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[1] en utilisant les données du « Health Survey for England (HSfE)» (Bajekal et coll., 2004; Pickering et coll., 2004; Pickering et coll., 2005).

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

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

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**		
	Biais	Variance	*EQM*	Biais	Variance	*EQM*
GSE	=	<	=	>	<	<

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[2] 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

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

ainsi permettre d'obtenir des estimations qui seraient moins biaisées que celles obtenues par la méthode de Chattopadyay et coll. (1999).

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 être créé 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., Scholes, S., Pickering, K. et Purdon, S. (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., Lahiri, P., Larsen, M. et Reimnitz, J. (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*, No F0207, INSEE, (Adresse Web : *www.insee.fr/fr/nom_def_met/methodes/doc_travail/docs_doc_travail/f0207.pdf*

Digaetano, R., Waksberg, J, Mackenzie, E, Hopkins U. et Yaffe, R (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 Conn, L. (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), 89-98.

Levy, P. S., et French, D. K. (1977), Synthetic estimation of state health characteristics based on the Health Interview Survey. *Vital and Health Statistics Series 2, No. 75*, U.S. Department of Health, Education & Welfare.

Malec, D., Davis, W. et Cao, X. (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 No. (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, No. 126.*

Pickering, K., Scholes, S. et Bajekal, M. (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., Scholes, S. et Bajekal, M. (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., Brock, D.B., Casady, R.J., et Schnack, G.A. (1979). Small area estimation: An empirical comparison and synthetic estimators for states. *Public Health Service Series 2-82 (No. 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), 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 Moon, G. (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), 71-81.

Langford, I. H., Leyland, A. H., Rasbash, J. et Goldstein, H. (1999), Multilevel modelling of the geographical distribution of diseases, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**(2), 253–268.

Malec, D., Sedransk, J. et Tompkins, L. (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., Sedransk, J., Moriarty, C. L. et Leclere, F. B. (1997), Small area inference for binary variables in the National Health Interview Survey, *Journal of the American Statistical Association*, **92**(439), 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), 201-207.

Vaish, A. K., Sathe, N. et Folsom, R. E. (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.