Estimations régionales des taux d'activité et d'emploi pour les personnes ayant une incapacité au Québec : application d'une méthode d'estimation pour petits domaines

Éric Gagnon
Jocelyne Camirand
Robert Courtemanche
Fanny Harvey
Institut de la statistique du Québec

Décembre 2009

Pour tout renseignement concernant l'ISQ et les données statistiques qui y sont disponibles, s'adresser à :

Institut de la statistique du Québec
200, chemin Sainte-Foy
Québec (Québec)
G1R 5T4
Téléphone : 418-691-2401

ou

Téléphone : 1-800-463-4090
(aucuns frais d'appel au Canada et aux États-Unis)

Site Web : www.stat.gouv.qc.ca

Citation suggérée :
GAGNON, Éric, Jocelyne CAMIRAND, Robert COURTEMANCHE et Fanny HARVEY (2009). *Estimations régionales des taux d'activité et d'emploi pour les personnes ayant une incapacité au Québec : application d'une méthode d'estimation pour petits domaines*, Québec, Institut de la statistique du Québec, 53 pages.

Publication produite et réalisée par :
Institut de la statistique du Québec

Étude subventionnée par :

Office des personnes handicapées du Québec
Ministère de l'Emploi et de la Solidarité sociale

Décembre 2009

Remerciements

Les auteurs tiennent à remercier Louis-Paul Rivest, professeur titulaire à l'Université Laval, qui a agi comme consultant dans le cadre de ce projet. Son expertise quant à l'estimation pour petits domaines a permis d'assurer le succès de ce projet. Des remerciements vont également à France Lapointe, Sylvain Végiard et Jean-François Cardin de l'ISQ pour leur soutien et leurs précieux commentaires. Finalement, nous remercions Valérie Bizier et Chantal Grondin de Statistique Canada qui, par l'entremise de nombreux échanges avec l'ISQ, ont contribué à l'avancement de ce projet.

Cette étude a été rendue possible grâce à la participation financière de l'Office des personnes handicapées du Québec (OPHQ), du ministère de l'Emploi et de la Solidarité sociale (MESS) et de l'Institut de la statistique du Québec.

Table des matières

1. Objectifs de l'étude

La présente étude a pour principal objectif de produire des estimations du taux d'activité et du taux d'emploi des personnes ayant une incapacité par région sociosanitaire au Québec. Ces estimations sont calculées à partir de l'Enquête sur la participation et les limitations d'activités (EPLA) réalisée par Statistique Canada en 2006[1].

Pour produire ces estimations, une méthode d'estimation pour petits domaines (EPD) a été utilisée. Puisque cette méthode n'a jamais été utilisée par l'Institut de la statistique du Québec (ISQ) dans le cadre des enquêtes de santé, ce projet vise à nous permettre non seulement d'obtenir des estimations, mais aussi de développer une méthode d'analyse, de la tester et d'évaluer la qualité des estimations produites.

Dans le présent rapport, nous exposons les méthodes d'EPD (section 2), notre choix de l'une d'entre elles (section 3) et son implantation dans le cadre de l'EPLA. La section 4 est consacrée à la présentation des résultats, tandis que la section 5 fournit une évaluation de leur qualité. Les résultats obtenus sont ensuite comparés à une source de données externe en guise de validation (section 6). Dans la section 7, nous expliquons de quelle façon les résultats devraient être interprétés. Finalement, la section 8 conclut le rapport.

2. Les méthodes d'EPD

De façon générale, les méthodes d'EPD découlent d'un modèle linéaire qui intègre aux données d'enquêtes des variables auxiliaires provenant de sources externes (par exemple recensement, données administratives). Plus la corrélation entre ces variables auxiliaires et la caractéristique étudiée est forte, meilleurs sont les estimateurs obtenus par la modélisation.

Ces méthodes s'appliquent particulièrement bien dans le cadre de l'EPLA 2006 puisqu'il existe une forte corrélation entre les questions du recensement sur l'incapacité et le taux d'incapacité obtenu à partir de l'EPLA 2006. De plus, le recensement inclut les questions permettant de définir la population active. Mentionnons que des estimations basées sur ces méthodes ont été publiées par Statistique Canada à partir des données de l'EPLA 2006. Elles portent sur les taux régionaux d'incapacité par âge et sexe (Bizier et autres, 2009).

Dans les prochaines sous-sections, nous décrivons d'abord la méthode d'estimation directe dans le cas de petits domaines. Nous donnons ensuite un aperçu de deux types de méthodes d'EPD, soit les estimateurs synthétiques et les estimateurs composites. Une mesure d'erreur permettant de comparer les différents estimateurs est également présentée. Des recommandations découlant d'un avis présenté par l'ISQ au ministère de la Santé et des Services sociaux (MSSS) sur les méthodes d'estimation pour petits domaines (Gagnon, 2008) terminent cette section.

1. Les résultats obtenus dans le présent rapport proviennent de données disponibles au Centre de données de recherche de Statistique Canada. Bien que la recherche et les analyses soient fondées sur des données de Statistique Canada, les opinions exprimées ne représentent pas celles de Statistique Canada.

2.1 Méthode d'estimation directe

L'estimateur direct est celui que l'on obtient en utilisant l'estimation pondérée habituelle avec les données d'une enquête. En raison des plans de sondage utilisés, il arrive souvent que cet estimateur ne soit pas stable (très variable) pour un petit domaine (par exemple une région). Même s'il est rarement stable pour un petit domaine, l'estimateur direct a, tout de même, la qualité d'être non biaisé. Enfin, un autre problème de cet estimateur est que s'il n'y a pas d'unité échantillonnée dans un petit domaine (par exemple s'il n'y a aucun répondant dans une région), il n'est pas possible de produire une estimation directe pour ce petit domaine.

Afin d'illustrer ce type d'estimateur, prenons l'exemple de l'EPLA. Pour produire une estimation directe du taux d'emploi des personnes ayant une incapacité, il suffit tout simplement de calculer une proportion pondérée en utilisant seulement l'information disponible pour les répondants de l'enquête. Cependant, si on calcule ce taux d'emploi pour un petit domaine (par exemple une région), l'estimation du taux d'emploi risque d'être très variable. Il se pourrait même qu'il n'y ait aucune unité échantillonnée dans un petit domaine et que, par conséquent, il ne soit pas possible d'obtenir une estimation directe du taux d'emploi pour ce domaine.

2.2 Méthode d'estimation synthétique pour petits domaines

Les estimateurs synthétiques pour petits domaines découlent, en général, d'un modèle qui intègre des variables auxiliaires provenant de sources externes à l'enquête, comme le recensement ou des données administratives. La présence de variables auxiliaires et l'adéquation du modèle vont permettre de produire des estimations plus stables (avec une plus petite variabilité) que les estimateurs directs.

Afin d'illustrer ce type d'estimateur, reprenons l'exemple de l'EPLA. Si on voulait produire une estimation synthétique du taux d'emploi des personnes ayant une incapacité, on pourrait par exemple se servir d'un modèle de régression utilisant les données de l'enquête et des variables auxiliaires du recensement. La variable dépendante serait le taux d'emploi chez les personnes ayant une incapacité calculé à l'enquête et les variables indépendantes pourraient provenir du recensement. Dans cette situation, les estimations synthétiques seraient tout simplement les valeurs prédites obtenues par le modèle pour chaque petit domaine.

Les estimateurs synthétiques ont cependant un inconvénient. Si les hypothèses du modèle ne tiennent pas pour un domaine, l'estimateur synthétique sera considéré comme biaisé au sens statistique, c'est-à-dire qu'il s'éloignera de la vraie valeur. Dans le cas de l'estimation d'un taux d'emploi, par exemple, une hypothèse du modèle pourrait être que ce taux est le même au niveau des petits domaines (par exemple des régions) qu'au niveau d'un grand domaine pour un groupe d'âge donné. Cependant, certains facteurs non inclus dans le modèle pourraient influer sur ce taux au niveau du petit domaine, le rendant différent de celui du grand domaine pour ce groupe d'âge. Dans cette situation, le taux obtenu par l'estimateur synthétique s'éloignerait de la vraie valeur du petit domaine pour se rapprocher de la valeur du grand domaine.

En résumé, les estimations synthétiques obtenues seraient probablement moins variables que celles obtenues par l'estimateur direct. Cependant, les estimations obtenues par la méthode synthétique seraient fort probablement biaisées.

2.3 Méthode d'estimation composite pour petits domaines

La méthode d'estimation composite pour petits domaines a été créée afin de combiner les avantages des estimateurs directs (non biaisés) avec ceux des estimateurs synthétiques (plus stables). Les estimateurs découlant de cette méthode résultent d'une combinaison linéaire de l'estimateur direct et de l'estimateur synthétique. Le poids assigné à chacun des deux estimateurs dans le calcul de l'estimateur composite peut être déterminé de différentes façons. Souvent, la partie de l'estimateur qui est plus stable et qui a un biais moindre se voit attribuer un plus grand poids dans le calcul de l'estimateur composite.

Dans l'exemple déjà présenté, on pourrait obtenir un estimateur composite en combinant tout simplement les estimateurs présentés aux sections 2.1 et 2.2. L'estimateur composite résultant serait sans doute plus stable que l'estimateur direct et moins biaisé que l'estimateur synthétique.

2.4 Comparaison des estimateurs

Le défi pour les méthodes d'estimation pour petits domaines (synthétiques et composites) est de trouver un estimateur qui sera plus stable (moins variable) que l'estimateur direct et qui n'aura pas un biais trop important. Il existe une mesure qui permet d'évaluer ces deux aspects simultanément – il s'agit de l'erreur quadratique moyenne ou *EQM* :

$$EQM = variance + (biais)^2$$

Dans cette équation représentant l'*EQM*, la variance correspond à la stabilité de l'estimateur. Globalement, un bon estimateur pour petits domaines devrait avoir une plus petite *EQM* que celle de l'estimateur direct. L'estimateur ayant la plus petite *EQM* est alors considéré comme le plus précis.

Il faut noter que le calcul du biais peut être assez complexe. En effet, le biais peut provenir de deux sources. Tout d'abord, il peut provenir du modèle utilisé pour le calcul de l'estimation pour petits domaines. Ensuite, il peut y avoir également un biais par rapport à la vraie valeur, comme nous l'avons mentionné à la section 2.2. Cette dernière partie est beaucoup plus difficile à calculer. Ainsi, la plupart du temps, le calcul de l'*EQM* incorpore seulement la partie du biais provenant du modèle. Pour avoir une idée de la seconde partie du biais, différents diagnostics existent; il en sera question à la section 5.

2.5 Recommandations pour l'utilisation des méthodes d'EPD

La présente section expose les recommandations découlant d'un avis soumis par l'ISQ au MSSS à l'automne 2008. Cet avis, reproduit à l'annexe 1, examine plusieurs méthodes d'estimation synthétiques et composites utilisées pour des enquêtes de santé dans différents pays.

Cet avis montre que toutes les méthodes examinées permettent d'obtenir des estimateurs plus stables que les estimateurs directs. En contrepartie, les estimations obtenues à l'aide de ces méthodes sont probablement biaisées. Ainsi, les grandes prévalences peuvent être sous-estimées et les petites prévalences peuvent être surestimées. Ce glissement des prévalences se traduit par une diminution de l'étendue des estimations obtenues. Ce phénomène implique alors une sous-estimation des différences inter-domaines.

Par contre, il est possible d'atténuer ce phénomène, c'est-à-dire de diminuer le biais des estimations, en recourant aux moyens suivants :

- utiliser un estimateur composite à la place d'un estimateur synthétique;
- construire l'estimateur à l'aide de variables auxiliaires fortement liées à l'estimation que l'on veut produire;
- modéliser les prévalences une à la fois.

La dernière solution proposée implique, toutefois, une quantité de travail non négligeable étant donné que le modèle doit être refait pour chaque caractéristique à estimer (pour la synthèse, voir tableau 2.5.1).

Tableau 2.5.1
Biais et quantité de travail selon des caractéristiques des estimateurs

	Modélisation d'une variable à la fois		**Utilisation d'information auxiliaire**		**Méthodes d'estimation**	
	Oui	Non	Moins	Plus	Synthétique	Composite
Biais de l'estimateur	↓	↑	↑	↓	↑	↓
Quantité de travail	↑	↓	↓	↑	↓	↑

Cet avis permet de conclure que malgré les problèmes liés au biais, les estimateurs pour petits domaines peuvent constituer une bonne solution, surtout si l'estimateur direct est très variable. De plus, il est certainement préférable d'utiliser une méthode d'EPD que d'approximer l'estimation d'un petit domaine par l'estimation d'un plus grand domaine. Cette dernière solution est sans doute la plus biaisée.

3. Choix et implantation de la méthode pour l'EPLA

Pour ce projet, différentes méthodes d'estimation pour petits domaines ont été considérées. La première méthode examinée consiste à obtenir des estimations pour petits domaines à l'aide de modèles utilisant les données individuelles pour prédire le taux d'emploi et le taux d'incapacité. De prime abord, cette avenue semblait très prometteuse étant donné que des données individuelles provenant du recensement sont disponibles. De plus, ces données sont très corrélées avec les variables à prédire. Cependant, le taux d'emploi et le taux d'activité chez les personnes avec incapacité sont considérés comme le ratio de deux quantités[2]. Or, dans la littérature actuelle, la modélisation utilisant le niveau individuel n'est pas utilisée pour la prédiction de ratios. De fait, la théorie permettant le calcul de la variance de telles estimations n'est pas encore développée. Étant donné que le développement d'une nouvelle théorie dépassait largement les objectifs de ce projet, l'ISQ a décidé de ne pas retenir cette avenue.

Puisqu'il n'est pas possible de recourir à la modélisation utilisant les données de niveau individuel, l'ISQ a décidé d'examiner la modélisation de niveau régional. Dans ce type de modélisation, les variables explicatives ne proviennent plus de données individuelles mais plutôt de statistiques régionales agrégées. Une telle modélisation permet l'estimation de ratios et de leur variance.

2. Le ratio du nombre de personnes ayant une incapacité en emploi ou actives sur le nombre de personnes ayant une incapacité.

Il existe différents types de modèles de niveau régional. Par exemple, Bizier et autres (2009) de Statistique Canada ont utilisé un modèle de liaison log-linéaire non apparié au modèle d'échantillonnage pour leurs travaux d'estimation du taux d'incapacité à partir des données de l'EPLA 2006. Pour obtenir les estimations à l'aide de ce type de modèle, l'approche hiérarchique bayesienne a été utilisée. Ce type d'approche a donné de très bons résultats pour les estimations canadiennes du taux d'incapacité. Un autre modèle de niveau régional fort connu est le modèle de Fay-Herriot (1979). Selon Louis-Paul Rivest, spécialiste de l'estimation pour petits domaines, ce dernier modèle peut très bien convenir pour la prédiction des taux visés par ce projet. En effet, le fait que le taux d'emploi et le taux d'activité prennent des valeurs oscillant entre 40 % et 60 % assure une distribution adéquate des données pour l'utilisation de ce modèle. De plus, de son avis, les résultats obtenus à partir de ce modèle devraient être très près de ceux qui auraient été obtenus à l'aide du type de modèle utilisé par Statistique Canada. Étant donné la complexité de ce dernier modèle et la difficulté à expliquer aux futurs utilisateurs le choix des distributions *a posteriori* utilisées pour cette modélisation, l'ISQ a décidé de ne pas utiliser le même modèle que Statistique Canada. C'est plutôt le modèle de Fay-Herriot qui a été retenu en raison de sa simplicité et aussi parce qu'il convient très bien pour la prédiction du type de données de ce projet.

L'utilisation du modèle de Fay-Herriot est en accord avec les recommandations formulées dans l'avis présenté au MSSS en 2008. En effet, l'estimateur obtenu à partir du modèle de Fay-Herriot est composite. De plus, ce modèle permet d'utiliser plusieurs variables auxiliaires corrélées avec les prévalences à estimer. Finalement, la modélisation sera faite de façon indépendante pour le taux d'emploi et le taux d'activité.

3.1 Modèle de Fay-Herriot

L'estimateur d'un petit domaine *i* obtenu à partir du modèle de Fay-Herriot se présente sous la forme suivante :

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

où $\hat{\theta}_i$ représente l'estimation directe obtenue pour le petit domaine *i* à l'EPLA, x_i représente un vecteur de variables auxiliaires pour le petit domaine *i*, β représente un vecteur des coefficients de régression associés aux variables auxiliaires ($x_i^{'}\beta$ représente donc l'estimation obtenue par modélisation), et où :

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2}$$

où σ_v^2 représente la variance provenant du modèle de régression et σ_i^2 représente la variance d'échantillonnage pour l'estimation directe du petit domaine *i*. De cette façon, l'estimateur de Fay-Herriot obtenu se situera entre l'estimateur direct provenant de l'enquête et l'estimateur obtenu par modélisation. De plus, l'estimateur de Fay-Herriot sera plus près de l'estimateur direct si celui-ci est très précis, sinon il sera plus près de l'estimateur obtenu par modélisation.

Afin d'obtenir l'estimateur de Fay-Herriot, il faut estimer σ_v^2 et β. Pour estimer ces paramètres, l'approche EBLUP[3] a été utilisée. Plusieurs algorithmes permettent l'estimation de σ_v^2 et β avec l'approche EBLUP. Dans le cadre de ce projet, trois algorithmes ont été testés. Il s'agit de la méthode du maximum de vraisemblance (ML), de la méthode du maximum de vraisemblance restreint (REML) et de la méthode itérative des moments de Fay-Herriot (FHI). Dans le cas qui nous occupe, seule la méthode FHI a permis d'obtenir des résultats adéquats[4]. L'utilisation de cette méthode permet d'obtenir de meilleures estimations de l'*EQM* de l'estimateur de petits domaines (Datta, Rao et Smith, 2005).

3.2 Lissage de variance

Ainsi que nous l'avons mentionné à la section 3.1, le calcul de l'estimateur de Fay-Herriot dépend en partie de la variance d'échantillonnage. Avec les données de l'EPLA, il est possible de calculer une estimation de la variance d'échantillonnage. Pour de très petites régions, il se peut que cette estimation ne soit pas de bonne qualité. Une telle situation peut affecter la qualité de l'estimateur de Fay-Herriot. Pour contourner ce problème, il existe une technique permettant de modifier les estimations de variance de l'enquête pour ensuite considérer ces variances comme connues. Cette technique, appelée lissage de variance, permet d'utiliser des variances plus uniformes lors du calcul de l'estimateur de Fay-Herriot. De cette façon, il est possible d'éviter les fluctuations de l'estimateur qui ne seraient dues qu'à une imprécision de l'estimateur de variance.

Pour ce projet, on a lissé la variance en utilisant un modèle de régression linéaire. Deux variables explicatives ont été retenues pour prédire la variance : l'inverse de la taille d'échantillon d'une région sociosanitaire à l'EPLA et la taille de l'échantillon provincial au recensement. Le modèle obtenu pour la variance du taux d'emploi affiche un R-carré de 0,89 et celui pour la variance du taux d'activité un R-carré de 0,84. Graphiquement, les variances lissées suivent la distribution des variances estimées, ce qui est très satisfaisant. Étant donné ces bons résultats, les variances prédites par ces modèles ont été utilisées lors du calcul de l'estimateur de Fay-Herriot.

3.3 Variables dépendantes

Deux modèles ont été développés : un où la variable dépendante est le taux d'emploi calculé à l'EPLA et l'autre où la variable dépendante est le taux d'activité calculé à l'EPLA.

Taux d'emploi : le taux d'emploi des personnes ayant une incapacité de 15-64 ans est le pourcentage de personnes avec incapacité en âge de travailler (15-64 ans) qui occupent un emploi.

Taux d'activité : le taux d'activité des personnes ayant une incapacité de 15-64 ans correspond au pourcentage des personnes avec incapacité de 15-64 ans qui occupent un emploi ou sont en chômage.

Ces deux variables dépendantes sont calculées par région sociosanitaire. Les unités d'analyse utilisées par les modèles sont régionales, ainsi que nous l'avons mentionné au début de la section 3.

3. EBLUP : Empirical best linear unbiased prediction.

4. Les méthodes ML, REML et FHI estiment σ_v^2 de façon itérative. Avec les modèles retenus, les méthodes ML et REML ont convergé vers une valeur négative de σ_v^2. Pour qu'une estimation adéquate de la variance des estimateurs soit produite, σ_v^2 doit prendre une valeur positive.

Pour calculer le taux d'emploi et le taux d'activité à l'EPLA, la variable « lf71 » est utilisée. Cette variable représente le statut d'emploi des répondants à l'enquête. Cette information a été obtenue à partir du recensement de la population canadienne de 2006.

3.4 Choix des variables auxiliaires et modélisation

La source de données externe utilisée pour la création de variables auxiliaires est le questionnaire long du recensement de 2006. Cette source de données est très riche et contient des variables qui sont fortement corrélées avec le taux d'emploi et le taux d'activité chez les personnes avec incapacité.

L'utilisation du modèle de Fay-Herriot nécessite que les variables auxiliaires soient calculées pour chaque petit domaine pour lequel une estimation est désirée. Dans ce projet, les petits domaines correspondent aux régions sociosanitaires québécoises. Cependant, le nombre de régions québécoises (n=17) n'est pas assez élevé pour obtenir une modélisation adéquate. Aussi, la décision a été prise d'effectuer la modélisation à partir de l'ensemble des régions sociosanitaires au Canada (n=120).

Pour déterminer quelles variables auxiliaires seraient examinées, des spécialistes de contenu de l'Office des personnes handicapées Québec (OPHQ), du ministère de l'Emploi et de la Solidarité sociale (MESS) et de l'ISQ ont été consultés. À la suite de cette consultation, les variables suivantes ont été retenues :

> l'âge, le sexe, la scolarité, le statut d'immigration, le statut de minorité visible, la langue maternelle parlée, la langue officielle à la maison, le fait de vivre ou non sous le seuil de pauvreté, la principale source de revenu, la valeur de la résidence, le revenu d'emploi, le statut d'emploi (lf71), le nombre d'heures travaillées, le fait de vivre seul ou non, le fait de vivre en milieu rural, la gravité des limitations et le fait d'avoir une limitation au travail.

Ces informations se rapportant à des individus sont utilisées pour créer des proportions et des moyennes par région sociosanitaire au Canada. Ces statistiques sont calculées pour la population âgée de 15 à 64 ans étant donné que le taux d'emploi et le taux d'activité portent sur ce groupe d'âge. En plus d'être calculées pour l'ensemble de la population de 15 à 64 ans, ces statistiques sont également calculées pour un sous-ensemble de la population constitué des personnes ayant répondu « oui » à au moins une des questions filtre du recensement portant sur l'incapacité.

Pour sélectionner les variables qui seront retenues dans le modèle final, des tests de corrélation entre les variables auxiliaires et les deux variables dépendantes (taux d'emploi et taux d'activité à l'EPLA) ont tout d'abord été effectués. Ces tests ont montré notamment que les variables auxiliaires calculées pour les personnes ayant répondu « oui » à au moins une des questions filtre du recensement étaient généralement plus corrélées que lorsqu'elles étaient calculées pour l'ensemble de la population. De plus, ces tests ont permis de déterminer quelles interactions devraient être testées dans les modèles de régression.

Pour compléter la sélection des variables pour le modèle final, la régression pondérée a été utilisée. Le poids utilisé pour la régression correspond au rapport de la taille d'échantillon de la région sociosanitaire sur la taille totale de l'échantillon à l'EPLA. Ainsi, plus d'importance a été accordée aux liens existants entre les variables auxiliaires et la variable d'intérêt dans les plus grandes régions que dans les plus petites régions[5]. En procédant de la sorte, on a développé deux modèles de régression pondérée : un où la variable dépendante est le taux d'emploi à l'EPLA et l'autre où la variable dépendante est le taux d'activité à l'EPLA.

Afin de déterminer les variables indépendantes à retenir dans chacun de ces deux modèles, diverses techniques statistiques de sélection de variables[6] ont été utilisées. Les variables retenues devaient, autant que possible, faire l'objet d'un consensus selon les différentes méthodes de sélection. De plus, pour faire les choix de variables, les diagnostics d'hétéroscédasticité, de normalité des résidus et de multicollinéarité des variables explicatives ont été examinés. Enfin, l'interprétation que l'on peut faire du modèle obtenu a également été considérée lors du choix des variables.

Pour le taux d'emploi, les variables suivantes ont été retenues dans le modèle final :

- taux d'emploi par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes en milieu rural par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- log des heures moyennes travaillées par région chez les personnes ayant répondu « oui » à au moins une des questions filtre recensement

- proportion de personnes n'étant pas de minorités visibles par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- interaction du log des heures moyennes travaillées et de la proportion de personnes n'étant pas de minorités visibles chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

Le R-carré de ce modèle est de 0,76. Les paramètres de ce modèle sont présentés à l'annexe 2.

Pour le taux d'activité, les variables suivantes ont été retenues dans le modèle final :

- taux d'activité par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes en milieu rural par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes dont la principale source de revenu est le gouvernement par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

5. Cette pondération permet d'éviter d'accorder trop d'importance aux estimations ayant une trop grande erreur d'échantillonnage.
6. Les techniques « stepwise », « forward », « backward », Cp de Mallows et R-carré ajusté ont été utilisées.

- proportion de personnes vivant seules par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes âgées de 25 à 34 ans par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

Le R-carré de ce modèle est de 0,69. Ce R-carré montre que le modèle est moins bien ajusté que celui créé pour le taux d'emploi. Cependant, lors de la création de ce modèle, une attention particulière a été accordée aux résidus pour le Québec. Cela a permis de retenir le modèle qui donnait les meilleurs résultats pour le Québec. Ainsi, les résultats québécois de ce modèle devraient être aussi valables que les résultats québécois pour le taux d'emploi. L'annexe 2 présente les paramètres de ce modèle.

Afin de valider si ces deux modèles sont adéquats pour le Québec, on a ajouté une variable explicative supplémentaire au modèle pour la tester. Cette variable indique si une région provient du Québec ou non; elle s'est révélée non significative dans les deux modèles. Cela confirme que les variables déjà présentes dans les modèles expliquent adéquatement le taux d'emploi et le taux d'activité pour le Québec.

4. Résultats

4.1 Taux d'emploi

Le tableau 4.1.1 présente les résultats pour le taux d'emploi. Les deux premières colonnes du tableau donnent le taux d'emploi calculé à l'EPLA, son intervalle de confiance à 95 % ainsi que son coefficient de variation (CV). Les deux colonnes suivantes présentent le taux d'emploi obtenu à partir du modèle de Fay-Herriot, son intervalle de confiance à 95 % ainsi que son CV[7]. La dernière colonne du tableau fournit la proportion du taux d'emploi de Fay-Herriot qui dépend du taux d'emploi de l'EPLA.

Un examen de ce tableau montre que le taux d'emploi de Fay-Herriot est plus précis que celui calculé à l'EPLA. En général, les deux taux sont assez rapprochés. Cependant, on remarque pour les régions Bas-St-Laurent, Côte-Nord, Chaudière-Appalaches et Laval, un écart important entre les deux taux d'emploi. La section 6 de ce rapport traite de ces écarts et explique, à l'aide d'une source de données externe, pourquoi le modèle s'éloigne autant des valeurs obtenues à l'EPLA. Malgré cet éloignement, notons que, sauf pour le Bas-St-Laurent, l'intervalle de confiance de l'estimation du taux d'emploi à l'EPLA contient toujours la valeur du taux d'emploi obtenu par la méthode de Fay-Herriot.

Le tableau 4.1.1 indique également que les taux d'emploi de Fay-Herriot se fondent principalement sur la prédiction du modèle. Sauf pour la région de Montréal et la Montérégie, l'estimation de l'EPLA est représentée à moins de 10 % dans le calcul du taux d'emploi de Fay-Herriot.

7. On calcule ce CV en prenant la racine carrée de l'*EQM* divisée par l'estimateur de Fay-Herriot. Il faut noter que l'*EQM* est calculée selon les équations 7.1.26 et 7.1.29 de Rao (2003), donc elle ne contient que le biais provenant du modèle.

Tableau 4.1.1

Taux d'emploi de l'EPLA et de Fay-Herriot par région sociosanitaire, Québec

Régions sociosanitaires[1]	Taux d'emploi de l'EPLA		Taux d'emploi de Fay-Herriot		
	% [IC]	CV (%)	% [IC]	CV (%)	% provenant de l'EPLA
01-Bas-Saint-Laurent	18,3 [4,1 ; 32,5]	39,7	34,6 [30,1 ; 39,1]	6,6	3,4
02-Saguenay–Lac-Saint-Jean	31,2 [16,2 ; 46,2]	24,6	33,2 [28,6 ; 37,8]	7,0	4,2
03-Capitale-Nationale	40,5 [29,3 ; 51,7]	14,1	41,7 [37,0 ; 46,4]	5,8	8,3
04-Mauricie et Centre-du-Québec	30,6 [18,2 ; 43,0]	20,7	34,3 [29,7 ; 38,9]	6,9	7,6
05-Estrie	40,6 [29,1 ; 63,0]	18,8	41,5 [36,7 ; 46,3]	5,9	5,0
06-Montréal	40,8 [33,7 ; 47,9]	8,9	40,5 [35,7 ; 45,3]	6,1	23,5
07-Outaouais	42,3 [30,0 ; 54,7]	14,9	46,3 [42,1 ; 50,5]	4,6	7,4
08-Abitibi-Témiscamingue	41,6 [22,8 ; 60,4]	23,1	37,0 [32,9 ; 41,1]	5,7	3,0
09-Côte-Nord	55,6 [25,2 ; 86,0]	27,9	37,4 [33,4 ; 41,4]	5,5	1,4
11-Gaspésie–Îles-de-la-Madeleine	27,7 [1,0 ; 54,4]	49,2	25,7 [21,0 ; 30,4]	9,3	1,4
12-Chaudière-Appalaches	57,7 [39,8 ; 75,6]	15,8	45,6 [41,2 ; 50,0]	4,9	4,7
13-Laval	33,9 [18,2 ; 49,6]	23,6	48,4 [44,2 ; 52,6]	4,4	4,1
14-Lanaudière	43,2 [31,7 ; 54,7]	13,6	40,8 [36,5 ; 45,1]	5,4	7,3
15-Laurentides	42,4 [30,4 ; 54,4]	14,4	44,5 [40,1 ; 48,9]	5,0	7,3
16-Montérégie	42,5 [35,1 ; 49,9]	8,9	45,0 [40,3 ; 49,7]	5,3	19,0
Province	40,3 [37,1 ; 43,5]	4,1			

1. Le taux d'emploi pour les régions Nord-du-Québec et Nunavik n'est pas diffusé pour des raisons de confidentialité et de qualité. Le taux d'emploi de Fay-Herriot pour les régions Nord-du-Québec et Côte-Nord regroupées apparaît à l'annexe 3. Il n'est pas inclus dans ce tableau puisqu'il ne peut pas être comparé au taux d'emploi à l'EPLA. Ce dernier n'est pas présenté pour des raisons de confidentialité.

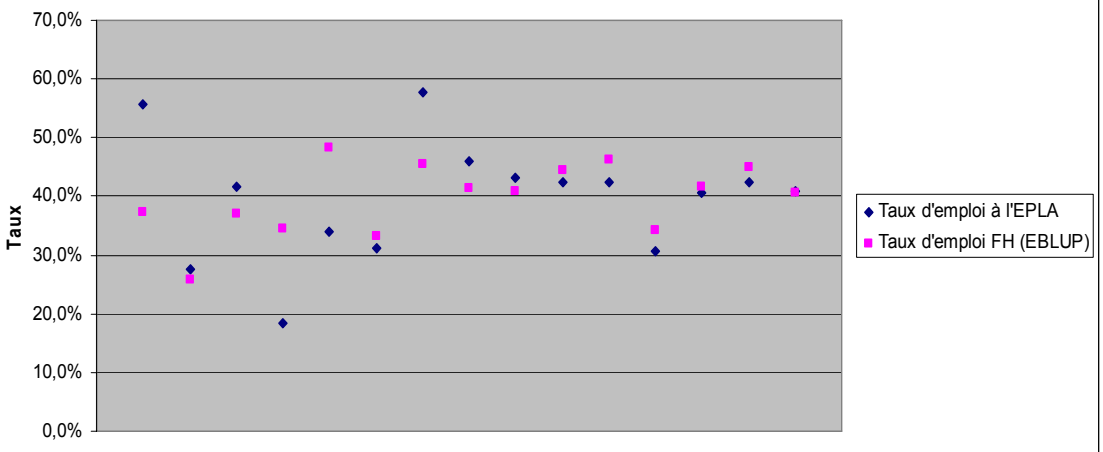
Source : Enquête sur la participation et les limitations d'activités de 2006 et Recensement de 2006.

Traitement : Institut de la statistique du Québec.

Notons que l'intervalle de confiance du taux d'emploi de Fay-Herriot a été calculé de la même façon que si le taux d'emploi avait été estimé à partir d'une enquête par échantillonnage comme l'EPLA. Cependant, dans le cas d'une estimation découlant d'un modèle, comme celle de Fay-Herriot, cette façon de faire se base sur l'hypothèse que le modèle est adéquat pour chacune des régions. On peut douter de cette hypothèse pour les plus petites régions. Ainsi, on ne peut garantir que le niveau de confiance des intervalles est vraiment de 95 %. Ces intervalles donnent tout de même une idée approximative de l'étendue des valeurs possibles de l'estimateur de Fay-Herriot. Ce problème existe également pour les intervalles de confiance obtenus pour le taux d'activité de Fay-Herriot.

Le graphique 1 superpose les estimations de l'EPLA à celles obtenues par le modèle de Fay-Herriot. On constate que plus la taille d'échantillon d'une région est grande, plus les deux estimations sont rapprochées. À l'inverse, les deux estimations ont tendance à s'éloigner lorsque la taille d'échantillon d'une région est petite. En général, on constate que les estimations de Fay-Herriot sont moins dispersées. En effet, les estimations de l'EPLA varient de 18,3 % à 57,7 % tandis que les taux de Fay-Herriot varient de 25,7 % à 48,4 %.

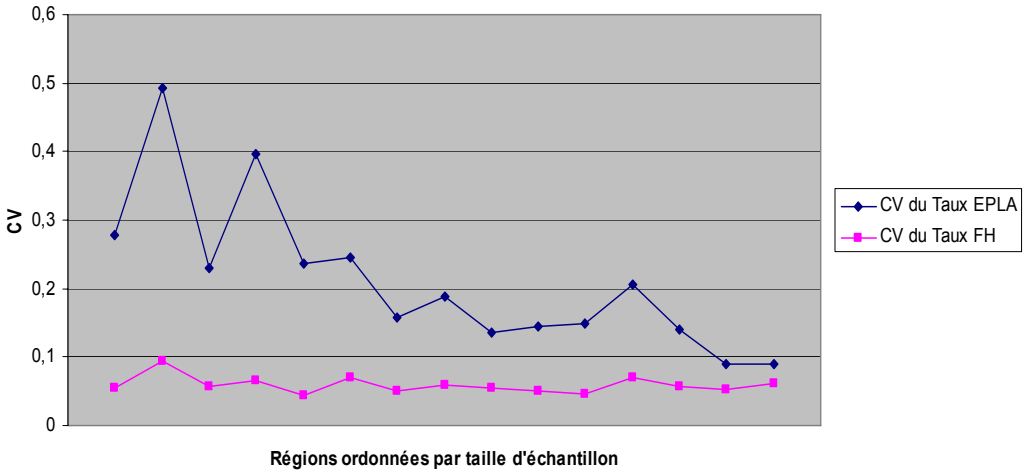
Graphique 1 : Distribution des taux d'emploi à l'EPLA et de Fay-Herriot selon les régions sociosanitaires québécoises



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

Le graphique 2 compare le CV des deux estimateurs. Lorsque la taille d'échantillon est grande, les deux CV sont très rapprochés. Inversement, plus la taille diminue, plus les deux CV s'éloignent. En fait, lorsque la taille d'échantillon est petite, le CV dépend presque exclusivement de la précision du modèle. On remarque que la précision du taux d'emploi de Fay-Herriot est constante peu importe la taille d'échantillon. Cela découle du fait que le taux de Fay-Herriot dépend principalement du modèle.

Graphique 2 : Distribution des coefficients de variation associés aux taux d'emploi à l'EPLA et de Fay-Herriot selon les régions sociosanitaires québécoises



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

4.2 Taux d'activité

Le tableau 4.2.1 présente les résultats obtenus pour le taux d'activité, soit le taux d'activité calculé à l'EPLA, le taux d'activité obtenu à partir du modèle de Fay-Herriot ainsi que leurs CV et intervalle de confiance respectifs. La dernière colonne du tableau présente, pour sa part, la proportion du taux d'activité de Fay-Herriot qui dépend du taux d'activité à l'EPLA.

Tableau 4.2.1

Taux d'activité de l'EPLA et de Fay-Herriot par région sociosanitaire, Québec

Régions sociosanitaires[1]	Taux d'activité à l'EPLA		Taux d'activité de Fay-Herriot		
	% [IC]	CV (%)	% [IC]	CV (%)	% provenant de l'EPLA
01-Bas-Saint-Laurent	24,1 [7,9 ; 40,4]	34,4	39,7 [32,9 ; 46,5]	8,7	9,8
02-Saguenay–Lac-Saint-Jean	34,1 [19,1 ; 49,1]	22,4	40,2 [33,5 ; 46,9]	8,5	11,8
03-Capitale-Nationale	43,1 [31,9 ; 54,3]	13,3	47,4 [40,9 ; 53,9]	7,0	21,0
04-Mauricie et Centre-du-Québec	36,2 [23,1 ; 49,3]	18,4	40,1 [33,6 ; 46,6]	8,3	19,6
05-Estrie	49,9 [32,8 ; 67,0]	17,5	46,6 [40,0 ; 53,2]	7,2	13,8
06-Montréal	52,5 [44,9 ; 60,1]	7,4	51,3 [45,6 ; 57,0]	5,7	43,1
07-Outaouais	49,2 [36,8 ; 61,6]	12,9	51,6 [45,6 ; 57,6]	5,9	19,1
08-Abitibi-Témiscamingue	47,5 [28,0 ; 67,1]	21,0	42,6 [36,2 ; 49,0]	7,7	8,7
09-Côte-Nord	61,5 [32,3 ; 90,7]	24,2	46,3 [40,1 ; 52,5]	6,8	4,3
11-Gaspésie–Îles-de-la-Madeleine	52,9 [23,9 ; 81,9]	28,0	37,4 [29,9 ; 44,9]	10,2	4,3
12-Chaudière-Appalaches	60,6 [43,5 ; 77,7]	14,4	50,3 [44,0 ; 56,6]	6,4	13,1
13-Laval	36,2 [20,3 ; 52,1]	22,4	50,2 [43,8 ; 56,6]	6,5	11,7
14-Lanaudière	51,2 [39,5 ; 62,9]	11,7	46,1 [40,1 ; 52,2]	6,7	19,0
15-Laurentides	50,1 [37,3 ; 62,9]	13,0	50,0 [44,0 ; 56,0]	6,1	19,0
16-Montérégie	46,1 [38,4 ; 53,8]	8,5	48,9 [43,5 ; 54,3]	5,6	38,0
Province	47,1 [43,7 ; 50,3]	3,6			

1. Le taux d'activité pour les régions Nord-du-Québec et Nunavik n'est pas diffusé pour des raisons de confidentialité et de qualité. Le taux d'activité de Fay-Herriot pour les régions Nord-du-Québec et Côte-Nord regroupées apparaît à l'annexe 3. Il n'est pas inclus dans ce tableau puisqu'il ne peut pas être comparé au taux d'activité à l'EPLA. Ce dernier n'est pas présenté pour des raisons de confidentialité.

Source : Enquête sur la participation et les limitations d'activités de 2006 et Recensement de 2006.

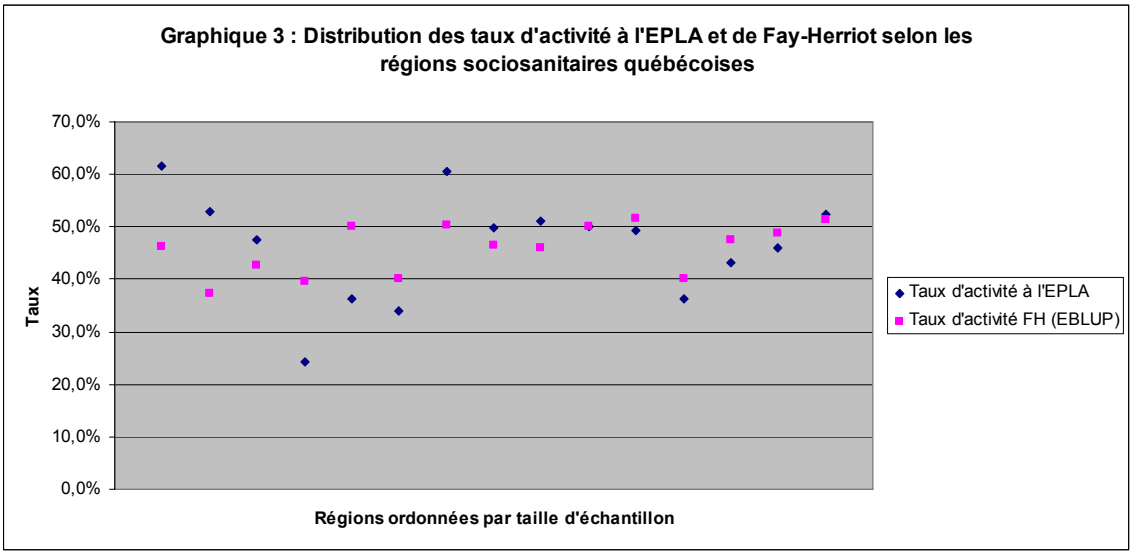
Traitement : Institut de la statistique du Québec.

Un examen de ce tableau montre que le taux d'activité de Fay-Herriot est encore une fois plus précis que celui calculé à l'EPLA. En général, les deux taux sont assez rapprochés. Cependant, on remarque un écart important entre les deux taux d'activité pour les régions Bas-St-Laurent, Côte-Nord, Gaspésie-Îles-de-la-Madeleine, Chaudière-Appalaches et Laval. Comme nous l'avons mentionné précédemment, la section 6 du présent rapport examinera pourquoi le modèle s'éloigne autant des valeurs obtenues à l'EPLA. Malgré cet éloignement, notons que l'intervalle de confiance de l'estimation du taux d'activité à l'EPLA contient toujours la valeur du taux d'activité obtenu par le modèle de Fay-Herriot.

Comme pour le taux d'emploi, les taux d'activité de Fay-Herriot se fondent principalement sur la prédiction du modèle. Cependant, la proportion représentée par l'estimation de l'EPLA est plus élevée qu'elle ne l'était pour le taux d'emploi. En effet, pour 11 régions, le taux d'activité direct représente plus de 10 % du calcul du taux d'activité de Fay-Herriot. Cela est certainement lié au fait que le modèle pour le taux d'activité est moins bien ajusté.

Le graphique 3 superpose les estimations de l'EPLA à celles obtenues par le modèle de Fay-Herriot. Les constats pour le taux d'activité sont similaires à ceux qui avaient été présentés pour le taux d'emploi. Ainsi, plus la taille d'échantillon d'une région est grande, plus les deux estimations sont rapprochées. À l'inverse, les deux estimations ont tendance à s'éloigner lorsque la taille d'échantillon d'une région est petite. De plus, les estimations de Fay-Herriot sont en général moins dispersées : les estimations de l'EPLA variant de 24,1 % à 61,5 % et celles de Fay-Herriot, de 37,4 % à 51,6 %.

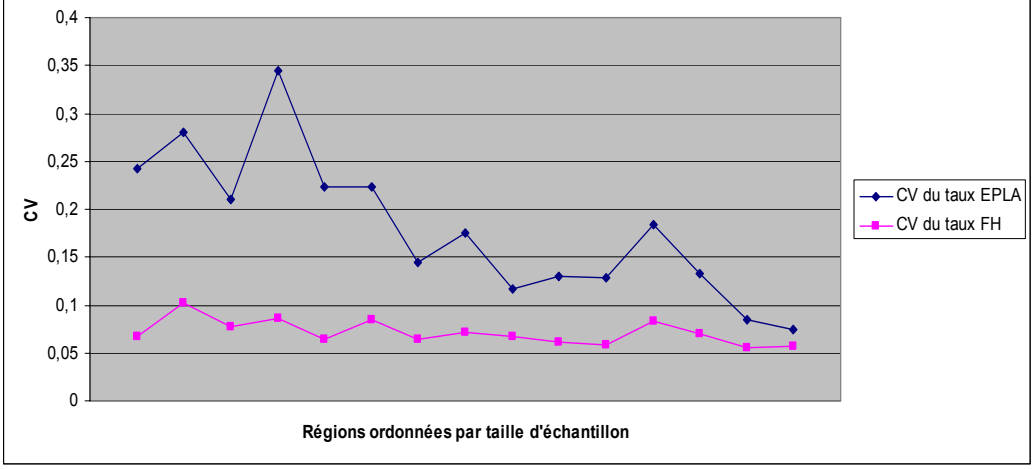
Graphique 3 : Distribution des taux d'activité à l'EPLA et de Fay-Herriot selon les régions sociosanitaires québécoises



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

Le graphique 4 compare le CV des deux estimateurs du taux d'activité. Encore une fois, les constats pour le taux d'activité vont dans le même sens que ceux concernant le taux d'emploi. Lorsque la taille d'échantillon est grande, les deux CV sont très rapprochés; et plus la taille diminue, plus les deux CV s'éloignent. En fait, lorsque la taille d'échantillon est petite, le CV dépend presque exclusivement de la précision du modèle. On remarque ici encore que la précision du taux d'activité de Fay-Herriot est constante peu importe la taille d'échantillon, ce qui découle du fait que le taux de Fay-Herriot dépend principalement du modèle.

Graphique 4 : Distribution des coefficients de variation associés aux taux d'activité à l'EPLA et de Fay-Herriot selon les régions sociosanitaires québécoises



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

4.3 Essai d'un modèle alternatif

La présente section relate l'essai d'un modèle alternatif pour la production d'estimations pour petits domaines. Par l'essai d'un tel modèle, l'ISQ visait à améliorer les estimations présentées aux deux sections précédentes. Si cela n'est pas possible, le nouveau modèle peut tout simplement confirmer la bonne qualité des modèles précédents.

Les résultats présentés aux sections précédentes montrent que les estimations obtenues pour les plus grandes régions sont de très bonne qualité. Dans une telle situation, les estimations de Fay-Herriot sont très près des estimations directes. Ce phénomène est normal étant donné que la pondération utilisée accorde plus d'importance à ces régions lors de la modélisation.

Par contre, du côté des régions ayant une plus petite taille d'échantillon, les écarts entre les estimations de l'EPLA et celles de Fay-Herriot sont plus grands. Est-ce que ces écarts sont liés seulement à l'imprécision des estimations directes? Est-ce que le fait de ne pas avoir assez accordé d'importance aux plus petites régions lors de la modélisation peut expliquer ces écarts? Pour tenter de répondre à ces questions, Louis-Paul Rivest a recommandé de procéder à l'essai d'une nouvelle modélisation qui accorderait un peu plus d'importance aux plus petites régions. Cela peut être réalisé en utilisant une pondération différente lors de la modélisation.

L'ISQ a suivi cette recommandation pour modéliser de nouveau le taux d'emploi. La nouvelle pondération utilisée se base sur l'inverse de la variance de Fay-Herriot. Cette pondération donne toujours plus d'importance aux plus grandes régions, mais permet d'atténuer quelque peu la différence entre les petites régions et les grandes. Les résultats obtenus avec cette nouvelle modélisation montrent que les variables retenues avec la première modélisation sont toujours significatives. Par ailleurs, un autre ensemble de variables significatives se distingue. Il s'agit des variables suivantes :

- taux d'emploi par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes en milieu rural par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes âgées de 15 à 24 ans par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes ayant dit avoir « souvent » des limitations parmi celles ayant répondu « oui » à au moins une des questions filtre du recensement

- proportion de personnes parlant l'une ou l'autre des langues officielles ayant répondu « oui » à au moins une des questions filtre du recensement

Ce nouveau modèle a été utilisé pour obtenir une deuxième série d'estimations pour le taux d'emploi. Après comparaison, il s'avère que les résultats sont très près des résultats obtenus à l'aide du premier modèle. Cependant, dans l'ensemble, ils sont un peu moins bons[8] que les résultats précédents, et ce, même pour les plus petites régions. Les résultats du modèle alternatif ne sont donc pas exposés dans le présent rapport. En conclusion, les résultats obtenus par le modèle alternatif appuient les résultats obtenus avec la première pondération. Pour cette raison, le modèle alternatif n'a pas été développé pour le taux d'activité.

5. Évaluation des modèles finaux

Ainsi que nous l'avons mentionné à la section 2.3, les estimateurs composites (comme celui utilisé pour ce projet) peuvent comporter un biais par rapport à la vraie valeur. Malheureusement, le biais intégré dans le calcul de la précision des estimateurs de Fay-Herriot présentés aux tableaux de la section 4 provient du modèle seulement. L'*EQM* est donc sous-estimée puisque le biais par rapport à la vraie valeur n'est pas quantifié. Les méthodes de diagnostic exposées dans la présente section nous permettent de porter un jugement sur l'existence d'un tel biais et d'en évaluer si possible l'ampleur.

5.1 Méthode d'évaluation du biais à l'aide de la pente de régression

Un premier diagnostic vise à examiner le biais des estimations de Fay-Herriot par rapport à la vraie valeur en utilisant les estimations directes de l'EPLA. Pour avoir une idée de ce biais, Brown et autres (2001) ont proposé d'effectuer une régression où la variable dépendante est ici le taux d'emploi à l'EPLA et où la variable explicative est le taux d'emploi obtenu par le modèle de Fay-Herriot. Si les estimations de Fay-Herriot sont sans biais par rapport à la vraie valeur, alors la pente

8. Afin de comparer la qualité des résultats, les critères présentés à la section 5 ont été utilisés.

de cette régression devrait être de « 1 » puisque les estimations de l'EPLA sont supposées sans biais. Plus la pente s'éloigne de « 1 », plus il y a de biais par rapport à la vraie valeur.

Les résultats de cette analyse montrent que pour le taux d'emploi, la pente est de 0,88. Cette pente n'étant pas significativement différente de « 1 », on ne peut conclure à la présence de biais (ou s'il y a un biais, il n'a pas été possible de le détecter à l'aide de cette procédure, ce qui est rassurant). Pour ce qui est du taux d'activité, la pente obtenue est de 0,93. Encore une fois, le nombre obtenu n'est pas significativement différent de « 1 ». Il est à noter que des résultats similaires sont obtenus pour les estimations des régions canadiennes.

5.2 Méthode de prédiction *a posteriori*

Un autre moyen d'évaluer la justesse des estimations de Fay-Herriot a été proposée par Meng (1994). Celui-ci suggère d'utiliser une méthode de prédiction *a posteriori* pour évaluer si la distribution des estimations de Fay-Herriot suit la même distribution que les estimations de l'EPLA, ces dernières étant sans biais par rapport à la vraie valeur. Cette méthode nécessite le calcul de la statistique suivante, qui suit une loi du Khi-carré :

$$T(\hat{\theta}_i^{FH}, \hat{\theta}) = \sum(\hat{\theta}_i - \hat{\theta}_i^{FH})^2 / \sigma_i^{\,2}$$

Si la valeur (« p-value ») associée à cette statistique est près de « 0,50 », alors cela signifie que les estimations de Fay-Herriot s'ajustent bien aux estimations de l'EPLA. La valeur (« p-value ») obtenue pour les estimations québécoises est de 0,56 pour le taux d'emploi et de 0,57 pour le taux d'activité. Ces valeurs étant très près de la valeur désirée, on ne peut conclure à la présence de biais. En terminant, il faut noter que les valeurs observées (« p-values ») pour les régions québécoises sont meilleures que celles obtenues pour les régions canadiennes.

5.3 Méthode des résidus en fonction des valeurs prédites

Cette méthode permet d'évaluer si les résidus du modèle de Fay-Herriot se distribuent uniformément selon les valeurs prédites. Si c'est le cas, on peut conclure que la variance du modèle est constante et que la qualité des résultats reste la même, quelles que soient les valeurs prédites. Des graphiques de résidus ont donc été réalisés. Ces graphiques suggèrent l'uniformité pour le taux d'emploi et le taux d'activité.

5.4 Méthode d'évaluation de la surestimation et de la sous-estimation

Dans la littérature, il est souvent mentionné que le biais se présente de la façon suivante : les petites proportions sont surestimées et les grandes proportions sont sous-estimées. Dans le présent projet, il est intéressant d'examiner le biais sous cet angle. Il faut déterminer tout d'abord ce qu'est une petite proportion et ce qu'est une grande proportion. Pour ce faire, le point de coupure utilisé ici est la moyenne du taux d'emploi canadien, soit 51 %. Ce qui est inférieur à ce chiffre est considéré comme une petite proportion et ce qui est supérieur comme une grande proportion. Treize régions québécoises sont sous ce seuil. La moyenne de leur estimation à l'EPLA est inférieure à la moyenne des taux obtenus par le modèle de Fay-Herriot (écart : 2,5 %). Cela semble indiquer une légère surestimation des petites proportions à l'EPLA. Par ailleurs, un examen des deux régions au-dessus du point de coupure indique également un écart. Cette fois-ci, la moyenne des estimations à l'EPLA est supérieure à la moyenne des taux de Fay-Herriot (écart : 15,2 %). Les estimations de l'EPLA pour ces deux régions (Côte-Nord et Chaudière-Appalaches) seraient donc sous-estimées. Toutefois, il serait hasardeux de considérer ces écarts uniquement comme des

biais. Ces écarts dépendent du biais et de l'erreur d'échantillonnage et il est difficile de départager ces deux quantités.

L'examen de la surestimation et de la sous-estimation a été répété pour le taux d'activité. Le point de coupure a été fixé cette fois-ci à 57,2 %. Les 13 régions québécoises qui sont sous ce seuil affichent un taux d'emploi moyen à l'EPLA inférieur au taux moyen obtenu par le modèle de Fay-Herriot (écart : 1,4 %). Cela semble indiquer une très faible surestimation. L'examen des deux régions supérieures au point de coupure montre que le taux moyen à l'EPLA est supérieur au taux moyen pour Fay-Herriot (écart : 12,8 %). Il semble donc y avoir une sous-estimation pour les régions de la Côte-Nord et de Chaudière-Appalaches. À titre de comparaison, au niveau des régions canadiennes, les écarts observés pour le taux d'emploi et le taux d'activité sont d'environ 5 % pour les petites et les grandes proportions.

Ces résultats montrent que les taux obtenus par le modèle de Fay-Herriot sont moins dispersés que ceux obtenus à l'EPLA. Ce phénomène semble confirmer la présence de biais par rapport à la vraie valeur. Les écarts observés dans la présente section peuvent être considérés comme des bornes supérieures pour ce biais.

5.5 Méthode de « couverture »

Finalement, un dernier diagnostic a été réalisé pour déterminer s'il existe un biais par rapport à la vraie valeur. Il s'agit d'une méthode qui évalue si les taux de Fay-Herriot sont « couverts » par les intervalles de confiance des taux obtenus à l'EPLA. En réalisant ce diagnostic, on peut constater que pour toutes les régions, sauf pour le Bas-St-Laurent et son taux d'emploi, les taux de Fay-Herriot se situent à l'intérieur d'un intervalle de confiance à 95 % construit à partir de l'estimation de l'EPLA. Étant donné que les estimations directes sont supposées sans biais de la vraie valeur, on peut penser que le biais associé aux taux de Fay-Herriot est faible sauf pour le Bas-St-Laurent et son taux d'emploi.

6. Validation externe des estimations modélisées

Les résultats présentés à la section 4 montrent que la plupart des taux de Fay-Herriot sont près des estimations de l'EPLA. En présence de faibles écarts, il ne semble pas difficile de faire confiance aux taux obtenus par le modèle de Fay-Herriot. Cependant, pour certaines régions, des écarts demeurent importants. Un moyen de réduire l'incertitude quant à la validité des estimations pour ces régions est de les soumettre à une validation externe, notamment en les comparant à d'autres sources de données. Le recensement canadien est la meilleure source permettant d'obtenir des données régionales sur ce sujet.

Le tableau 6.1 présente de nouveau les taux d'emploi à l'EPLA et les taux d'emploi obtenus à partir du modèle de Fay-Herriot. De plus, il présente deux taux d'emploi calculés à partir du recensement de 2006. Le premier de ces deux taux est calculé pour les personnes de 15 à 64 ans ayant répondu « oui » à au moins une des questions filtre du recensement sur l'incapacité. On se rappellera que ce taux a été retenu comme variable explicative dans le modèle servant à prédire le taux d'emploi. Le second taux présenté est le taux d'emploi calculé pour l'ensemble de la population de 15 à 64 ans (avec ou sans incapacité). Ce dernier taux s'éloigne nettement des trois premiers étant donné que son calcul porte sur une population très différente. Il représente une source de validation externe de qualité.

Tableau 6.1
Taux d'emploi de l'EPLA, de Fay-Herriot et du recensement de 2006 par région sociosanitaire, Québec

Régions sociosanitaires	Taux d'emploi à l'EPLA	Taux d'emploi de Fay-Herriot	Taux d'emploi[1] au recensement (groupe « oui »)	Taux d'emploi au recensement
			%	
01-Bas-Saint-Laurent	18,3	34,6	35,4	64,3
02-Saguenay–Lac-Saint-Jean	31,2	33,2	35,6	63,4
03-Capitale-Nationale	40,5	41,7	43,9	73,3
04-Mauricie et Centre-du-Québec	30,6	34,3	36,3	67,3
05-Estrie	46,0	41,5	40,8	69,5
06-Montréal	40,8	40,5	45,5	67,8
07-Outaouais	42,3	46,3	48,1	71,9
08-Abitibi-Témiscamingue	41,6	37,0	39,6	65,6
09-Côte-Nord	55,6	37,4	40,4	64,2
11-Gaspésie–Îles-de-la-Madeleine	27,7	25,7	27,9	53,6
12-Chaudière-Appalaches	57,7	45,6	44,0	74,3
13-Laval	33,9	48,4	52,1	74,2
14-Lanaudière	43,2	40,8	42,0	71,2
15-Laurentides	42,4	44,5	46,5	72,4
16-Montérégie	42,5	45,0	48,1	73,5

1. La valeur des taux présentés dans ce tableau diffère très légèrement de la valeur des taux utilisés pour la modélisation. Des taux pondérés sont présentés dans ce tableau, alors que, pour la modélisation, des taux non pondérés ont été utilisés. De plus, pour préserver la confidentialité, les taux fournis dans le présent rapport ont été arrondis.

Source : Enquête sur la participation et les limitations d'activités de 2006 et Recensement de 2006.

Traitement : Institut de la statistique du Québec.

Un examen du tableau montre que, pour 10 régions, le taux de Fay-Herriot se situe entre le taux à l'EPLA et le taux au recensement chez les personnes du groupe « oui ». Il est intéressant en particulier d'examiner les régions où il y a un écart important entre le taux à l'EPLA et le taux de Fay-Herriot. Par exemple, pour le Bas-St-Laurent, le taux de Fay-Herriot est très près du taux au recensement chez les « oui » (34,6 % et 35,4 %, respectivement). Il en va de même pour les régions de la Côte-Nord (37,4 % et 40,4 %, respectivement), de Chaudière-Appalaches (45,6 % et 44,0 %, respectivement) et de Laval (48,4 % et 52,1 %, respectivement). On note également que le taux d'emploi à l'EPLA pour le Bas-St-Laurent est nettement inférieur aux taux des autres régions. Par contre, pour les trois autres taux, c'est la Gaspésie–Îles-de-la-Madeleine qui obtient le taux le plus bas.

Le tableau 6.2 présente les taux d'activité calculés selon les mêmes méthodes. L'examen montre que, pour huit régions, les taux de Fay-Herriot se situent entre le taux à l'EPLA et le taux au recensement chez les « oui ». Encore une fois, il est intéressant d'examiner les régions qui se distinguent par des écarts importants entre les taux à l'EPLA et les taux de Fay-Herriot. Les taux de

Fay-Herriot se sont beaucoup rapprochés des taux calculés chez les « oui » au recensement pour les régions Bas-St-Laurent (39,7 % et 42,4 %, respectivement), Côte-Nord (46,3 % et 50,2 %, respectivement), Gaspésie–Îles-de-la-Madeleine (37,5 % et 37,8 %, respectivement) et Chaudière-Appalaches (50,3 % et 48,5 %, respectivement). Pour ce qui est de la région de Laval, le taux de Fay-Herriot s'est rapproché un peu du taux chez les « oui » au recensement (50,2 % et 58,7 %, respectivement). Le taux de Fay-Herriot pour Laval se situe donc dans les plus élevés au Québec, ce qui correspond aux deux taux calculés au recensement.

Tableau 6.2
Taux d'activité à l'EPLA, de Fay-Herriot et au recensement de 2006 par région sociosanitaire, Québec

Régions sociosanitaires	Taux d'activité à l'EPLA	Taux d'activité de Fay-Herriot	Taux d'activité[1] au recensement (groupe « oui »)	Taux d'activité au recensement
			%	
01-Bas-Saint-Laurent	24,1	39,7	42,4	73,1
02-Saguenay–Lac-Saint-Jean	34,1	40,2	42,9	72,1
03-Capitale-Nationale	43,1	47,4	50,0	78,7
04-Mauricie et Centre-du-Québec	36,2	40,1	43,2	74,1
05-Estrie	49,9	46,6	47,2	76,3
06-Montréal	52,5	51,3	54,5	76,3
07-Outaouais	49,2	51,6	54,2	78,2
08-Abitibi-Témiscamingue	47,5	42,6	46,1	74,2
09-Côte-Nord	61,5	46,3	50,2	74,2
11-Gaspésie–Îles-de-la-Madeleine	52,9	37,4	37,8	67,3
12-Chaudière-Appalaches	60,6	50,3	48,5	79,1
13-Laval	36,2	50,2	58,7	80,0
14-Lanaudière	51,2	46,1	48,2	76,9
15-Laurentides	50,1	50,0	52,4	78,3
16-Montérégie	46,1	48,9	54,4	79,1

1. La valeur des taux présentés dans ce tableau diffère très légèrement de la valeur des taux utilisés pour la modélisation. Les raisons énumérées pour le taux d'emploi s'appliquent également pour le taux d'activité.

Source : Enquête sur la participation et les limitations d'activités de 2006 et Recensement de 2006.

Traitement : Institut de la statistique du Québec.

Ces résultats sont rassurants et permettent de faire davantage confiance aux taux de Fay-Herriot lorsqu'il y a des écarts importants avec les taux à l'EPLA. Ainsi, malgré les inquiétudes soulevées à la section précédente, il n'y a pas lieu de croire que la sous-estimation des taux d'emploi et d'activité pour la Côte-Nord et Chaudière-Appalaches est aussi importante que ce que suggère la comparaison des taux à l'EPLA et de Fay-Herriot. De plus, dans le cas de la région du Bas-St-Laurent, il semble que ce soit plutôt le taux à l'EPLA qui soit une valeur aberrante.

Bien que ces résultats nous amènent à conclure que les estimations de Fay-Herriot sont de bonne qualité, nous recommandons qu'elles soient soumises à une autre étape de validation externe. Ainsi, il nous apparaît important que des experts et des futurs utilisateurs (au premier chef les organismes gouvernementaux concernés comme l'OPHQ et le MESS) examinent les taux de Fay-Herriot obtenus et les valident à la lumière de l'information dont ils disposent au niveau régional. En Grande-Bretagne (Haworth et Cruddas, 2003), une telle étape de validation a été jugée nécessaire avant la diffusion des données comme statistiques officielles (auparavant, ces données étaient présentées comme exploratoires). Nous recommandons une démarche similaire.

7. Interprétation

La présente section présente un exemple d'analyse des taux calculés à partir du modèle de Fay-Herriot (taux FH) et fournit un guide pour leur interprétation.

Pour une interprétation juste des taux calculés à partir du modèle de Fay-Herriot, il convient de rappeler que ces taux doivent être considérés comme des prédictions provenant d'un modèle. Étant donné la faible représentativité des données observées à l'EPLA dans le calcul des taux FH, ceux-ci ne peuvent être considérés comme des estimations provenant d'une enquête. Contrairement aux estimations habituelles qui ne comportent pas de biais, les taux FH sont biaisés. Bien que ce biais ne semble pas très important, il est tout de même présent, particulièrement pour les plus petites régions. Lors de la modélisation, il était difficile d'expliquer les facteurs faisant varier le taux d'emploi et le taux d'activité dans les plus petites régions étant donné la présence d'une grande erreur d'échantillonnage. Malgré la présence de biais, les résultats sont présentés avec l'hypothèse que le modèle est adéquat pour chaque région, même les plus petites. Il est donc important de toujours rappeler la distinction entre les estimations provenant d'une enquête et les taux FH lorsque ces derniers sont présentés dans un rapport. Les utilisateurs pourront alors être conscients de la portée et des limites liées à de tels résultats. De la même manière, il faudra expliquer que les CV ou les tests effectués sont obtenus à partir d'une modélisation.

À titre d'exemple d'analyse, la présente section compare les taux d'emploi et d'activité régionaux de Fay-Herriot au taux provincial obtenu à l'EPLA. Précisons d'abord quelques aspects méthodologiques. Lorsqu'on est en présence d'estimations directes, la comparaison d'une estimation régionale à une estimation provinciale revient à comparer l'estimation régionale à celle de l'ensemble des autres régions de la province. Cette façon de faire simplifie beaucoup les calculs puisqu'on peut négliger la corrélation entre les deux estimations. Dans la situation présente, on procède de la même façon. Afin de comparer l'estimation régionale FH à l'estimation directe provinciale, on compare l'estimation régionale FH avec l'estimation directe de l'ensemble des autres régions de la province. Cependant, même si cette façon de faire diminue grandement la corrélation entre les deux estimations, elle ne l'élimine pas complètement. En effet, la partie synthétique de l'estimateur FH a été obtenue à partir de l'ensemble des régions du Canada. Donc, il existe une certaine corrélation entre l'estimation régionale FH et l'estimation directe de l'ensemble des autres régions de la province. On pense que cette corrélation est très petite et positive. Toutefois, l'ISQ a décidé de négliger ces corrélations dans les comparaisons effectuées ici. Il en découle que les résultats des tests de différences entre les régions sont quelque peu conservateurs puisqu'on surestime la variance de la différence.

En utilisant la méthodologie présentée au paragraphe précédent, on a effectué des tests de comparaison tout d'abord pour le taux d'emploi. Rappelons d'abord que, selon l'EPLA, le taux d'emploi au Québec au sein de la population ayant une incapacité est de 40,3 % en 2006. Selon les taux régionaux modélisés, les régions Bas-Saint-Laurent (34,6 %), Saguenay–Lac-Saint-Jean (33,2 %), Mauricie et Centre-du-Québec (34,3 %) et Gaspésie–Îles-de-la-Madeleine (25,7 %) affichent un

taux d'emploi significativement plus bas que le taux d'emploi des autres régions du Québec, chez les personnes ayant une incapacité. Par ailleurs, les régions Outaouais (46,3 %), Chaudière-Appalaches (45,6 %) et Laval (48,4 %) ont un taux d'emploi modélisé significativement supérieur au taux d'emploi de l'ensemble des autres régions du Québec.

De plus, il faut souligner que la région de la Montérégie (45,0 %) semble avoir un taux d'emploi modélisé supérieur à celui des autres régions. Bien que cette différence ne soit pas significative au seuil de 5 %, il est utile de le mentionner étant donné que la méthode de comparaison utilisée est conservatrice et qu'une différence aurait été détectée pour cette région à un seuil qui est très près de 5 %.

Pour ce qui est du taux d'activité, il est selon l'EPLA de 47,1 % chez les personnes ayant une incapacité au Québec. Selon les taux modélisés, les régions Bas-Saint-Laurent (39,7 %), Mauricie et Centre-du-Québec (40,1 %) et Gaspésie–Îles-de-la-Madeleine (37,4 %) affichent un taux significativement inférieur au taux d'activité de l'ensemble des autres régions du Québec. Il faut souligner également que la région Saguenay–Lac-Saint-Jean (40,2 %) semble avoir un taux d'activité modélisé inférieur à celui des autres régions. Cette différence n'est pas significative au seuil de 5 % en raison de l'approche conservatrice retenue, mais aurait pu l'être à un seuil très près de 5 %.

Il est à noter également qu'aucune région n'a un taux d'activité modélisé significativement supérieur au taux d'activité global des autres régions. Il y a moins de différences significatives détectées pour le taux d'activité étant donné que les prédictions de FH pour le taux d'activité sont moins précises que celles obtenues pour le taux d'emploi.

En terminant, précisons que les tests évoqués dans la présente section comportent certaines limites sur le plan strictement statistique. En plus d'être conservateurs, ces tests sont effectués comme si les taux comparés provenaient seulement d'une enquête par échantillonnage. En réalité, la comparaison porte en partie sur des taux modélisés (taux FH), ce dont les tests ne tiennent pas compte. Malgré cette mise en garde, ces tests constituent tout de même une approximation acceptable pour la détection de différences significatives.

8. Utilité de la méthode et conclusion

Les estimations pour petits domaines exposées dans le présent rapport pour le taux d'emploi et le taux d'activité ont été obtenues à l'aide du modèle de Fay-Herriot. Les multiples validations effectuées aux sections précédentes montrent que le modèle est de très bonne qualité. Il n'en demeure pas moins que les taux obtenus proviennent principalement de la modélisation et donc qu'ils ne peuvent pas être considérés comme des estimations directes provenant d'une enquête. Par conséquent, il faut les utiliser en mentionnant clairement qu'ils proviennent d'un modèle et qu'ils sont des prédictions de ce modèle. De plus, les taux obtenus par modélisation reposent sur l'hypothèse que le modèle est adéquat pour chacune des régions. Toutefois, ces taux modélisés peuvent être d'une grande utilité, car ils permettent de combler le manque d'information au niveau régional sur le statut d'emploi des personnes ayant une incapacité au Québec.

Il serait souhaitable que la diffusion des taux obtenus par modélisation soit précédée d'une étape de validation externe auprès d'experts ou d'utilisateurs ciblés, dont l'OPHQ et le MESS au premier chef. Il est important selon nous que les utilisateurs examinent les taux diffusés dans le présent rapport à l'aide des sources d'information dont ils disposent dans leur milieu. Des résultats allant dans le même sens que ceux figurant dans le présent rapport permettraient aux utilisateurs de se sentir en confiance avec la méthodologie et les résultats. Ceux qui auront à utiliser ces données au

quotidien doivent être convaincus de leur utilité. Si une telle étape de validation est concluante, les taux régionaux obtenus par modélisation pourraient être utilisés officiellement pour l'analyse du statut d'emploi chez les personnes ayant une incapacité au Québec.

D'autres travaux de ce type pourraient être réalisés si cette étape est concluante. Dans le futur, il serait possible d'appliquer la méthodologie utilisée dans ce projet aux données d'une éventuelle EPLA (par exemple en 2011) pour produire de nouveau des taux d'emploi et d'activité par région sociosanitaire. Ces travaux devraient également comprendre une validation interne et externe de la qualité des estimations. On pourrait aussi envisager d'utiliser les taux modélisés pour étudier l'évolution dans le temps (par exemple de 2006 à 2011) des taux régionaux d'emploi et d'activité pour la population avec incapacité.

S'il y a un intérêt pour cette information, on pourrait aussi étudier la possibilité de procéder à la prédiction du taux d'emploi et du taux d'activité pour des sous-régions à l'intérieur de régions sociosanitaires populeuses comme Montréal. Toutefois, les prédictions obtenues seraient encore moins fondées sur les estimations directes de l'EPLA et dépendraient encore plus du modèle.

Par ailleurs, sur le plan strictement statistique, il serait intéressant de tester d'autres modèles que celui de Fay-Herriot avec les mêmes données : par exemple, le modèle de liaison log-linéaire non apparié au modèle d'échantillonnage de Statistique Canada. Des résultats semblables obtenus par ce modèle pourraient confirmer davantage les conclusions du présent rapport.

De plus, lors de travaux futurs, il serait intéressant de corriger le problème de sous-dispersion des données. À cet effet, Rao (2003) propose de reproduire les estimations pour petits domaines en accordant plus d'importance aux estimations directes. De cette manière, les résultats obtenus peuvent posséder une dispersion qui semble plus près de la réalité.

D'autres travaux devraient être menés également afin d'obtenir le calcul exact des tests de comparaison des taux FH avec les taux globaux des autres régions. Cela permettrait sûrement de détecter d'autres différences significatives.

En terminant, il faut se demander si la méthodologie développée ici peut s'appliquer à d'autres statistiques que le taux d'emploi et le taux d'activité. Existe-t-il, pour ces autres statistiques, des variables auxiliaires de sources externes aussi corrélées avec ce qu'on veut prédire que celles utilisées pour ce projet? Si de telles variables existent dans le fichier du recensement par exemple, alors un tel projet serait réalisable. Cependant, il faudrait s'assurer que tous les facteurs importants liés à ces statistiques sont quantifiables et disponibles dans le fichier du recensement.

9. Références

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.

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.

GAGNON, É. (2008). *Avis sur les méthodes d'estimation pour petites régions dans le cadre des enquêtes de santé*, Institut de la statistique du Québec, octobre 2008.

HAWORTH, M., et M. CRUDDAS (2003). « Developing small area estimates in the UK – A practitioners' perspective », Proceedings of Statistics Canada Symposium 2003.

MENG, X.L. (1994). « Posterior predictive *p*-values », *The Annals of Statistics*, vol. 22, n° 3, p. 1142-1160.

RAO, J.N.K. (2003). *Small Area Estimation,* New York, John Wiley & Sons, 344 p.

Annexe 1

Avis sur les méthodes d'estimation pour petites régions dans le cadre des enquêtes de santé

Éric Gagnon
DMDES
Institut de la statistique du Québec

Novembre 2008

Table des matières

1. Contexte

Les besoins d'estimation de statistiques sur la santé pour de petites régions québécoises sont grandissants. Bien souvent, dans les enquêtes de santé menées au Québec, le plan de sondage ne permet pas d'obtenir d'estimation précise pour ces petites régions; par exemple, les tailles d'échantillon peuvent ne pas être suffisantes. Dans une telle situation, l'estimateur direct (celui obtenu en utilisant l'estimation pondérée habituelle) se révèlera souvent de faible précision. Afin d'améliorer la qualité des estimations régionales, on retrouve dans la littérature différentes méthodes d'estimation pour petites régions faisant appel à de l'information auxiliaire tels le recensement ou les données administratives. Ces méthodes permettent de produire des estimations précises pour des territoires ou domaines plus petits que ceux prévus par le plan de sondage et les estimateurs directs. En raison des limites des estimateurs traditionnels, le ministère de la Santé et des Services sociaux a demandé à l'Institut de la statistique du Québec (ISQ) de dresser un portrait des différentes méthodes d'estimation pour petites régions et d'évaluer la possibilité d'utiliser une de ces méthodes pour améliorer la qualité des estimations produites pour de petites régions québécoises.

Ce document présente les résultats de l'étude menée par l'ISQ concernant l'estimation pour petites régions. La section 2 de ce document présente brièvement la revue de la littérature effectuée par l'ISQ. La section 3 donne une brève description de la méthode traditionnellement utilisée (méthode directe) pour la production d'estimations à partir de petites régions. La section 4 présente la méthode d'estimation synthétique pour petites régions. Dans cette section, des exemples provenant de trois enquêtes utilisant des méthodes synthétiques sont présentés. La section 5 présente la méthode d'estimation composite pour petites régions. Dans cette section, trois exemples d'application des méthodes composites sont présentés. Finalement, la section 6 présente notre conclusion et décrit quelles sont les possibilités d'application de l'une ou l'autre de ces méthodes, le tout dans le but d'améliorer la qualité des estimations québécoises.

2. Revue de la littérature

Pour réaliser ce mandat, l'ISQ a procédé à une revue des articles traitant de méthodes d'estimation pour petites régions dans le cadre d'enquêtes de santé. Cette revue de littérature a montré qu'au début des années 1980, la recherche concernant les méthodes d'estimation pour petites régions était assez limitée dans le domaine de la santé (Statistique Canada, 1983; Statistique Canada, 1986). Par contre, plusieurs documents ont été trouvés concernant ce type d'estimation pour des enquêtes à caractère économique. De plus, les quelques articles recensés et portant sur des enquêtes de santé se rattachaient presque exclusivement au « National Center for Health Statistics » des États-Unis et à son enquête le « National Health Interview Survey (NHIS) » qui s'appelait à l'époque « Health Interview Survey » (NCHS, 1968 ; NCHS, 1977a; NCHS, 1977b; Levy et French, 1977; NCHS, 1978 ; Schaible et coll., 1979 ; Digaetano et coll., 1980).

Dans les années qui ont suivi, des statisticiens responsables d'autres enquêtes de santé ont commencé à s'intéresser à ces méthodes d'estimation. Des articles plus récents provenant de plusieurs pays tels que les États-Unis, le Canada, l'Angleterre, la France et Taïwan ont été répertoriés. Ces articles montrent qu'il existe deux types de méthode d'estimation pour petites régions : les méthodes synthétiques et les méthodes composites. Un résumé des articles les plus intéressants portant sur ces méthodes est présenté dans les sections 4 et 5 de ce document.

3. Méthode d'estimation directe pour petites régions

L'estimateur direct est celui obtenu en utilisant l'estimation pondérée habituelle. En raison des plans de sondage utilisés, il arrive souvent que cet estimateur soit de faible précision pour une petite région. À la limite, s'il n'y a pas d'unité échantillonnée dans une petite région, il n'est pas

possible de produire d'estimation directe pour celle-ci. En revanche, s'il est possible de produire une estimation directe pour une petite région, il est certain que cette estimation est non biaisée.

Le défi avec les méthodes d'estimation pour petites régions est de trouver un estimateur qui aura une meilleure précision que l'estimateur direct et qui n'aura pas un biais trop important. Il existe une mesure qui permet de combiner ces deux aspects : il s'agit de l'erreur quadratique moyenne ou *EQM* :

$$EQM = variance + biais^2$$

Dans cette équation représentant l'*EQM*, la partie représentée par la variance correspond à la précision de l'estimateur. Dans le cas de l'estimateur direct, l'*EQM* est égale à la variance de l'estimateur seulement puisque celui-ci n'est pas biaisé.

4. Méthode d'estimation synthétique pour petites régions

Les estimateurs indirects ou synthétiques pour petites régions dépendent, en général, d'un modèle qui intègre des variables auxiliaires provenant de sources externes à l'enquête telles que le recensement ou les données administratives. L'intégration de variables auxiliaires et l'adéquation du modèle vont permettre de produire des estimations de meilleure précision que les estimateurs directs. Si les hypothèses du modèle ne tiennent pas pour une région, l'estimateur synthétique sera biaisé pour cette région. Des détails théoriques concernant cette méthode d'estimation peuvent être trouvés dans Rao (2003). La suite de cette section porte sur des exemples de méthodes synthétiques appliquées à des petites régions en France, en Angleterre et aux États-Unis.

4.1 Méthode d'estimation synthétique pour petites régions utilisée pour l'enquête « Handicaps-Incapacités-Dépendance (HID) » (France)

L'enquête HID a été réalisée auprès de 16 945 individus sélectionnés à partir des 360 000 ayant répondu à la pré-enquête VQS (Vie Quotidienne et Santé) associée au recensement français. Pour cette enquête, la production d'estimation de prévalence des handicaps de population pour 8 régions et 91 départements était souhaitée. Comme l'estimateur direct de l'échantillon de l'enquête HID ne permettait pas d'obtenir des estimations précises, un estimateur synthétique a été utilisé (Couet, 2002). L'hypothèse de base de l'estimateur utilisé suppose que le comportement moyen dans un département à l'intérieur d'un sous-groupe est identique au comportement moyen national pour ce même sous-groupe. Les sous-groupes, définis indépendamment des prévalences estimées, sont formés à partir des variables suivantes : le sexe, la classe d'âge, la tranche d'unité urbaine et le groupe VQS. En d'autres termes, le modèle est identique indépendamment de la variable étudiée. Par la suite, pour chaque localité, l'échantillon national a été pondéré pour représenter la répartition de leur population selon les sous-groupes. Les localités pouvaient, par la suite, obtenir des estimations locales avec cette pondération. Pour réaliser cette pondération, il était évidemment nécessaire de disposer des comptes du recensement (ou d'estimations provenant de l'enquête VQS) concernant le nombre de personnes dans chacun des sous-groupes pour chacune des localités.

L'utilisation d'un estimateur synthétique permet, tel qu'il est mentionné précédemment, d'obtenir une meilleure précision que l'estimateur direct. Cependant, cette amélioration ne doit pas se faire au prix d'une augmentation trop importante du biais. C'est pourquoi, dans le cadre de ce projet, une étude a été menée afin d'évaluer le biais lié à cet estimateur. Cet examen a été mené pour une région disposant d'une taille d'échantillon suffisante pour obtenir un estimateur direct de qualité. De cette façon, l'estimateur synthétique a pu être comparé à l'estimateur direct. Le

constat découlant de cette comparaison est que l'estimateur synthétique affichait un biais pour toutes les variables analysées. L'estimateur synthétique obtenu était toujours plus grand que l'estimateur direct.

Afin de diminuer le biais de cet estimateur, les Français ont pensé à différentes solutions. Malheureusement, ces solutions n'ont pas amené les résultats escomptés. Une des solutions proposées était d'augmenter le nombre de variables auxiliaires pour créer les sous-groupes. Effectivement, ceci a permis de diminuer le biais. Cependant, ce gain a été possible au détriment de la précision de l'estimateur. Celle-ci devenait moins bonne étant donné l'augmentation du nombre de sous-groupes créés pour la pondération et la baisse des effectifs au niveau national pour chacun de ces sous-groupes.

Notons, en terminant, que cette méthode utilisée par la France ressemble à la méthode australienne des « small area predictors of disability ». Notons également que les Australiens ont fait l'examen d'autres estimateurs qui permettraient d'obtenir des estimations moins biaisées que cet estimateur synthétique. Plus de détails à ce sujet sont donnés dans Elazar et Conn (2004).

4.1.1 Forces et faiblesses de l'estimateur synthétique pour l'enquête HID

Forces de l'estimateur synthétique pour l'enquête HID:

- Meilleure précision que l'estimateur direct.
- L'estimateur est obtenu même pour des régions qui n'ont pas été échantillonnées.
- L'estimateur est simple à utiliser.

Faiblesses de l'estimateur synthétique pour l'enquête HID:

- Il faut disposer des comptes du recensement ou d'estimations provenant de l'enquête VQS (Vie Quotidienne et Santé) menée auprès d'un échantillon de plus grande taille que celui utilisé pour l'enquête HID. Ces comptes sont nécessaires pour connaître la répartition régionale selon les sous-groupes.
- L'estimateur est biaisé. Il est toujours plus grand que l'estimateur direct pour la région étudiée.
- L'estimateur sera le même pour deux régions qui ont la même distribution de population selon les sous-groupes utilisés, et ce, même si les deux régions ont d'autres caractéristiques démographiques très différentes.
- Les estimations régionales ont tendance à se regrouper autour de l'estimation nationale. L'étendue des estimations régionales possibles s'amenuise.

4.2 Méthode d'estimation synthétique utilisée pour le Health Survey for England (Angleterre)

En 2004, le « National Centre for Social Research (NatCen) » a été mandaté par le « Department of Health » en Angleterre pour produire des estimations sur des comportements de santé au niveau des petites régions anglaises[9] en utilisant les données du « Health Survey for England (HSfE)» (Bajekal et coll., 2004; Pickering et coll., 2004; Pickering et coll., 2005).

9. Des estimations seront produites pour environ 8 000 petites régions anglaises appelées « ward ».

Deux méthodes d'estimation synthétique ont été retenues par le NatCen pour obtenir de telles estimations :

1) Estimation synthétique à l'aide de modèles multiniveaux et avec des variables auxiliaires caractérisant des régions.
2) Estimation synthétique à l'aide de modèles multiniveaux et avec des variables auxiliaires caractérisant des régions et des variables auxiliaires caractérisant les individus.

La première méthode consiste à prédire ou à estimer la prévalence d'une caractéristique de santé en utilisant seulement de l'information caractérisant les régions. Des informations du recensement ou d'autres sources administratives telles que le revenu moyen, l'espérance de vie et la proportion de personnes à faible revenu dans une région peuvent être utilisées. Ces informations peuvent être disponibles pour les petites régions ou bien pour d'autres régions plus grandes. La présence d'informations disponibles pour différents niveaux de région implique une approche de modélisation multiniveau.

Pour créer un modèle multiniveau, les données du HSfE sont utilisées. La variable dépendante est la prévalence de la caractéristique de santé, et les différentes informations caractérisant les régions constituent les variables explicatives du modèle. Une fois le modèle ajusté, celui-ci est utilisé afin de prédire la prévalence de santé pour toutes les petites régions étudiées. Cette procédure doit être reprise pour chaque prévalence de santé à prédire.

La seconde méthode employée pour estimer la prévalence de santé consiste à utiliser, en plus des variables auxiliaires de région, des variables auxiliaires caractérisant les individus. Des informations telles que l'âge de l'individu, le sexe ou son état matrimonial peuvent être utilisées. Cette méthode permet donc de tenir compte d'un effet régional et d'un effet individuel pour la prédiction. La présence d'informations disponibles au niveau individuel et pour différents niveaux régionaux implique une approche de modélisation multiniveau.

Encore une fois, pour créer un modèle multiniveau, les données du HSfE sont utilisées. On retient la prévalence de la caractéristique de santé étudiée comme variable dépendante et les différentes informations individuelles et régionales constituent les variables explicatives du modèle. Une fois le modèle ajusté, celui-ci est utilisé afin de prédire la prévalence de santé pour chaque individu présent dans les petites régions. Les prévalences prédites peuvent être ensuite combinées pour obtenir la prévalence pour la petite région. Comme pour la première méthode, cette procédure doit être reprise pour chaque prévalence de santé à prédire.

Cette seconde méthode comporte une difficulté importante liée à l'étape de la prédiction de la prévalence. En effet, pour effectuer la prédiction de la prévalence pour les individus, il faut évidemment disposer, pour chacun des individus, de l'information individuelle utilisée dans le modèle. Cette information doit donc être disponible pour le recensement. La quantité d'information de niveau individuel provenant du recensement est très limitée, ce qui implique finalement que très peu de variables de niveau individuel peuvent être retenues dans le modèle.

Par ailleurs, les deux méthodes retenues pour ce projet sont biaisées. Une recherche antérieure (Twigg et Moon, 2002) a montré que de tels estimateurs donnaient, pour de petites prévalences, des estimations de 20 % supérieures aux estimations directes et pour de grandes prévalences, des estimations de 10 % inférieures aux estimations directes. Malgré ce biais, le NatCen pense que ces deux estimateurs peuvent être très bons pour effectuer des classements de régions par exemple.

Il est à noter que la méthode utilisée par la France pour l'enquête HID a également été testée dans le cadre de ce projet. Cependant, celle-ci n'a pas été retenue étant donné qu'elle est davantage biaisée que les deux autres méthodes.

Malgré les différentes faiblesses des deux estimateurs présentés dans cette section, le NatCen se sent à l'aise de recommander leur utilisation pour l'estimation de prévalence de santé pour de petites régions anglaises. En fait, les chercheurs en sont venus à la conclusion que les deux méthodes amenaient des résultats équivalents. Étant donné que la méthode utilisant le niveau région seulement est plus simple et que celle-ci est déjà utilisée par l'« Office for National Statistics (ONS) » en Angleterre pour d'autres estimations pour petites régions, le NatCen recommandera probablement cette dernière approche pour des estimations officielles de prévalence de santé.

4.2.1 Forces et faiblesses des estimateurs synthétiques utilisés pour le HSfE

Forces de la méthode de niveau région seulement :

- Meilleure précision que l'estimateur direct.
- L'estimateur est obtenu même pour des régions qui n'ont pas été échantillonnées.
- Il n'y a pas de contrainte liée à la disponibilité des données individuelles.
- Cette méthode est moins biaisée que d'autres méthodes synthétiques comme celle proposée pour l'enquête HID en France.

Faiblesses de la méthode de niveau région seulement :

- L'estimateur est biaisé. L'estimateur sous-estime les grandes prévalences et surestime les petites prévalences.
- Cette approche ne permet pas de produire des estimations par sous-groupe démographique à l'intérieur des petites régions.
- L'application de cette méthode peut nécessiter beaucoup de travail. En effet, un modèle doit être créé pour chaque prévalence de santé à prédire.

Forces de la méthode utilisant le niveau individuel et régional :

- Meilleure précision que l'estimateur direct.
- L'estimateur est obtenu même pour des régions qui n'ont pas été échantillonnées.
- Cette méthode est moins biaisée que d'autres méthodes synthétiques comme celle proposée pour l'enquête HID en France.
- Cette approche permet d'obtenir des estimations par sous-groupe démographique à l'intérieur des petites régions.

Faiblesses de la méthode utilisant le niveau individuel et régional :

- L'estimateur est biaisé. L'estimateur sous-estime les grandes prévalences et surestime les petites prévalences.
- La quantité d'information individuelle qui peut être incluse dans le modèle est assez limitée.
- L'application de cette méthode peut nécessiter beaucoup de travail. En effet, un modèle doit être créé pour chaque prévalence de santé à prédire.

4.3 Méthode d'estimation synthétique utilisée pour le NHANES (États-Unis)

Dans le cadre du « National Health and Nutrition Examination Survey III (NHANES III) » aux États-Unis, des estimations de prévalence d'obésité pour petites régions ont été produites en utilisant un estimateur synthétique (Malec et coll., 1996). Cet estimateur est obtenu à l'aide d'un modèle hiérarchique. Ce modèle s'apparente au modèle multiniveau du HSfE utilisant de l'information individuelle et régionale.

Pour effectuer la modélisation, de l'information individuelle est utilisée (sexe, ethnie en 3 catégories, phase de l'enquête, âge en groupe de 5 ans) ainsi que de l'information régionale. Pour la modélisation, on compte 30 variables régionales utilisées dans le modèle. Ces variables sont reliées à des thèmes tels que l'éducation (ex : % de personnes avec un diplôme collégial), l'économie (ex : taux de pauvreté), la démographie (ex : % de la population vivant en milieu rural), la population active (ex : % de la population travaillant dans le domaine de la construction) et les soins de santé (ex : nombre d'hôpitaux). Ces informations régionales proviennent du « Area Resource File » détenu par le « U.S. Department of Health and Human Services ».

Avant d'utiliser le modèle pour prédire les résultats pour les petites régions, Malec et coll. (1996) ont validé le modèle en utilisant les estimations nationales produites par l'estimateur direct. La comparaison entre les données prédites par le modèle et les estimations directes a été effectuée pour 78 sous-groupes démographiques. Ces comparaisons ont montré qu'au niveau national les estimations obtenues par le modèle étaient équivalentes aux estimations obtenues directement. Ce qui signifie qu'au niveau national, le biais est négligeable.

4.3.1 Forces et faiblesses de l'estimateur synthétique utilisé pour le NHANES III

Forces de l'estimateur synthétique pour le NHANES III :

- Meilleure précision que l'estimateur direct.
- L'estimateur est obtenu même pour des régions qui n'ont pas été échantillonnées.
- Cette approche permet d'obtenir des estimations par sous-groupe démographique à l'intérieur des petites régions.

Faiblesses de l'estimateur synthétique pour le NHANES III :

- Même s'il n'y a rien d'écrit à ce sujet dans l'article consulté, on peut penser que cet estimateur est biaisé lorsque des estimations pour petites régions sont produites.
- Cette approche nécessite de disposer d'un certain nombre de caractéristiques individuelles provenant du recensement, ce qui peut limiter le nombre de caractéristiques incluses dans le modèle.
- L'application de cette méthode peut nécessiter beaucoup de travail. En effet, un modèle doit être créé pour chaque prévalence de santé à prédire.

5. Méthode d'estimation composite de petites régions

La méthode d'estimation composite de petites régions a été créée afin de combiner les avantages des estimateurs directs (non biaisés) et des estimateurs synthétiques (bonne précision). Les estimateurs découlant de cette méthode résultent d'une combinaison linéaire de l'estimateur direct et de l'estimateur synthétique. Le poids assigné à chacun des deux estimateurs dans le calcul de l'estimateur composite peut être déterminé de différentes façons. À cet effet, Rao (2003) présente différentes méthodes pour la création de ces poids. En pratique, il

arrive souvent que le poids de chacun des estimateurs est choisi de façon à être inversement proportionnel à son *EQM*. Par exemple, si l'estimateur synthétique obtenu à partir du modèle a une *EQM* deux fois moins élevée que celle de l'estimateur direct, l'estimateur composite sera beaucoup plus près de l'estimateur synthétique que de l'estimateur direct.

En bref, l'estimateur composite devrait être plus précis que l'estimateur direct et aussi moins biaisé qu'un estimateur synthétique. Globalement, un bon estimateur composite devrait avoir une meilleure *EQM* que l'estimateur direct.

5.1 Méthode d'estimation composite analysée pour le NHIS (États-Unis)

Le « National Health Interview Survey (NHIS) » est une enquête annuelle américaine qui a été réalisée la première fois en 1957. Tel que mentionné à la section 2, cette enquête fait l'objet de recherche sur les estimateurs pour petites régions depuis quelques décennies déjà. Un document intitulé : « National Health Interview Survey : Research for the 1995-2004 Redesign (1999) » fait état des dernières recherches à ce sujet pour le remaniement décennal de cette enquête.

Dans ce document, de nouveaux estimateurs composites développés pour le remaniement sont présentés. Ces estimateurs se basent sur la théorie de Bayes pour attribuer l'importance respective des estimateurs direct et synthétique dans la construction de l'estimateur composite. En fait, pour ces nouveaux estimateurs composites, l'estimateur direct est utilisé pour représenter les données échantillonnées de la petite région et l'estimateur synthétique est utilisé pour représenter les données non échantillonnées. Puisque la partie synthétique de ces estimateurs est importante, ceux-ci ont été appelés « Generalized Synthetic Estimator (GSE) » (Marker, 1993).

Il est intéressant de noter que la partie synthétique de ces estimateurs est composée de deux parties : l'une calculée à partir des données de l'enquête et l'autre calculée à partir des données de l'enquête antérieure. La première partie provenant de l'enquête est calculée en utilisant une méthode similaire à celle utilisée pour l'enquête HID de la France. C'est-à-dire que l'on utilise l'estimateur national que l'on pondère par la répartition régionale selon différents sous-groupes démographiques. Ces sous-groupes sont définis à l'aide des variables suivantes : groupes d'âge, sexe, ethnie et type de milieu urbain. Pour la création de la seconde partie, la même méthodologie est utilisée avec l'estimateur national de l'année précédente. Notons, finalement, que l'un des GSE présenté est construit à partir de 16 sous-groupes démographiques et l'autre à partir de 32 sous-groupes. L'ajout de sous-groupes dans le deuxième GSE doit permettre d'obtenir un estimateur moins biaisé. En contrepartie, l'effectif échantillonné à l'intérieur de chacun de ces sous-groupes peut faire défaut, ce qui détériorera la précision de ce deuxième estimateur.

Ces GSE ont été comparés aux estimateurs direct et synthétique habituels en utilisant les données du « NHIS » de 1988. Les comparaisons ont montré que les résultats des GSE ressemblaient beaucoup aux résultats obtenus pour l'estimateur synthétique habituel (voir **tableau 1**). Ceci n'est pas surprenant, puisque la partie synthétique occupe une part importante des GSE. Comme pour l'estimateur synthétique habituel, l'étendue des valeurs d'estimation de prévalences pour les petites régions a tendance à se rétrécir pour se rapprocher de l'estimation nationale. De plus, mentionnons que les GSE donnent des estimations de meilleures précisions que l'estimateur synthétique habituel. Cependant, le biais obtenu pour les GSE est équivalent au biais de l'estimateur synthétique. Le biais étant beaucoup plus important que la variance pour ces estimateurs, cela fait en sorte que l'*EQM* des GSE est presque égale à l'*EQM* de l'estimateur synthétique habituel. En terminant, mentionnons que l'*EQM* des GSE est plus petite que l'*EQM* de l'estimateur direct, ce qui constitue un avantage des GSE sur l'estimateur direct.

Tableau 1
Comparaison du GSE avec les estimateurs synthétique et direct

	Estimateur synthétique pour NHIS			Estimateur direct pour NHIS		
GSE	Biais	Variance	*EQM*	Biais	Variance	*EQM*
	=	<	=	>	<	<

5.1.1 Forces et faiblesses des estimateurs composites utilisés pour le NHIS

Forces :

- Les estimateurs affichent une meilleure précision que l'estimateur direct.
- Les estimateurs possèdent une *EQM* plus petite que celle de l'estimateur direct.
- Les estimateurs sont obtenus même pour des régions qui n'ont pas été échantillonnées.

Faiblesses :

- Les estimateurs ont tendance à se rétrécir autour de l'estimation nationale.
- Les estimateurs sont biaisés.
- Lorsque l'effectif échantillonné dans une petite région est nul ou quasi-nul, ces estimateurs deviennent à toutes fins utiles des estimateurs synthétiques.
-

5.2 Méthode d'estimation composite utilisée pour l'ESCC (Canada)

Statistique Canada (SC) a utilisé, pour le cycle 1.1 de son enquête sur la santé des collectivités canadiennes (ESCC), la méthode composite de Chattopadhyay et coll. (1999) pour l'estimation de prévalences de santé pour de petites régions de l'Île-du-Prince-Édouard. Cette méthode, comme les autres méthodes composites, permet d'ajuster l'estimateur direct de manière à obtenir un estimateur plus précis.

Pour employer cette méthode, il faut poser différentes hypothèses. Tout d'abord, la première hypothèse est que les prévalences régionales suivent une loi uniforme centrée sur la prévalence provinciale. Le choix de cette loi permet une certaine variabilité des données, contrairement à une loi normale qui concentre plus fortement les données autour de la moyenne. Par exemple, la moyenne des fumeurs pour les 65 ans et plus des régions de l'Île-du-Prince-Édouard suit une loi uniforme centrée sur la moyenne de la province des fumeurs de 65 ans et plus.

Ensuite, la seconde hypothèse est que la variabilité des statistiques entre les cinq régions de l'Île-du-Prince-Édouard et la donnée de cette province était la même qu'entre les régions des différentes provinces et la moyenne provinciale. À partir de ces deux hypothèses, on peut ainsi obtenir des estimations plus précises des prévalences régionales. L'estimateur composite devient alors une combinaison de l'estimateur direct de la région et de l'estimateur de la province. Cette combinaison est effectuée pour chacun des sous-groupes d'âge et de sexe à l'intérieur d'une petite région. Toutes ces estimations régionales sont ainsi glissées vers l'estimation provinciale.

L'avantage de cette méthode est qu'elle produit des estimations de prévalences régionales qui sont plus précises que les estimations directes. Cependant, ces estimations sont vraisemblablement biaisées étant donné qu'elles ne tiennent pas compte de plusieurs caractéristiques régionales, autres que l'âge et le sexe, ayant un effet sur la prévalence à estimer. Malheureusement, lors de la production de ces estimations composites, SC n'a pas

mesuré le biais de ces estimateurs. Dans l'article de Chattopadhyay et coll. (1999), une méthode de calcul de l'*EQM* est présentée. Cependant, cette méthode n'a pas été utilisée par SC. L'organisme statistique national du Canada s'est limité à produire une mesure de la variance pour l'estimateur à l'aide des poids *bootstrap* disponibles pour cette enquête.

SC est conscient des problèmes liés à cet estimateur. D'ailleurs, un projet de recherche est en cours à SC dans le but de trouver quelle serait la meilleure méthodologie pour la production d'estimation pour des petites régions[10] dans le cadre de l'ESCC. Un rapport préliminaire a été rédigé à ce sujet mais celui-ci ne peut pas être diffusé pour le moment. Nous savons, pour l'instant, que ce rapport proposerait une méthode améliorée qui utiliserait de l'information auxiliaire. Il reste à savoir si, dans la pratique, ce nouveau modèle pourrait s'appliquer facilement.

5.2.1 Forces et faiblesses de l'estimateur composite utilisé pour l'ESCC (cycle 1.1)

Forces :

- L'estimateur affiche une meilleure précision que l'estimateur direct.
- L'estimateur est obtenu même pour des régions qui n'ont pas été échantillonnées.
- Cette approche permet d'obtenir des estimations par sous-groupe démographique à l'intérieur des petites régions.
- L'estimateur affiche un plus petit biais qu'un estimateur synthétique.

Faiblesses :

- L'estimateur est biaisé.
- L'estimateur a tendance à se rétrécir pour se rapprocher de la valeur provinciale. Plus la taille de l'échantillon est petite, plus le glissement vers la statistique provinciale est prononcé.
- Il faut disposer des comptes du recensement pour les sous-groupes utilisés.
- Lorsque la taille de l'échantillon est très petite dans une région, l'estimateur dépend presque exclusivement de sa partie synthétique et donc de la distribution selon l'âge et le sexe de cette sous-région. Donc, si deux sous-régions ont une distribution selon l'âge et le sexe qui se ressemble et si elles possèdent une petite taille d'échantillon équivalente, alors l'estimation composite résultante pour ces deux sous-régions sera pratiquement la même. Les autres facteurs influençant la prévalence de santé (par exemple : le niveau de pauvreté d'une sous-région) ne seront donc pas pris en considération dans une telle situation.

5.3 Méthode d'estimation composite de Larsen (États-Unis)

Cette méthode a été développée par Larsen (2003) dans le but d'améliorer la méthode composite présentée à la section précédente en y ajoutant des variables auxiliaires. L'ajout de variables auxiliaires s'effectue par le biais d'un modèle de régression où les paramètres sont estimés, en quelque sorte, de manière à minimiser un écart entre les estimations directes et les estimations obtenues par le modèle. Cette nouvelle méthodologie devrait ainsi permettre d'obtenir des estimations qui seraient moins biaisées que celles obtenues par la méthode de Chattopadyay et coll. (1999).

10. Cette méthodologie servirait également pour la production d'estimations pour de petites populations (petites prévalences) dans le cadre de l'ESCC.

Les variables auxiliaires utilisées par Larsen permettent de caractériser les petites régions. Pour construire l'estimateur, neuf variables de ce type ont été analysées: le taux de non-emploi, le pourcentage de maisons vacantes, le pourcentage de 18 ans ou moins, le pourcentage de la population appartenant à une minorité, le pourcentage d'enfants pauvres, le pourcentage d'enfants vivant dans une famille monoparentale, le pourcentage de jeunes ayant commis un crime, le pourcentage de familles recevant l'assistance fédérale pour les familles à faible revenu qui ont des enfants (AFDC) et le pourcentage de naissances qui ont nécessité des soins prénataux dans les trois premiers mois de grossesse. Ces variables proviennent du recensement américain et du « Kid Count 1993 survey ».

Une telle approche utilisant des variables auxiliaires de niveau régional a été utilisée également pour le HSfE (voir section 4.2) et pour NHANES III (voir section 4.3). Cependant, ce qui distingue ce nouvel estimateur des deux autres est qu'il ne contient pas seulement une partie synthétique mais aussi une partie directe. De plus, sa partie synthétique est calculée de manière à se rapprocher le plus possible de la partie directe, ce qui permet d'obtenir un estimateur moins biaisé.

Pour savoir si la méthodologie utilisée pour son estimateur est adéquate, Larsen a testé celle-ci sur les données d'une enquête menée auprès des ménages par la firme Gallup. Dans le cadre de ces tests, Larsen voulait calculer la prévalence de personnes qui sont dépendantes de l'alcool pour chacune des petites régions.

Ces tests ont mené Larsen à créer deux estimateurs : le premier avec une variable auxiliaire seulement (pourcentage d'enfants pauvres) et l'autre avec deux variables auxiliaires (pourcentage d'enfants pauvres et pourcentage d'enfants vivant dans une famille monoparentale). L'ajout de variables auxiliaires supplémentaires dans le modèle ne permettait pas d'améliorer substantiellement l'estimateur. Les résultats obtenus pour ces deux estimateurs montrent que ceux-ci tendent à être moins biaisés que l'estimateur de Chattopadhyay et coll. En effet, les deux estimateurs de Larsen semblent être plus près de l'estimateur direct lorsque celui-ci est non-nul. De plus, la même constatation est effectuée lorsque l'estimateur direct est nul.

Cependant, les tests effectués montrent également que la variance des estimateurs de Larsen est plus élevée que la variance de l'estimateur de Chattopadhyay et coll. Et finalement, l'*EQM* des estimateurs de Larsen est un peu plus élevée que l'*EQM* de l'estimateur de Chattopadhyay et coll. Ce qui fait dire à Larsen que ce sont les variables choisies pour la construction des sous-groupes démographiques qui ont le plus d'impact sur les estimations produites. Larsen mentionne également qu'un meilleur choix de variables auxiliaires aurait pu produire une meilleure *EQM* pour ses estimateurs.

5.3.1 Forces et faiblesses de l'estimateur composite de Larsen

Forces :

- L'estimateur affiche une meilleure précision que l'estimateur direct.
- L'estimateur a un biais un peu moins élevé que l'estimateur de Chattopadhyay et coll.
- L'estimateur est obtenu même pour des régions qui n'ont pas été échantillonnées.
- Cette approche permet d'obtenir des estimations par sous-groupe démographique à l'intérieur des petites régions.

Faiblesses :

- L'estimateur est biaisé.
- L'estimateur a une *EQM* un peu plus élevée que celle de l'estimateur de Chattopadhyay et coll. Cependant, l'utilisation de variables auxiliaires davantage corrélées avec l'estimation à produire pourrait peut-être amener une meilleure *EQM*.

- Il faut disposer des comptes du recensement pour les sous-groupes utilisés et de différentes statistiques caractérisant les régions.
- L'estimateur a tendance à se rétrécir pour se rapprocher de la valeur provinciale. Plus la taille de l'échantillon est petite, plus le glissement vers la statistique provinciale est prononcé. On peut supposer que ce phénomène est moins prononcé que dans le cas de l'estimateur de Chattopadhyay et coll.
- L'application de cette méthode peut nécessiter beaucoup de travail. En effet, un modèle doit êtrecréé pour chaque prévalence de santé à prédire.

6. Conclusion et possibilité d'application pour des estimations québécoises

6.1 Conclusion

Plusieurs constats peuvent être formulés au sujet de l'ensemble des méthodes d'estimation présentées dans cet avis (voir **tableau 2**). De prime abord, toutes ces méthodes permettent d'obtenir une meilleure précision pour les estimations de prévalences de santé que les estimateurs directs.

Tableau 2
Caractéristiques principales des méthodes répertoriées

Nom de l'enquête	Type de méthode	Modélisation d'une variable à la fois	Information auxiliaire utilisée autre que l'âge et le sexe
Enquête HID (France)	Synthétique	Non	Oui
HSfE (Angleterre)	Synthétique	Oui	Oui
NHANES (États-unis)	Synthétique	Oui	Oui
NHIS (États-unis)	Composite	Non	Oui
ESCC (Canada)	Composite	Non	Non
Larsen (États-unis)	Composite	Oui	Oui

En contrepartie, les estimations produites sont biaisées. Les grandes prévalences sont sous-estimées et les petites prévalences sont surestimées. Ceci entraîne un déplacement des estimations régionales vers l'estimation provinciale. Ce phénomène implique une sous-estimation des différences interrégionales. Évidemment, il est possible d'atténuer ce phénomène en utilisant un estimateur composite à la place d'un estimateur synthétique, en incorporant dans l'estimateur des variables auxiliaires fortement liées à l'estimation que l'on veut produire et en modélisant les prévalences une à la fois. La dernière solution proposée implique toutefois une quantité de travail non négligeable étant donné que le modèle doit être refait pour chaque caractéristique de santé à estimer (voir tableau 3 pour la synthèse). Malgré les problèmes liés au biais, les estimateurs pour petites régions peuvent constituer une bonne solution surtout si l'estimateur direct a une faible précision. De plus, il vaut certainement mieux utiliser l'un des estimateurs présentés dans ce document que d'approximer l'estimation d'une petite région par l'estimation d'une plus grande région. Cette dernière solution est sans doute la plus biaisée.

Tableau 3
Biais et quantité de travail selon des caractéristiques des estimateurs

	Modélisation d'une variable à la fois		Utilisation d'information auxiliaire		Méthodes d'estimation	
	Oui	Non	Moins	Plus	Synthétique	Composite
Biais de l'estimateur	↓	↑	↑	↓	↑	↓
Quantité de travail	↑	↓	↓	↑	↓	↑

Finalement, pour se guider dans le choix d'un estimateur pour petites régions, l'*EQM* devrait être utilisée. Tel que mentionné précédemment, cette mesure permet de tenir compte simultanément de la précision et du biais de l'estimateur. Un estimateur sera généralement considéré meilleur qu'un autre estimateur si son *EQM* est plus petite que celle de l'autre estimateur. Il faut noter, cependant, que le calcul de l'*EQM* est plus ardu que le calcul de précision habituel étant donné la difficulté à mesurer le biais de l'estimateur.

6.2 Possibilité d'application pour des estimations québécoises

Afin de déterminer quelle méthode d'estimation pour petites régions pourrait être utilisée pour la production d'estimations québécoises, des tests devraient être effectués sur différentes enquêtes. Une première possibilité serait d'utiliser les données de l'ESCC. Pour cette enquête, deux types d'estimations pourraient être examinées :

1. Production d'estimations pour des régions plus petites que celles habituellement diffusées pour l'ESCC.
2. Production d'estimations pour des sous-groupes démographiques (exemple: par groupe âge-sexe) à l'intérieur des régions habituellement diffusées pour l'ESCC.

Une autre possibilité serait d'utiliser les données de l'enquête sur la participation et les limitations d'activité (EPLA) de Statistique Canada. Des tests, menés à partir de cette enquête, seraient motivés, d'une part, puisqu'il semble y avoir de l'intérêt pour la production d'estimations régionales pour cette enquête. D'autre part, des questions du recensement portant sur les limitations pourraient être utilisées comme variables auxiliaires pour la méthode d'estimation retenue.

Quelle que soit la possibilité retenue pour les tests, il nous semble que les méthodes d'estimation composite devraient être privilégiées par rapport aux méthodes synthétiques, lorsque c'est possible, étant donné qu'elles sont moins biaisées. De plus, l'utilisation d'une méthode composite utilisant un modèle avec variable auxiliaire dans sa partie synthétique permettrait sans doute d'obtenir des estimations moins biaisées. Cependant, ce type d'approche peut nécessiter un temps non négligeable de modélisation pour chaque estimation à produire. Si l'on ne veut pas assumer une telle charge de travail, une méthode composite sans modélisation serait à privilégier.

Par ailleurs, il serait important de valider les méthodes d'estimation qui seront retenues pour les tests. Pour ce faire, la meilleure façon est de comparer les résultats de l'estimateur retenu avec les résultats de l'estimateur direct. Évidemment, cette comparaison peut seulement être faite dans une région ayant une taille d'échantillon suffisante. Un suréchantillonnage pour certaines régions pourrait permettre de telles comparaisons. Ces validations peuvent permettre, par exemple, de déceler si l'estimateur retenu est biaisé et, si oui, d'indiquer dans quel sens va ce biais. Les sur-échantillons disponibles pour certaines régions de l'ESCC pourraient permettre ce genre de validation. Toutefois, l'absence de suréchantillons pour l'EPLA pourrait rendre difficile de telles validations.

Il ne faudrait pas pour autant écarter la possibilité de faire des tests à partir des données de l'EPLA. Il faut se rappeler que la disponibilité d'information du recensement se rapportant aux limitations pourrait servir à développer un estimateur de qualité pour cette enquête. Si on voulait valider la qualité de l'estimateur créé pour l'EPLA, en l'absence de sur-échantillons, il faudrait peut-être utiliser des regroupements de petites régions qui permettraient d'obtenir un estimateur direct adéquat pour la comparaison.

En terminant, le choix final d'une méthode d'estimation à tester pour le Québec devrait s'inspirer des résultats obtenus lors des derniers travaux de SC sur le sujet. Les travaux menés par SC permettront certainement l'obtention d'une méthodologie améliorée par rapport à ce qui avait été utilisé la dernière fois pour l'ESCC. Enfin, il serait important de recevoir l'avis d'un chercheur universitaire (spécialiste dans le domaine) avant de finaliser le choix de cette méthode.

7. Références

BAJEKAL, M., S. SCHOLES, K. PICKERING et S. PURDON (2004), *Synthetic estimation of healthy lifestyles indicators: Stage 1 report.* (Unpublished) – Adresse web: www.natcen.ac.uk/smu_reports05/Synthetic_Estimation_Stage_1_Report.pdf

CHATTOPADHYAY, M., P. LAHIRI, M. LARSEN et J. REIMNITZ (1999), Estimation composite de la prévalence des drogues pour des zones infraétats, *Techniques d'enquêtes*, **25**(1), 91-97.

COUET, C. (2002), Estimations locales dans le cadre de l'enquête HID, *Document de travail*, N° F0207, INSEE, (Adresse web : www.insee.fr/fr/nom_def_met/methodes/doc_travail/docs_doc_travail/f0207.pdf

DIGAETANO, R., J. WAKSBERG, E. MACKENZIE, U. HOPKINS et R. YAFFE (1980), Synthetic Estimates for Local Areas from the Health Interview Survey, *Proceedings of the Section on Survey Research Method*, American Statistical Association, pp. 46-55.

ELAZAR, D. et L. CONN (2004), Small Area Estimation of Disability in Australia, *Research Paper 1351.0.55.006 Australian Bureau of Statistics.*

LARSEN, M.D. (2003), Estimation of small-area proportions using covariates and survey data, *Journal of Statistical Planning and Inference* **112** (2003), pp. 89-98.

LEVY, P. S., et D. K. FRENCH (1977), Synthetic estimation of state health characteristics based on the Health Interview Survey. *Vital and Health Statistics Series 2, N°. 75*, U.S. Department of Health, Education & Welfare.

MALEC, D., W. DAVIS, et X. CAO (1996), Small Area Estimates of Overweight Prevalence using the Third National Health And Nutrition Examination Survey (NHANES III), *Proceedings of the Section on Survey Research Method*, American Statistical Association, pp. 326-331.

MARKER, D.A. (1993), Small Area Estimation for the U.S. National Health Interview Survey, *Proceedings of the Section on Survey Research Method*, American Statistical Association, pp. 11-20.

NATIONAL CENTER FOR HEALTH STATISTICS (1968), *Synthetic State Estimates of disability*, P.H.S. Publication 1759, Washington, DC: U.S. Government Printing Office.

NATIONAL CENTER FOR HEALTH STATISTICS (1977a), *Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey*, D.H.E.W Publication N°. (PHS) 78-1349, U.S. Government Printing Office, Washington, D. C.

NATIONAL CENTER FOR HEALTH STATISTICS (1977b), *State Estimates of Disability and Utilization of Medical Services, 1969-1971*, D.H.E.W Publication No. (PHS) 77-1241, U.S. Government Printing Office, Washington, D. C.

NATIONAL CENTER FOR HEALTH STATISTICS (1978), *State Estimates of Disability and Utilization of Medical Services, 1974-1976*, D.H.E.W Publication No. (PHS) 78-1241, U.S. Government Printing Office, Washington, D. C.

NATIONAL CENTER FOR HEALTH STATISTICS (1999), National Health Interview Survey: Research for the 1995-2004 redesign. *Vital and Health Statistics Series 2, N°. 126.*

PICKERING, K., S. SCHOLES, et M. BAJEKAL (2004), *Synthetic estimation of healthy lifestyles indicators: Stage 2 report.* (Unpublished) – Adresse web: www.natcen.ac.uk/smu_reports05/Synthetic_Estimation_Stage_2_Report.pdf

PICKERING, K., S. SCHOLES, et M. BAJEKAL (2005), *Synthetic estimation of healthy lifestyles indicators: Stage 3 report.* (Unpublished) – Adresse web: www.natcen.ac.uk/smu_reports05/Synthetic_Estimation_Stage_3_Report.pdf

RAO, J.N.K (2003), *Small Area Estimation*, Wiley Series in Survey Methodology.

SCHAIBLE, W.L., D. B. BROCK, R. J. CASADY, et G. A. SCHNACK (1979). Small area estimation: An empirical comparison and synthetic estimators for states. *Public Health Service Series 2-82 (N°. 80-1356)*, NCHS, D.H.E.W., U.S. Government Printing Office, Washington, D. C.

STATISTIQUE CANADA (1983). Une bibliographie pour l'estimation pour les petites régions, *Techniques d'enquêtes*, **9**(2), pp. 267-287.

STATISTIQUE CANADA (1986). *Small area statistics, an international symposium '85*. Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University/University of Ottawa, August 1986.

TWIGG, L. et G. MOON (2002), Predicting small area health-related behavior: a comparison of multilevel synthetic estimation and local survey data, *Social Science and Medicine* **54**(6), 931-937.

8. Autres Références

CHANG, H.-Y. (2004), Exploring the Feasibility of Using Small-area Estimation to Estimate Health Behaviors in Remote Areas in Taiwan, *Proceedings of the Section on Survey Research Method*, American Statistical Association.

CONGDON, P. (2006), Estimating diabetes prevalence by small area in England, *Journal of Public Health*, **28**(1), pp. 71-81.

LANGFORD, I. H., A. H. LEYLAND, J. RASBASH, et H. GOLDSTEIN (1999), Multilevel modelling of the geographical distribution of diseases, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**(2), pp. 253–268.

MALEC, D., J. SEDRANSK, et L. TOMPKINS (1993), Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey, *Case Studies in Bayesian Statistics*, C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Singpurwalla (Eds.), New York: Springer-Verlag, pp. 377-389.

MALEC, D., J. SEDRANSK, C. L. MORIARTY et F. B. LECLERE (1997), Small area inference for binary variables in the National Health Interview Survey, *Journal of the American Statistical Association*, **92**(439), pp. 815-826.

MARKER, D. A. (2001), Production d'estimations régionales d'après les données d'enquêtes nationales : Méthodes visant à réduire au minimum l'emploi d'estimateurs indirects, *Techniques d'enquêtes*, **27**(2), pp. 201-207.

VAISH, A. K., N. SATHE, et R. E. FOLSOM (2004), Small Area Estimates of Diabetes and Smoking Prevalence in North Carolina Counties: 1996-2002 Behavioral Risk Factor Surveillance System, *Proceedings of the Section on Survey Research Method*, American Statistical Association, pp. 4535-4544.

Annexe 2

Variables retenues pour le modèle de Fay-Herriot sur le taux d'emploi :

- taux d'emploi par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (taux_emploi)

- proportion de personnes en milieu rural par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (rural)

- log des heures moyennes travaillées par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (log_heures)

- proportion de personnes n'étant pas de minorités visibles par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (minorité_non)

- interaction du log des heures moyennes travaillées et de la proportion de personnes n'étant pas de minorités visibles chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (heureXminorité)

Paramètres estimés par le modèle et leurs erreurs-type

Variables	Paramètres	Erreurs-type
ordonnée	-3,60	1,83
taux_emploi	1,14	0,08
rural	0,09	0,03
log_heures	0,96	0,50
minorité_non	5,34	2,17
heureXminorité	-1,48	0,59

Variables retenues pour le modèle de Fay-Herriot sur le taux d'activité :

- taux d'activité par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (taux_activité)

- proportion de personnes en milieu rural par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (rural)

- proportion de personnes dont la principale source de revenu est le gouvernement par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (gouvernement)

- proportion de personnes vivant seules par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (seule)

- proportion de personnes âgées de 25 à 34 ans par région chez les personnes ayant répondu « oui » à au moins une des questions filtre du recensement (25_34_ans)

Paramètres estimés par le modèle et leurs erreurs-type

Variables	Paramètres	Erreurs-Type
ordonnée	0,26	0,26
taux_activité	0,55	0,31
rural	0,12	0,04
gouvernement	-0,47	0,29
seule	0,21	0,14
25_24_ans	0,56	0,46

Annexe 3

Tableau 1
Taux d'emploi de Fay-Herriot pour la région regroupée de Côte-Nord et Nord-du-Québec

Région sociosanitaire	Taux d'emploi de Fay-Herriot		
	% [IC]	CV (%)	% provenant de l'EPLA
Côte-Nord et Nord-du-Québec	37,4 [33,5 ; 41,3]	5,3	1,7

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

Tableau 2
Taux d'activité de Fay-Herriot pour la région regroupée de Côte-Nord et Nord-du-Québec

Région sociosanitaire	Taux d'emploi de Fay-Herriot		
	% [IC]	CV (%)	% provenant de l'EPLA
Côte-Nord et Nord-du-Québec	46,8 [40,6 ; 53,0]	6,7	5,2

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