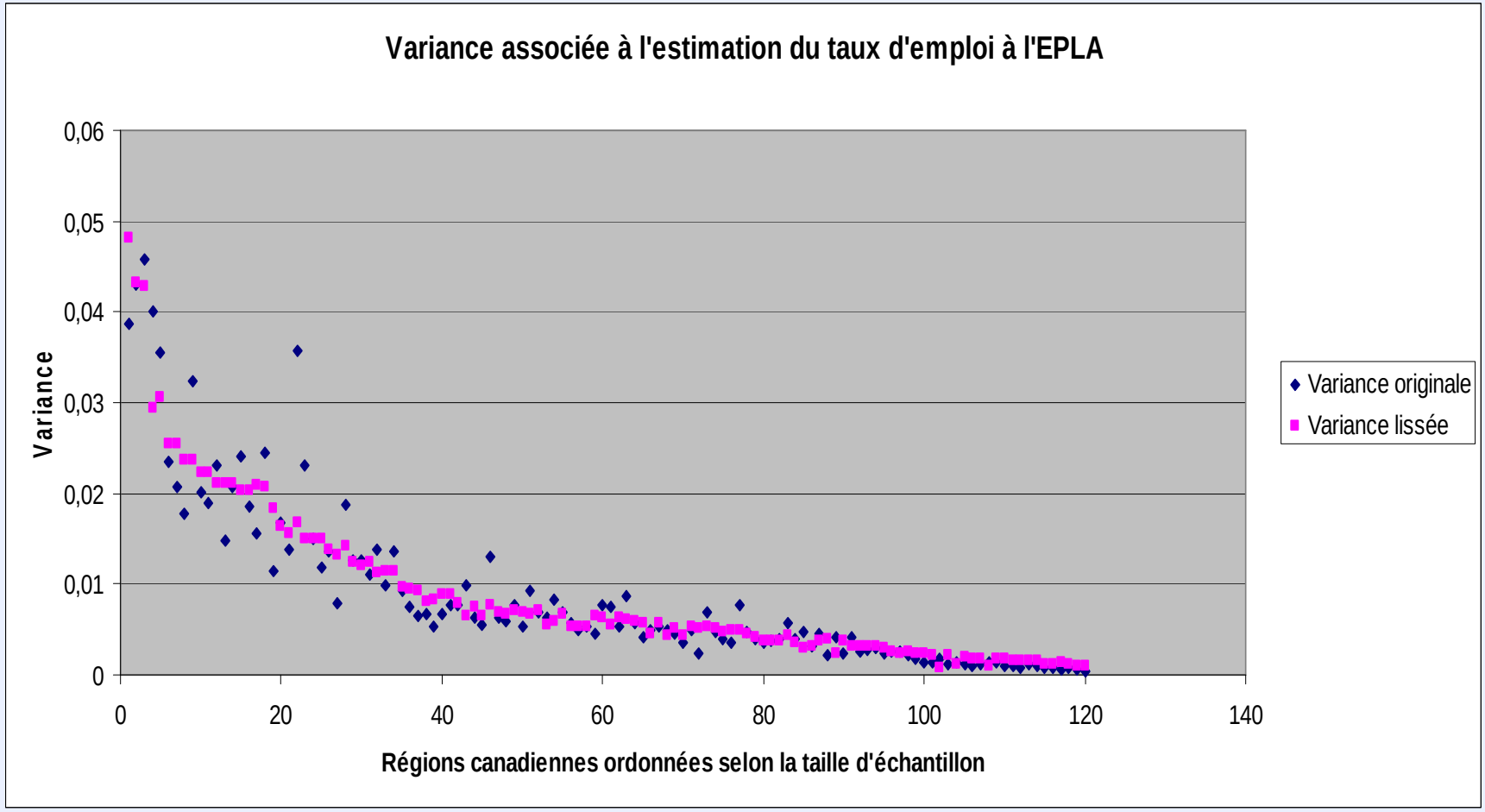
Institut de la statistique Québec

Lissage de la variance

- Le calcul de $\hat{\theta}_i^{FH}$ dépend en partie de σ_i^2.
  - σ_i^2 devrait être la vraie valeur de la variance
  - variance d'échantillonnage obtenue à l'EPLA est plutôt estimée:
    - peut donc s'éloigner beaucoup de la vraie valeur pour certaines petites régions => impact sur $\hat{\theta}_i^{FH}$
- Solution $\Rightarrow$ lissage de $\hat{\sigma}_i^2$:
  - prédire la variance d'échantillonnage à l'aide de la modélisation
  - variable explicative retenue:
    - inverse de la taille d'échantillon d'une région sociosanitaire et nombre de répondants au recensement par province
  - R^2 obtenu de 0,89.
  - valeurs lissées (ou prédites par le modèle) peuvent être utilisées comme σ_i^2 dans $\hat{\theta}_i^{FH}$

Institut de la statistique Québec

Lissage de la variance



Institut de la statistique Québec

Choix des variables auxiliaires

- Consultation de spécialistes sur l'emploi et sur les limitations d'activité pour le choix initial
- Variables retenues dans le modèle final:
  - x_1 : taux d'emploi chez les personnes avec limitations au recensement
  - x_2 : proportion de personnes en milieu rural (parmi limitations au rec.)
  - x_3 : log des heures moyennes travaillées (parmi limitations au rec.)
  - x_4: prop. de personnes n'étant pas de minorités visibles(…limitations…)
  - x_5 : interaction de x_3 et x_4
- Modèle final et paramètres estimés:

$$\theta_{prédit} = -3{,}60 + 1{,}14x_1 + 0{,}09x_2 + 0{,}96x_3 + 5{,}34x_4 - 1{,}48x_5$$

- R^2 de 0,76

Institut de la statistique Québec

Présentation des résultats

Régions sociosanitaires	Taux d'emploi EPLA (%)	CV (%)	Taux d'emploi F-H (%)	CV (%)	% de EPLA
Bas-Saint-Laurent	**18,3**	*39,7*	**34,6**	*6,6*	3,4
Saguenay–Lac-Saint-Jean	31,2	*24,6*	33,2	*7,0*	4,2
Capitale-Nationale	40,5	*14,1*	41,7	*5,8*	8,3
Mauricie-Centre-du-Québec	30,6	*20,7*	34,3	*6,9*	7,6
Estrie	40,6	*18,8*	41,5	*5,9*	5,0
Montréal	40,8	*8,9*	40,5	*6,1*	23,5
Outaouais	42,3	*14,9*	46,3	*4,6*	7,4
Abitibi-Témiscamingue	41,6	*23,1*	37,0	*5,7*	3,0

Source : Enquête sur la participation et les limitations d'activités de 2006 et Recensement de 2006.
Traitement : Institut de la statistique du Québec.

Institut de la statistique Québec

Présentation des résultats

Régions sociosanitaires	Taux d'emploi EPLA (%)	CV (%)	Taux d'emploi F-H (%)	CV (%)	% de EPLA
Côte-Nord	**55,6**	*27,9*	**37,4**	*5,5*	1,4
Gaspésie–Îles-de-la-Madeleine	27,7	*49,2*	25,7	*9,3*	1,4
Chaudière-Appalaches	**57,7**	*15,8*	**45,6**	*4,9*	4,7
Laval	**33,9**	*23,6*	**48,4**	*4,4*	4.1
Lanaudière	43,2	*13,6*	40,8	*5,4*	7,3
Laurentides	42,4	*14,4*	44,5	*5,0*	7,3
Montérégie	42,5	*8,9*	45,0	*5,3*	19,0

Source : Enquête sur la participation et les limitations d'activités de 2006 et Recensement de 2006.
Traitement : Institut de la statistique du Québec.

Institut de la statistique Québec

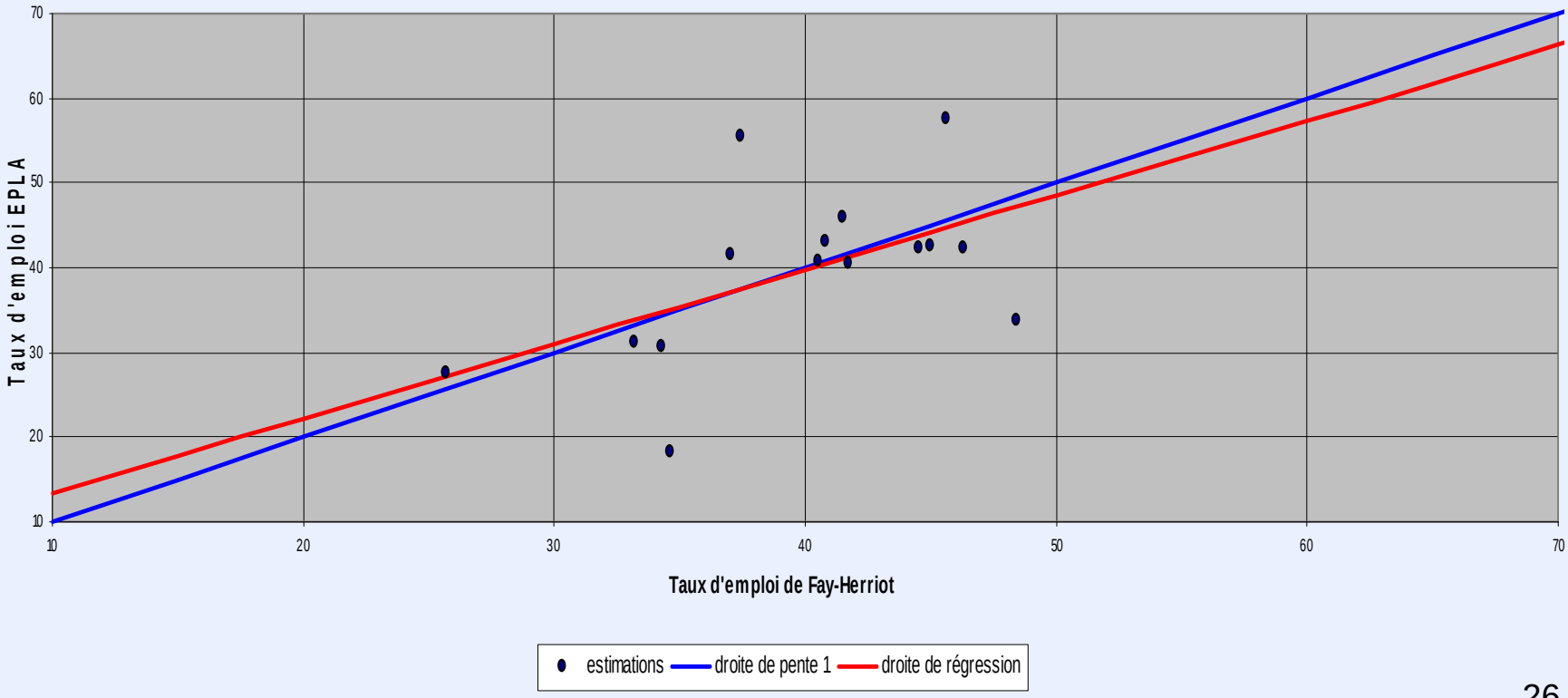
Comparaison avec la variable explicative la plus importante

Régions sociosanitaires	Taux d'emploi à l'EPLA	Taux d'emploi de Fay-Herriot	Taux d'emploi au recensement (avec limitation)
01-Bas-Saint-Laurent	18,3	34,6	35,4
02-Saguenay-Lac-Saint-Jean	31,2	33,2	35,6
03-Capitale-Nationale	40,5	41,7	43,9
04-Mauricie et Centre-du-Québec	30,6	34,3	36,3
05-Estrie	46,0	41,5	40,8
06-Montréal	40,8	40,5	45,5
07-Outaouais	42,3	46,3	48,1
08-Abitibi-Témiscamingue	41,6	37,0	39,6
09-Côte-Nord	55,6	37,4	40,4
11-Gaspésie-Îles-de-la-Madeleine	27,7	25,7	27,9
12-Chaudière-Appalaches	57,7	45,6	44,0
13-Laval	33,9	48,4	52,1
14-Lanaudière	43,2	40,8	42,0
15-Laurentides	42,4	44,5	46,5
16-Montérégie	42,5	45,0	48,1

Institut de la statistique Québec

Validation - Diagnostics

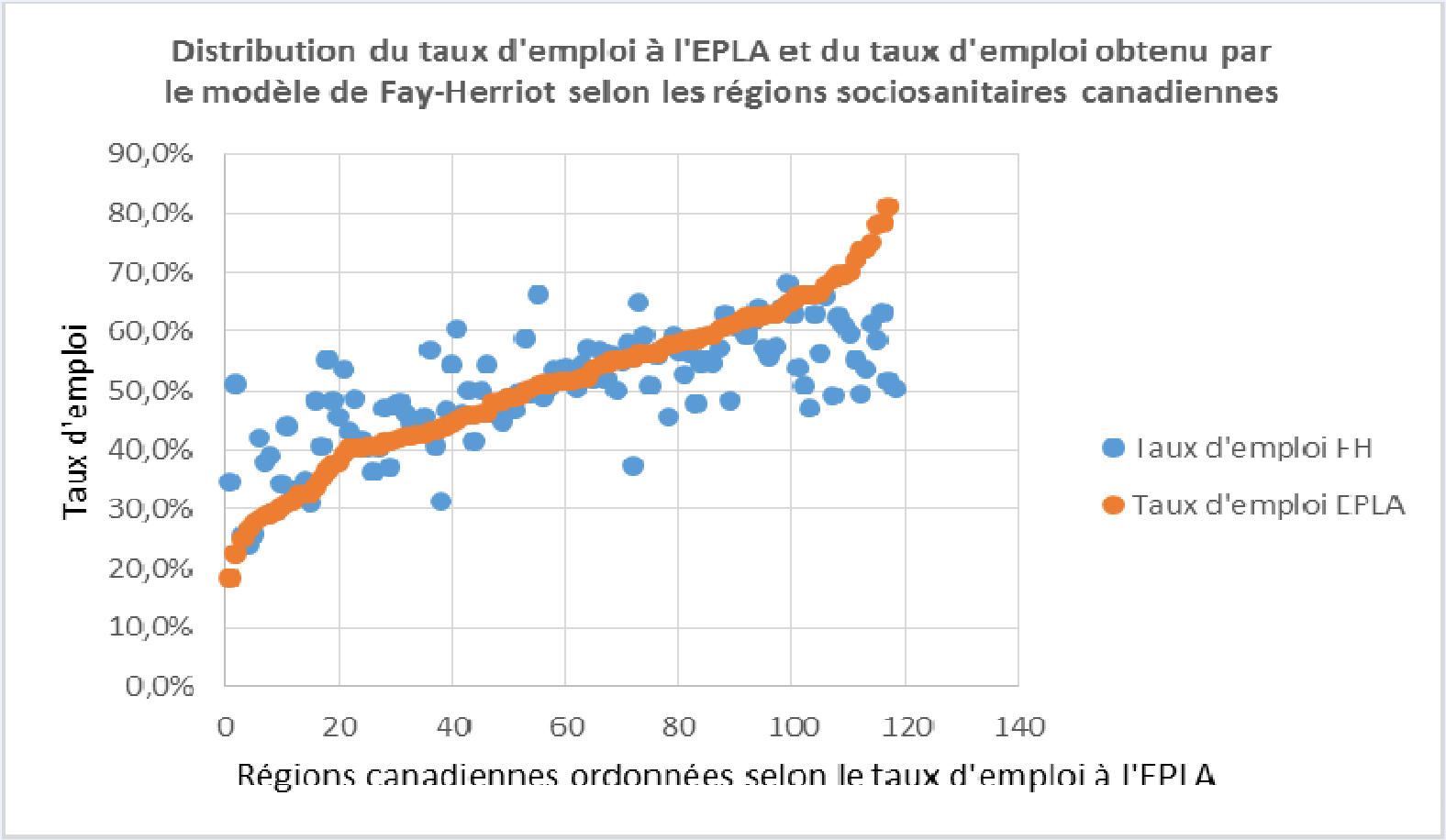
- Évaluation du biais par rapport à la vraie valeur à l'aide de la pente de régression. (Brown et al., 2001)



Institut de la statistique Québec

Validation – Diagnostics

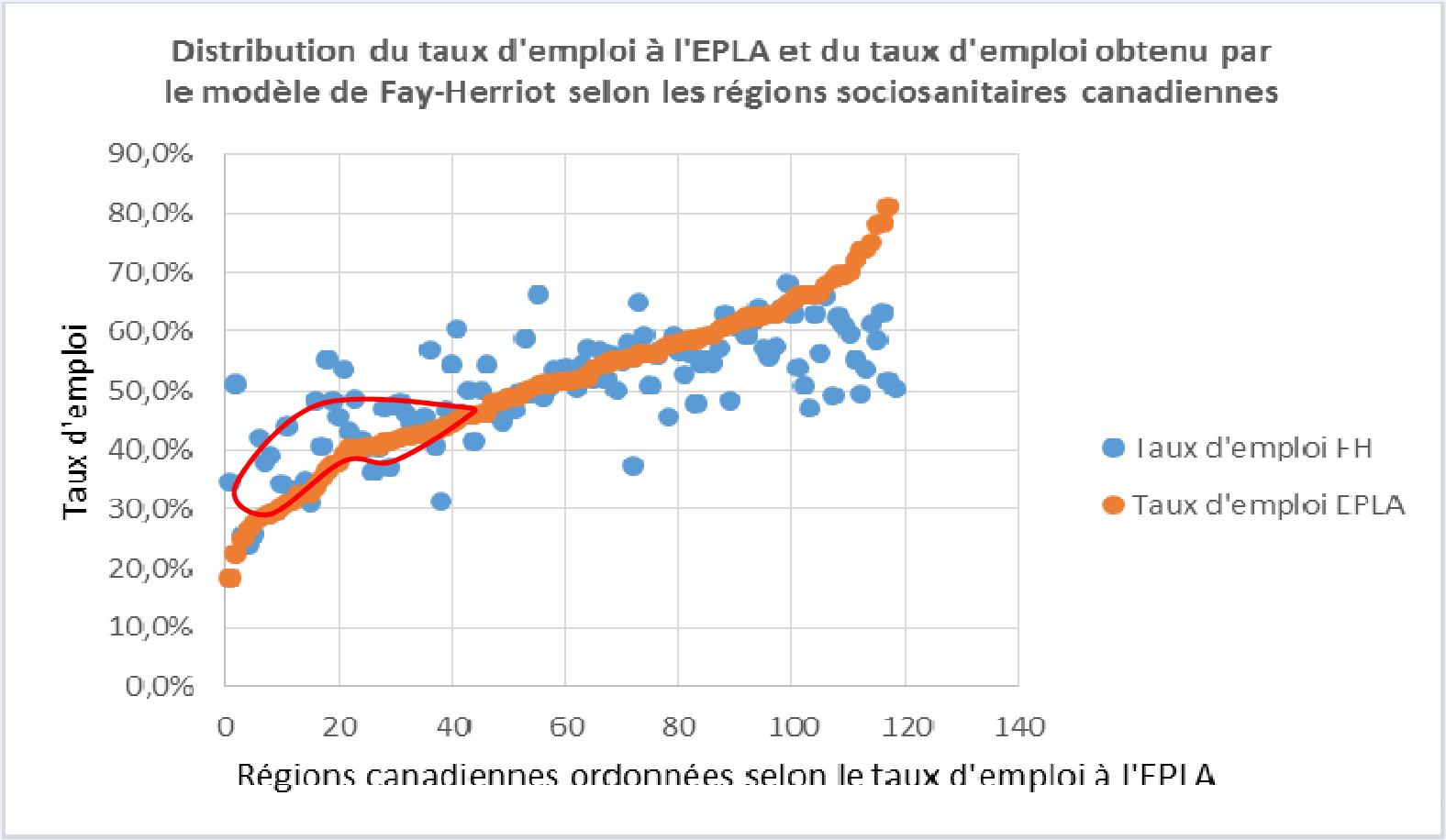
- Évaluation de la surestimation et de la sous-estimation:



Institut de la statistique Québec

Validation – Diagnostics

- Évaluation de la surestimation et de la sous-estimation:



Institut de la statistique Québec

Validation – Diagnostics

- Autres diagnostics:
  - Évaluation de l'ajustement à l'aide de la méthode de prédiction *a posteriori* (Meng, 1994):
  - Évaluation de la couverture (Brown et al., 2001)
  - Évaluation de la normalité
- Validation externe:
  - Auprès d'experts et de futurs utilisateurs
  - Taux d'emploi de FH utilisés malgré les limites

Institut de la statistique Québec

Conclusion

- Recommandations aux utilisateurs:
  - Taux de Fay-Herriot doivent être considérés comme des prédictions d'un modèle
  - Taux de Fay-Herriot sont légèrement biaisés
  - Être prudent avec les intervalles de confiance
- Suivre l'évolution du taux d'emploi dans le temps:
  - Attention: ENM a remplacé le questionnaire long du recensement
  - Meilleurs outils: logiciel développé par SC

(Estevao,Hidiroglou,You, 2012)

Institut de la statistique Québec

Références

- Baillargeon, S., L.-P. Rivest et M. Ferland (2007). « Stratification en enquêtes entreprises: une revue et quelques avancées », *Proceedings of the Survey Methods Section, Statistical Society of Canada.*
- Baillargeon, S. et L.-P. Rivest (2011). « Élaboration de plans stratifiés en R à l'aide du programme *stratification* », *Technique d'enquête*, vol. 37, p. 53-65.
- Bizier, V., Y. You, L. Veilleux et C. Grondin (2009). *Une approche expérimentale fondée sur un modèle d'estimation des comptes et des taux d'incapacité pour les adultes dans les petites régions au moyen des données de l'Enquête sur la participation et les limitations d'activités de 2006*, [En ligne] : http://www.statcan.gc.ca/dli-ild/meta/pals-epla/2006/pals-epla2006sae-fra.doc
- Brown, G., R. Chambers, P. Heady et D. Heasman (2001). « Évaluation des méthodes d'estimation régionale dans leur application aux estimations du chômage tirées de l'enquête sur la population active au Royaume-Uni », *recueil du Symposium 2001 de Statistique Canada*.
- Datta, G.S., J.N.K. Rao et D.D. Smith (2005). « On measuring the variability of small area estimators under a basic areal level model », *Biometrika*, vol. 92, n° 1, p. 183-196.
- Estevao, V.M., M.A. Hidiroglou, Y. You (2012). *Small area estimation – Area Level – FH Model – EBLUP Estimation – Methodology Specifications.pdf.* Document de Statistique Canada.
- Fay, R.E., et R.A. Herriot (1979). « Estimation of income for small places: An application of James-Stein procedures to census data », *Journal of the American Statistical Association*, vol. 85, p. 398-409.
- Ghosh, M. et J.N.K. Rao (1994). « Small area estimation: an appraisal », *Statistical Science,* vol. 9, n° 1, p. 55-93.

Institut de la statistique Québec

Références

- Lavallée, P. et M.A. Hidiroglou (1988). « Sur la stratification de populations asymétriques », *Techniques d'enquête*, vol. 14, p. 35-45.
- Meng, X.L. (1994). « Posterior predictive *p*-values », *The Annals of Statistics*, vol. 22, nº 3, p. 1142-1160.
- Rao, J.N.K. (2003). *Small Area Estimation,* New York, John Wiley & Sons, 344 p
- Rivest, L.-P. (1995). «A composite estimator for provincial under-coverage in the Canadian census » *Recueil 1995 de la Section des méthodes d'enquête, Société statistique du Canada.* P. 33-38
- Rivest, L.-P. et E. Belmonte (2000). « Une erreur quadratique moyenne conditionnelle des estimateurs régionaux », *Techniques d'enquête*, vol. 26, p. 79-90.
- Rivest, L.-P. et N. Vandal (2003). « Mean squared error estimation for small areas when the small area variances are estimated », *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Laboratory for research in Statistics and Probability, Carleton University, Ottawa, ON, Canada.
- Royce, D (1992). « Une comparaison d'estimateurs d'un ensemble de totaux de population», *Techniques d'enquête*, vol. 18, p. 121-138.